

DZstats v.2.2

J. E. Saylor
K. E. Sundell
Department of Earth and Atmospheric Sciences
University of Houston
Houston TX

TABLE OF CONTENTS

Table of Contents.....	2
1. INTRODUCTION AND DESCRIPTION	4
1.1. Selection of PDPs, KDEs, or LA-KDEs	4
1.2. Module overview.....	4
1.1.1. Two Sample Compare	4
1.1.2. Intersample Compare.....	4
1.1.3. Subsample Compare	4
1.3. Calculation of Probability Density Functions	5
1.4. Kolmogorov-Smirnov test.....	6
1.5. Kuiper test	7
1.6. Similarity.....	8
1.7. Cross-correlation	8
1.8. Likeness.....	8
2. TWO SAMPLE COMPARE.....	9
2.1. Data input format	9
2.2. Import data	9
2.3. Specify the minimum and maximum interval of interest and the step interval of computation.....	9
2.4. Plotting individual samples	10
2.5. Plotting two overlaid PDPs or CDFs.....	10
2.6. Calculating similarity statistics	11
2.7. Exporting statistical data	11
2.8. Exporting graphs	12
3. INTERSAMPLE COMPARE	13
3.1. Data import format.....	13
3.2. Import data	13
3.3. Specify the minimum and maximum interval of interest and the step interval of computation.....	14
3.4. Plotting multiple samples.....	14
3.5. Apply individual statistical tests to all analyses from multiple samples.....	15
3.6. Apply all statistical tests to all analyses from multiple samples.....	15
3.7. Exporting PDPs, KDEs, LA-KDEs or Bandwidths to spreadsheets.....	15
3.8. Exporting graphs	16
3.9. Analysis bandwidths	16

4. SUBSAMPLE COMPARE	17
4.1. Data import format	17
4.2. Import data	17
4.3. Specify the minimum and maximum interval of interest and the step interval of computation.....	17
4.4. Specify the number of ages to be randomly drawn from each sample	18
4.5. Specify the number of trials to conduct	18
4.6. Plot trial 1	18
4.7. Apply individual statistical tests to multiple subsets of analyses from multiple samples	19
4.8. Apply all statistical tests to multiple subsets of analyses from multiple samples.....	20
4.9. Exporting graphs	20
5. REFERENCES	21

1. INTRODUCTION AND DESCRIPTION

DZstats is designed to provide rapid, easily implemented statistical assessment of the similarity of multiple large detrital geochronological or thermochronological data sets. It is organized into three modules; “Two Sample Compare”, “Intersample Compare”, and “Subsample Compare” which are accessed from the DZstats main screen (Fig. 1).

1.1. Selection of PDPs, KDEs, or LA-KDEs

Prior to selecting the module in which to compare geochronological data sets, the user must select whether probability density plots (PDPs), adaptive kernel density estimates (KDEs), or locally adaptive kernel density estimates (LA-KDEs) will be used in the comparison metric. **The selection is made on the DZstats main page (Fig. 1).** The algorithms used in calculating PDPs, KDEs, and LA-KDEs are presented in section 1.3 below.

PDPs: Probability density plots are calculated using an algorithm developed for this software package and incorporate grain age uncertainties from the data input files. The PDP option can also be used to implement a uniform bandwidth KDE (e.g., Vermeesch and Garzanti, 2015) by replacing the grain age uncertainties by the desired bandwidth (in Myr) in the input data files.

KDEs: Adaptive kernel density estimates are calculated using the algorithm of Botev et al. (2010) which optimizes the bandwidth and applies that bandwidth over the entire sample space.

LA-KDEs: Locally adaptive kernel density estimates are calculated using the algorithm of Shimazaki and Shinomoto (2010) which calculates a variable bandwidth which is inversely proportional to the local data density.

1.2. Module overview

1.1.1. Two Sample Compare

The Two Sample Compare module provides rapid assessment of the similarity of two samples. This module is not intended as the primary analysis tool in DZstats but rather is provided for convenience and troubleshooting applications.

1.1.2. Intersample Compare

The Intersample Compare module is designed to provide rapid comparison of similarity of multiple samples. The module calculates the statistical measures of similarity once for each sample pair using all analyses from each sample. It can handle large numbers of analyses ($n > 1 \times 10^6$), multiple data sets, and data sets of different sizes. It also calculates and exports the probability density functions (PDFs) based on finite mixture distribution (probability density plots, PDPs), kernel density estimates (KDEs), or locally adaptive kernel density estimates (LA-KDEs) for external plotting and analysis.

1.1.3. Subsample Compare

The Subsample Compare module randomly selects **n** analyses from each of the data sets in the import data file, where **n** is the user-specified number of analyses. It calculates all of the statistical measures of similarity for each sample pair. It then repeats this process **t** times, where **t** is the user-specified number of trials. This provides **t** estimates of the similarity between each sample pair, from which the module calculates and outputs the mean and standard deviation for each statistical test for each sample pair, subsampled **n** times.

Because n subsamples are drawn from all data sets, n cannot be greater than the number of analyses in the smallest data set.

Figure 1.1. The main menu of DZstats facilitates selection of the method of probability density function computation by selecting the radio buttons for “Probability Density Plot”, “Kernel Density Estimate”, or “Locally Adaptive Kernel Density Estimate” on the top left of the screen. The user can then access the desired module by selecting the associated button on the bottom left of the screen.



1.3. Calculation of Probability Density Functions

The Similarity, Likeness, and Cross-correlation coefficients are based either on a finite mixture distribution of the probability density functions (PDFs) or KDEs of the sample ages. Mixture distributions are a commonly used approach to model a population composed of two or more sub-populations and are used in a variety of disciplines including the physical sciences, medicine, economics, engineering, and the social sciences (Smith and Bartlet, 1961; Behboodan, 1970; Everitt and Hand, 1981; Titterington et al., 1985; Lo et al., 2001; Everitt, 2005; Pearson, 2011; Martin, 2012; Miller, 2014). The mixture distribution ($f(x)$), calculated from n observations is given as

$$f(x) = \sum_{i=1}^n w_i f_i(x), \quad (\text{eq. 1})$$

where w_i , the mixing proportion, is typically $1/n$ and must satisfy the relationship

$$\sum_{i=1}^n w_i = 1. \quad (\text{eq. 2})$$

In these expressions, $f_i(x)$ is the PDF,

$$f_i(x) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{x - \mu_i}{\sigma_i} \right)^2 \right], \quad (\sigma > 0) \quad (\text{eq. 3})$$

where the mean (μ) is the mean grain age and the standard deviation (σ) is based on analytical uncertainty (Fig. 1A). “The final distribution, $f(x)$, is a legitimate probability distribution in its own right,” (Miller, 2014) “called the *mixture density*” (Pearson, 2011). We note that this is broadly the same procedure introduced into the geochronological literature in the 1980s (Jessberger et al., 1980; Hurford et al., 1984; Dodson et al., 1988) and widely adopted to produce PDPs (alternatively termed Probability Density Distributions, Brandon, 1996; Fedo et al., 2003; Sircombe, 2004; Gehrels, 2012). We retain the term probability density plot due to its popularity

within the geochronological literature and a desire to minimize introduction of multiple terms with the same referent.

