Error Assessment in Serial Morphometry: Application to Multiple Sclerosis

Mehul P. Sampat Dominik S. Meier Brian Healy Annika Berger
Charles R.G. Guttmann

Center for Neurological Imaging
Brigham and Women's Hospital
Boston, MA

6th September, 2008

Outline

Outline

Background

- The measurement of brain atrophy has become an established method of estimating MS disease severity and progression.
- The atrophy is quantified by MRI morphometric measurements such as brain parenchymal fraction (*BPF*).
- Atrophy is a slow process and thus requires high level of measurement precision/sensitivity.

Background

- The reliability of *BPF* and other morphometric measurements is frequently measured with scan-rescan experiments.
- In scan-rescan experiments two MRI's of the patient are obtained within a short interval (e.g., 30 minutes).
- Between the two imaging sessions patients may be asked to exit and re-enter the scanner area
- In the absence of true biological/structural change, any differences in the measurements from the two scans is attributed as measurement precision.

Sources of error

Principal sources of error are:

1. Patient re-positioning error
2. Instrument variability (e.g., magnet and/or receiver coil)
3. Algorithm specific errors (e.g., stochastic algorithms or nonlinear algorithm transfer function)

Outline

Problem Statement

Questions:

- Is scan-rescan error a reliable indicator of real-life error?
- Can we analyze reliability of the scan-rescan error ?
- For a given image segmentation algorithm, how much change in BPF indicates “real change” ?
- Various techniques exist for error assessment: which one has greatest practical utility ?

Problem Statement

Questions:

- Is scan-rescan error a reliable indicator of real-life error?
- Can we analyze reliability of the scan-rescan error ?
- For a given image segmentation algorithm, how much change in BPF indicates “real change” ?
- Various techniques exist for error assessment: which one has greatest practical utility ?

Problem Statement

Questions:

- Is scan-rescan error a reliable indicator of real-life error?
- Can we analyze reliability of the scan-rescan error ?
- For a given image segmentation algorithm, how much change in BPF indicates “real change” ?
- Various techniques exist for error assessment: which one has greatest practical utility ?

Problem Statement

Questions:

- Is scan-rescan error a reliable indicator of real-life error?
- Can we analyze reliability of the scan-rescan error ?
- For a given image segmentation algorithm, how much change in BPF indicates “real change” ?
- Various techniques exist for error assessment: which one has greatest practical utility ?

Outline

Datasets Used

- 20 MS patients
- 5 MRI per patient (baseline MRI plus 4 weekly MRIs each) .
- Axial dual-echo PDw/T2w protocol on a 1.5T
- Acquisition parameters
  - TE 30/80 ms, TR = 3000 ms
  - resolution: 0.93-by-0.93-by-3 mm
- None of the patients had received any form of therapy or steroids.

Image analysis

- Template-driven segmentation algorithm (TDS) used for image segmentation
- Segmentated classes: White matter (*WM*), Gray matter (*GM*), Cerebro-spinal fluid (*CSF*).
- Brain Parenchymal Fraction (BPF) was computed as:

$$BPF = \frac{WM + GM}{WM + GM + CSF} = 1 - \frac{CSF}{ICC}$$

Outline

Error assessment techniques

We compare the following three methods to analyze measurement error:

1. Standard error of measurement
2. Voxel-wise classification consistency
3. Intra-class correlation coefficient: a measure of agreement

Standard error of measurement (*SEM*)

$$SEM = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\sigma_i^2}, \text{ where } \sigma_i = \sqrt{\frac{1}{M-1}\sum_{j=1}^{M}(x_{i,j}-\mu_i)^2}$$

M : no. of observers, N : no. of subjects

$x_{i,j}$: measurement made by observer j for subject i

Given the *SEM*, the threshold value (δ), that represents 95% of change beyond chance is:

$$\delta = \sqrt{2} * 1.96 * SEM$$

δ is the threshold's which signify "real change" (with 95% certainty).

