Hippocampal segmentation with MAGeT

Jon Pipitone

December 5, 2012

Abstract

Neuroimaging research often relies on automated anatomical segmentations of MR images of the brain. Current multi-atlas based approaches provide accurate segmentations of brain images by propagating manually derived segmentations of specific neuroanatomical structures to unlabelled data. These approaches often rely on a large number of such manually segmented atlases that take significant time and expertise to produce. We present an algorithm for the automatic segmentation of the hippocampus that minimizes the number of atlases needed while still achieving similar accuracy to multi-atlas approaches.

TODO Finish this...

1 Introduction

The hippocampus is of particular interest to many researchers because it is implicated in forms of brain dysfunction such as Alzheimer's disease and schizophrenia, and has functional significance in cognitive processes such as learning and memory. For many research questions involving magnetic resonance imaging (MRI) data accurate identification of the hippocampus and its subregions is a necessary first step to better understand the individual neuroanatomy of subjects.

Currently, the gold standard for neuroanatomical segmentation is manual delineation by an expert human rater. This is problematic for segmentation of the hippocampus for several reasons. First, manual segmentation takes a significant investment of time and expertise [**?**] which may not be readily available to researchers or clinicians. Second, the amount of data produced in neuroimaging experiments increasingly exceeds the capacity for identification of specific neuroanatomical structures by an expert manual rater. Third, the true definition of hippocampal anatomy in MR images is disputed [**?**], as evidenced by efforts to create an unified segmentation protocol [**?**].

Compounding each of these problems is the significant neuroanatomical variability in the hippocampus throughout the course of aging, development, and neuropsychiatric disorders [**?**]. Additionally, it may be necessary to use several different hippocampal definitions or, in fact, make specific modifications in the course of research. For example, Poppenk et al. [**?**] found that subdividing the hippocampus into anterior and posterior regions resulted in a predictive relationship between volume difference of those regions and recollection memory performance. Thus, while manual segmentation of the hippocampus is a necessary technique, to researchers or clinicians who do not have access to the needed human expertise its use may be infeasible.

Automated segmentation techniques overcome the need for human expertise by performing segmentations computationally. A popular class of automated methods, *multi-atlas-based segmentation*, rely on a set of expertly labeled neuroanatomical atlases. Each atlas is warped to fit a subject's neuroanatomy using nonlinear registration techniques [**? ?**]. Atlas labels are then transformed by this warping and a *label fusion* technique, such as voxel-wise voting, is used to merge the competing label definitions from each atlas into a final segmentation for a subject.

Many descriptions of multi-atlas-based segmentation algorithms report relying on an atlas library containing anywhere between 30 and 80 expertly labeled brains [**? ? ? ? ?**]. As noted, the production of an atlas library requires significant manual effort, and is limited since the choice of atlases or segmentation protocol may not reflect the underlying neuroanatomical variability of the population under study or be suited to answer the research questions at hand.

In this paper we propose an automated segmentation method to address the above issues of existing multi-atlas-based methods. Principally, our method aims to dramatically reduce the number of manually labelled atlases needed (under 10). This is achieved by using the small atlas library to boot-strap a much larger "template library", which is then used to segment the subjects in a similar fashion to basic multi-atlas segmentation. This approach has the additional advantage of using the unique subject population on hand to initialize the segmentation process and improve accuracy.

The essential insight of generating a template library is not new. Heckemann [**?**] compared generating a template library from a single atlas to standard multi-atlas segmentation and found poor performance and so deemed the approach as inviable. The LEAP algorithm [**?**] proceeds by iteratively segmenting the unlabelled image most similar to the atlas library images and then incorporating the now-labelled image into the atlas library, but requires 30 starting atlases. The novelty of our method is to demonstrate the possibility of producing comparable segmentation accuracy to these and other multi-atlas-based methods while using significantly fewer manually created atlases.

In our previous work [**?**], we applied MAGeT brain to segmentation of the human striatum, globus pallidus, and thalamus using a single histologically-derived atlas. The main contribution of this paper is to extend our approach to the human hippocampus and perform a thorough validation over a range of atlas and template library sizes, which was not done in our previous work. Due to the small number of atlases required, our method can easily accommodate different hippocampal definitions. Our aim is not to improve on segmentation accuracy beyond existing methods, but instead to provide a method that trades off manual segmentation expertise for computational processing time while providing sufficient accuracy for clinical and research applications.

2 Methods

2.1 MAGeT Brain Algorithm

TODO Make sure we explain what MAGeT stands for

TODO reference Chakravarty2012

In this paper, we use the term *atlas* to mean any manually segmented MR image, and the term *atlas library* to mean a set of such images. We use the term *template* to refer to any MR image, and associated labelling, used to segment another image, and the term *template library* to refer to a set of such images. An atlas library may be used as a template library but, as we will discuss, a template library may also be composed of images with computer generated labellings.

The segmentation approach we propose is best understood as an extension of basic multi-atlas segmentation [**?**]. In multi-atlas segmentation, an atlas library and unlabelled MR images are given as input. Every atlas image is nonlinearly registered to each unlabelled image, and then each atlas' labels are propagated via the resulting transformations. These labels are then fused to produce a single, definitive segmentation by some label fusion method (e.g. voxel-wise majority vote).

Our extension adds a preliminary stage in which a template library is constructed from input images, and used in place of an atlas library in the standard multi-atlas-based method. To create the template library, labels from each atlas image are propagated to each template library image via the transformation resulting from a non-linear registration between pair of images. As a result, each template library image has a label from each atlas. Basic multi-atlas segmentation is then used to produce segmentations for the entire set of unlabelled images (including those images used in the template library).

