Data Quality Review
November 2012
Ben Postlethwaite

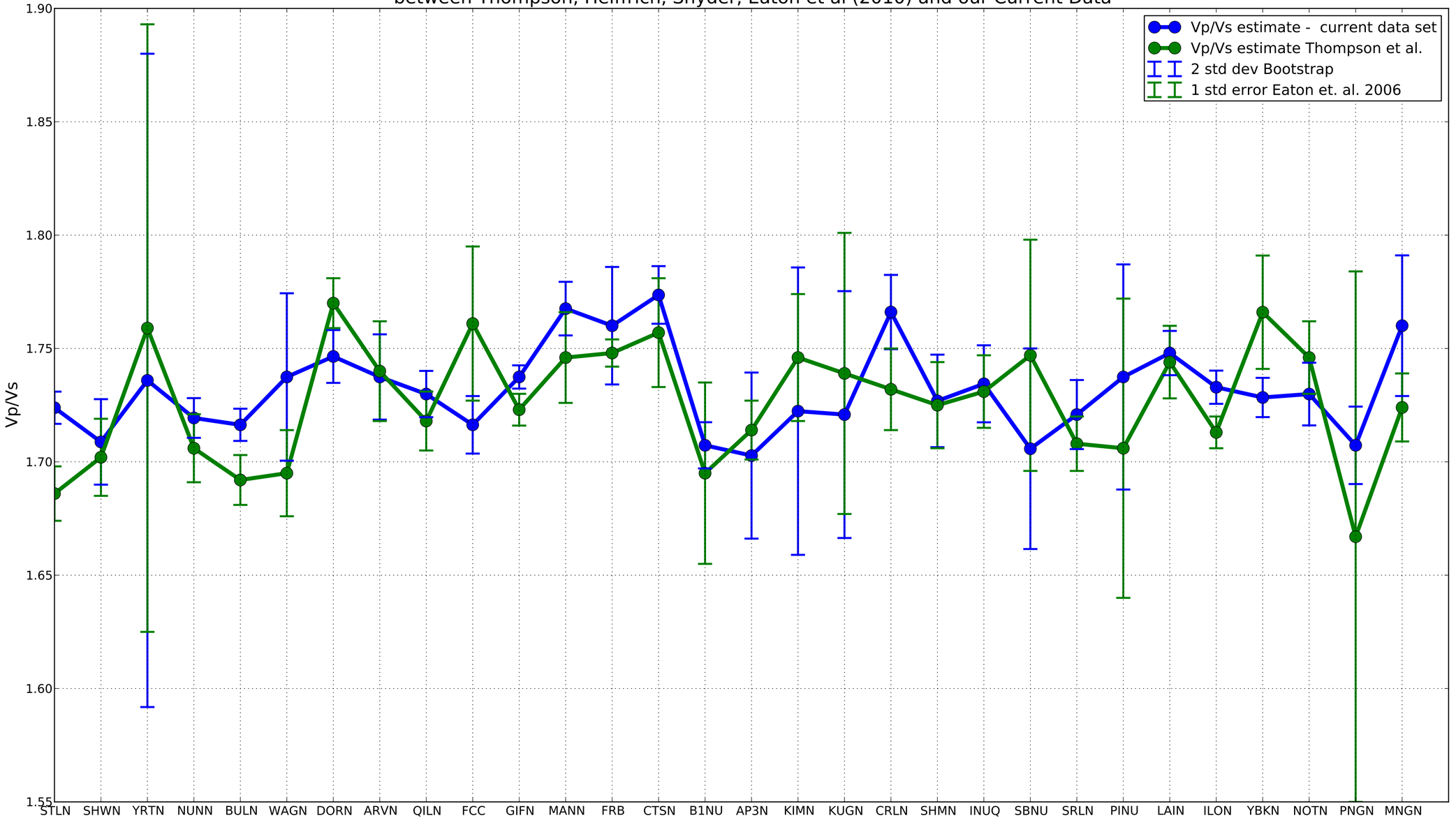
This is a quick summary of a few tests and comparisons performed to gauge the quality of the data I have accumulated and processed thus far. Following this quality review, sometime next week, I will email out the published histograms and maps which I hope to include in my AGU poster. This review has three short sections corresponding to three avenues of inquiry into the data quality I have made over the last few days.

Section 1: Third-party data comparison

I have plotted my data against data taken from paper *Precambrian crustal evolution: Seismic constraints from the Canadian Shield* by Thompson, Bastow, Snyder et al. (2010). Overall there is general agreement between the results with most of the points falling within the given error bounds. Several still show good correlation but are outside the error bounds. I had originally assigned a 1 sigma error estimate but it was immediately evident that this was too low so I resigned it to 2 * standard deviation. This appears to be closer in alignment with the error estimates in the Thompson et al. paper. My calculation uses a different method than that of Thompson et al. who base their method on that of Eaton et al. (2006).