Kernel density estimation is a non-parametric method of estimating a sample's probability density function. The KDE is defined as

$$\hat{f}_h(x) = \sum_{i=1}^n \frac{1}{nh} K\left(\frac{x - x_i}{h}\right) \quad (\text{eq. 4})$$

where $\hat{f}_h(x)$ is the density estimate, h is the bandwidth (also called window width or smoothing parameter), K is the kernel function, and x_i is the mean grain age (Silverman, 1986). Kernel estimates are a special case of a general mixture density composed of n , typically identical (but see exceptions below) component kernels (Scott, 1992). The kernel function can be any of a number of functions, including for example boxcar, triangular, or, most commonly, Gaussian kernels. Selection of the kernel function is not as critical as selection of the appropriate bandwidth, h . If h is too large, the KDE will be oversmoothed, resulting in loss of resolution. Alternatively, if h is too small, the KDE will be artificially rough and feature too many modes. Silverman (1986) notes that, "if the estimates are smoothed sufficiently to deal with [spurious noise in the tails of the estimates], then essential detail in the main part of the distribution is masked." KDEs also discard variability in uncertainties associated with data acquisition (heteroscedastic uncertainties); a feature that can lead to either over- or undersmoothing of the resultant KDE. Several approaches have been developed to deal with the issues raised by KDEs including bandwidth optimization algorithms, locally adaptive, variable-bandwidth KDEs, and deconvolution techniques to account for heteroscedastic uncertainties (e.g., Delaigle and Meister, 2008; Staudenmayer et al., 2008; Carroll et al., 2009; Botev et al., 2010; Shimazaki and Shinomoto, 2010; McIntyre and Stefanski, 2011). We use two variable bandwidth KDEs in our calculation of the Cross-correlation, Similarity, and Likeness coefficients. The first is a locally adaptive, variable-bandwidth KDE (LA-KDE; Shimazaki and Shinomoto, 2010). In this model the local bandwidth is inversely proportional to the data density over the local sample space, resulting in reduced smoothing in intervals with high data density and increased smoothing over intervals with lower data densities. The second is the diffusion-based adaptive bandwidth model of Botev et al. (2010). This model estimates the optimal bandwidth which is then applied uniformly across the sample space.

DZstats applies the five tests that have formed the most common basis for statistical analysis by the detrital geochronology and thermochronology community.

1.4. Kolmogorov-Smirnov test

The non-parametric two-sample K-S test tests the null hypothesis that two samples are drawn from parent populations with the same distribution, or informally, that the variation between two age populations is within the expected variation assuming random sampling of a parent population. It is based on the K-S statistic (D), which is the maximum difference between the empirical cumulative distribution functions (CDF) of the two samples (Fig. 1E), and returns a p-value that is inversely proportional to the confidence level at which that the two samples fail the hypothesis. The D -value is calculated as

$$D_{1,2} = \sup_x |F_1(x) - F_2(x)|, \quad (\text{eq. 5})$$

where F_1 and F_2 are the CDFs of the two samples, constructed from n_1 and n_2 observations, respectively. The probability (p) that the observed D -value is greater than the expected D -value for samples drawn from the same population is inversely proportional to the significance level at which the null hypothesis can be rejected and is calculated by Stephens (1970) as,

$$p(D_{\text{observed}} > D_{\text{critical}}) = Q_{KS}(\lambda) = 2 \sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 \lambda^2} \quad (\text{eq. 6})$$

where

$$\lambda = \left(\sqrt{n_e} + 0.12 + \frac{0.11}{\sqrt{n_e}} \right) D \quad (\text{eq. 7})$$

and

$$n_e = \frac{n_1 n_2}{n_1 + n_2} \quad (\text{eq. 8})$$

with limiting values of

$$Q_{KS}(0) = 1 \text{ and } Q_{KS}(\infty) = 0 \quad (\text{eq. 9})$$

Thus, for example, a p -value < 0.05 correlates to a $> 95\%$ confidence level that the two samples are not drawn from the same parent population.

1.5. Kuiper test

A commonly used alternative to the two-sample K-S test is the two-sample Kuiper test (Kuiper, 1960; Press et al., 2007). Like the K-S test, the Kuiper test tests the null hypothesis that two samples are drawn from parent populations with the same distribution. This variant of the K-S test guarantees equal sensitivities for the entire cumulative distribution functions of the two samples, whereas the K-S test tends to be more sensitive near the median and relatively insensitive to the distribution tails. The Kuiper statistic (V) is calculated from two CDFs F_1 and F_2 , each constructed from n_1 and n_2 observations, respectively,

$$V(x) = \max_{-\infty < x < \infty} [F_1(x) - F_2(x)] + \max_{-\infty < x < \infty} [F_2(x) - F_1(x)] \quad (\text{eq. 10})$$

with a probability that level (p) that $V_{\text{observed}} > V_{\text{critical}}$ is given by Stephens (1970) as,

$$p(V_{\text{observed}} > V_{\text{critical}}) = Q_{KP}(\lambda) = 2 \sum_{i=1}^{\infty} (4i^2 \lambda^2 - 1) e^{-2i^2 \lambda^2} \quad (\text{eq. 11})$$

where

$$\lambda = \left(\sqrt{n_e} + 0.155 + \frac{0.24}{\sqrt{n_e}} \right) V \quad (\text{eq. 12})$$

and n_e is defined as in equation 4.

As with the K-S test, this is subject to the limiting conditions

$$Q_{KP}(0) = 1 \text{ and } Q_{KP}(\infty) = 0 \quad (\text{eq. 13})$$

Similar to the K-S test, a p -value < 0.05 correlates to a $> 95\%$ confidence that the two samples are not drawn from the same parent population.

1.6. Similarity

The Similarity coefficient measures whether samples have similar modal sub-intervals as well as similar proportions of components in each of the modes (Fig. 1B). Gehrels et al. (2000) define it as

$$S = \sum_{i=1}^n \sqrt{f(i)g(i)} \quad (\text{eq. 14})$$

where $f(x)$ and $g(x)$ are the PDPs or KDEs and n is the interval of interest. A value of 1 indicates samples that are perfectly matched both in the modes and modal proportions, while a value of 0 indicates that the two samples share no modes or that modal proportions vary between the samples.

1.7. Cross-correlation

The Cross-correlation coefficient (“PDF-crossplot R^2 value”) is the coefficient of determination of a cross-plot of PDPs or KDEs of two samples for the same age intervals (Saylor et al., 2012; 2013). This is similar to a Q-Q or P-P plot (Wilk and Gnanadesikan, 1968) with the exception that PDPs or KDEs, rather than CDFs, are used in the cross-plot. For each sample, the relative probability for each 1 Myr interval for 0–4000 Ma was calculated and cross-plots were generated by plotting the relative probability for the same ages (Fig. 1D). The result is that the cross-plot is sensitive not only to the presence or absence of age peaks but also changes in the relative magnitude or shape of the peaks. For samples that share the same age peaks, peak shapes, and peak magnitudes (i.e., identical age spectra), the R^2 value of the cross-plot will be 1 while for those that share no age peaks, the R^2 value will be 0. For samples that share either some, but not all peaks, or have peaks of different magnitudes or shapes the R^2 value of the cross-plot will be between 0 and 1 with a higher value for samples that are more similar. They are also more sensitive to differences between samples than traditional Q-Q plots because the relative probability is not a monotonically increasing function.

1.8. Likeness

Likeness (Satkoski et al., 2013) is the complement of the area mismatch (Fig. 1C, Amidon et al., 2005a; 2005b). The area mismatch (M) is calculated as

$$M = (\sum_{i=1}^{i=n} |f(i) - g(i)|)/2 \quad (\text{eq. 15})$$

where $f(x)$ and $g(x)$ are the PDPs or KDEs of sample one and two, respectively, and n is the interval of interest. Likeness, L , is then