Source code can be found at `http://github.com/pipitone/MAGeTbrain`.

TODO Note that algorithm does not specify a registration algorithm or fusion method[**?**]. Performance on these is test in experiments

Algorithm 1 Pseudocode for the MAGeT Brain algorithm

```
function MultiAtlas(Templates, Subjects)
    for all subject do
        for all template do
            propagate all labels for template to subject space
            store subject labels
        end for
        fuse subject labels
    end for
end function

function MAGeTBrain(Subjects, Atlases, n)
    for i = 1 → n do
        choose a subject to be used as a template
        propagate labels from each atlas to template space
        store the template with all of its labels
    end for
    MultiAtlas(Templates, Subjects)
end function
```

Table 1: ADNI-1 1.5T Screening demographics

	N	CN $N = 414$	EMCI $N = 279$	LMCI $N = 560$	AD $N = 278$	Combined $N = 1531$
Age at baseline Years	1531	70.9 74.0 78.4	66.0 71.0 76.2	69.4 74.3 79.4	70.9 75.9 80.9	69.6 74.0 79.2
Sex : Female	1531	50% (206)	44% (123)	38% (215)	45% (125)	44% (669)
Education	1531	14.0 16.0 18.0	14.0 16.0 18.0	14.0 16.0 18.0	12.0 16.0 17.8	14.0 16.0 18.0
Ethnicity : Unknown	1531	1% (4)	1% (2)	1% (4)	1% (3)	1% (13)
Not Hisp/Latino		96% (398)	95% (265)	96% (540)	96% (267)	96% (1470)
Hisp/Latino		3% (12)	4% (12)	3% (16)	3% (8)	3% (48)
CDR-SB	1531	0.0 0.0 0.0	0.5 1.0 1.5	1.0 1.5 2.0	3.5 4.5 5.0	0.0 1.0 2.5
ADAS 13	1511	6.0 9.0 12.0	8.0 12.0 16.0	14.3 18.7 23.0	24.0 29.0 34.0	10.0 15.3 23.0
MMSE	1531	29.0 29.0 30.0	28.0 29.0 30.0	26.0 27.0 29.0	21.2 23.0 25.0	26.0 28.0 29.0

a b c represent the lower quartile a, the median b, and the upper quartile c for continuous variables.
N is the number of non–missing values.
Numbers after percents are frequencies.

2.2 Subjects

2.2.1 ADNI-1 1.5T Screening

TODO - image characteristics

TODO - describe SNT manual segmentations

TODO - baseline/cross-validation dataset demographics?

TODO Check table only includes 1.5T from the Screening

2.2.2 SZ First Episode Patients

TODO image characteristics, demographics

2.3 Image pre-processing

Before images were registered, the N3 algorithm [16] is first used to minimize the intensity nonuniformity in each of the atlases and unlabeled subject images.

2.4 Registration

2.4.1 Automatic Normalization and Image Matching and Anatomical Labeling (ANIMAL)

The ANIMAL algorithm carries out Image registration is in two phases. In the first, a 12-parameter linear transformation (3 translations, rotations, scales, shears) is estimated between images using an algorithm that

Table 2: Schizophrenia First Episode Patient Demographics

	N	FEP $N = 81$
Age	80	21 23 26
Gender : M	81	63% (51)
Handedness : ambi	81	6% (5)
left		5% (4)
right		89% (72)
Education	81	11 13 15
SES : lower	81	31% (25)
middle		54% (44)
upper		15% (12)
FSIQ	79	88 102 109

a b c represent the lower quartile *a*, the median *b*, and the upper quartile *c* for continuous variables.
N is the number of non–missing values.
Numbers after percents are frequencies.

maximizes the correlation between blurred MR intensities and gradient magnitude over the whole brain [**?**]. In the second phase, nonlinear registration is completed using the ANIMAL algorithm [**?**]: an iterative procedure that estimates a 3D deformation field between two MR images. At first, large deformations are estimated using blurred version of the input data. These larger deformations are then input to subsequent steps where the fit is refined by estimating smaller deformations on data blurred with a Gaussian kernel with a smaller FWHM. The final transformation is a set of local translations defined on a bed of equally spaced nodes that were estimated through the optimization of the correlation coefficient.

For the purposes of this work we used the regularization parameters optimized in Robbins et al. [**?**].

TODO link to MNI website here

TODO command line

2.4.2 Automatic Normalization Tools (ANTS)

ANTs is a diffeomorphic registration algorithm which provides great flexibility over the choice of transformation model, objective function, and the consistency of the final transformation. The transformation is estimated in a hierarchical fashion where the MRI data is subsampled, allowing large deformations to be estimated and successively refined at later hierarchical stages (where the data is subsampled to a finer grid). The deformation field and the objective function are regularized with a Gaussian kernel at each level of the hierarchy. The ANTs algorithm is freely available `http://www.picsl.upenn.edu/ANTS/`. We used an implementation of the ANTS algorithm compatible with the MINC data format, mincANTS `https://github.com/vfonov/mincANTS`.

We used the following command line when running the ANTS command,

```
mincANTS 3 -m PR[target_file.mnc,source_file.mnc,1,4]
 --number-of-affine-iterations 10000x10000x10000x10000x10000
 --affine-gradient-descent-option 0.5x0.95x1.e-4x1.e-4
 --use-Histogram-Matching --MI-option 32x16000
 -r Gauss[3,0] -t SyN[0.5] -i 100x100x100x20
 -o transformation.xfm
```

These settings were adapted from the "reasonable starting point" given in the ANTS manual.