My error estimates are calculated in a bootstrap approach after filtering the receiver functions for quality. The receiver functions are only chosen if the simultaneous deconvolution algorithm finds a parameter beta which minimizes a cross correlation function. If no beta is found within a given range I discard the resulting receiver function. The functions are then normalized and tested for impulsiveness by measuring and comparing peak heights with the worst offenders being discarded. The parameter gridsearch for Vp/Vs and H is performed 1024 times with randomly chosen data - permitting multiples - and 2 * the standard deviation from these runs in calculated.

Some of my earlier published histograms seemed to indicate lower Vp/Vs values than we expected for the Canadian Shield area and the comparison with the *Precambrian crustal evolution* data shows that this is indeed real. The comparison dataset actually has a slightly lower mean than the mean of my data, less than one percent, well within aggregate error.


Comparison for given Can. Shield Stations
between Thompson, Helffrich, Snyder, Eaton et al (2010) and our Current Data
Vp/Vs estimate - current data set
Vp/Vs estimate Thompson et al.
2 std dev Bootstrap
1 std error Eaton et. al. 2006
Vp/Vs
1.90
1.85
1.80
1.75
1.70
1.65
1.60
1.55
STLN SHWN YRTN NUNN BULN WAGN DORN ARVN QILN FCC GIFN MANN FRB CTSN B1NU AP3N KIMN KUGN CRLN SHMN INUQ SBNU SRLN PINU LAIN ILON YBKN NOTN PNGN MNGN

Section 2: Extreme value check

In reviewing some of my data with Michael and Nik, several anomalously low values in both Vp/Vs estimates and Mooney's Vp demanded further attention. In the case of the Vp/Vs estimates two stations have been removed for being too noisy, while one has been judged as clean enough to stand as a deviation. The low Vp estimates in the Mooney database were, in all but one cases, data taken from Continental Shelf and Oceanic Plate environments, which have now been removed from future calculations. All of the values were being parsed and computed correctly from the database. New averages will be computed with this refined data set.

Section 3: Split by Azimuthal Cluster

M. Bostock asked that the data be separated by source region and independently calculated to see how close the source region estimates are to the aggregate. I split each station data into two k-mean clusters centred over Japan and Chile and reprocessed. The results show strong inconsistencies in more data than I expected.

Figuring out what to do with this information is difficult so I have written a logical query against this data to provide a first draft of how I should be using it.

| japanR - R | < 2 * std(R)	**(1)**
| chileR - R | < 2 * std(R)	**(2)**
| japanR - chiliR | < 0.055	**(3)**

Where R = Vp/Vs calculated for the dataset as a whole and japanR and chiliR and the regional estimates, std(R) is my calculated bootstrap error estimate. I am using two standard deviation since I believe my error estimates to be low.

The value 0.055 is my arbitrary cutoff number, Vp/Vs values with a standard deviation higher than this are not included in my calculations, so I have used the same number here for consistency.

Total stations examined: 133
For stations failing **(1)** or **(2)** and **(3)** there is a likely significant lateral heterogeneity or some processing flaw and the station may be marked as unusable.
number of stations: 37

For stations failing **(1)** or **(2)** but passing **(3)** the data may be reasonable but my bootstrap error estimate is much too optimistic. Readjustment of error required.
n*umber of stations: 44*

