Tina Memo No. 2016-003
Internal.

Volumetric Accuracy and Parameter dependencies of MR Brain Tissue Model

S. V Notley and N. A. Thacker

Last updated
27 / 2 / 2016

Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

Abstract

Previous work has investigated the accuracy of estimating the pure tissue mean parameters [7] and described a method of improving parameter estimates based on regional partitioning of the images. In this current work, we wish to gain insight into the accuracy of volumetric estimates of each tissue class based on the known accuracies of the full parameter set of the pure tissue distributions. The predominant dependency between errors on each tissue class volume estimate and the errors in the model parameters is identified.

Introduction

Medical image segmentation is a fundamental tool for quantitative data analysis. However, accurate classification of tissues using MR datasets has proven difficult, with assessment of commonly available software tools generating as much as 10% differences between classifications and ground truth [4, 11, 6] across large anatomical areas. Given the quantitative information available (even in a single voxel), this seems shockingly poor. Unfortunately, in order to use MR data as a reliable quantitative tool in clinical investigation, far greater accuracy is probably needed. In particular, identifying regions of brain data as pathology and quantifying subtle changes (tumour heterogeneity) will not be reliable unless normal tissue is reliably modelled to begin with. **Can we expect to achieve any better perfomance than 10% volume errors?**

The first issue we need to appreciate is how performance figures are obtained. Any comparison with a human generated ground truth will increase estimated error levels, even though reproducibility might be more important in clinical decision support. In addition, commonly used evaluation metrics are not entirely reliable [9] and do not provide us with enough information to identify and rank all of the possible sources of errors (noise, parameter errors, biases unmodelled image behaviour etc.). A better approach might put less emphasis on subjective definitions and involve more direct analysis of how errors accumulate during analysis. This becomes more important when seeking to improve performance, in order to focus on the issues which make the largest contributions.

Classification is expected to work best when using a data density model and Bayes Theorem. When using this approach several sources of inaccuracy can be suggested; poor quality data, inappropriate models, or simply poor software design. In principle, based upon conventional models for statistical measurement such as Multinomial distributions (or their Poisson approximation), the volumes of data processed (many thousands of voxels) are capable of supporting much greater precision in volume estimation (V) than is generally observed. Levels of around one percent might even be possible for volumes above 10,000 voxels, provided the estimated data density model is appropriate and accurate [1].

Obtaining accurate measurements from analysis starts with well behaved data. In this work we have taken steps to ensure that data are of the best quality available. In particular we do not use multi-coil acquisition [5], which has both spatially varying signal **and** noise characteristics, making data density distributions significantly more difficult (perhaps impossible) to model. Data quality is further enhanced by using a multi-image analysis, i,e, multi-dimensional data density distributions, which enhance the available information when compared to using a single image. In addition, our models are more complex than the standard Gaussian mixture approach, incorporating partial volume distributions (again multi-dimensional) which better describe observed distributions and accord with the MR (partial volume) image formation process. Finally, we use analysis of errors on resulting image segmentations to confirm that the errors in estimated parameters accord with observed performance as a validation of software and theoretical integrity [8]. In doing so we confirm that our results are the best we can expect to obtain for this model, associated assumptions and data quality. Taking the time to do this also has other potential benefits. We would expect that knowledge of errors on computed results will also have the unusual important role in scientific and medical applications.