TODO reference manual?

2.5 Label Fusion

Label fusion is term given to the process of combining the information from several candidate labellings for an MR image into a single labelling. In this paper we explore the benefits of three different fusion methods.

2.5.1 Voxel-wise Majority Vote

Labels are propagated from all template library images to a subject. Each output voxel is given the most frequent label at that voxel location amongst all candidate labellings. Ties are broken arbitrarily.

2.5.2 Cross-correlation Weighted Majority Vote

An optimal combination of subjects from the template library has previously been shown to improve segmentation accuracy [**?** **?**]. In this method, each template library image is ranked in similarity to each unlabelled image by the normalized cross-correlation (CC) of image intensities after linear registration, over a region of interest (ROI) generously encompassing the hippocampus. Only the top ranked template library image labels are used in a voxel-wise majority vote. The ROI is heuristically defined as the extent of all atlas labels after linear registration to the template, dilated by three voxels [**?**]. The number of top ranked template library image labels is a configurable parameter.

The `xcorr_vol utility` from the ANIMAL toolkit is used to calculate the cross-correlation similarity measure.

2.5.3 Normalised Mutual Information Weighted Majority Vote

This method is similar to cross-correlation weighted voting except that image similarity is calculated by the normalised mutual information score over the region of interest.

TODO reference?

The `itk_similarity` utility from the EZMinc toolkit is used to calculate the normalised mutual information measure.

2.6 Goodness-of-fit

Each segmentation was evaluated against the 'gold-standard' manual segmentation from the dataset using the Dice Kappa (κ) overlap metric:

$$\kappa = \frac{2a}{2a + b + c}$$

where a is the number of voxels common to the candidate segmentation and the gold standard and $b + c$ is the sum of the voxels uniquely identified by either the automatically generated candidate segmentation or the gold-standard.

TODO I would do this after you define the experiments

2.7 Experiments

Experiments were performed to assess the performance of MAGeT brain with various parameter settings as well as on diverse datasets. In each experiment we contrast the performance of MAGeT brain with that standard single- and multi-atlas segmentations derived from the same atlas library.

2.7.1 ADNI-1 cross-validation

To test the accuracy of the MAGeT brain algorithm with different parameter settings, repeated random sub-sampling cross-validation (RRSCV) was performed on a subset of the ADNI-1 dataset.

Dataset evaluated. 69 1.5T images were randomly selected from the *ADNI1:Screening 1.5T* standardized dataset. Demographics for this subset are shown in Table X

TODO FIXME: table reference

Table 3: ADNI-1 cross-validation subset demographics

	CN $N = 23$	LMCI $N = 23$	AD $N = 23$	Combined $N = 69$
Age at baseline Years	72.2 75.5 78.5	71.0 77.1 81.4	71.7 77.8 81.8	71.5 76.6 81.3
Sex : Female	43% (10)	43% (10)	43% (10)	43% (30)
Education	16.0 16.0 18.0	15.0 16.0 18.0	12.0 16.0 16.5	14.0 16.0 18.0
Ethnicity : Unknown	0% (0)	0% (0)	0% (0)	0% (0)
Not Hisp/Latino	100% (23)	100% (23)	100% (23)	100% (69)
Hisp/Latino	0% (0)	0% (0)	0% (0)	0% (0)
CDR-SB	0.00 0.00 0.00	0.75 1.50 1.50	4.00 4.50 5.00	0.00 1.50 4.00
ADAS 13	4.67 5.67 12.34	14.34 16.00 20.50	23.83 29.00 31.66	10.00 16.00 25.33
MMSE	28.5 29.0 30.0	25.0 27.0 28.0	21.0 23.0 24.0	24.0 27.0 29.0

a b c represent the lower quartile a, the median b, and the upper quartile c for continuous variables.
Numbers after percents are frequencies.

.

Atlas and template libary. Atlases consisted of images taken from the dataset, with corresponding manual labels provided by SNT. Atlas library size was varied from 3 to 9 images. The remaining images were segmented, with the template library size varying from 3 to 20 images. Template library images were selected randomly from the images to be segmented.

Registration method. Both the ANTS and ANIMAL registration methods were used.

TODO In ANIMAL, images aligned to TAL space

Label fusion. Majority vote, Cross-correlation weighted majority vote, and Normalized Mutual Information weighted majority vote are used. With the weighted majority vote fusion methods, the number of top labels used in the fusion was varied from 3 to 20 images.

Evaluation. Repeated random sub-sampling cross-validation (RRSCV) consists of repeated trials in which items from the dataset are randomly assigned to a training set or validation set. In each trial, performance on the validation set is measured, and then averaged across all trials.

We performed RRSCV on each combination of parameters listed above: atlas library size, template library size, registration method, and label fusion method. We performed 10 trials per parameter combination. In each validation trial, the training set consisted of the images used as atlases, and the validation set consisted of the images to be segmented. The MAGeT brain algorithm and the basic multi-atlas segmentation procedure were applied to segment the images in the validation set. Additionally, in each trial, the single-atlas segmentation was obtained for each atlas-template.

The gold-standard for the segmentation accuracy of images in the validation set was the SNT manual labels.

TODO number of registrations/comparisions

2.7.2 ADNI-1 Screening Validation

To test the accuracy of MAGeT brain on a real-world task we segment the entire ADNI-1 dataset using an atlas set that is not representative of the subject set.