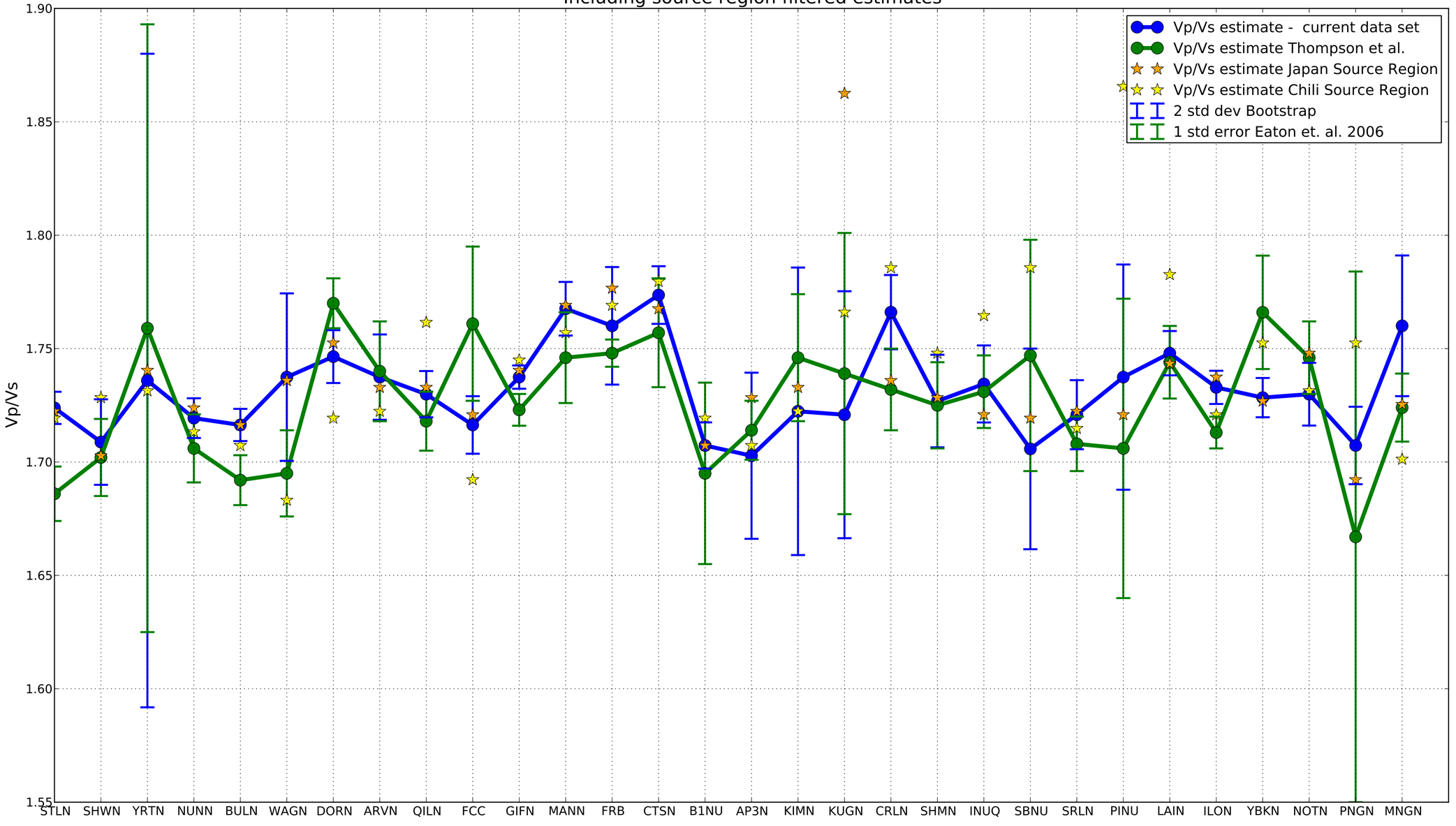
For stations passing **(1)** and **(2)** and **(3),** Great.
number of stations: 50

For stations passing **(1)** and **(2)** but failing **(3)** the regional Vp/Vs difference may be above the arbitrary cut-off of +\- 0.055 but could be included provided the std Vp/Vs error is not above some other arbitrarily chosen cap (Vp/Vs +/- 0.1) and there are visible reflected phases.
number of stations: 2

These numbers raise some questions. Can I report error estimates for the data which is lower than the difference between the Japan source region Vp/Vs estimate and the Chili/Peru source region Vp/Vs estimate? Removing the 37 stations listed above which fail either condition **(1)** or **(2)** and also fail **(3)** is going to damage the representativeness of my data for some Geological Provinces and for coverage of Canada as a whole.

The last figure (shown below) is the same as the first but this time I have included the regional Vp/Vs estimates and increased my error bars to 2 standard deviations. You can see that most show strong correlation with the Vp/Vs estimates I have calculated though some fall outside 2 standard deviations.

I would like to thank you for taking the time to read this review. In the next few days I will publish and email out histograms and maps showing the state of my data so far. Any recommendations, comments and criticism are most welcome. Especially some comment on how to treat my low error estimates and on how to deal with the information provided by the source region separated data test.


Comparison for given Can. Shield Stations
between Thompson, Helffrich, Snyder, Eaton et al (2010) and our Current Data.
Including source region filtered estimates
Vp/Vs estimate - current data set
Vp/Vs estimate Thompson et al.
Vp/Vs estimate Japan Source Region
Vp/Vs estimate Chili Source Region
2 std dev Bootstrap
1 std error Eaton et. al. 2006
Vp/Vs
1.90
1.85
1.80
1.75
1.70
1.65
1.60
1.55
STLN SHWN YRTN NUNN BULN WAGN DORN ARVN QILN FCC GIFN MANN FRB CTSN B1NU AP3N KIMN KUGN CRLN SHMN INUQ SBNU SRLN PINU LAIN ILON YBKN NOTN PNGN MNGN