Research Paper

GEOSPHERE

GEOSPHERE; v. 12, no. 1

doi:10.1130/GES01237.1

12 figures; 2 tables; 4 supplemental files

CORRESPONDENCE: jesaylor@uh.edu

CITATION: Saylor, J.E., and Sundell, K.E., 2016, Quantifying comparison of large detrital geochronology data sets: Geosphere, v. 12, no. 1, p. 203–220, doi:10.1130/GES01237.1.

Received 23 July 2015
Revision received 20 October 2015
Accepted 17 November 2015
Published online 7 January 2016

Quantifying comparison of large detrital geochronology data sets

Joel E. Saylor and Kurt E. Sundell

Department of Earth and Atmospheric Sciences, University of Houston, Science & Research 1 Building, 3507 Cullen Boulevard, Room 312, Houston, Texas 77204, USA

ABSTRACT

The increase in detrital geochronological data presents challenges to existing approaches to data visualization and comparison, and highlights the need for quantitative techniques able to evaluate and compare multiple large data sets. We test five metrics commonly used as quantitative descriptors of sample similarity in detrital geochronology: the Kolmogorov-Smirnov (K-S) and Kuiper tests, as well as Cross-correlation, Likeness, and Similarity coefficients of probability density plots (PDPs), kernel density estimates (KDEs), and locally adaptive, variable-bandwidth KDEs (LA-KDEs). We assess these metrics by applying them to 20 large synthetic data sets and one large empirical data set, and evaluate their utility in terms of sample similarity based on the following three criteria. (1) Similarity of samples from the same population should systematically increase with increasing sample size. (2) Metrics should maximize sensitivity by using the full range of possible coefficients. (3) Metrics should minimize artifacts resulting from sample-specific complexity. K-S and Kuiper test p-values passed only one criterion, indicating that they are poorly suited as quantitative descriptors of sample similarity. Likeness and Similarity coefficients of PDPs, as well as K-S and Kuiper test D and V values, performed better by passing two of the criteria. Cross-correlation of PDPs passed all three criteria. All coefficients calculated from KDEs and LA-KDEs failed at least two of the criteria.

As hypothesis tests of derivation from a common source, individual K-S and Kuiper p-values too frequently reject the null hypothesis that samples come from a common source when they are identical. However, mean p-values calculated by repeated subsampling and comparison (minimum of 4 trials) consistently yield a binary discrimination of identical versus different source populations. Cross-correlation and Likeness of PDPs and Cross-correlation of KDEs yield the widest divergence in coefficients and thus a consistent discrimination between identical and different source populations, with Cross-correlation of PDPs requiring the smallest sample size. In light of this, we recommend acquisition of large detrital geochronology data sets for quantitative comparison. We also recommend repeated subsampling of detrital geochronology data sets and calculation of the mean and standard deviation of the comparison metric in order to capture the variability inherent in sampling a multimodal population.

These statistical tools are implemented using DZstats, a MATLAB-based code that can be accessed via an executable file graphical user interface. It implements all of the statistical tests discussed in this paper, and exports the results both as spreadsheets and as graphic files.

INTRODUCTION

Over the past two decades detrital geochronology has emerged as a standard tool for sediment provenance analysis, largely due to the advent of in situ analytical techniques yielding rapid, precise U-Pb analyses (Fedo et al., 2003; Gehrels et al., 2008; Shaulis et al., 2010). Applications of detrital geochronology encompass questions of sediment provenance, sediment budgets, correlations between sedimentary units, and determinations of depositional age (e.g., Gehrels, 2012; Vermeesch, 2012). These applications have benefitted from growing samples sizes, which have allowed robust intersample comparison, identification of small subpopulations, and quantification of relative subpopulation proportions (Vermeesch, 2004; Andersen, 2005; Pullen et al., 2014).

The increase in detrital geochronological data presents challenges to existing approaches to data visualization and comparison (Vermeesch, 2012, 2013; Vermeesch and Garzanti, 2015). It also highlights the need for quantitative techniques able to evaluate and compare multiple large data sets while still representing the uncertainty present in the data sets and the variability inherent in sampling a multimodal population (Andersen, 2005; Gehrels, 2012; Satkoski et al., 2013). Visual comparison of age distributions as probability density plots (PDPs), cumulative distribution functions (CDFs), cross-plots of CDF pairs (Q-Q plots), or kernel density estimates (KDEs) has been and will likely continue to be foundational for data interpretation. However, these visual methods become increasingly cumbersome with large sample numbers or sample size. For example, although large data sets are essential for continental-scale or high-resolution correlation, interpretation of these extensive data sets often relies on visual comparison of full page or multipage age distribution plots. Visual inspection also yields no quantitative comparison between samples, hampering application of forward mixing models and increasing the possibility of subjectivity or approximation in interpretation.