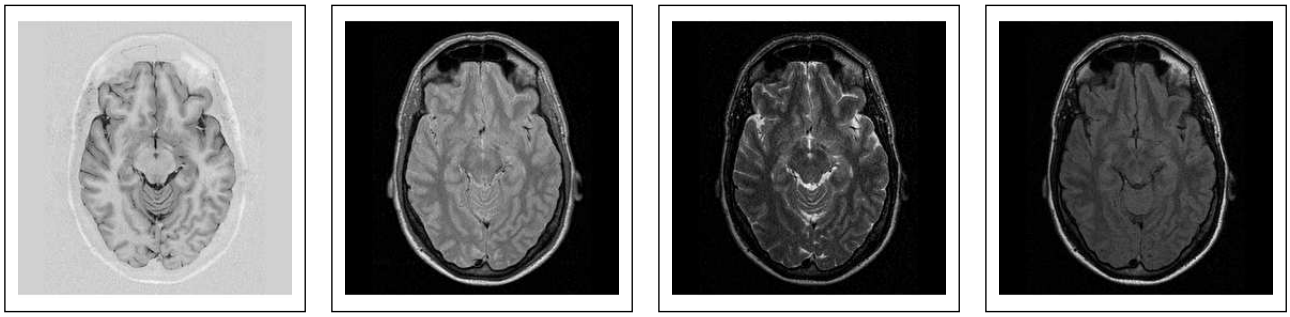
A recent extension to the methods has involved modifying the parameter estimation process in order to reduce parameter error. However, knowing the expected error on the parameter is only a first step in understanding final performance. In any clinical use, a more objective assessment of performance will be given by the ability to estimate volumes of tissue. It is therefore important to understand how errors on parameters affect subsequent errors in quantities. As in all such cases this issue can be approached via the method of error propagation [12]. The intention of this work was largely to show, at least in one case, some typical figures and the level of conformity of predictions from the method with results in real data.

[1] $var(V) \approx V$, so that for $V = 10,000$ $SD(V)/V \approx \sqrt{V}/V = 1.0\%$

Methodology

Data Sets

The data set used in this paper consists of 4 MR images taken from the the same 38 year old male. The image types are (a) Inversion Recovery Turbo Spin Echo (IRTSE), (b) Variable Echo - Proton Density (VE-PD), (c) Variable Echo - T2 (VE-T2), and (d) Fluid-Attenuated Inversion Recovery (FLAIR). The voxel dimensions used were 1mm x 1mm x 5mm with each image having a resolution of 256 x 256 pixels.



(a) (b) (c) (d)

Figure 1: *The MR data set used. (a) Inversion Recovery Turbo Spin Echo (IRTSE), (b) Variable Echo - Proton Density (VE-PD), (c) Variable Echo - T2 (VE-T2), and (d) Fluid-Attenuated Inversion Recovery (FLAIR)*

The estimated noise levels (in grey levels) on each image is given in the table below:

Image	IRTSE	VE(PD)	VE (T2)	FLAIR
σ	80	70	74	80

Table 1. Estimated noise level on images

Error Propagation Analysis

For a system with a single degree of freedom, the standard deviation of the system output may be estimated, to first order as:

$$\sigma_f = |\frac{df}{d\theta}|\sigma_\theta \tag{1}$$

where f is the system transfer function, and θ is the parameter of the transfer function. For systems with N degrees of freedom, the errors due perturbations of the individual parameters add in quadrature to give:

$$\sigma_f^2 = \mathbf{D}^T \mathbf{C}_f \mathbf{D} \tag{2}$$

where D is a N dimensional vector of partial differentials of the transfer function with respect to the parameters:

$$\mathbf{D} = (\frac{\partial f}{\partial \theta_1}, \frac{\partial f}{\partial \theta_2}, \cdots, \frac{\partial f}{\partial \theta_N})^T$$

where θ_n is the n-th parameter of the transfer function f. The matrix $\mathbf{C}_f$ is an NxN symmetrical matrix of parameter covariances:

$$\mathbf{V}_f = \begin{pmatrix} \sigma_{\theta_1}^2 & \sigma_{\theta_1}\sigma_{\theta_2} & \sigma_{\theta_1}\sigma_{\theta_3} & \cdots & \sigma_{\theta_1}\sigma_{\theta_N} \\ \sigma_{\theta_2}\sigma_{\theta_1} & \sigma_{\theta_2}^2 & \sigma_{\theta_2}\sigma_{\theta_3} & \cdots & \sigma_{\theta_2}\sigma_{\theta_N} \\ \vdots & & \ddots & & \vdots \\ \sigma_{\theta_N}\sigma_{\theta_1} & \sigma_{\theta_N}\sigma_{\theta_2} & \cdots & \cdots & \sigma_{\theta_N}^2 \end{pmatrix}$$

Covariance Matrix Estimation

The model [3, 10] is composed of an additive mixture of pure Gaussian tissue distributions (one for each tissue class) and partial volume distributions between pairs of pure tissue distributions. Since the partial volumes distribution