Voxel-wise classification consistency

Standard error of measurement (*SEM*)

$$SEM = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\sigma_i^2}, \text{ where } \sigma_i = \sqrt{\frac{1}{M-1}\sum_{j=1}^{M}(x_{i,j}-\mu_i)^2}$$

M : no. of observers, N : no. of subjects

$x_{i,j}$: measurement made by observer j for subject i

Given the *SEM*, the threshold value (δ), that represents 95% of change beyond chance is:

$$\delta = \sqrt{2} * 1.96 * SEM$$

δ is the threshold's which signify "real change" (with 95% certainty).

Voxel-wise classification consistency

Standard error of measurement (*SEM*)

$$SEM = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\sigma_i^2}, \text{ where } \sigma_i = \sqrt{\frac{1}{M-1}\sum_{j=1}^{M}(x_{i,j}-\mu_i)^2}$$

M : no. of observers, N : no. of subjects

$x_{i,j}$: measurement made by observer j for subject i

Given the *SEM*, the threshold value (δ), that represents 95% of change beyond chance is:

$$\delta = \sqrt{2} * 1.96 * SEM$$

δ is the threshold's which signify "real change" (with 95% certainty).

Voxel-wise classification consistency

Intra-class correlation coefficient

$$ICCoeff = \frac{\text{variability in the BPF measurements of patients}}{\text{Total variability (from all sources)}}$$

$$ICCoeff = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n} \cdot (MS_C - MS_E)}$$

n denotes the number of images, MS_R is the mean square error between images, MS_E is the residual mean square error, MS_C is the mean square error between observers.

Table: Interpretation of ICCoeff

ICCoeff value	Strength of Agreement
ICCoeff < 0.4	Indicates poor reproducibility
$0.4 <$ ICCoeff < 0.75	Indicates fair to good reproducibility
ICCoeff > 0.75	Indicates excellent reproducibility

Intra-class correlation coefficient

$$ICCoeff = \frac{\text{variability in the BPF measurements of patients}}{\text{Total variability (from all sources)}}$$

$$ICCoeff = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E + \frac{k}{n} \cdot (MS_C - MS_E)}$$

n denotes the number of images, MS_R is the mean square error between images, MS_E is the residual mean square error, MS_C is the mean square error between observers.

Table: Interpretation of ICCoeff

ICCoeff value	**Strength of Agreement**
ICCoeff < 0.4	Indicates poor reproducibility
$0.4 <$ ICCoeff < 0.75	Indicates fair to good reproducibility
ICCoeff > 0.75	Indicates excellent reproducibility

Outline

Pair-wise comparisons of MR scans

Table: *SEM* and δ values for *BPF*, *WM* and *GM* for each of the four pair-wise comparisons. bMR denotes the baseline MRI.

		Pair of time-points compared			
		bMR & week 1	bMR & week 2	bMR & week 3	bMR & week 4
BPF	*SEM*	0.0040	0.0048	0.0057	0.0055
	δ	0.0111	0.0134	0.0157	0.0154
WM	*SEM*	0.0046	0.0050	0.0053	0.0054
	δ	0.0128	0.0138	0.0146	0.0149
GM	*SEM*	0.0045	0.0055	0.0042	0.0044
	δ	0.0126	0.0151	0.0116	0.0122