The analysis above points to a growing need for data assessment beyond visual inspection alone. The scientific community responded to the need for quantitative comparison metrics through application of statistical tests intended to determine whether samples were drawn from the same parent population using hypothesis tests such as the Kolmogorov-Smirnov (K-S) test (e.g., DeGraaff-Surpless et al., 2003; Fedo et al., 2003; Weislogel et al., 2010; Lawrence et al., 2011). These methods have been applied to determine if two samples may have been drawn from the same parent population, and to determine the degree of similarity between sample populations. Alternative metrics such as the Likeness/Mismatch, Similarity, and Cross-correlation (cross-plot

R^2 values) coefficients lack a clear hypothesis test, but assess the degree of similarity between samples (Gehrels, 2000; Amidon et al., 2005a; Saylor et al., 2012; 2013; Satkoski et al., 2013).

In this paper we evaluate the limits of these quantitative approaches when applied to large detrital geochronological data sets by applying them to 20 synthetic data sets as well as a published empirical detrital zircon U-Pb data set. Our goal is not to define a single statistical metric or data set size that should be applied in all cases, but rather to evaluate the sensitivity of proposed metrics over a range of data set sizes. We envision the methods discussed in this paper as complimentary to, rather than replacing, the now-standard approaches of data visualization and analysis.

METHODS

Quantitative Methods

We used five methods to evaluate the similarity between synthetic data sets and a large empirical data set published by Pullen et al. (2014). We applied the K-S and Kuiper tests to each pair of data sets and calculated Similarity, Likeness, and Cross-correlation coefficients (Fig. 1) based on the sample PDPs, adaptive KDEs, and locally adaptive, variable-bandwidth KDEs (LA-KDEs). We also compared the effect size of the K-S and Kuiper tests (D and V values, respectively) to determine whether these provide a more robust comparison than their derivative p-values. This yields a total of 13 metrics for each of the subsamples of the populations described in the following.

Kolmogorov-Smirnov Test

The nonparametric 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 CDFs 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

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

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 was calculated by Stephens (1970) as

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

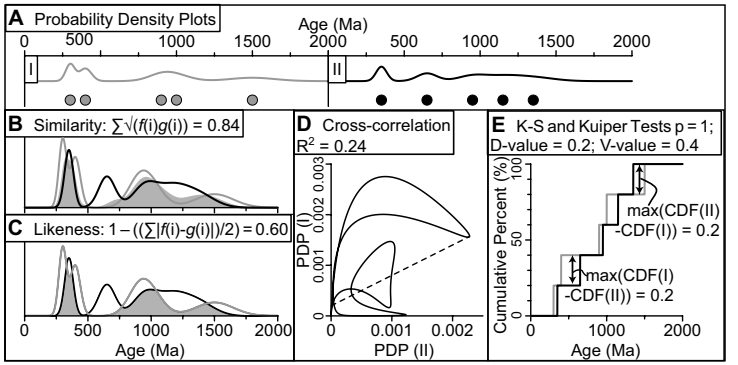


Figure 1. Schematic illustration of the methods discussed in the text. (A) Probability density plots (PDPs) of age distributions with 5 modal ages, each with an associated uncertainty of 10%. (B) Similarity calculation and coefficient for PDPs I and II. (C) Likeness calculation and coefficient for PDPs I and II. (D) Cross-correlation coefficient (coefficient of determination, dashed line) of cross-plots of PDPs I and II. (E) Kolmogorov-Smirnov (K-S) D value and Kuiper V values of cumulative distribution functions (CDF) based on the distributions shown in A. K-S and Kuiper tests do not incorporate uncertainties or bandwidth associated with each modal age.

where

$$\lambda = \left(\sqrt{n_e} + 0.12 + \frac{0.11}{\sqrt{n_e}}\right)D, \quad (3)$$

and

$$n_e = \frac{n_1 n_2}{n_1 + n_2}, \quad (4)$$

with limiting values of

$$Q_{KS}(0) = 1 \text{ and } Q_{KS}(\infty) = 0. \quad (5)$$

Thus, for example, a p-value <0.05 corresponds to a >95% confidence level that the 2 samples are not drawn from the same parent population. The K-S test requires relatively large sample sizes to accurately reject the null hypothesis in part due to distortion of the distributions introduced by random sampling of a large population.

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 (Fig. 1E),

$$V(x) = \max_{-\infty<x<\infty}[F_1(x) - F_2(x)] + \max_{-\infty<x<\infty}[F_2(x) - F_1(x)] \quad (6)$$

with a probability that level (p) that $V_{observed} > V_{critical}$, calculated by Stephens (1970),

$$p(V_{observed} > V_{critical}) = Q_{KP}(\lambda) = 2\sum_{i=1}^{\infty}(i^2\lambda^2 - 1)e^{-2i^2\lambda^2}, \quad (7)$$

where

$$\lambda = \left(\sqrt{n_e} + 0.155 + \frac{0.24}{\sqrt{n_e}}\right)V, \quad (8)$$

and n_e is defined as in Equation 4.

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

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

Similar to the K-S test, a p-value <0.05 corresponds to a >95% confidence that the two samples are not drawn from the same parent population. Also like the K-S test, the Kuiper test requires relatively large sample sizes to accurately reject the null hypothesis.

