Large data sets in Petrology: Recommended Best Practices

Prepared for CIDER 2017 by Prof. Adam Kent (Oregon State University) & the students of the "Big Data in Petrology" OSU graduate class.

Application of data science techniques – the use of large amounts of data to obtain insight not available from smaller individual data sets – is transforming research in many scientific fields. Petrology and geochemistry – being inherently data driven anyway – are no different. Compiling and examining multiple smaller datasets into a "big" dataset or *data compendium*, allows us to address many questions beyond those for which the data was originally collected. As a result a slow revolution is taking over the field of igneous petrology and related disciplines as large and very large data sets become increasingly available.

Recommendations/Best Practices

The availability of large datasets is a powerful tool, but care is required for effective use of large data sets to minimize bias and error. The following describes some recommendations for best practices compiled from a recent graduate class "*Big Data in Petrology*" at Oregon State University.

Compiling and Filtering Large Datasets

An overwhelming amount of data is now accessible, almost all online, but not all of it is useful, complete, valid, precise and/or correct. When compiling data, be mindful of the overarching question to be answered before choosing the appropriate data to include. Certain geologic regions, samples of varying ages, or data from older analytical methods may be more or less appropriate for inclusion in a dataset. Not all data is of the same quality, and once a dataset is compiled, filtering of spurious data must be undertaken. Poor data may be the result of analytical quality issues, data entry errors and inadequate or inaccurate metadata (*metadata* is the data that describes the data and provides context and includes things such as sample locations, sample material, analytical methods etc.). Various datasets may require different methods for removing erroneous or outlying data. It is very important to clearly define the procedures undertaken in each circumstance. One common method of removing outliers is to reject any value that lies greater than 1.5 times the interquartile range (IQR) outside of the first or third quartile. Le Chauvenet's criteria, based on a z-score approach, can also be useful. However, in some cases extreme but valid data points may exist within the dataset, and manual treatment of outlying data may be more robust. In all instances, one should clearly define what an outlier is, how it was determined, and why it was disregarded. Testing the sensitivity of different filtering mechanisms can increase the robustness of conclusions drawn from the data. Critically, all methods used for compiling and filtering the dataset should be thoroughly described and should be reproducible by other workers.

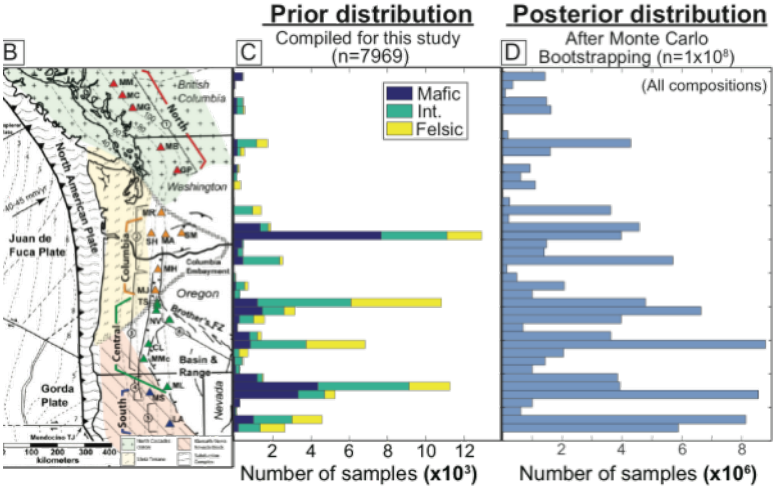

Figure 1. Samples available ("Prior distribution") and bootstrap corrected ("Posterior Distribution") for volcanic rocks from the Cascade Arc.

Minimizing Sampling Bias

In geology/petrology research, certain regions of the Earth are more fully studied than others. This may be the result of the ease of access to the samples (e.g. subaerial samples are more numerous than subaqueous) or the relative excitement or popularity of a certain volcano or location (e.g. Mt. St. Helens is heavily sampled compared to the rest of the Cascade Arc). For example, if one were to calculate the

average composition of the Cascade Arc, it would be heavily skewed towards Mt. St. Helens (Figure 1) The common approach of compiling histograms of number of available analyses can exacerbate these issues if significance is assigned to frequency.