$$L = 1 - M \quad (\text{eq. 16})$$

2. TWO SAMPLE COMPARE

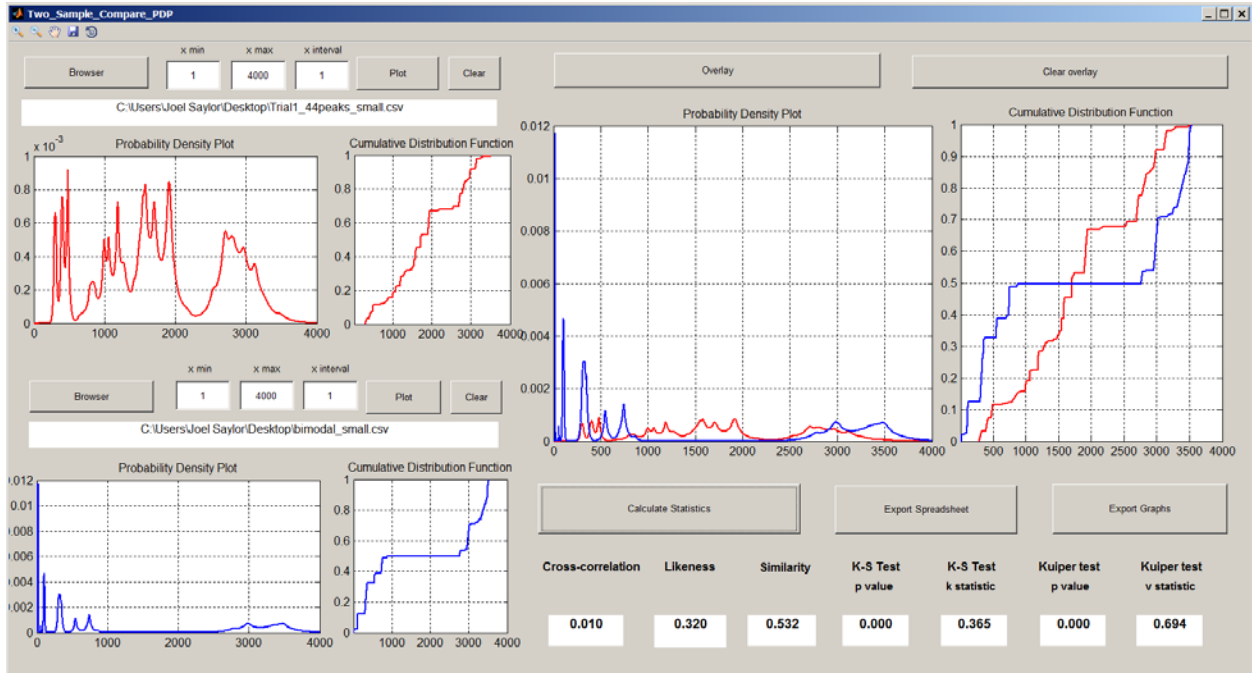


Figure 2.1. Overview of the Two Sample Compare window.

2.1. Data input format

Data should be organized into two comma-delimited .csv files without header rows. In each file, the first column is the grain mean ages, and the second column is the 1σ or 2σ uncertainty, or user specified bandwidth associated with each age.

2.2. Import data

Use the “Browser” buttons on the left side of the Two Sample Compare window to browse for, and import, the data for each sample (Fig. 2.1).

2.3. Specify the minimum and maximum interval of interest and the step interval of computation

Enter the minimum age of the interval of interest in field “x min.”

Enter the maximum age of the interval of interest in the field “x max.”

Enter the step interval of computation in the field “x interval.”

Interval fields are located to the right the “Browser” buttons for each sample (Fig. 2.1). A larger interval of computation allows more rapid computation, but decreases resolution.

For example, in the figures below, one sample was plotted with a step interval of 1 Myr (red) or 100 Myr (blue) (Fig.2.2).

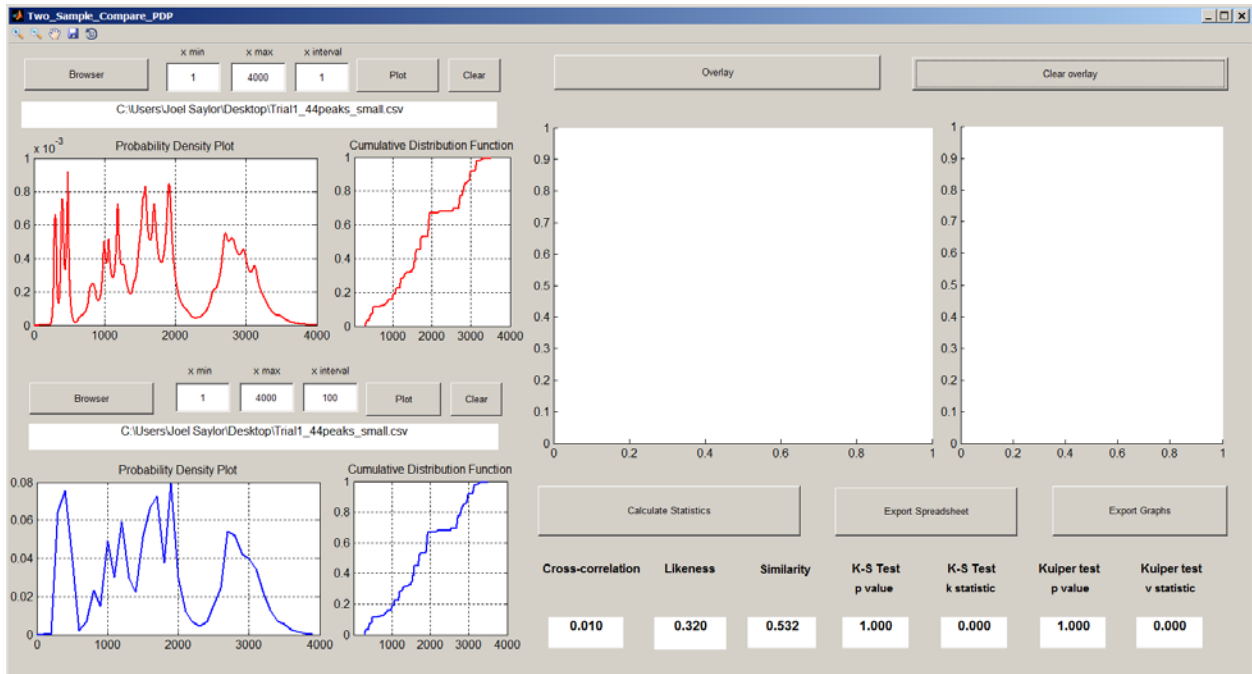


Figure 2.2. Comparison of the same sample plotted with a step interval of 1 Myr (red, top) or 100 Myr (blue, bottom).

2.4. Plotting individual samples

Individual samples can be plotted using the “Plot” button above that samples file pathway (Fig.2.3).

2.5. Plotting two overlaid PDPs or CDFs

The “Overlay” button will overlay the sample PDPs or KDEs, and CDFs with error in the three fields below the button. Note that the scale for the overlay graph is automatically adjusted to match the desired interval of interest (Fig. 2.3).

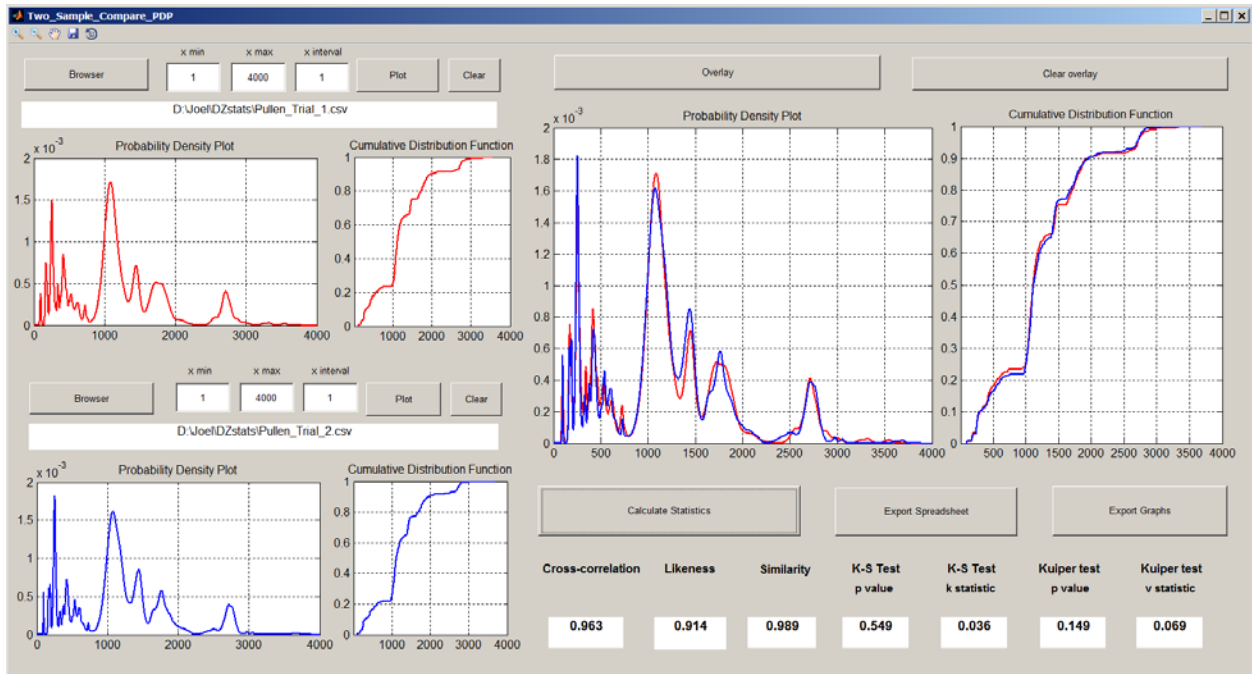


Figure 2.3. Two Sample compare showing the two samples “Pullen_Trial_1.csv” and “Pullen_Trial_2.csv” plotted individually and overlaid.