Mixture Distributions and Kernel Density Estimations

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 subpopulations and are used in a variety of disciplines including the physical sciences, medicine, economics, engineering, and social sciences (Smith and Bartlet, 1961; Behboodian, 1970; Everitt and Hand, 1981; Titterington et al., 1985; Lo et al., 2001; Everitt, 2005; Pearson, 2011; Martin, 2012; Miller, 2014). The discrete mixture distribution [$f(x)$], calculated from n observations is given as

$$f(x) = \sum_{i=1}^{n} w_i f_i(x), \quad (10)$$

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

$$\sum_{i=1}^{n} w_i = 1. \quad (11)$$

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

$$f_i(x) = \frac{1}{\sigma_i\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2\right], \ (\sigma > 0), \quad (12)$$

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, p. 99) "called the *mixture density*" (Pearson, 2011, p. 470). 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 an alternative method of estimating a sample's PDF. KDEs are nonparametric in the sense that they are applicable regardless of the shape of the PDF and do not require a particular type of parameter in the population elements (i.e., do not require a normal Gaussian distribution in population elements). The KDE is defined as

$$\hat{f}_h(x) = \sum_{i=1}^{n}\frac{1}{nh}K\left(\frac{x-x_1}{h}\right), \quad (13)$$

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 for locally adaptive KDEs) component kernels (Scott, 1992). The kernel function (K) 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, p. 18) noted that, "[b]ecause the window width is fixed across the entire sample, there is a tendency for spurious noise to appear in the tails of the estimates; if the estimates are smoothed sufficiently to deal with this, 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 oversmoothing or undersmoothing of the resultant KDE. Several approaches have been developed to deal with the issues raised by KDEs, including bandwidth optimization algorithms, LA-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 the 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.

Cross-Correlation Coefficient

The Cross-correlation coefficient 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. The cross-plot is sensitive to the presence or absence of age peaks and to changes in the relative magnitude or shape of the peaks. For samples with identical age peaks, peak shapes, and peak magnitudes (i.e., identical age spectra), the R^2 value of the cross-plot will be 1; for those that share no age peaks, the R^2 value will approach 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.

Similarity Coefficient

The Similarity coefficient measures whether samples have overlapping modes as well as similar proportions of components in each of the modes (Fig. 1B). Gehrels (2000) defined it as

$$S = \sum_{i=1}^{n} \sqrt{f(i)g(i)}, \tag{14}$$

where $f(i)$ and $g(i)$ are the PDPs or KDEs and i are ages between 1 and n. 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.

Likeness Coefficient

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

$$M = \left(\sum_{i=1}^{i=n} \left| f(i) - g(i) \right|\right) / 2, \tag{15}$$

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

$$L = 1 - M. \tag{16}$$

Synthetic Data Sets

We created 20 synthetic data sets intended to produce hypothetical empirical detrital geochronology data sets and their associated uncertainties (Supplemental File 1[1]). Data sets were produced by first specifying the number of modal ages in the population (between 4 and 150 age modes, Fig. 2). Unless otherwise specified, age modes were randomly distributed between 10 and 3500 Ma so that all sample ages drawn from it would have geologically reasonable ages (Fig. 3). In the bimodally distributed population, modal ages where excluded from the central half of the interval from 0 to 3500 Ma and age modes were randomly drawn from values younger than 875 and older than 2625 Ma. In the centrally distributed population, modal ages were constrained to the central half of the interval from 0 to 3500 Ma (i.e., age modes were randomly drawn from between 875 and 2625 Ma). Centrally distributed and bimodally distributed samples were included to ensure that at least two samples have no overlapping age modes. Modal abundances for 10 of the distributions were specified a priori, while the other 10 were allowed to vary randomly (Fig. 2). The standard deviation about each mean modal age was also randomly assigned