There are multiple strategies that can be employed to reduce the effect of sampling bias, and careful thought should be given to the overall goal of a study when choosing the appropriate. This can include calculating individual averages (beware average averages though!) or assigning equal weight to individual locations before comparing them. Bootstrap resampling (Figure 1), where synthetic data sets are produced by sampling a known distribution with replacement is another effective technique, and can be used to infer different parameters, such as variance or average of a large dataset by taking smaller samples, calculating the mean and variance and repeating the process many times. The step of randomly resampling many times (often >10,000) is referred to as Monte Carlo analysis of bootstrap resampling, and ensures that the parameters calculated from the samples are robust and accurately represent the overall population. Boostrapping can also be done using a weighting scheme such that regions, volcanoes, or eruptive units that have been relatively undersampled have a higher chance of being selected for one of the bootstrap sample sets. In doing so, the data from undersampled locations are being "pulled up by their bootstraps" to the level of sampling completed at other locations. This process creates a "posterior distribution" that is more uniform that the original "prior distribution" (Figure 1).

Data Interpretation and Visualization

Data interpretation must be approached with caution. By nature, large datasets can make it easy to see what we want to see due to the breadth and type of information available (many degrees of freedom). Care should be taken to avoid making overarching interpretations based on single variables. Similarly, avoid characterizing of large portions of the earth with single variable or averages. Multivariate and related approaches can be very powerful with all this data.

Visualization of large datasets also present additional challenges. Traditional techniques such as bivariate plots are inadequate when large amounts of data are plotted. Some general graphing suggestions include the use of histograms or kernel density plots, box and whisker plots, and density contour graphs. Many of these techniques require binning of the data, or grouping of a continuous range of data into small number intervals. In this case the interval chosen must be such that it is not too large or too small to clearly show variations in the data. Statistical criteria exist to estimate correct binning.

You Can Help!

If you contribute data to an online data base (which you should!) or even if you publish data (as most published data is added to online databases anyway) it is critical to provide correct metadata, such as tectonic setting, sample location, latitude and longitude, rock type, material type (whole rock, mineral, glass, etc.), and age (if available). Careful attention during the editing process will safeguard against errors in the data or metadata that could cause confusion or propagation of incorrect. It is the responsibility of authors and the petrologic community as a whole to produce and report high quality data and metadata.

Some Useful References

Gale. A, Dalton. C.A, Langmuir. C.H, Su. Y, Schilling. J-G, 2013, The mean composition of ocean ridge basalts, Geochemitsry, Geophysics, Geosystems, v.14, n.3, p. 489- 517, doi:10.1029/2012GC004334, ISSN: 1525-2027.

Gazel. E, Hayes. J.D, Hoernel. K, Kelemen. P, Everson. E, Holbrook. W.S, Hauff. F, van den Bogaard. P, Vance. E.A, Chu. S, Calvert. A.J, Carr. M.J, Yogodzinski. G.M, 2015, Continental crust generated in oceanic arcs, Nature Geoscience, v. 8, p.321-327, doi: 10.1038/NGEO2392.

Keller, C.B, Schoene. B, 2012, Statistical geochemistry reveals disruption in secular lithospheric evolution about 2.5 Gyr ago, Nature, v.485, p.490-492, doi: 10.1038/nature11024.

Keller, C.B, Schoene. B, Barboni. M, Samperton. K.M, Husson. J.M, 2015, Volcanic-plutonic parity and the differentiation of the continental crust, Nature, v.523, p 301-305, doi:10.1038/nature14584.

Turner. S.J, Langmuir. C.H, 2015, The global chemical systematics of arc front stratovolcanoes: Evaluating the role of crustal processes, Earth and Planetary Science Letters, v. 422, p. 182-193.

Turner. S.J, Langmuir. C.H, 2015, What processes control the chemical compositions of arc front stratovolcanoes? Geochemitsry, Geophysics, Geosystems, p. 1865-1893, doi: 10.1002/2014GC005633