2.6. Calculating similarity statistics

Apply all of the statistical tests and populate all of the fields by selecting the “Calculate Statistics” button.

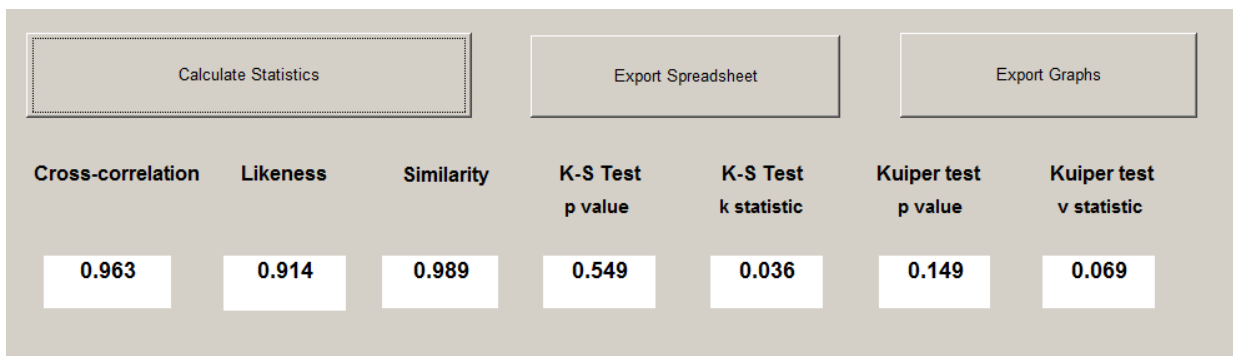


Figure 2.4. Results of statistical analyses appear in windows below the respective test names. Results can be exported as a table using the “Export Spreadsheet” button (see Fig. 2.5.).

2.7. Exporting statistical data

Export all statistical data to a table by selecting the “Export Spreadsheet” button. This opens a dialogue window to save the calculated statistics as an Excel spreadsheet.

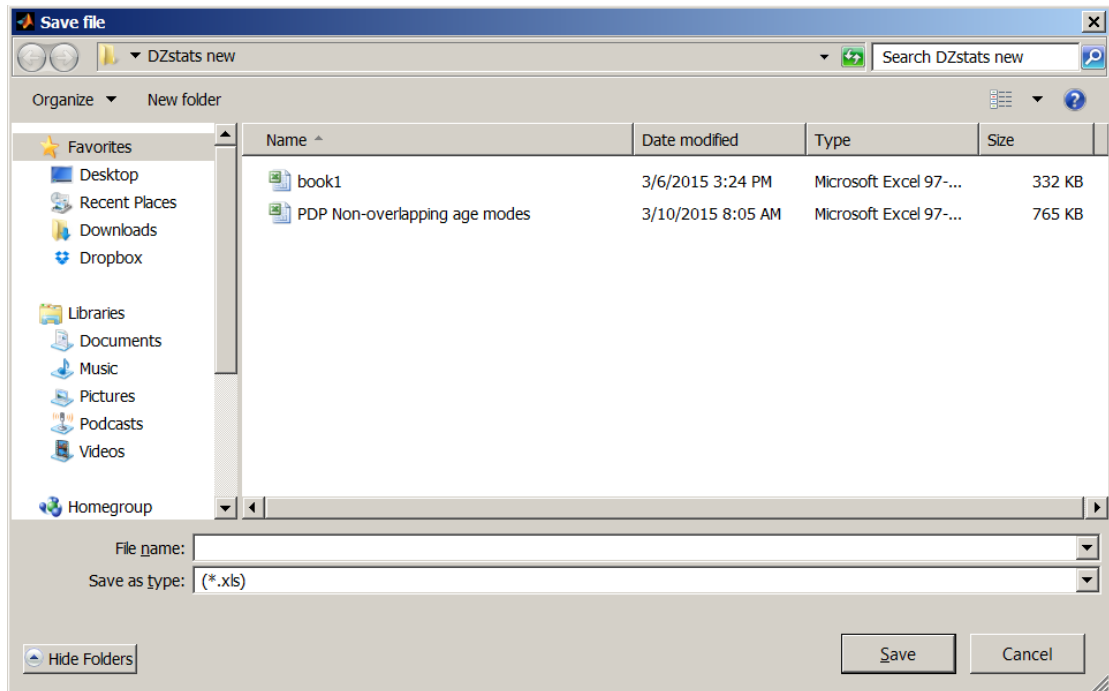


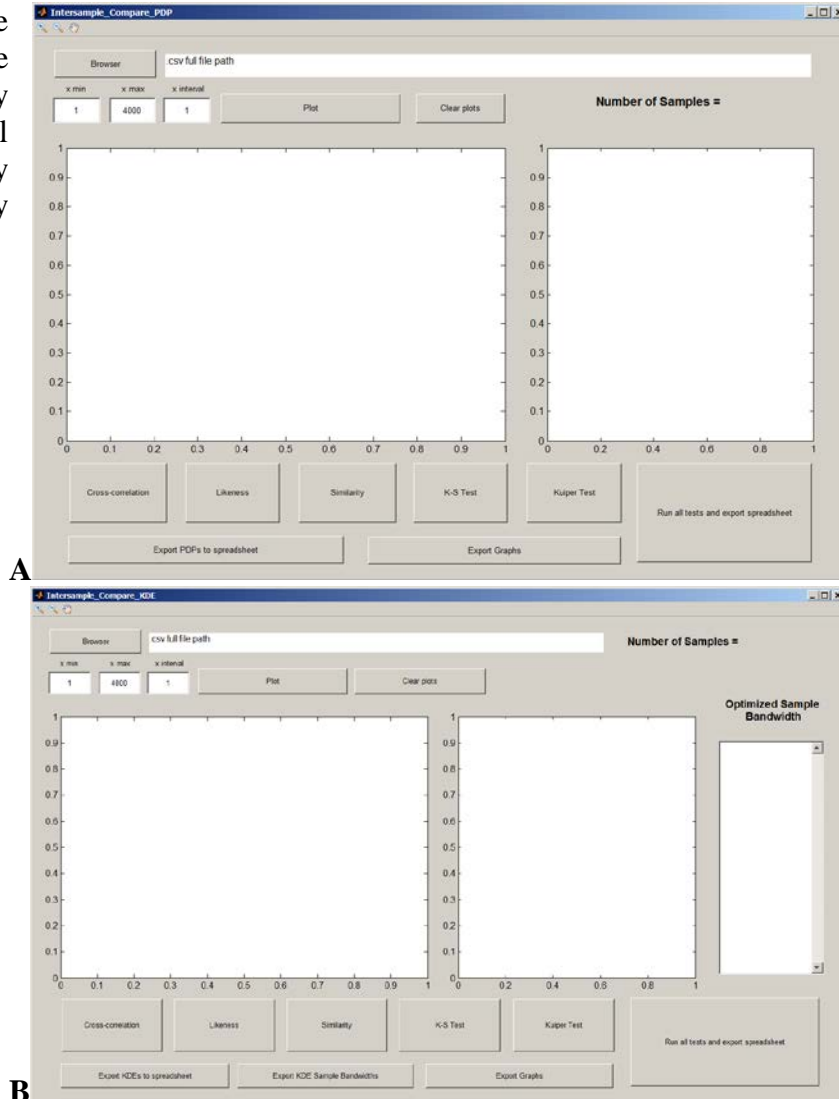
Figure 2.5. Results of the “Export Spreadsheet” function. Results can be saved as a spreadsheet for further analysis.