Note for scan-rescan experiment: $SEM = 0.0020$, $\delta = 0.0056$

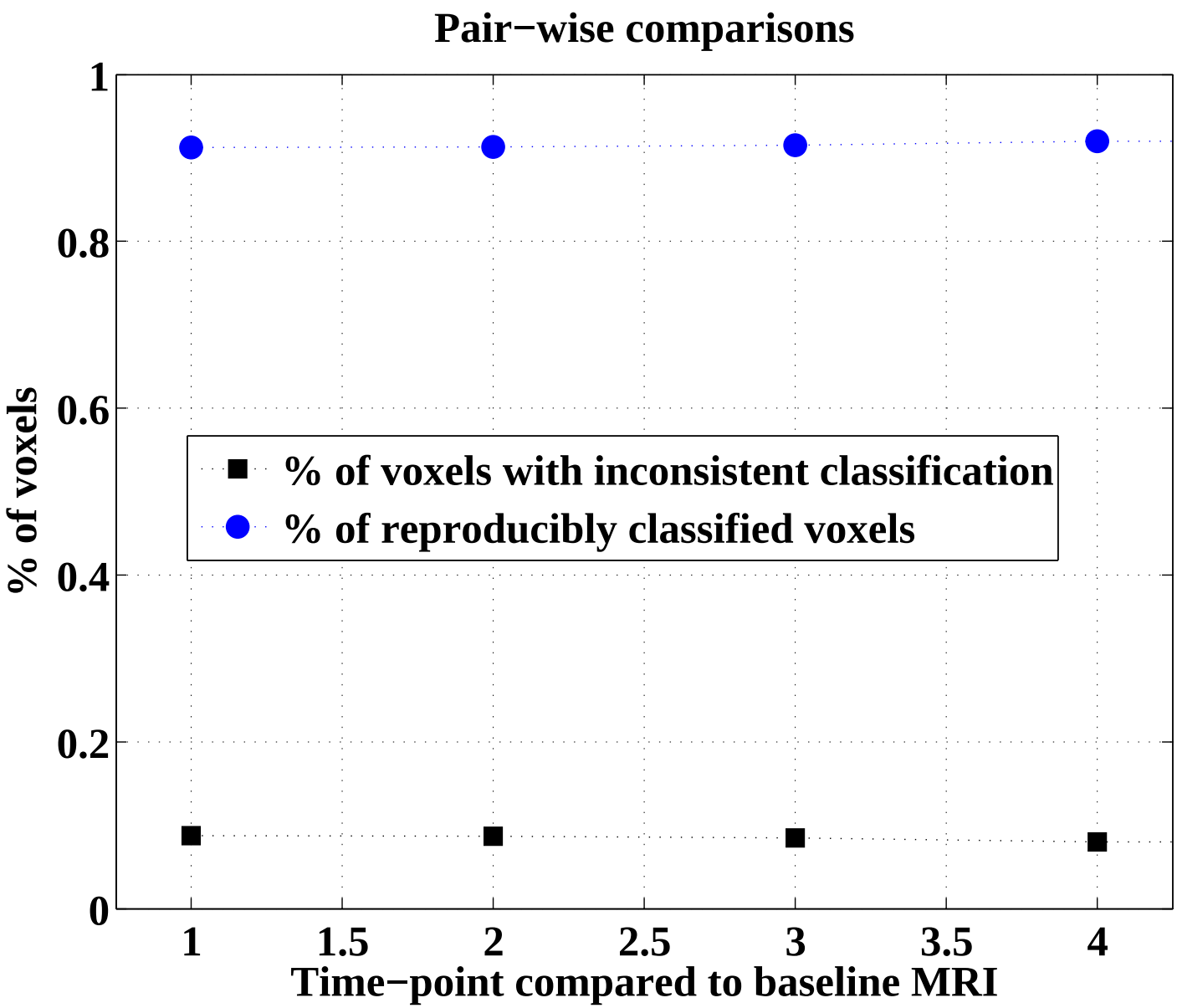
Comparison across multiple MR scans