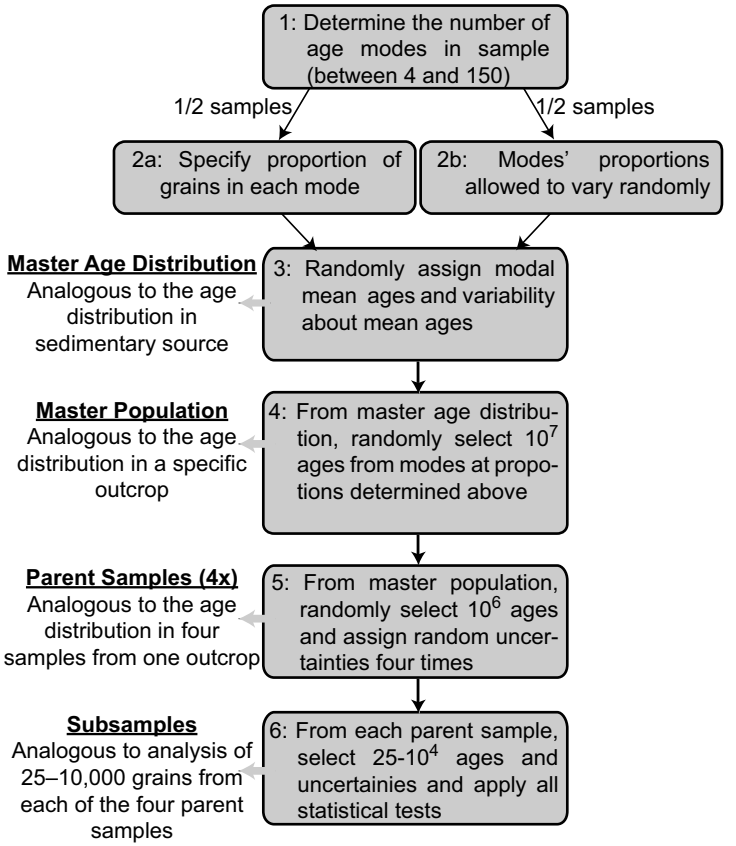
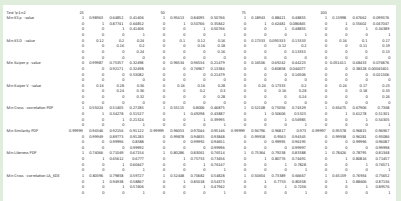


Figure 2. Flow chart of steps taken to produce the synthetic data sets. The text to the left of the boxes indicates the terms applied to the data sets produced in each step as well as the real-world analog for each of the steps.



[1]Supplemental File 1. Synthetic data sets used in this research. Please visit http://dx.doi.org/10.1130/GES01237.S1 or the full-text article on www.gsapubs.org to view Supplemental File 1.

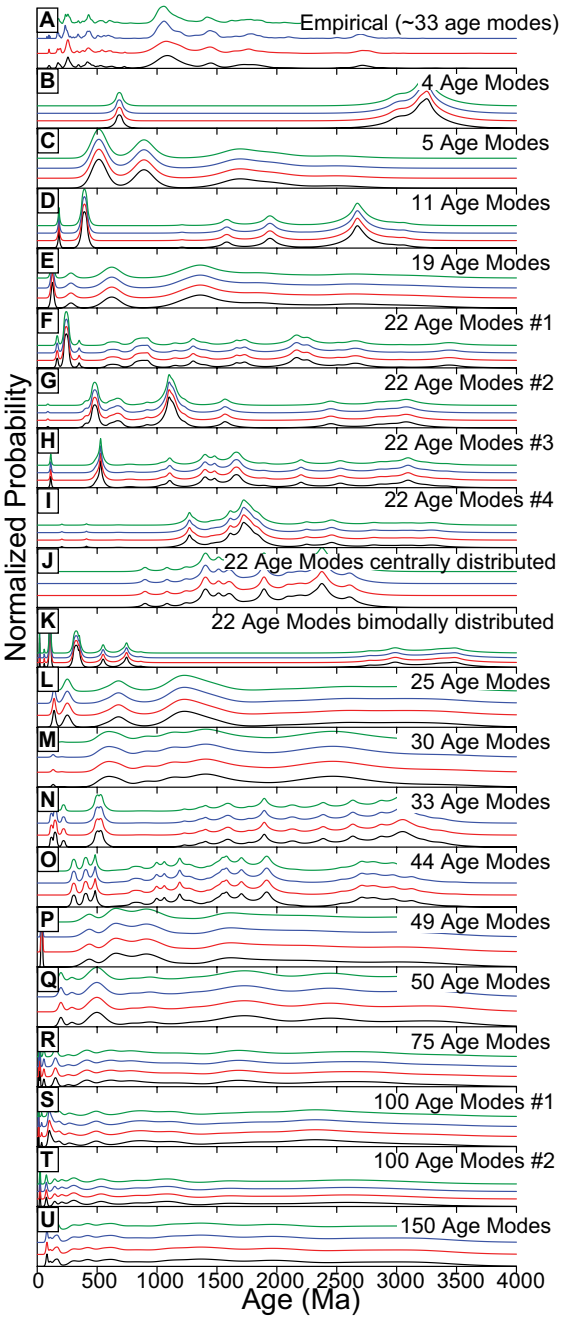


Figure 3. Normalized probability density plots of the empirical and 20 synthetic parent samples. Colors (black, red, blue, or green) correspond to the four samples drawn from parent populations in step 5 (parent samples) of Figure 2.

as much as 2% of the mean age for ages older than 900 Ma, and as much as 10% of the mean age for ages younger than 900 Ma. Randomization in these three steps was conducted using the random number generator, RAND( ), included in Microsoft Excel 2010, based on the algorithm of Matsumoto and Nishimura (1998), and generates pseudorandom numbers with replacement with a period of 2^{19937} (L'Ecuyer and Simard, 2007). See Mélard (2014) for a discussion of the statistical robustness of the RAND() function in Excel 2010. This procedure produces a continuous distribution that is analogous to the theoretical distribution of mineral ages in a hypothetical sediment source area (Fig. 2). In empirical studies, as in our model, it is impossible to fully reconstruct the continuous distribution from the discrete samples drawn from it. The closest we can achieve is a very good approximation.