2.8. Exporting graphs

Export the graphs by selecting the “Export Graphs.” Graphs can then be saved in a variety of formats, including jpg, tiff, bmp, ai, pdf, png etc.

3. INTERSAMPLE COMPARE

Figure 3.1. Overview of the Intersample Compare window for A) Probability Density Plots or B) Kernel Density Estimates or Locally Adaptive Kernel Density Estimates.



3.1. Data import format

Data should be organized into a single comma-delimited .csv file without header rows. In each file, each sample is represented by a pair of columns where the first column is the grain mean ages, and the second column is the 1σ or 2σ uncertainty, or user specified bandwidth associated with each age.

EXAMPLE: To compare four samples, the import .csv file should contain eight columns with columns 1, 3, 5, and 7 containing the mean grain ages, and columns 2, 4, 6, and 8 containing the grain 1σ uncertainty. Sample 1 would then occupy columns 1 and 2, Sample 2 would occupy columns 3 and 4, Sample 3 would occupy columns 5 and 6, and Sample 4 would occupy columns 7 and 8.

3.2. Import data

Use the “Browser” button to browse for, and import, the data.

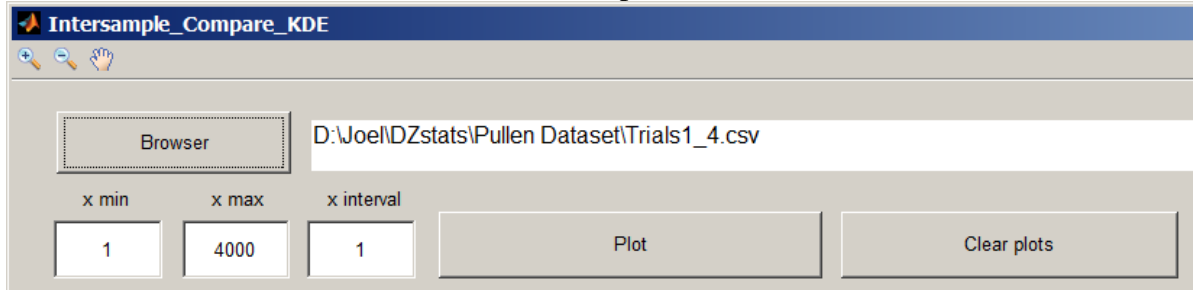


Figure 3.2. The “Browser” button can be used to find and import the data file which is automatically inserted into the import data file line.

3.3. Specify the minimum and maximum interval of interest and the step interval of computation

Enter the minimum age of the interval of interest in field “x min.”

Enter the maximum age of the interval of interest in the field “x max.”

Enter the step interval of computation in the field “x interval.” A larger interval of computation allows more rapid computation, but decreases resolution.

3.4. Plotting multiple samples

Individual samples can be plotted using the “Plot” button below that samples file pathway.

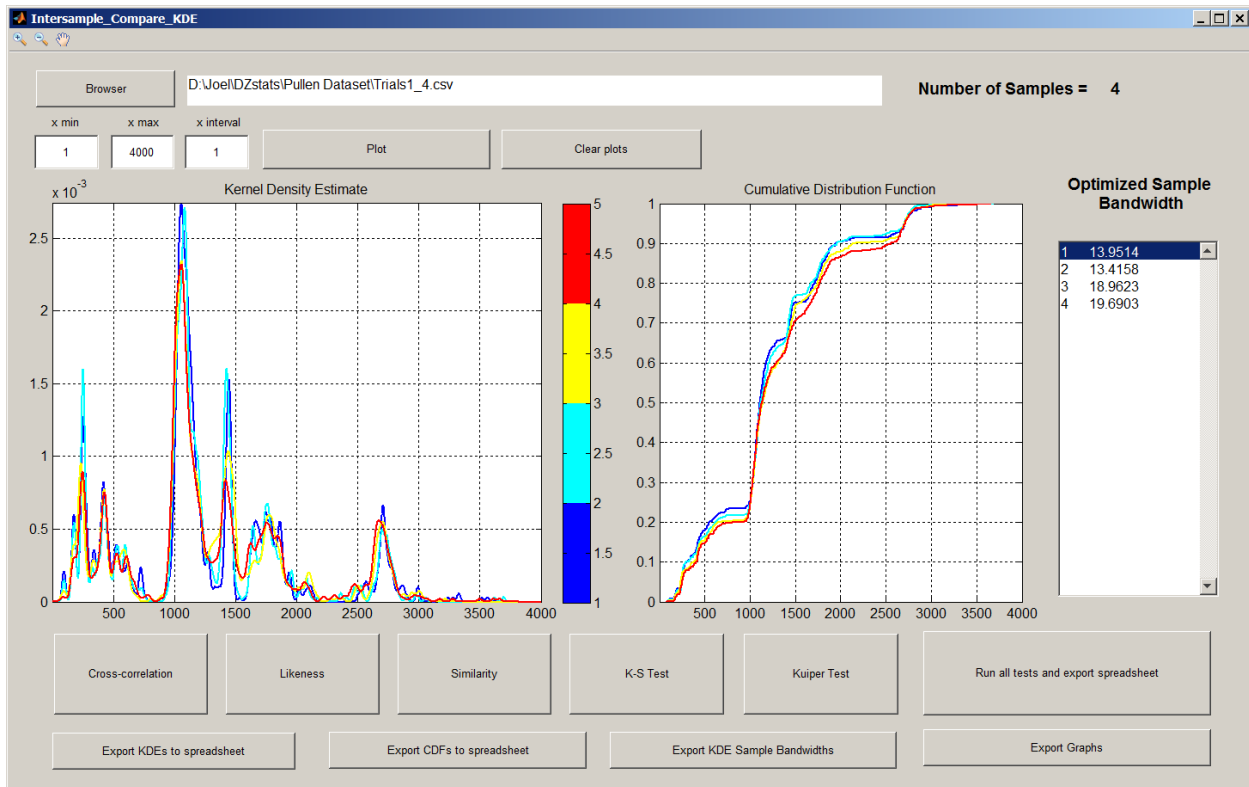


Figure 3.3. Result of importing 4 samples and plotting them (using the “Plot” button) over the range 1-3,000 (x-min = 1, x-max = 3,000) with a step interval of 1 Myr (x-interval = 1).

3.5. Apply individual statistical tests to all analyses from multiple samples

Each test can be applied to all samples by selecting the button labelled with the test name. This also opens and populates a table with the results of the test. Data can be copied and pasted to a standard spreadsheet software package such as Excel.

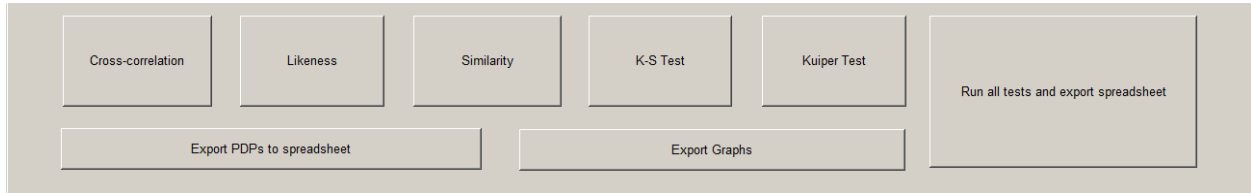
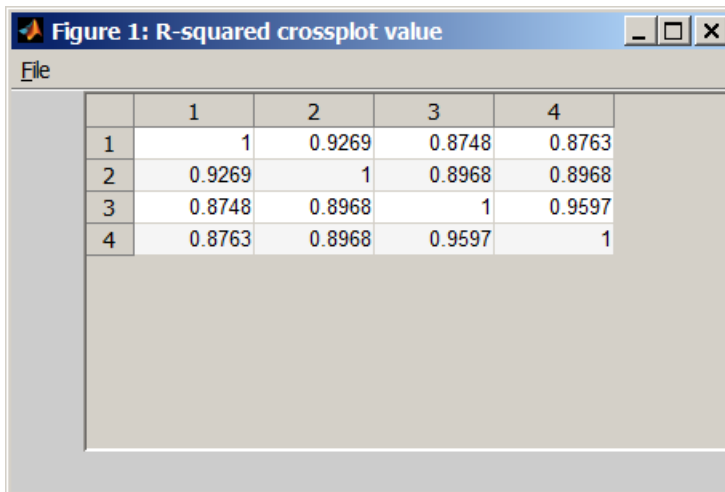


Figure 3.4. Individual statistical tests and export functions are selected and applied using the buttons at the bottom of the Intersample Compare window (PDP Intersample Compare shown here).