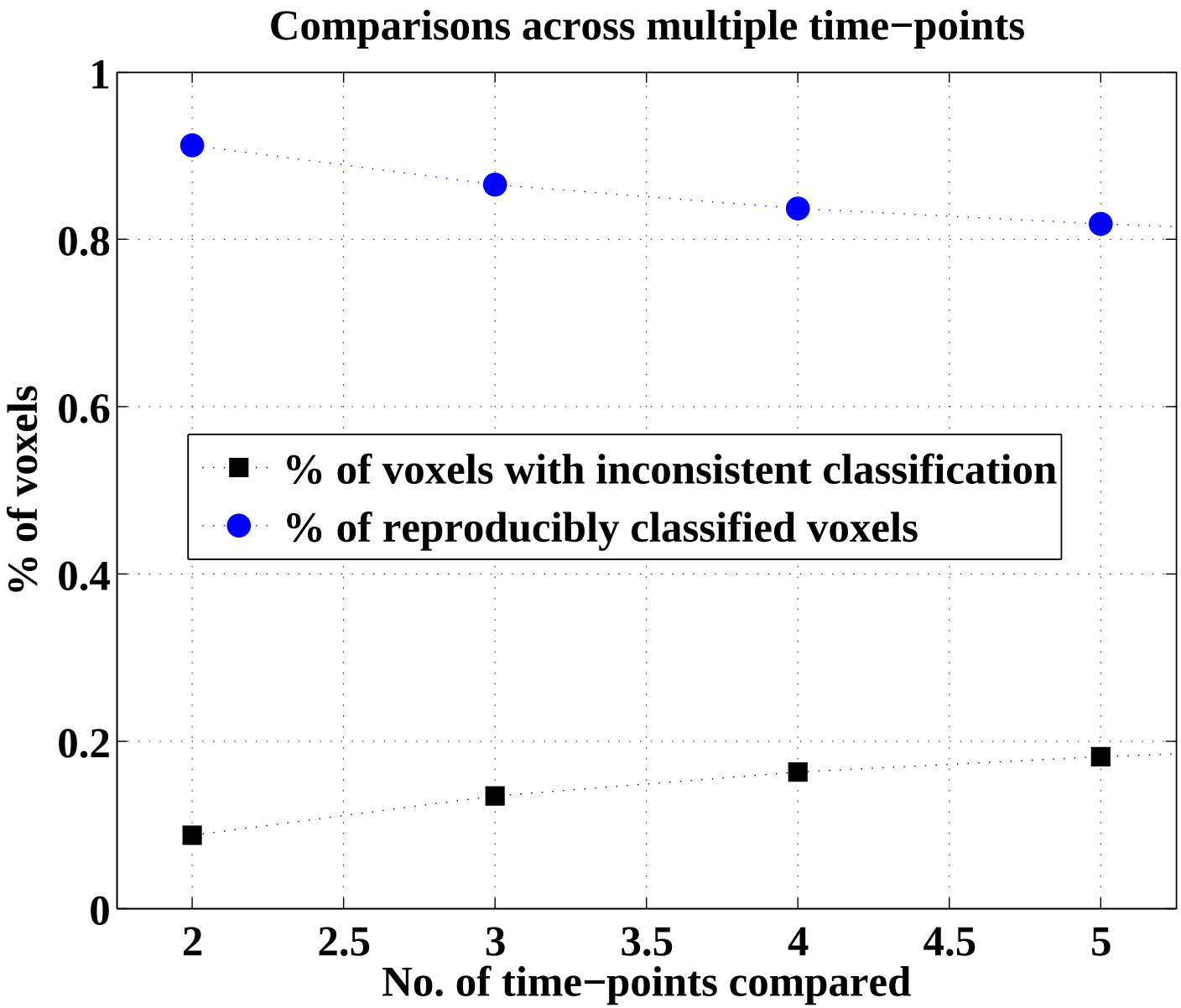
Table: *SEM* and δ values for *BPF*, *WM* and *GM* variables obtained by comparing the measurements of these variables from multiple MR scans.