Dataset evaluated. All images from the *ADNI1:Screening 1.5T* standardized dataset.

Atlas and template libary. The atlas library consisted of the entire Winterburn atlas set. The Winterburn atlases are digital segmentations of the hippocampus in five in-vivo 300u isotropic T1-weighted MR scans, and include subfield segmentations for the cornus ammonis (CA) 1, CA4, dentate gyrus, subiculum, and CA 2 and 3 combined. Subjects in the Winterburn atlases range in age from 29-57 years (mean age of 37), and include two males and three females.

The template library consisted of 21 randomly selected images from the ADNI1 data dataset (7 healthy, MCI and AD subjects).

Registration method. - ANTS as it performed best in the cross-validation experiment.

TODO how to reference results whilst still in Methods?

Label fusion. Majority vote, as it is simpliest to run and performed equally well in cross-validation experiment.

Evaluation.

Since hippocampal segmentation protocols differ between the ADNI labels and Winterburn atlases, this poses a problem for direct similarity comparisons between labels produced by MAGeT brain and the ADNI labels.

TODO explain why we did(n't) resegment the ADNI images with a the low-res protocol.

To evaluate the performance of MAGeT brain, we correlate our segmentation volumes with manual segmentation volumes, as well as with hippocampal volumes of established automated segmentation methods.

Additionally, we compared classification accuracy of subjects by diagnosis based on hippocampal volume using both the SMT labels and our produced labels.

TODO description of validation – rms `validate` or t-test: contrast with QDA or LDA (Coupe 2011) used in LOOCV

2.7.3 SZ First Episode Patient Validation

To validate that MAGeT's performance generalises to other diseases. Measure performance using best parameter settings on a different disease dataset.

Dataset evaluated. SZ First Episode Patients

Atlas and template libary. two different atlas sets: a manual hippocampal segmentation of patients, and Winterburn atlas set.

Registration method. ANTS.

Label fusion. Majority vote.

Evaluation. We validate the FEP-atlas segmentations using Dice's Kappa, and the Winterburn-atlas segmentations by correlating volumes.

2.7.4 Winterburn Atlases Validation

Dataset evaluated. T1 BRAVO scans of the same subjects included in the Winterburn atlas set. These scans are taken within .... weeks of the scans for the Winterburn atlases.

Atlas and template libary. Atlas library is Winterburn T1 atlases. Template library consists of all five T1 BRAVOs, plus 15 T1 healthy control images.

Registration method. ANTS

Label fusion. Majority vote.

Evaluation. Leave one out cross-validation (LOOCV) in which all five subjects are segmented in separate runs of MAGeT brain. In each run, the subject to be segmented is excluded from the Atlas library (so only four atlases are used).

Segmentation accuracy is judged by difference in hippocampal volume.

2.8 Results

2.8.1 ADNI-1 Cross-Validation

- find significant improvement over multi-atlas performed with the same parameters. Also, find smoothed performance is monotonically increasing but asymptotic in size of both template and atlas library, with peak performance reached after 15 templates.

TODO Can we statistically capture "peak" performance? The point at which gains become statistically insignificant?

merge	Atlases	ANTS	ANIMAL
1	3.00	0.81	0.76
2	4.00	0.79	0.75
3	5.00	0.82	0.79
4	6.00	0.82	0.78
5	7.00	0.83	0.80
6	8.00	0.83	0.79
7	9.00	0.84	0.80

Number of Atlases	ANIMAL Kappa (Jaccard)	ANTS Kappa (Jaccard)
3	0.76 (0.63)	0.80 (0.69)
4	0.75 (0.62)	0.79 (0.67)
5	0.79 (0.66)	0.82 (0.71)
6	0.78 (0.65)	0.82 (0.70)
7	0.80 (0.67)	0.83 (0.72)
8	0.79 (0.67)	0.83 (0.72)
9	0.80 (0.68)	0.84 (0.73)

Multi-atlas means:

- more atlases -¿ better performance - larger template library -¿ better performance, but tails off around 10-15 templates - no significant difference between majority or weighted vote methods (haven???t tested this statistically though). - consistently performs better than average naive performance by XXX - using ANTS, with a large enough template library (¿12) MAGeT brain performs better than the average multi-atlas approach with the same number of atlases. using ANIMAL, 5 or more atlases needed before boost seen. - more atlases -¿ smaller template library required to improve on average multi-atlas performance - discuss variance? best/worst case? –how often do we expect random template library selection to work decently

TODO cost (in registrations) / benefit trade off graph: show number of registrations per Kappa? or hours of manual labour per Kappa?)

2.8.2 Winterburn Atlas Segmentation of ADNI Baseline Images

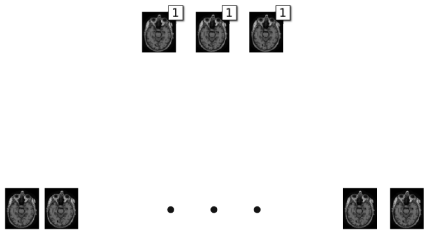
- A2A shows that if atlas population strongly(?) represents subject set variability, then free choice from atlas population will produce improvements (we know this b/c of extensive validation trials). - what about in the case where atlas population doesn???t strongly represent subject set variability (e.g. a priori atlas set)? then, we can use atlas selection to refine atlas set?