EXAMPLE: The sample entered in the import data file in columns 1 and 2 (“Sample 1”) is in row 1 and column 1 in the output table (Fig. 3.5). The sample entered in columns 3 and 4 (“Sample 2”) is in row 2 and column 2, etc.



	1	2	3	4
1	1	0.9269	0.8748	0.8763
2	0.9269	1	0.8968	0.8968
3	0.8748	0.8968	1	0.9597
4	0.8763	0.8968	0.9597	1

Figure 3.5. Result of pressing the “Cross-correlation” button (see Fig. 3.4.).

3.6. Apply all statistical tests to all analyses from multiple samples

To obtain the results of all statistical tests select the, “Run all tests and save as Excel spreadsheet,” button. Save the .xls file in the desired location using the pop-up browser window. The spreadsheet will contain the output of each of the statistical tests applied to all sample pairs in a single worksheet.

3.7. Exporting PDPs, KDEs, LA-KDEs or Bandwidths to spreadsheets

Probability density functions, kernel density estimates, and kernel band widths can all be exported as Excel spreadsheets by selecting the appropriate buttons.

EXAMPLE: In the exported spreadsheets, Ages (x-axis in figure 3.3) are in column 1, while the relative probability (y-axis in figure 3.3) for Sample 1 is in column 2, Sample 2 is in column 3, etc.

3.8. Exporting graphs

Export the graphs by selecting the “Export Graphs.” Graphs can then be saved in a variety of formats, including jpg, tiff, bmp, ai, pdf, png etc.

3.9. Analysis bandwidths

The KDE and LA-KDE Intersample Compare modules will also calculate and display the bandwidths used in the kernel density estimation (Fig. 3.3, 3.6). The KDE Intersample Compare will display each of the bandwidths used for each sample (Fig. 3.6A), while the LA-KDE Intersample Compare module will display a scrolling window with the bandwidth applied at each of the selected x-intervals (Fig. 3.6B)

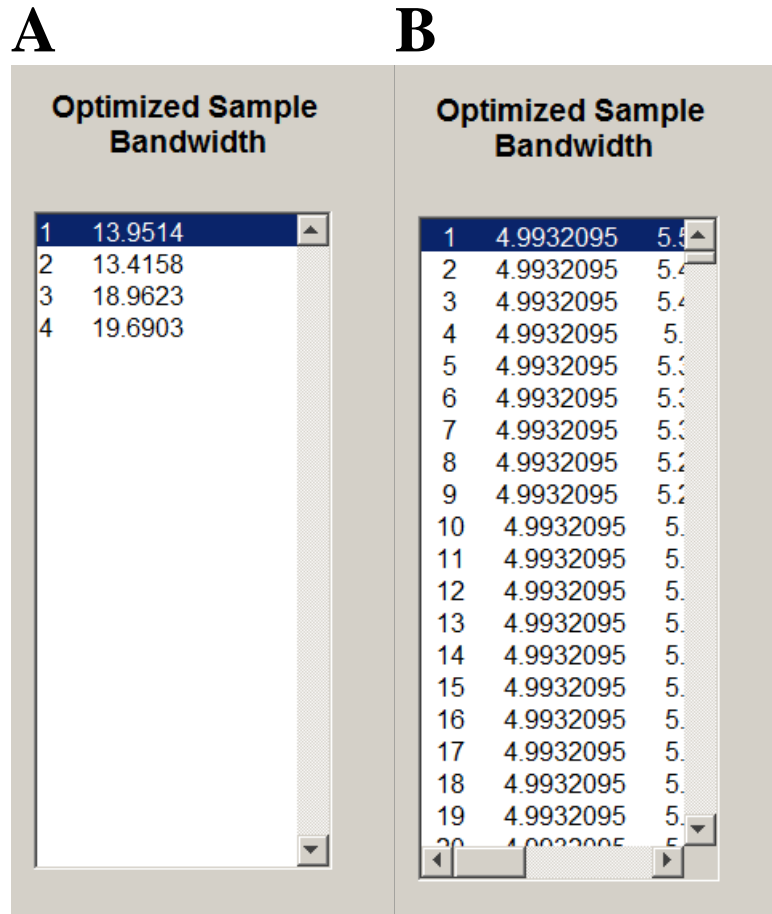


Figure 3.6 A) Optimized kernel bandwidths for the kernel density estimates of the four samples shown in figure 3.3. B) Local optimized kernel bandwidths for each Myr (1-19 shown in the figure) for each of the four samples (only sample 1 shown in the figure) shown in figure 3.3.

4. SUBSAMPLE COMPARE

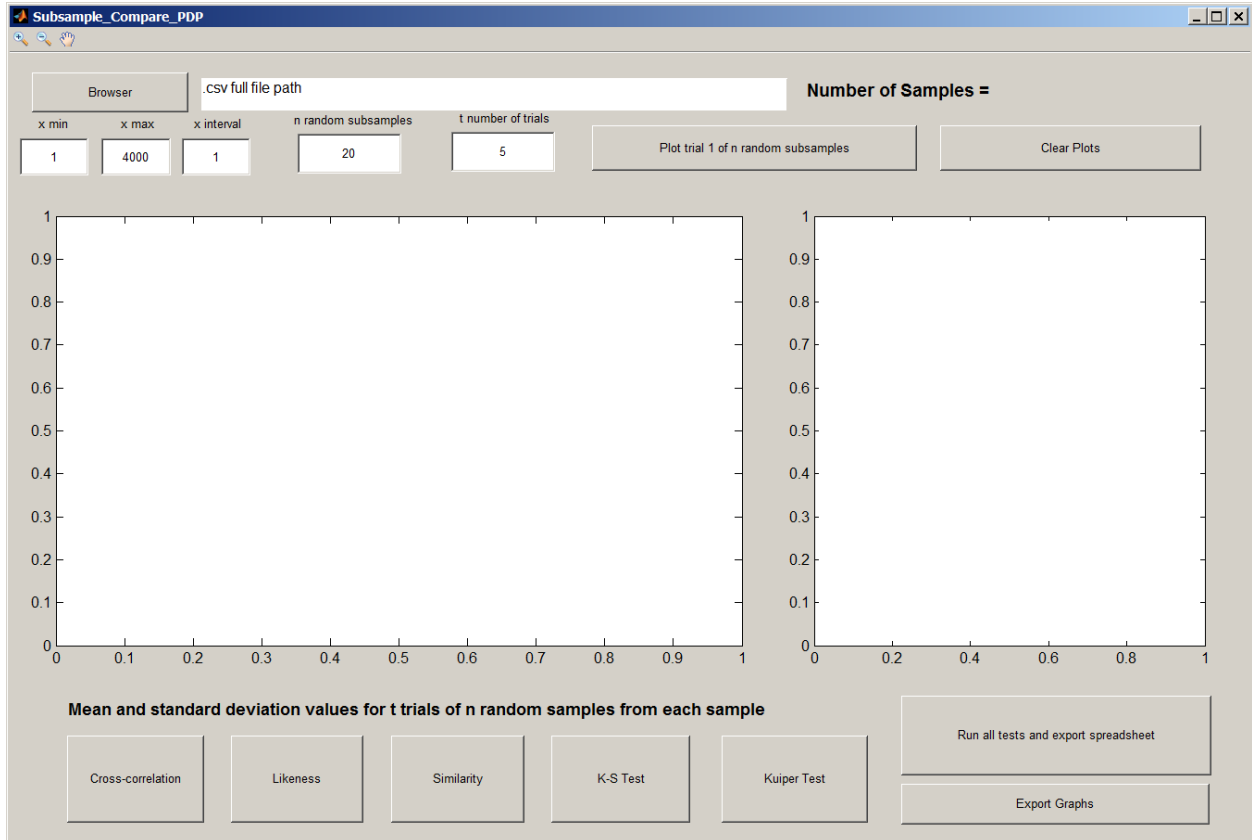


Figure 4.1. Overview of the Subsample Compare module window showing the import data pathway. This setup will apply the statistical tests to analyses between 1 and 4000 Ma. It is setup to draw 20 random subsamples from each of the four datasets in the import file and repeat the process 5 times.