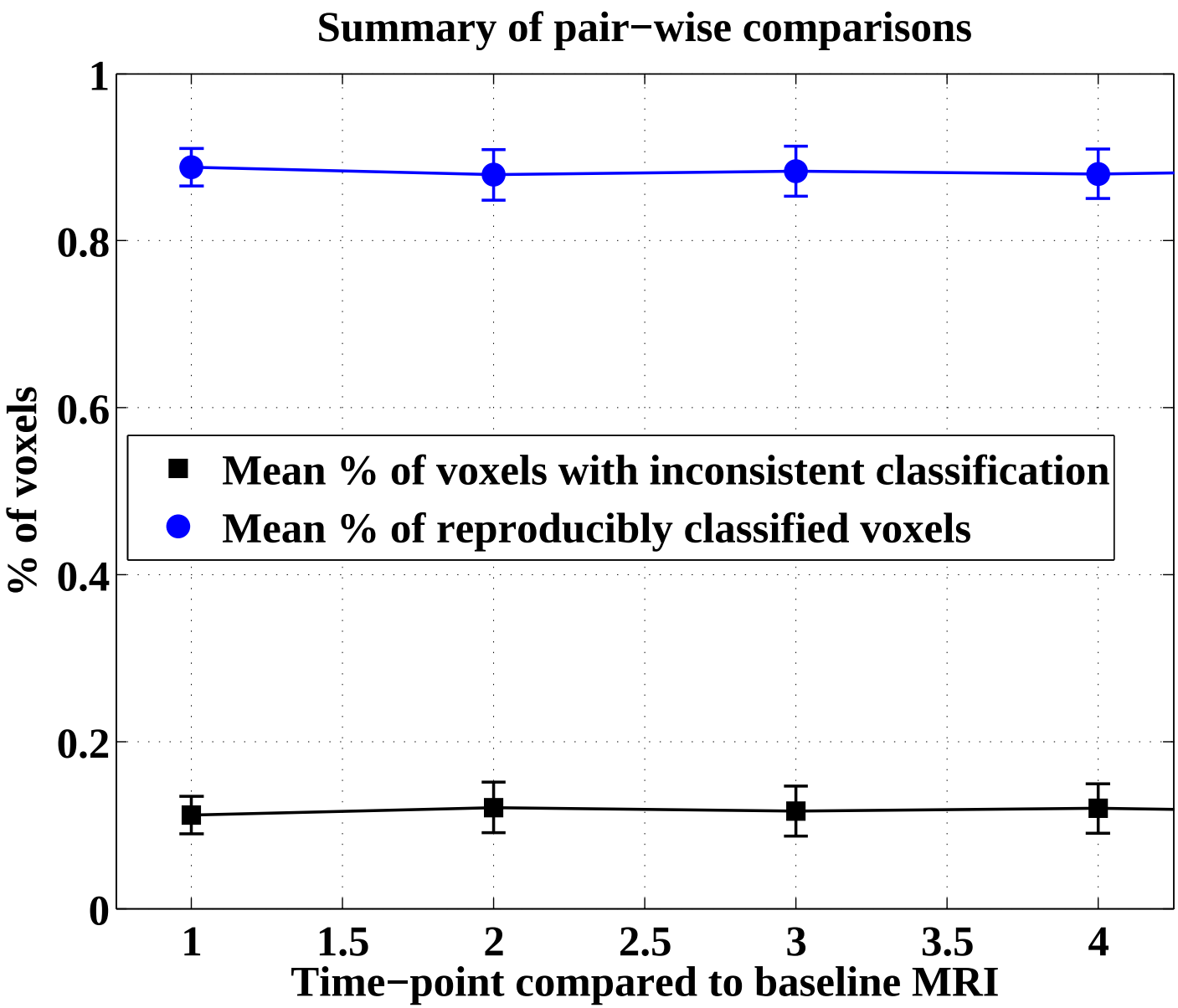
		No. of MRIs compared			
		2	3	4	5
BPF	*SEM*	0.0040	0.0045	0.0047	0.0048
	δ	0.0111	0.0125	0.0130	0.0133
WM	*SEM*	0.0046	0.0048	0.0046	0.0047
	δ	0.0128	0.0133	0.0127	0.0130
GM	*SEM*	0.0045	0.0049	0.0043	0.0042
	δ	0.0126	0.0135	0.0120	0.0115