Using the specified distributions and proportions of modal ages, we generated 10^7 ages and assigned uncertainties randomly between 0.5% and 10% of the randomly generated ages. These data sets would be analogous to the mineral ages in a particular outcrop (Fig. 2). Uncertainties assigned at this stage are analogous to the final reported age uncertainty, which incorporates both random and systematic uncertainties, and are intended to reflect uncertainties reported for multiple geochronometric systems. Hence the mean assigned uncertainty is slightly higher than is typically reported for zircon U-Pb ages, but slightly lower than typical (U-Th)/He or fission track ages. For this and subsequent steps randomization was conducted using the pseudorandom rand function in MATLAB. The rand function implements the ziggurat algorithm of Marsaglia and Tsang (1984) to generate matrices of pseudorandom, uniformly distributed values with replacement.

Four samples of size $n_2 = 10^6$ were then randomly drawn from the ($n_1 = 10^7$) population without replacement for further comparison using the randperm function in MATLAB. The randperm function implements the same algorithm as rand to create a randomly arranged vector matrix containing integers one through n_1 (inclusive). The first n_2 integers of the vector matrix were then used as indices to sample the population. These samples are analogous to four sandstone samples taken from an outcrop (Fig. 2). The size of the samples was selected so that a subsample of a size that could be reasonably produced during geochronological analysis would be <1% of the sample size. Four subsamples of between 25 and 10,000 ages and associated uncertainties were then randomly drawn without replacement from each sample using the randperm function. Each statistical method was applied to pairs of the subsamples, yielding a total of six comparisons for each test for every subsample size. Subsamples are analogous to the grains actually analyzed during geochronological analysis.

We applied each of the 13 metrics to 6 combinations of 38 subsamples of the 4 samples from each of the 20 synthetic populations for a total of 59,280 analyses. We also applied the comparison metrics to pairs from all unrelated data sets for a total of 190 unique combinations. We then applied the 13 metrics to these 190 combinations of 25 subsamples of 20 populations for a total of 61,750 analyses of different populations.

Empirical Data Set

Empirical data sets are from Pullen et al. (2014), who produced 4 large detrital zircon U-Pb data sets by analyzing between 962 and 1067 grains from a fluvial quartz arenite capping the upper Cretaceous Wahweap Formation (sample CP40 of Dickinson and Gehrels, 2008) in 4 separate trials (Fig. 3A). Analysis of a combined PDP incorporating all ages from the 4 trials indicates that it has 33 major age modes, identified as local maxima. We randomly drew 25 subsamples from each of the 4 data sets and applied each of the 13 metrics to all 6 pairs of the subsamples for a total of 1950 analyses of the empirical data sets.

RESULTS

We present the results of 4 representative synthetic populations with 5, 30, 50, and 100 age modes and the empirical data sets in detail below (Figs. 4–8), followed by a summary of the results from all tests (Fig. 9). The detailed results for the remainder of the populations are presented Supplemental File 2[2].

Identical Synthetic Populations

As expected for sampling an essentially infinite population, K-S and Kuiper mean p-values for samples of the same population show no trend with increasing sample size (Figs. 9A, 9B). However, unexpectedly, mean p-values calculated from multiple samples of the same population have high standard deviations for all subsample sizes between n = 25 and 10,000 (Figs. 4E, 4F, 5E, 5F, 6E, 6F, 7E, 7F). These two traits suggest that the p-values are extremely sensitive to minor variations in samples of the same population and so will be unreliable quantitative metrics of the similarity of two samples. For $\alpha = 0.05$, the standard null hypothesis would be rejected for as much as 73% and 58% of the sample pairs for at least one subsample size for the K-S and Kuiper tests, respectively (see minimum p-values in supplemental Figs. S1–S16 in Supplemental File 2). Because only 5% of the tests should be significant with $\alpha = 0.05$, the fact that as many as 73% of the tests are significant suggests that the K-S and Kuiper tests are too powerful.

Conversely, the Cross-correlation, Likeness, and Similarity coefficients of sample PDPs (Fig. 9C), and the D and V values (Figs. 9A, 9B) behave systematically, with each approaching their predicted asymptotes of 1 and 0, respectively, and decreasing standard deviation.

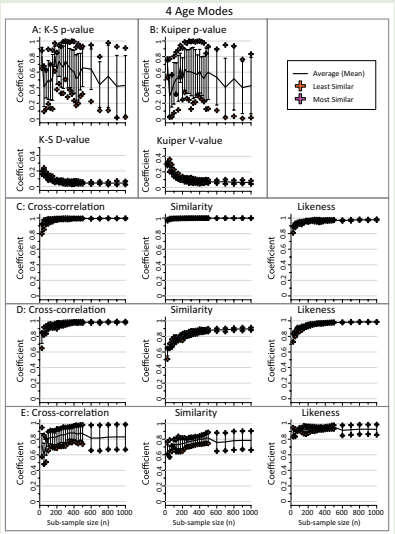
Counterintuitively, there is a nadir in Cross-correlation, Similarity, and Likeness coefficients between n = 50 and n = 400 when used to compare the KDE or LA-KDE (Figs. 4G–4I, 5G–5I, 6G–6I, 7G–7I, 9D, 9E). The nadir becomes more pronounced with increasing number of age modes, but is evident even with as few as five age modes (Fig. 4G). The minimum coefficient of the nadir as well as the subsample size at which coefficients begin to rise again are proportional to the number of age modes in the sample.