4.1. Data import format

The data import format is the same as for “Intersample Compare.” Data should be organized into a single comma-delimited .csv file without header rows. In each file, each sample is represented by a pair of columns where the first column is the grain mean ages, and the second column is the 1σ or 2σ uncertainty, or user specified bandwidth associated with each age.

EXAMPLE: To compare four samples, the import .csv file should contain eight columns with columns 1, 3, 5, and 7 containing the mean grain ages, and columns 2, 4, 6, and 8 containing the grain 1σ uncertainty. Sample 1 would then occupy columns 1 and 2, Sample 2 would occupy columns 3 and 4, Sample 3 would occupy columns 5 and 6, and Sample 4 would occupy columns 7 and 8.

4.2. Import data

Use the “Browser” button to browse for, and import, the data.

4.3. Specify the minimum and maximum interval of interest and the step interval of computation

Enter the minimum age of the interval of interest in field “x min.”

Enter the maximum age of the interval of interest in the field “x max.”

Enter the step interval of computation in the field “x interval.” A larger interval of computation allows more rapid computation, but decreases resolution.

EXAMPLE: In figure 4.1 the lower bound on the interval of interest is 1 Ma (“x min”), the upper bound is 4000 Ma (“x max”), and the step interval is 1 Myr (“x interval”).

4.4. Specify the number of ages to be randomly drawn from each sample

Enter the number of ages (“n”) to be randomly drawn from each sample in the field labelled, “n random subsamples.” Because the same number of ages will be drawn from each sample, THIS NUMBER CANNOT EXCEED THE MAXIMUM SIZE OF THE SMALLEST SAMPLE SET.

EXAMPLE: In figure 4.1 the number of analyses to be randomly drawn from each sample is 20 (“n random subsamples”).

4.5. Specify the number of trials to conduct

Enter the number of times that n grains will be drawn in the field labelled, “t number of trials.”

EXAMPLE: In figure 4.1 the number of analyses to be randomly drawn from each sample is 5 (“t number of trials”).

4.6. Plot trial 1

Plot the first trial (of n trials) to ensure that data was loaded correctly prior to applying statistical tests.

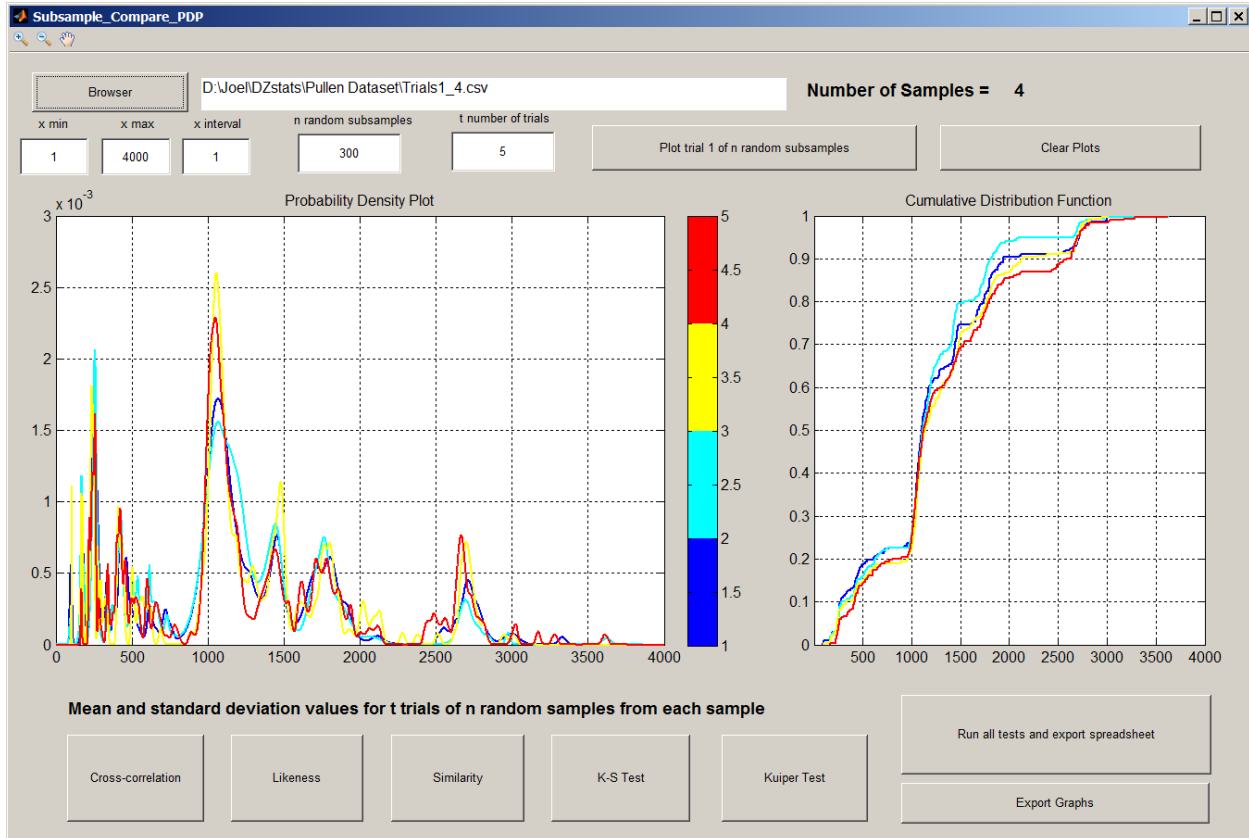


Figure 4.2. Plot of 1 trial of subsampling 300 random grain ages from the 4 detrital datasets loaded in figure 4.1. These plots provide confirmation that data was loaded and subsampled correctly before applying the statistical tests which can take several minutes depending on the number of subsamples and trials.

4.7. Apply individual statistical tests to multiple subsets of analyses from multiple samples

Each test can be applied to all samples by selecting the button labelled with the test name. This also opens and populates a table with the results of the test. Data can be copied and pasted to a standard spreadsheet software package such as Excel. For each test, the mean and standard deviation of the metrics will be returned based on applying the statistical metric to t trials of n random age draws.

EXAMPLE: The sample entered in the import data file in columns 1 and 2 (“Sample 1”) is in row 1 and column 1 in the output table. The sample entered in columns 3 and 4 (“Sample 2”) is in row 2 and column 2, etc.

	1	2	3	4
1	n			
2	300			
3	trials			
4	5			
5				
6	Mean Cross Correlation Coefficient			
7	1	0.9086	0.7573	0.8006
8	0.9086	1	0.7503	0.7883
9	0.7573	0.7503	1	0.8078
10	0.8006	0.7883	0.8078	1
11				
12	Standard deviation Cross Correlation Coefficient			
13	0	0.0223	0.0545	0.0135
14	0.0223	0	0.0522	0.0453
15	0.0545	0.0522	0	0.0740
16	0.0135	0.0453	0.0740	0

Figure 4.3. Example of the output of one of the statistical tests as the result of application of statistical tests 5 times to 200 random subsamples of four samples. The output includes both the mean statistical values and their standard deviation.

4.8. Apply all statistical tests to multiple subsets of analyses from multiple samples

To obtain the results of all statistical tests select the, “Run all tests and export spreadsheet,” button. This will apply all statistical tests and open a “Save file” dialogue box when finished. Save the .xls file in the desired location using the pop-up browser window. The spreadsheet will contain the mean, standard deviation, minimum, and maximum of each of the statistical tests applied to all sample pairs in a single worksheet.

4.9. Exporting graphs

Export the graphs by selecting the “Export Graphs.” Graphs can then be saved in a variety of formats, including jpg, tiff, bmp, ai, pdf, png etc.

5. REFERENCES

We are deeply indebted to the following people and publications whose work is included in this application or provided inspiration for this application.