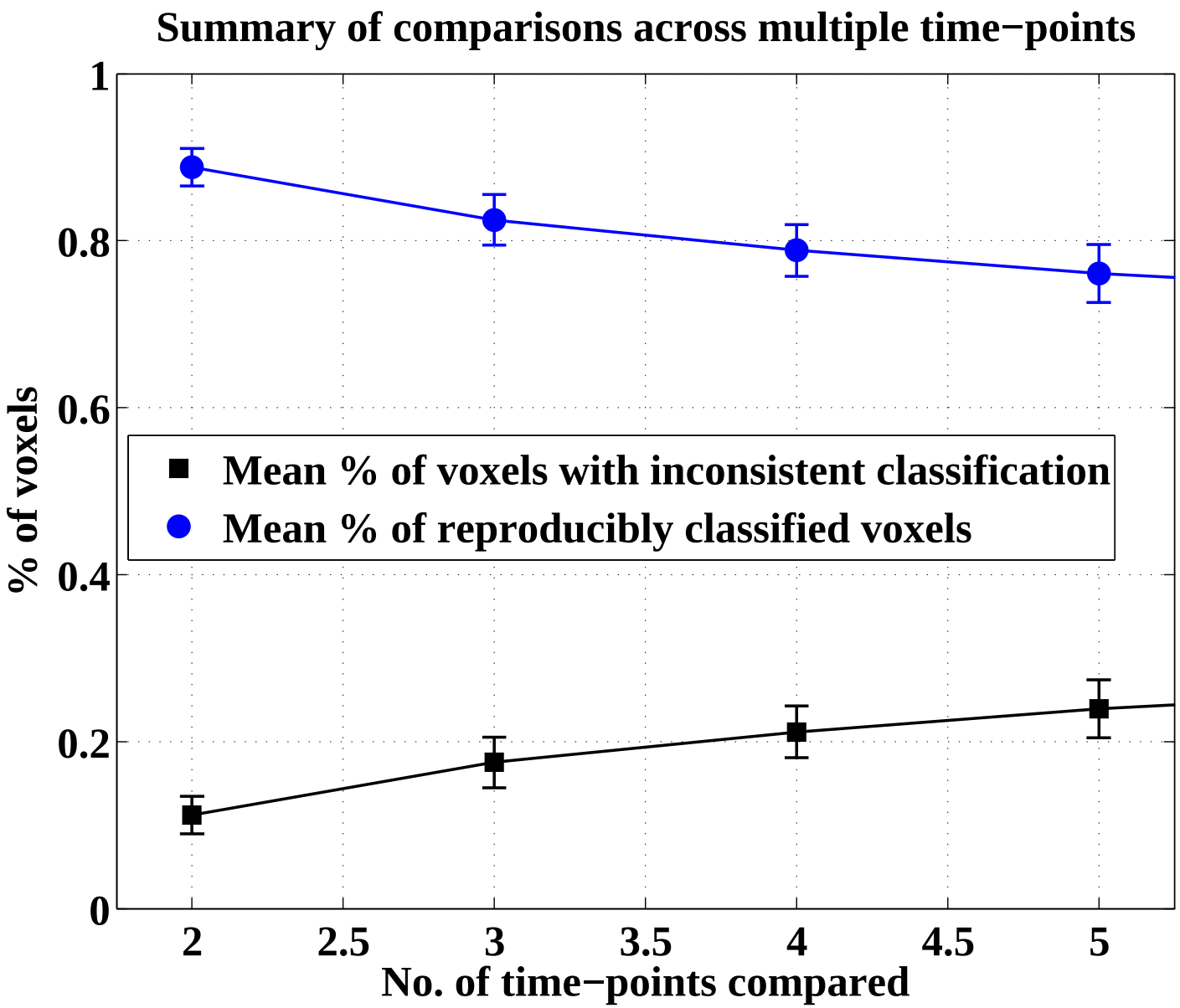
Note for scan-rescan experiment: $SEM = 0.0020$, $\delta = 0.0056$