TODO Kappa against our manual rater is low

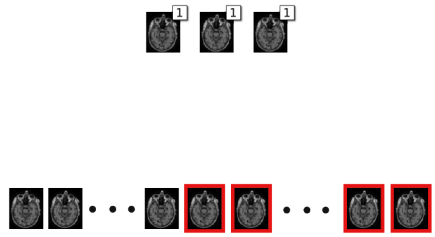
2.8.3 First Episode Schizophrenic patients

High volume correlation between Winterburn segmentation volumes and ground truth. (High-ish?) Kappa when using manual segmentations as Atlases.

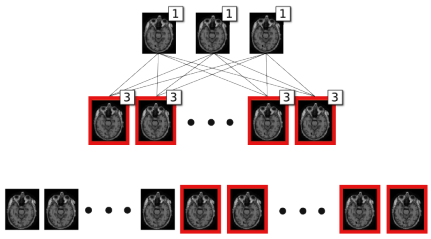
Some other text here...



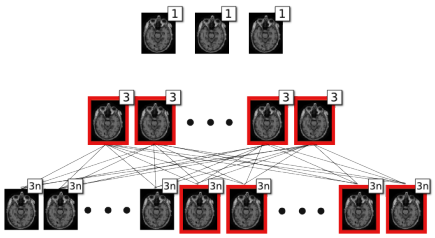
Top row: atlases (with indicators showing that each atlas image has one set of labels).
Bottom row: subject images.



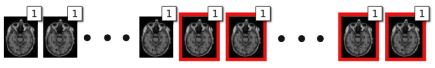
STAGE 1: n template images (in red) are selected arbitrarily from the subjects.



STAGE 3: Labels from the atlases are propagated to the template library.



STAGE 4: Labels from the template library are propagated to the subject.
This is much like standard Multi Atlas propagation, except that each
template has multiple labelings (one from each atlas).



STAGE 5: Subject labels are fused to give a single label for each subject.

Figure 1: Diagram of the MAGeT Brain algorithm

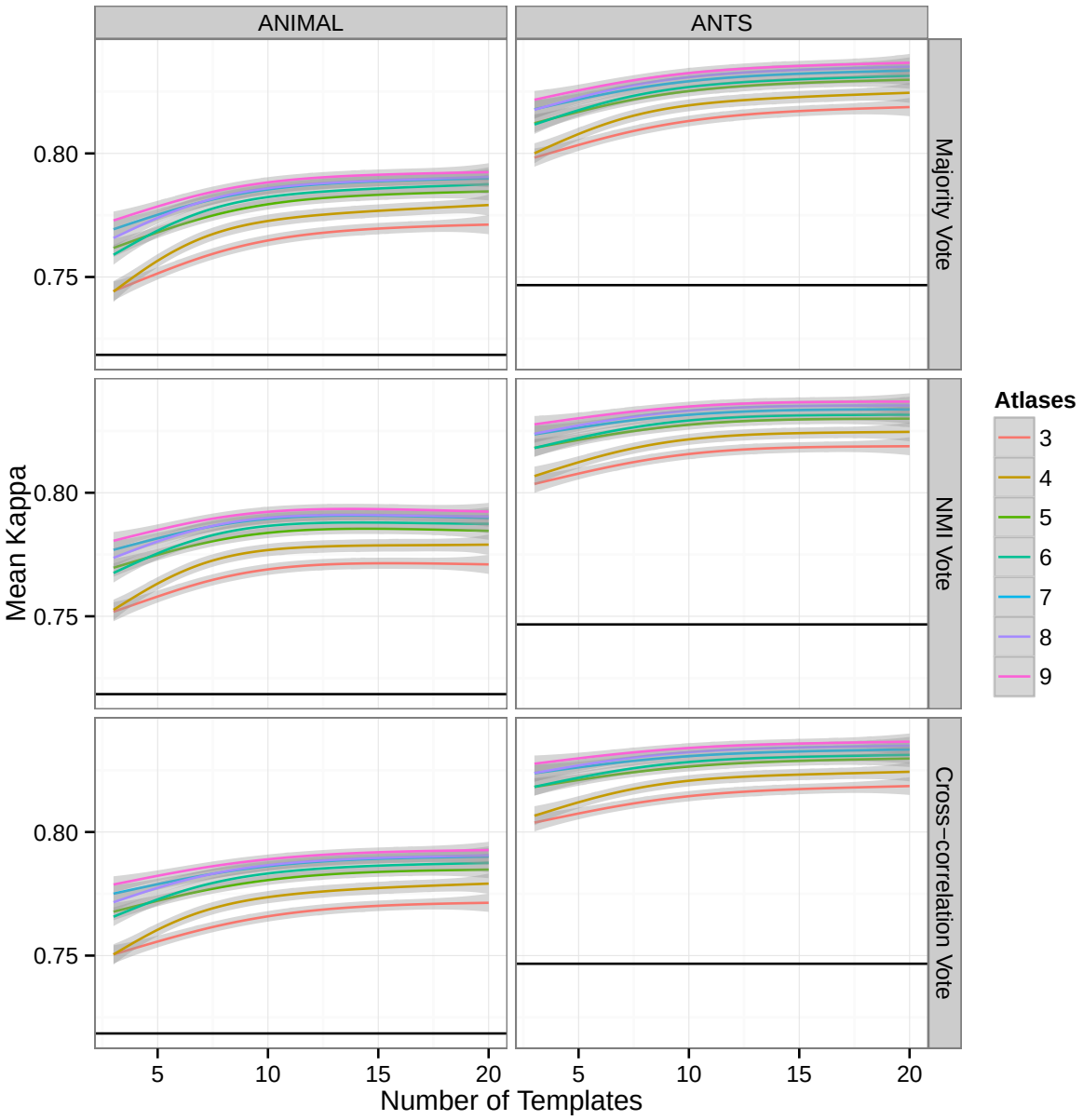


Figure 2: Comparison of MAGeT performance on ADNI-1 subset. Smoothing line fitted using GAM (generalised additive model) from R with defaults from ggplot2 (formula: y s(x, bs = "cs"))