- Amidon, W. H., Burbank, D. W., and Gehrels, G. E., 2005a, Construction of detrital mineral populations: insights from mixing of U-Pb zircon ages in Himalayan rivers: *Basin Research*, v. 17, p. 463-485.
- Amidon, W. H., Burbank, D. W., and Gehrels, G. E., 2005b, U-Pb zircon ages as a sediment mixing tracer in the Nepal Himalaya: *Earth and Planetary Science Letters*, v. 235, no. 1-2, p. 244-260.
- Andersen, T., 2005, Detrital zircons as tracers of sedimentary provenance: limiting conditions from statistics and numerical simulation: *Chemical Geology*, v. 216, no. 3-4, p. 249-270.
- Behboodiani, J., 1970, On a Mixture of Normal Distributions: *Biometrika*, v. 57, no. 1, p. 215-217.
- Botev, Z. I., Grotowski, J. F., and Kroese, D. P., 2010, Kernel density estimation via diffusion: *Annals of Statistics*, v. 38, no. 5, p. 2916-2957.
- Brandon, M. T., 1996, Probability density plot for fission-track grain-age samples: *Radiation Measurements*, v. 26, no. 5, p. 663-676.
- Carroll, R. J., Delaigle, A., and Hall, P., 2009, Nonparametric Prediction in Measurement Error Models: *Journal of the American Statistical Association*, v. 104, no. 487, p. 993-1003.
- Delaigle, A., and Meister, A., 2008, Density estimation with heteroscedastic error: *Bernoulli*, v. 14, no. 2, p. 562-579.
- Dodson, M. H., Compston, W., Williams, I. S., and Wilson, J. F., 1988, A search for ancient detrital zircons in Zimbabwean sediments: *Journal of the Geological Society of London*, v. 145, p. 977-983.
- Everitt, B., and Hand, D. J., 1981, *Finite mixture distributions*: New York Chapman and Hall.
- Everitt, B. S., 2005, *Finite Mixture Distributions*, *Encyclopedia of Statistics in Behavioral Science*, John Wiley & Sons, Ltd.
- Fedo, C. M., Sircombe, K. N., and Rainbird, R. H., 2003, Detrital zircon analysis of the sedimentary record, *in* Hanchar, J. M., and Hoskin, P. W. O., eds., *Zircon*, Volume 53, p. 277-303.
- Gehrels, G., 2000, Introduction to detrital zircons studies of Paleozoic and Triassic strata in western Nevada and northern California, *in* Soreghan, M. J., and Gehrels, G. E., eds., *Paleozoic and Triassic paleogeography and tectonics of western Nevada and northern California*, Volume 347: Boulder, Geological Society of America Special Paper, p. 1-17.
- Gehrels, G., 2012, Detrital zircon U-Pb geochronology: current methods and new opportunities, *in* Busby, C., and Azor, A., eds., *Tectonics of Sedimentary Basins: Recent Advances*, Blackwell Publishing Ltd. , p. 47-62.
- Gehrels, G. E., Stewart, J. H., and Ketner, K. B., 2002, Cordilleran-margin quartzites in Baja California - implications for tectonic transport: *Earth and Planetary Science Letters*, v. 199, no. 1-2, p. 201-210.

- Hurford, A. J., Fitch, F. J., and Clarke, A., 1984, Resolution of the age structure of the detrital zircon populations of two Lower Cretaceous sandstones from the Weald of England by fission track dating: *Geological Magazine*, v. 121, no. 04, p. 269-277.
- Jessberger, E. K., Dominik, B., Staudacher, T., and Herzog, G. F., 1980, $^{40}\text{Ar}/^{39}\text{Ar}$ ages of Allende: *Icarus*, v. 42, no. 3, p. 380-405.
- Kuiper, N. H., 1960, Tests concerning random points on a circle: *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen*, v. 63, p. 38-47.
- Lo, Y. T., Mendell, N. R., and Rubin, D. B., 2001, Testing the number of components in a normal mixture: *Biometrika*, v. 88, no. 3, p. 767-778.
- Martin, B. R., 2012, *Statistics for Physical Sciences: An Introduction*, Waltham, Academic Press, 301 p.:
- McIntyre, J., and Stefanski, L., 2011, Density Estimation with Replicate Heteroscedastic Measurements: *Annals of the Institute of Statistical Mathematics*, v. 63, no. 1, p. 81-99.
- Miller, M. B., 2014, *Mathematics and Statistics for Financial Risk Management*, Hoboken, John Wiley & Sons, Inc. , 336 p.:
- Pearson, R. K., 2011, *Exploring data in engineering, the sciences, and medicine*, New York, Oxford University Press.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P., 1997, *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge, Cambridge University Press.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P., 2007, *Numerical Recipes: The Art of Scientific Computing*, New York, Cambridge University Press.
- Pullen, A., Ibanez-Mejia, M., Gehrels, G. E., Ibanez-Mejia, J. C., and Pecha, M., 2014, What happens when $n=1000$? Creating large- n geochronological datasets with LA-ICP-MS for geologic investigations: *Journal of Analytical Atomic Spectrometry*, v. 29, no. 6, p. 971-980.
- Satkoski, A. M., Wilkinson, B. H., Hietpas, J., and Samson, S. D., 2013, Likeness among detrital zircon populations-An approach to the comparison of age frequency data in time and space: *Geological Society of America Bulletin*, v. 125, no. 11-12, p. 1783-1799.
- Saylor, J. E., Knowles, J. N., Horton, B. K., Nie, J. S., and Mora, A., 2013, Mixing of Source Populations Recorded in Detrital Zircon U-Pb Age Spectra of Modern River Sands: *Journal of Geology*, v. 121, no. 1, p. 17-33.
- Saylor, J. E., Stockli, D. F., Horton, B. K., Nie, J., and Mora, A., 2012, Discriminating rapid exhumation from syndepositional volcanism using detrital zircon double dating: Implications for the tectonic history of the Eastern Cordillera, Colombia: *Geological Society of America Bulletin*, v. 124, no. 5-6, p. 762-779.
- Scott, D. W., 1992, *Multivariate density estimation: Theory, practice, and visualization*, New York, John Wiley & Sons, Inc. , 317 p.:
- Shimazaki, H., and Shinomoto, S., 2010, Kernel bandwidth optimization in spike rate estimation: *Journal of Computational Neuroscience*, v. 29, no. 1-2, p. 171-182.
- Silverman, B. W., 1986, *Density Estimation for Statistics and Data Analysis*, London, Chapman and Hall, 175 p.:

- Sircombe, K. N., 2004, AgeDisplay: an EXCEL workbook to evaluate and display univariate geochronological data using binned frequency histograms and probability density distributions: Computers & Geosciences, v. 30, no. 1, p. 21-31.**
- Smith, D. M., and Bartlet, J. C., 1961, Calculation of the Areas of Isolated or Overlapping Normal Probability Curves: Nature, v. 191, no. 4789, p. 688-689.**
- Staudenmayer, J., Ruppert, D., and Buonaccorsi, J. R., 2008, Density estimation in the presence of heteroscedastic measurement error: Journal of the American Statistical Association, v. 103, no. 482, p. 726-736.**
- Stephens, M. A., 1970, Use of the Kolmogorov-Smirnov, Cramer-Von Mises and Related Statistics Without Extensive Tables: Journal of the Royal Statistical Society. Series B (Methodological), v. 32, no. 1, p. 115-122.**
- Titterton, D. M., Smith, A. F. M., and Makov, U. E., 1985, Statistical analysis of finite mixture distributions, Chichester, John Wiley & Sons, Ltd., 243 p.:**
- Vermeesch, P., 2012, On the visualisation of detrital age distributions: Chemical Geology, v. 312, p. 190-194.**
- Vermeesch, P., 2013, Multi-sample comparison of detrital age distributions: Chemical Geology, v. 341, p. 140-146.**
- Vermeesch, P., and Garzanti, E., 2015, Making geological sense of 'Big Data' in sedimentary provenance analysis: Chemical Geology, v. 409, p. 20-27.**
- Wilk, M. B., and Gnanadesikan, R., 1968, Probability plotting methods for the analysis of data: Biometrika, v. 55, no. 1, p. 1-17.**