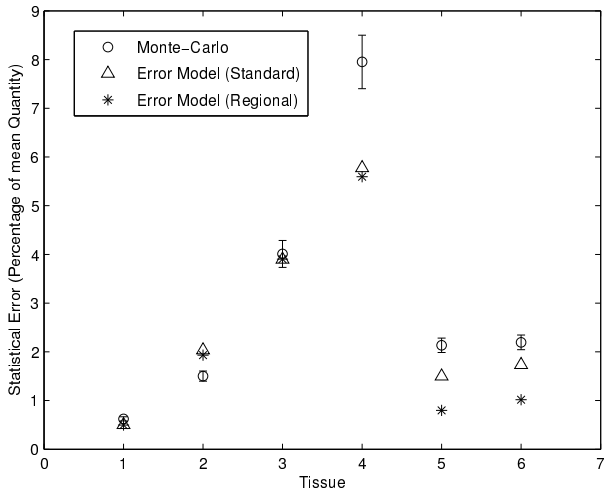


Figure 2: *Error propagation results comparing:(a) Monté-Carlo data (circles), (b) Predicted Results with standard approach (Triangles), (c) Predicted results with regional approach (asterisk)*

parameters are derived from the pure tissue distribution parameters we only model the variations in the pure tissue parameters. For, T, tissue classes and, N, images there will be T x N mean and T x N variance parameters (one for each image in each tissue class). Errors in covariance parameters (covariance across images) lead to rotations in the pure tissue distributions and are significant only for long extended distributions. Given the noise levels above, the distributions are expected to be compact and thus for tractability we do not consider the covariance parameters in the following analysis.

The covariance matrix $\mathbf{C}_f$ was estimated from Monté-Carlo simulations. To facilitate this, the standard deviation of the noise in each of the images of the data set were estimated. Based on these noise estimates, a surrogate data set was created by adding random noise fields (Gaussian and independent across pixels) to each image. For each image, the level of noise introduced was half the standard deviation of the original estimated image noise on each respective image. Maximum likelihood fits of the model parameters were then estimated on each surrogate data set using a modified expectation maximisation algorithm (12 iterations) [3]. The process was repeated 100 times to give the statistical power to obtain an estimate of the variance and covariance of the model parameters. The quantities for each tissue were also calculated for each set of parameters. The estimated true standard deviation of the parameters and tissue quantities given the image noise was estimated by doubling these values; accounting for the level of additional noise being half the image noise levels.

Derivative Estimation

Assuming the parameters of the model are at the minimum volume error, i.e. converged, the partial derivatives of equation (2) may be numerically estimated by shifting the respective parameters relative to the minimum. The partial derivative estimates are given by the central difference approximation:

$$\frac{\partial V}{\partial \theta_x} \approx D_v(\theta_x) = \frac{f(\theta_x + \Delta\theta_x) - f(\theta_x - \Delta\theta_x)}{2\Delta\theta_x} \tag{3}$$

Results

Figure 2 show the results of applying the error propagation analysis to estimate the error on tissue quantities. Results are shown for the standard approach and those expected by doubling the statistical accuracy of the grey and white matter parameters (as found with the regional approach to parameter estimation). The Monté-Carlo results for the statistical accuracy of the tissue quantities are also shown for comparison. As previously stated, for tractability, the error model only contains statistical errors related to the model mean and variance parameters.

Errors in the other model parameters (co-variance and partial volume parameters) and the initial parameter values

are not accounted for in the error model. As a results of the truncated error model, systematic differences between the error model results and the Monté-Carlo results are expected. The error model trend systematically follows the Monté-Carlo results.

For a doubling in the accuracy of the gray and white matter parameters the error model approximately predicts a doubling in the accuracy of the tissue quantity estimates. From this we would expect the statistical error on the grey and white matter volumes to be around the 1% level.