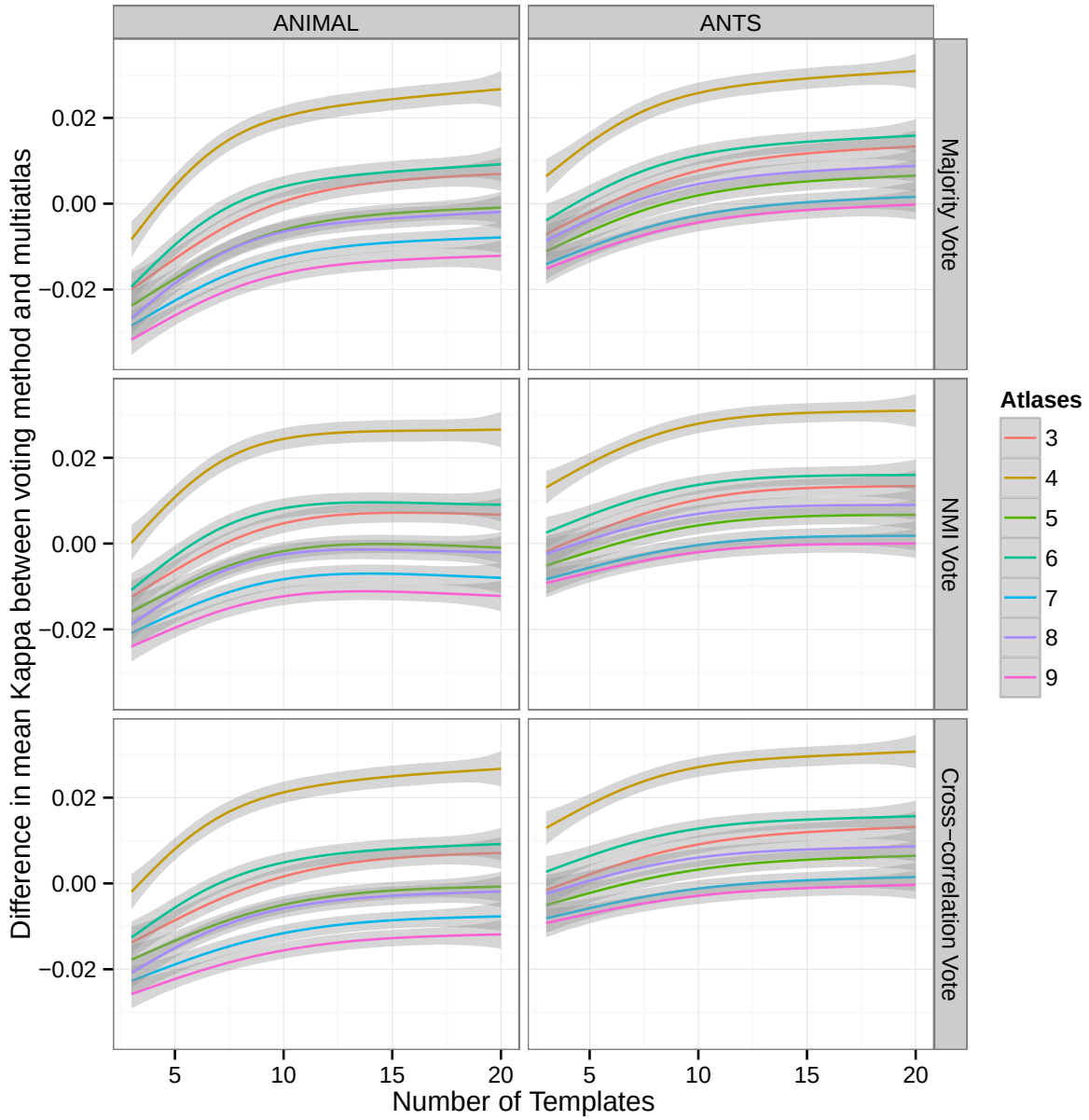


Figure 3: Difference in mean Kappa between MAGeT brain and multi-atlas

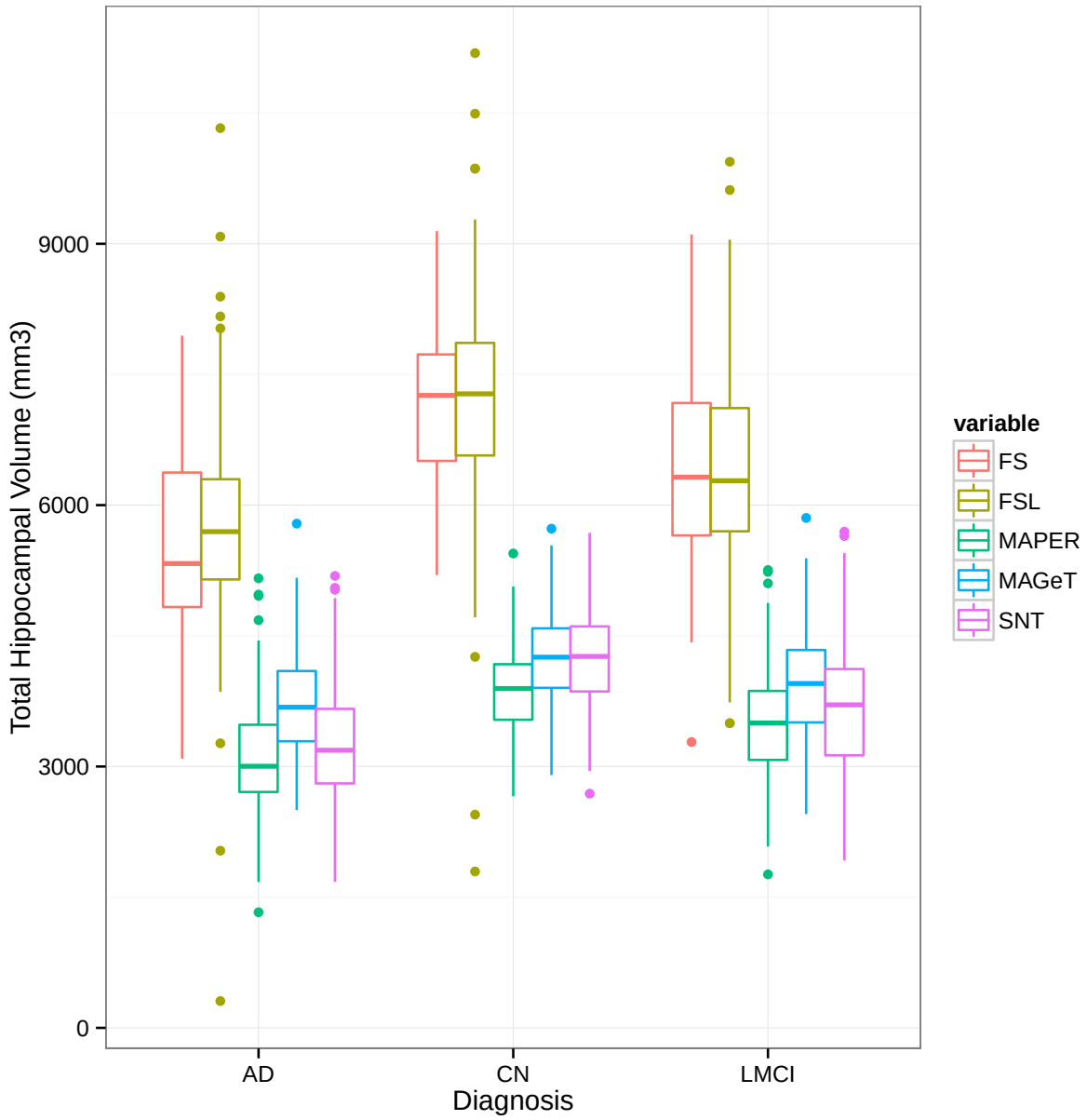


Figure 4: Comparison of HC volumes by FreeSurfer (FSF), MAGeT brain (MAGeT), MAPER, and manual (SNT).