Empirical Samples

K-S and Kuiper tests of the four empirical samples yield an inverse relationship between the mean p-values and data set size with large standard deviations for all values of n (Figs. 8E, 8F). In contrast, mean Similarity, Likeness, and Cross-correlation coefficients consistently increase, and measurements from comparison of individual sample pairs converge with increasing data set size (Figs. 8G, 8H, 8I). However, for Cross-correlation, the absolute increase in the coefficient and decrease in standard deviation is more dramatic than in any of the alternatives.

Different Synthetic Populations

K-S and Kuiper test p-values for different populations decrease with increasing data set size (Figs. 10A, 10B). Mean p-values rapidly drop below 0.05 for n > 75. However, maximum p-values remain above 0.05 for n < 1000 and n < 475 for the K-S and Kuiper tests, respectively (Figs. 10A, 10B). In contrast, the mean, maximum, and minimum D and V values show little change with increasing subsample size and yield mean values of ~0.35 and 0.5, respectively, and a range that spans almost the entire possible range of coefficients. The PDP Cross-correlation, Likeness, and Similarity coefficients are also relatively constant with increasing n. However, PDP Cross-correlation yields maximum values between 0.5 and 0.65 for n > 100, and mean values < 0.1 (Fig. 10C). PDP Similarity and Likeness yield coefficients that span almost the entire range of possible coefficients and have means of 0.7–0.8 and 0.5–06, respectively. Maximum Cross-correlation, Similarity, and Likeness coefficients for KDEs or LA-KDEs decrease dramatically with increasing n (Figs. 10D, 10E). As with PDP Cross-correlation, the KDE and LA-KDE Cross-correlation yields mean coefficients that are typically <0.1. Mean KDE and LA-KDE Similarity coefficients decrease from ~0.7 to between 0.3 and 0.4 between n = 25 and n = 200 with little change for larger subsample sizes. Mean KDE and LA-KDE Likeness coefficients decrease from 0.4 to 0.5 for n = 25 to ~0.2 for n = 200 and do not change significantly for larger subsamples sizes.



[2]Supplemental File 2. Plots of detailed results of 16 samples not presented in Figures 4–7. MATLAB files, executable file, and User Manual for DZstats. Please visit http://dx.doi.org/10.1130/GES01237.S2 or the full-text article on www.gsapubs.org to view Supplemental File 2.