Outline


Pair−wise comparisons
% of voxels
1
0.8
0.6
0.4
0.2
0
% of voxels with inconsistent classification
% of reproducibly classified voxels
1
1.5
2
2.5
3
3.5
4
Time−point compared to baseline MRI


Comparisons across multiple time−points
% of voxels
No. of time−points compared
% of voxels with inconsistent classification
% of reproducibly classified voxels
0
0.2
0.4
0.6
0.8
1
2
2.5
3
3.5
4
4.5
5


Summary of pair−wise comparisons
% of voxels
Time−point compared to baseline MRI
Mean % of voxels with inconsistent classification
Mean % of reproducibly classified voxels
0
0.2
0.4
0.6
0.8
1
1
1.5
2
2.5
3
3.5
4


Summary of comparisons across multiple time−points
% of voxels
No. of time−points compared
Mean % of voxels with inconsistent classification
Mean % of reproducibly classified voxels
0
0.2
0.4
0.6
0.8
1
2
2.5
3
3.5
4
4.5
5

Outline

Results

Table: *ICCoeff* for *BPF*, *WM* and *GM*, for each of the four pair-wise comparisons. bMR denotes the baseline MRI.

	Pair of time-points compared			
	bMR & week 1	bMR & week 2	bMR & week 3	bMR & week 4
BPF	0.9929	0.9898	0.9851	0.9860
WM	0.9933	0.9926	0.9916	0.9912
GM	0.9373	0.9195	0.9501	0.9448

Outline

Scan-rescan experiments under-estimate measurement error:

- Measurement error in scan-rescan
- Previous work from our group (Wei, 2004) has shown that
  - $SEM = 0.0020$, $\delta = 0.0056$ (Scan-rescan)
  - $SEM = 0.0040$, $\delta = 0.0111$ (two MRI taken a week apart)
  - the *SEM* from weekly measurements is 2 times greater than that from scan-rescan .
- Utility of SEM
  - By comparing various methods to quantify measurement variability we find that the *SEM* demonstrates highest practical utility.
  - SEM is simple to compute and allows one to compute threshold's which signify real biological change (with 95% certainty).

Discussion

Table: *ICCoeff* for *BPF* for each of the four pair-wise comparisons. bMR denotes the baseline MRI.

	Pair of time-points compared			
	bMR & week 1	bMR & week 2	bMR & week 3	bMR & week 4
BPF	0.9929	0.9898	0.9851	0.9860

- In our example, the scan-rescan *ICCoeff* was 0.998,
- Recall, an *ICCcoeff* > 0.75 denotes excellent agreement.
- Using only the *ICCcoeff* implies that there is no difference in the measurement error computed from a scan-rescan setting versus that quantified from weekly MRI.
- We have two methods which produce very contrasting results.

Outline

Future work

- Obtain or collect a weekly MRI data-set of healthy individuals
- Test the performance of multiple algorithms
- Limitations: Loss of spatial information.

Table: *SEM* and δ values for *BPF*, *WM* and *GM* for each of the four pair-wise comparisons. bMR denotes the baseline MRI.

		Pair of time-points compared			
		bMR & week 1	bMR & week 2	bMR & week 3	bMR & week 4
BPF	*SEM*	0.0040	0.0048	0.0057	0.0055
	δ	0.0111	0.0134	0.0157	0.0154
WM	*SEM*	0.0046	0.0050	0.0053	0.0054
	δ	0.0128	0.0138	0.0146	0.0149
GM	*SEM*	0.0045	0.0055	0.0042	0.0044
	δ	0.0126	0.0151	0.0116	0.0122

Note for scan-rescan experiment: $SEM = 0.0020$, $\delta = 0.0056$