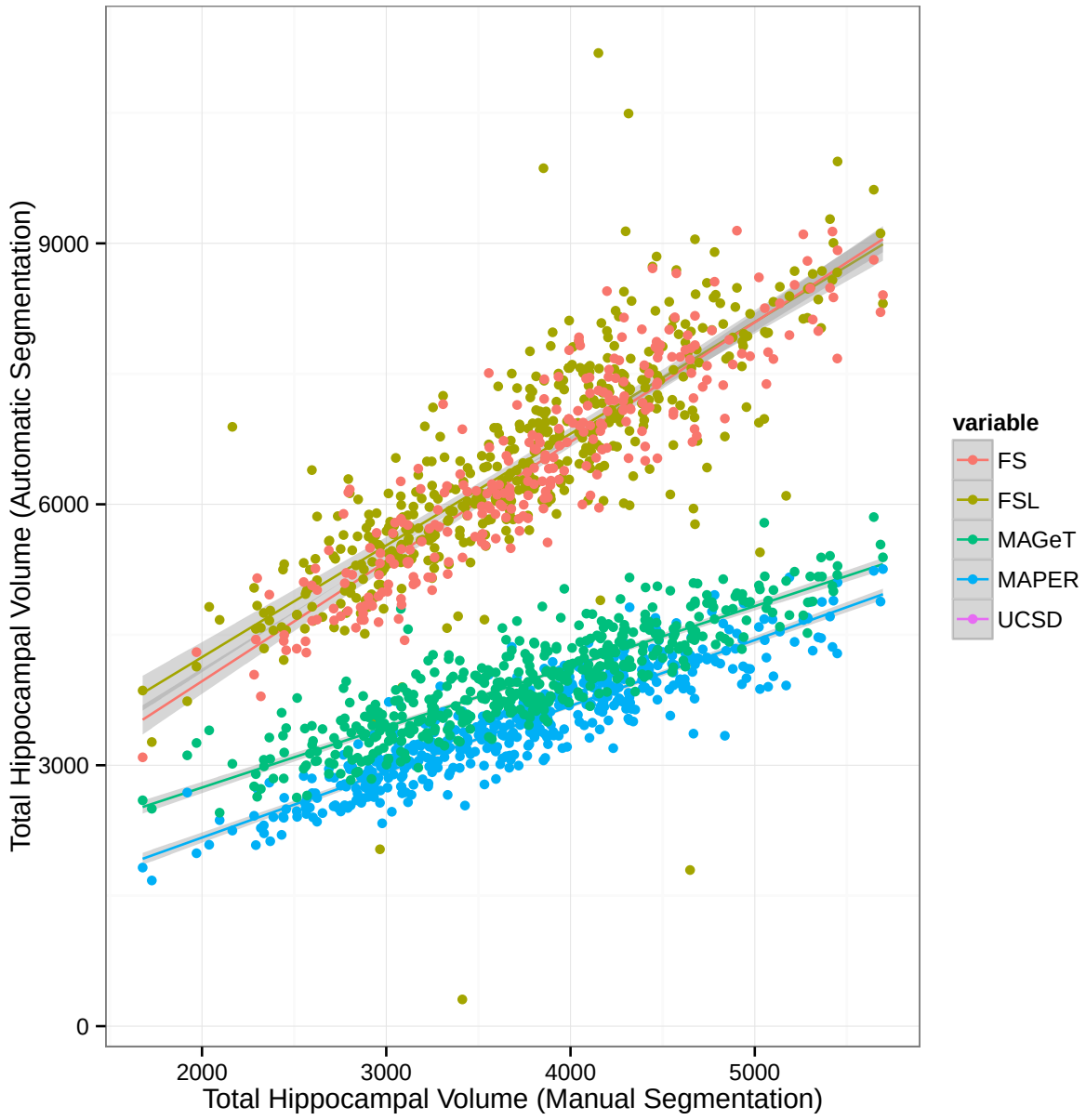


Figure 5: **ADNI Baseline cohort.** Comparison of HC volumes by FreeSurfer (FSF), MAGeT brain (MAGeT), MAPER, and manual (SNT).