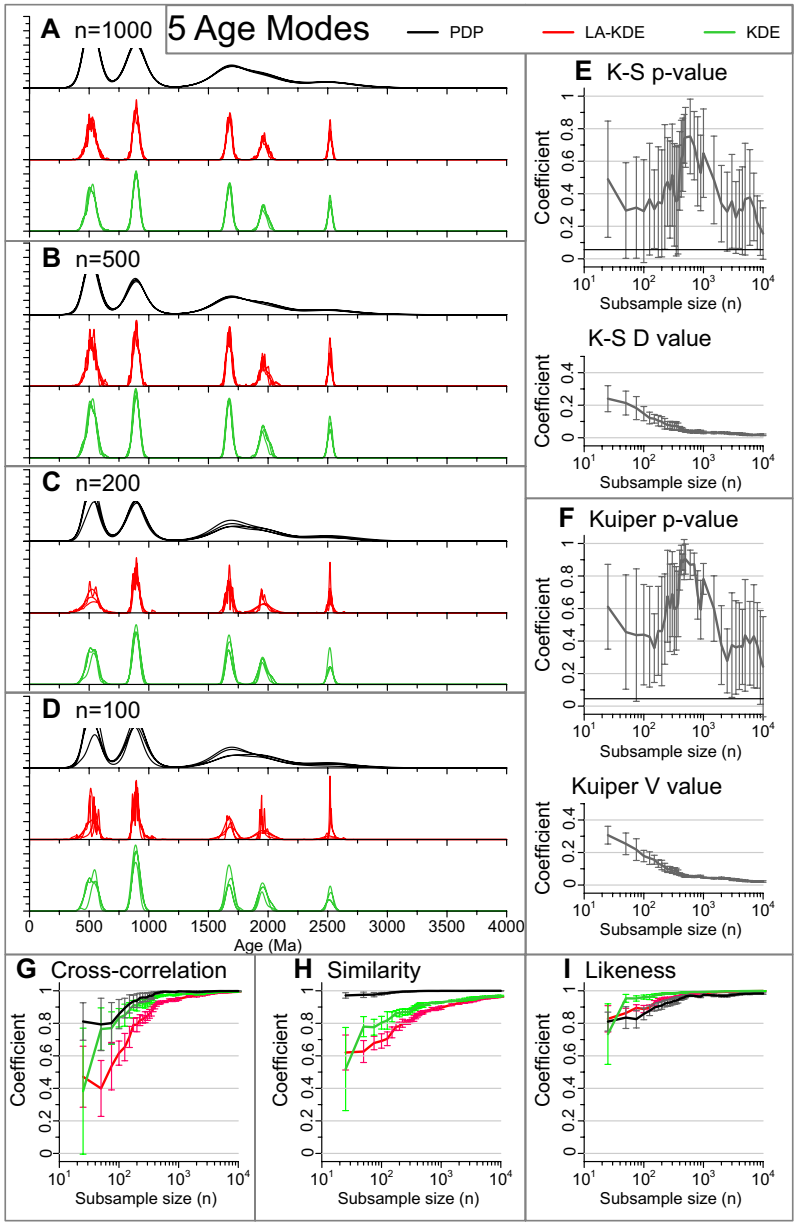


Figure 4. Summary of data and comparison for the synthetic population with 5 age modes. Probability density plots (PDPs, in black), adaptive Kernel density estimates (KDEs, in green), and locally adaptive (LA) KDEs (in red) are shown for subsamples. (A) For n = 1000. (B) For n = 500. (C) For n = 200. (D) For n = 100. X-axis for density distributions is as the lowest sample. (E) Mean and standard deviation for Kolmogorov-Smirnov (K-S) test p-values and D values. (F) Mean and standard deviation for Kuiper test p-values and V values. (G) Coefficients of PDPs, KDEs, and LA-KDEs calculated using Cross-correlation. (H) Using Similarity. (I) Using Likeness. 1σ error bars calculated from repeated subsampling of the four samples. K-S and Kuiper tests are calculated independently of either PDPs or KDEs.