Mean Parameter	Image No.	Contribution
CSF(σ^2)	1	0.19
CSF(σ^2)	4	0.17
CSF(μ)	4	0.14
CSF(μ)	1	0.12
CSF(μ)	3	0.11
CSF(σ^2)	3	0.10
CSF(σ^2)	2	0.09

Table 2: Components accounting for approximately 90% of total variance of CSF

Table 2 shows the dominant sources of error contributing to approximately 90% of the errors on the CSF quantity. The table clearly shows that it is errors on the CSF parameters alone that account for over 90% of the CSF quantity error and that it is the parameters related to images 1 & 4 (IRTSE and FLAIR) with the greatest influence. However, errors propagated from the mean and variance parameters of the other two images are also significant around the 10% level. Observation of scatter plots for these images show that the distribution of the CSF voxels are well separated from the other tissues and would be expected to give the most information related to class separability.

Mean Parameter	Image No.	Contribution
WM(μ)	1	0.20
CSF(μ)	4	0.14
CSF(μ)	3	0.13
GM(σ^2)	3	0.13
GM(σ^2)	1	0.12
WM(μ)	2	0.08
GM(μ)	3	0.06
CSF(σ^2)	4	0.05
GM(σ^2)	2	0.05

Table 3: Components accounting for approximately 90% of total variance of GM

Table 3 shows the dominant sources of error contributing to approximately 90% of the errors on the grey matter quantity. From this we can see that errors on parameters related to CSF, grey and white matter all contribute significantly to the error on the grey matter quantity. It may also be noted that all images have a significant contribution to this error. Observation of the distribution of the grey matter voxel intensities show that, in intensity space, the grey matter voxels are proximal to the CSF and white matter distributions in all images. Thus, errors on any of these parameter distributions can be expected to cause errors in the quantity estimate on the grey matter.

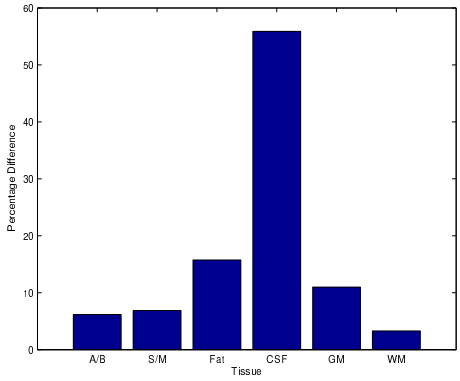


Figure 3: *Systematic differences between using partial volumes and non-partial volumes*

Mean Parameter	Image No.	Contribution
GM(μ)	1	0.17
WM(σ^2)	1	0.15
WM(μ)	1	0.11
WM(σ^2)	2	0.09
GM(μ)	2	0.06
Fat(μ)	2	0.05
WM(μ)	2	0.05
WM(σ^2)	4	0.04
CSF(μ)	3	0.03
GM(μ)	3	0.03
GM(σ^2)	1	0.03
Fat(μ)	3	0.02
GM(μ)	4	0.02
WM(σ^2)	3	0.02

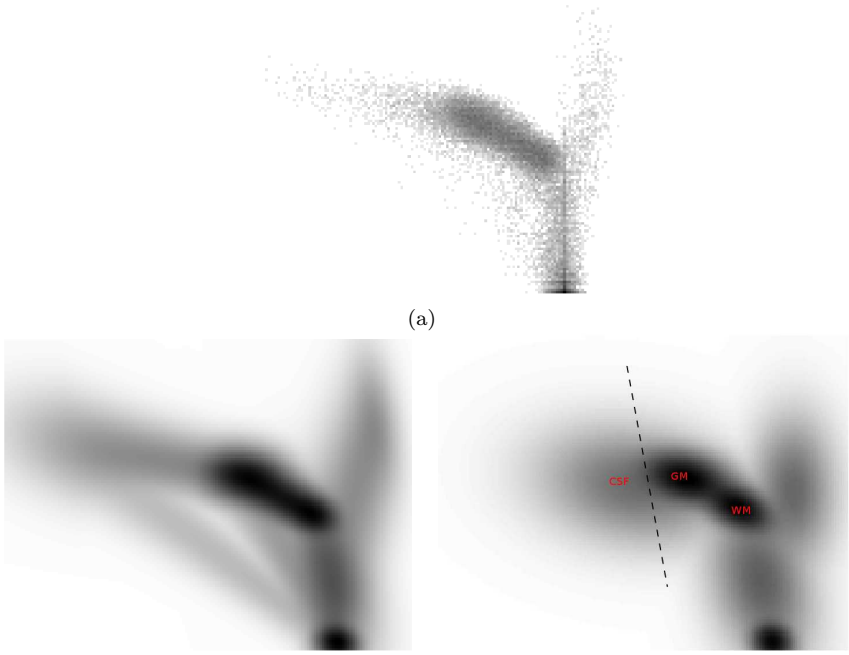
Table 4: Components accounting for approximately 90% of total variance of WM