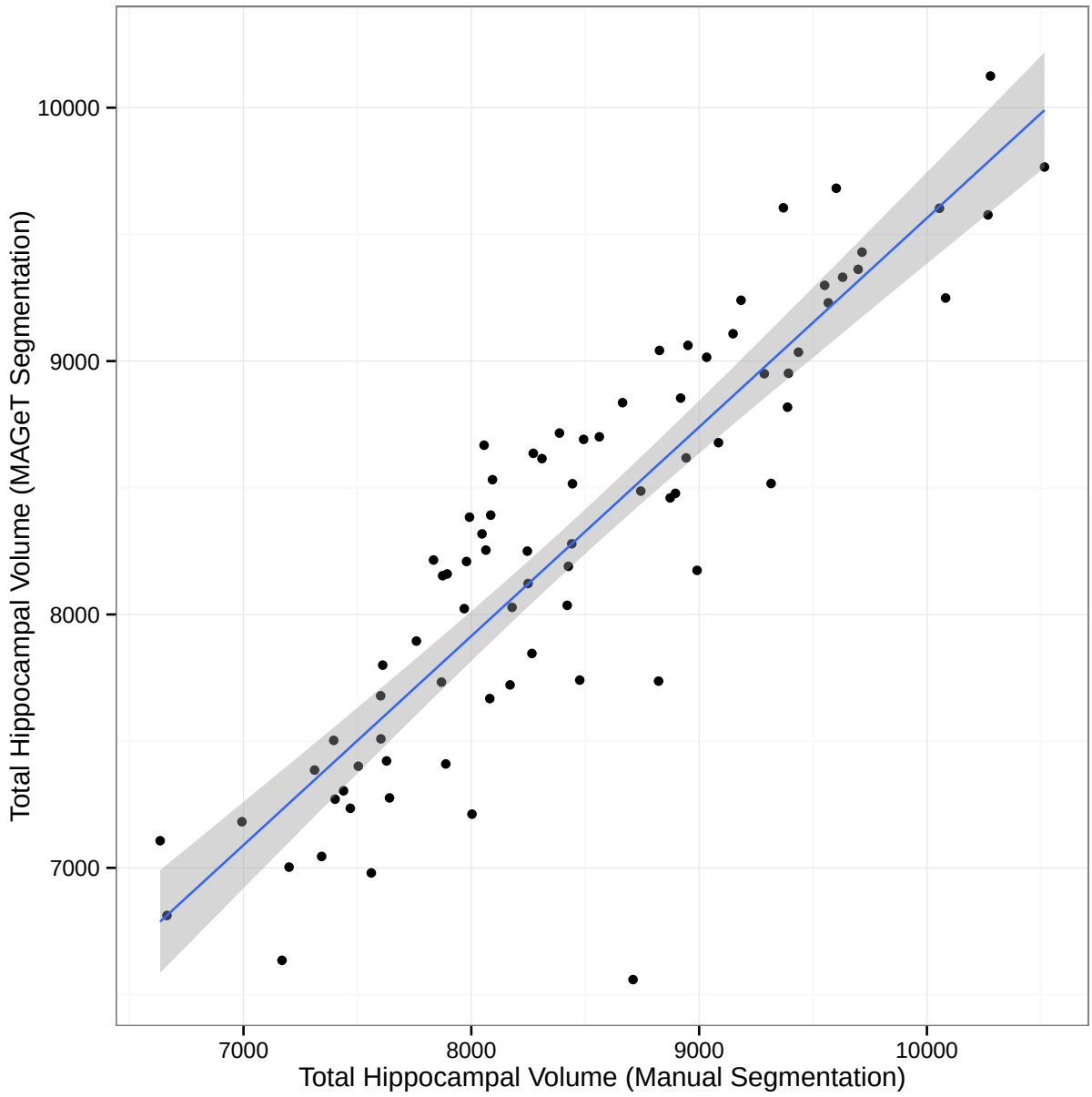


Figure 6: **First Episode Schizophrenic Patients.** Comparison of total HC volumes for MAGeT against manually rated volumes of