Table 4 shows the dominant sources of error contributing to approximately 90% of the errors on the white matter quantity. In this case, it can be seen that the predominant sources of error are in the grey matter and white matter parameter estimates related to images 1 and 2 (IRTSE and VE-PD). These observations are consistent with the expected relative proximities of the the mean tissue grey levels.

The error model analysis may be used to predict the improvements in statistical error on the quantities due to improvements in the statistical accuracy of the model parameters. Many of the widely available and commonly used segmentation algorithms found in the literature [13, 2, 1] use simpler Gaussian mixture modelling to model the pure tissue distributions and ignoring partial volume effects. To investigate the systematic differences between using the full partial volume method and a pure Gaussian mixture model Monté-Carlo simulations were also run under both circumstances. Figure 3 shows the results of this experiment and shows the difference as a percentage of the quantities estimated using the full partial volume method. Figure 4 shows that there are significant systematic differences between the description of data density using the two methods. When working with only single images, these distribution errors will still be present but are far more difficult to identify. For the white matter there is around a 4% difference; for air/bone and skin/muscle this rises to between 5% and 10%; Grey Matter and fat tissue classes have between 10% and 20% differences; the CSF shows the most dramatic change with over a 50% difference.

Conclusions

Monté-Carlo simulations, based on estimated image noise levels, show that for a 256x256 MR image datasets the statistical error on grey and white matter tissue quantities are around 2%. An error model using these accuracies has been shown to produce quantity errors on the different tissue classes that systematically follow the results in



(a)

(b)

(c)

Figure 4: *(a) Scatter plot between IRTSE and VD-PD images, (b) full model fit, (c) model fit with no partial volume distributions*

Monté-Carlo. For tractability, not all sources of error are included in the error model and some differences between Monteé-Carlo results and the error model may have been expected. None-the-less predictions from the theory were within measurment errors.

Previously, it has been shown that using the regional approach to fitting model parameter gave around 2 two fold increase in the accuracy of the model parameters for grey and white matter distributions. Inclusion of this improvement in parameter accuracies into the error model demonstrates that the method produces a two fold improvement in the quantity estimates on these tissues.

Although agreement between theory and data is shown in only one dataset (and one image slice), the purpose of this work was not to provide an unequivical empirical generality of this result. Rather it was to gain some insight as to the parameter-to-volume error process and show that in such cases (where data volumes are large enough to make statistcal errors due to noise and sample size insignificant in comparison to those from paramter estimates), the functional form of error propegation correctly predicts an observable linear correlation between parameter and volume measurement accuracies. As a single slice is already a relatively small quantity of data, this approximation is expected only to get better for most applications of segmentation. From this we can conclude that the error model is representative of the way errors are likely to be propagated in a real system.

The errors modelled here seem negligible in comparison to those seen in ground truth evaluations. However, for our work these figures are significant because we have already made significant improvements in performance. In order to demonstrate this, a final experiment shows that even exclusion of the partial volume distributions from the model introduces large systematic changes in quantity estimates. A simple visual comparison of the scatter plots (figure 4), for the first two images and the model distributions, show that the full model is more representative of the true distribution. We do not claim that this is a new finding, such conclusions have been in the literature for over a decade, our results are presented here for completeness. We also do not claim that this observation is of relevance to single image multi-coil acquisitions, (which would start with less reliable information). We claim only that without partial volume terms we would not expect to significantly improve segmentation reliability, even with carefully controlled multi-spectral single coil acquisitions.

Unmodelled partial volume data produces systematic shifts in mean tissue parameters. In turn, the means of the tissue class distributions furthest from the means found with the full model produce large systematic difference in tissue quantities. The level of systematic error (bias) introduced by excluding partial volume distributions is of the same order as those found in independent reviews in the literature [4, 11, 6] and are also consistent with our previous observation regarding large biases in CSF volume estimation [3, 10]. This leads to the conclusion that observed errors in popular software packages could be explained by biases introduced by the assumed density

model, rather than simply data quality and sample size (as already argued in the introduction).

Published evaluations, based on Jaccard or DICE overlaps, do not provide the appropriate results with which to confirm our findings, as the errors on segmentation results would have needed to be evaluated separately in terms of both bias and variance. For future studies, an investigation of error in these terms might allow researchers to understand the contributions from both and then test improvements (such as better model assumptions) to address them.

In answer to the question posed at the beginning of this document; **Yes, given the best data and approach, we really might expect to do much better than 10% segmentation errors.** This in turn may make it possble to understand subtle changes in the heterogeneity of brain tumours. However, ensuring that the data conforms to the data density model (as required), does not allow any room for compromise in MR proocols. Special tools will also be neccesary to ensure model to data conformity (quality control) during analysis. Initially, we believe these quality control methods will not be automatic and will require expert operation.

References

[1] Segmentation and structural analysis. *http://fsl.fmrib.ox.ac.uk/fslcourse*, 2014.

[2] J. Ashburner and K.J. Friston. Image segmentation. In R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, K.J. Friston, C.J. Price, S. Zeki, J. Ashburner, and W.D. Penny, editors, *Human Brain Function*. Academic Press, 2nd edition, 2003.

[3] P. A. Bromiley and N. A. Thacker. Multi-dimensional Medical Image Segmentation with Partial Volume and Gradient Modelling. *Annals of the BMVA*, 2008(2):1–23, 2008.

[4] K. A. Clark, R. P. Woods, D. A. Rottenburg, A. W. Toga, and J. C. Mazziotta. Impact of Aquisition Protocols and Processing Streams on Tissue Segmentation of T1 Weighted MR Images. *Neuroimage*, 29:185–202, 2006.

[5] A.H. Herlihy G.A. Coutts I.R. Young D.J.Larkman, J.V. Hajnal and G. Ehnholm. Use of multicoil arrays for separation of signal from multipl-slices simultaneously excited. *Journal of Magnetic Resonance Imaging*, 13:313–317, 2001.

[6] K. Kazemi and N. Noorizadeh. Quantitative Comparison of SPM, FSL, and Brainsuite for Brain MR Image Segmentation. *J. Biomed Phys Eng*, 4(1):13–26, 2014.

[7] S. V. Notley and N. A. Thacker. Improving the stability of mri parameter estimates using regional sub-sampling. *Tina-Memo*, (2015-001), 2015.

[8] et. al. R.M. Haralick. On the use of error propagation for statistical validation of computer vision software. *IEEE Pattern Analysis and Machine Intelligence*, 27(10):1603–1614, 2005.

[9] T. Rohlfing. Image similarity and tissue overlaps as surrogates for image registration accuracy: Widely used but unreliable. *IEEE Trans Med Imaging*, 31(2):153–163, Feb 2012.

[10] N. A. Thacker, P. A. Bromiley, and D. C. Williamson. Multi-dimensional Medical Image Segmentation with Partial Volume and Gradient Modelling. *Tina-Memo*, (2004-009), 2006.

[11] O. Tsang, A. Gholipaur, K. Gopinath, R. Briggs, and I. Panahi. Comparison of Tissue Segmentation Algorithms in Neuroimage Analysis Software Tools. In *30th Annual International IEEE EMBS Conference*, pages 3924–3928, 2008.

[12] R. M. Haralick V. Ramesh. Random perturbation models and performance characterization in computer vision. *In Proceedings of Conference on Computer Vision and Pattern Recognition*, CVPR92:521527, 1992.

[13] M. W. Woolrich, S. Jbadi, B. Patenaude, M. Chappell, S. Makni, T. Behrens, C. Beckmann, M. Jenkinson, and S. M. Smith. Bayesian Analysis of Neuroimaging Data in FSL. *Neuroimage*, 45:173–186, 2009.