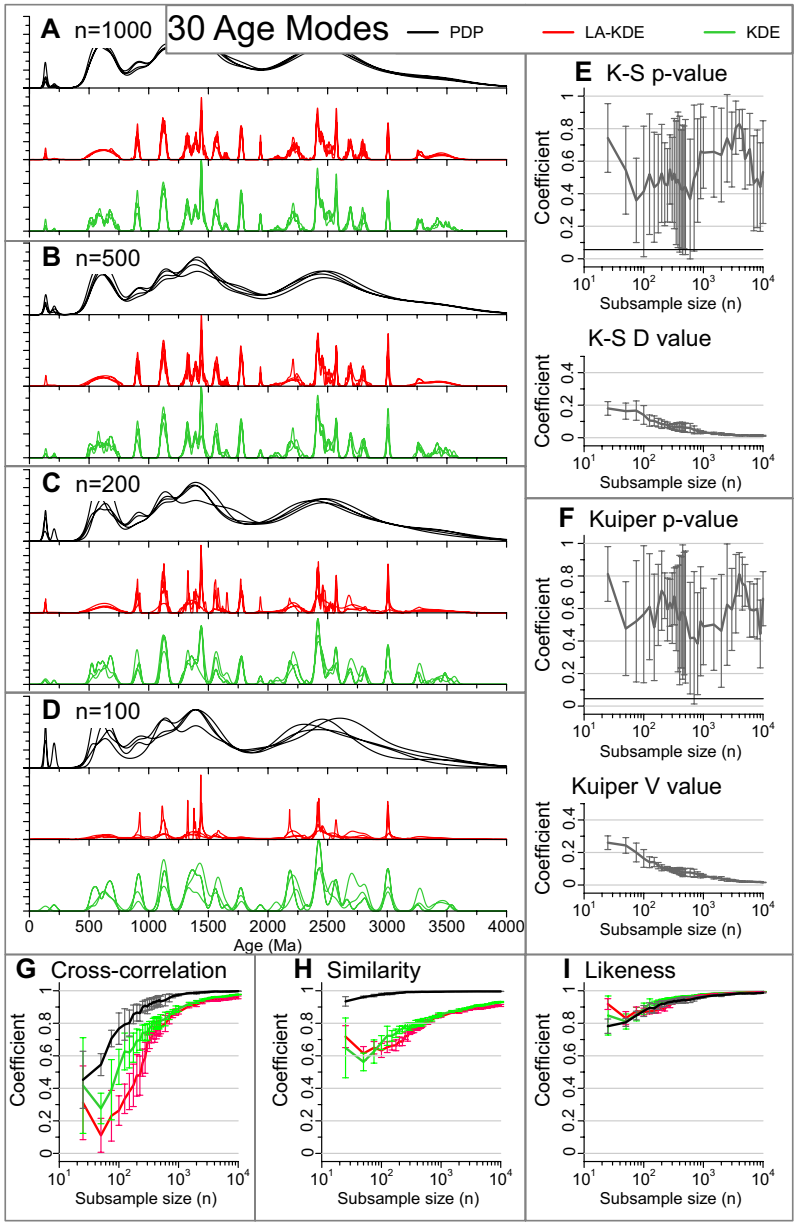


Figure 5. Summary of data and comparison for the synthetic population with 30 age modes. Probability density plots (PDPs in black), adaptive Kernel density estimates (KDEs, in green) and locally adaptive (LA) KDEs (in red) are shown for subsamples. (A) For n = 1000. (B) n = 500. (C) n = 200. (D) n = 100. X-axis for density distributions is as the lowest sample. (E) Mean and standard deviation for Kolmogorov-Smirnov (K-S) test p-values and D values. (F) Mean and standard deviation for Kuiper test p-values and V values. (G) Coefficients of PDPs, KDEs, and LA-KDEs calculated using Cross-correlation. (H) Calculated using Similarity. (I) Calculated using Likeness. 1σ error bars are calculated from repeated subsampling of the four samples. K-S and Kuiper tests are calculated independently of either PDPs or KDEs.

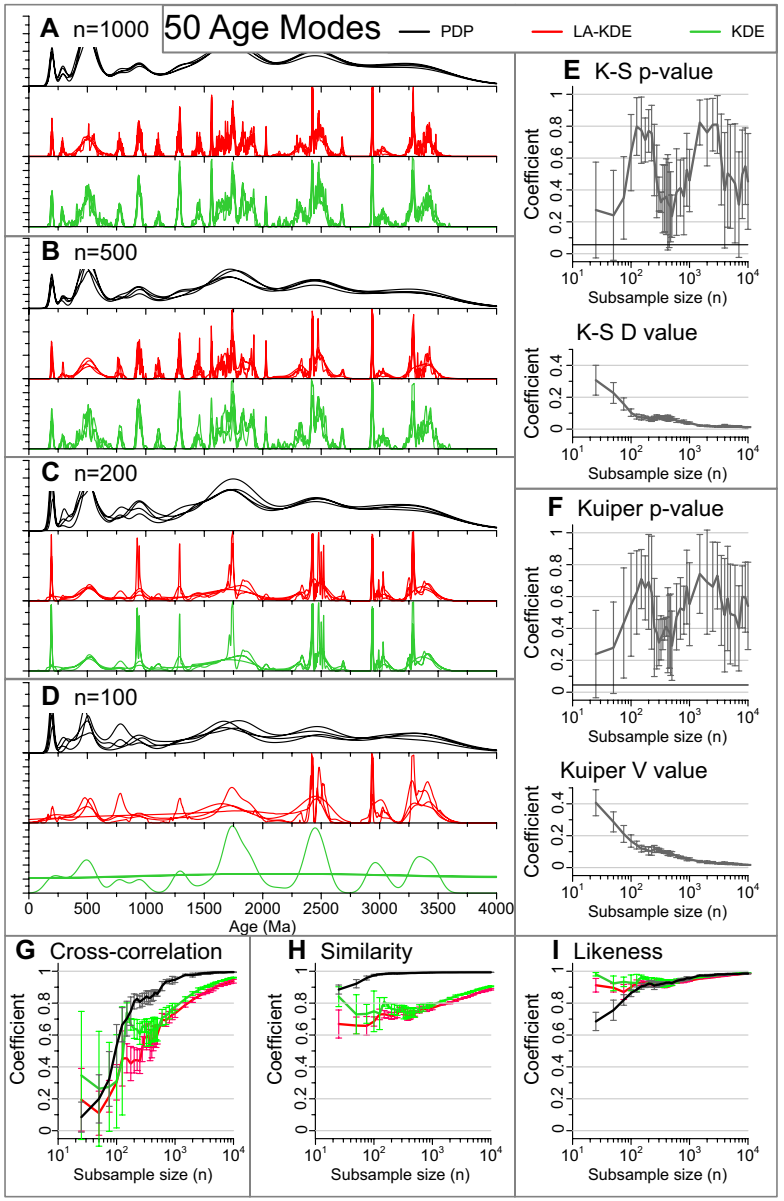


Figure 6. Summary of data and comparison for the synthetic population with 50 age modes. Probability density plots (PDPs in black), adaptive Kernel density estimates (KDEs, in green) and locally adaptive (LA) KDEs (in red) are shown for subsamples. (A) For n = 1000. (B) For n = 500. (C) For n = 200. (D) For n = 100. X-axis for density distributions is as the lowest sample. The tendency of the KDE algorithm to oversmooth the distribution is particularly evident in D. (E) Mean and standard deviation for Kolmogorov-Smirnov (K-S) test p-values and D values. (F) Mean and standard deviation for Kuiper test p-values and V values. (G) Coefficients of PDPs, KDEs, and LA-KDEs calculated using Cross-correlation. (H) Calculated using Similarity. (I) Calculated using Likeness. 1σ error bars calculated from repeated subsampling the four samples. K-S and Kuiper tests are calculated independently of either PDPs or KDEs.

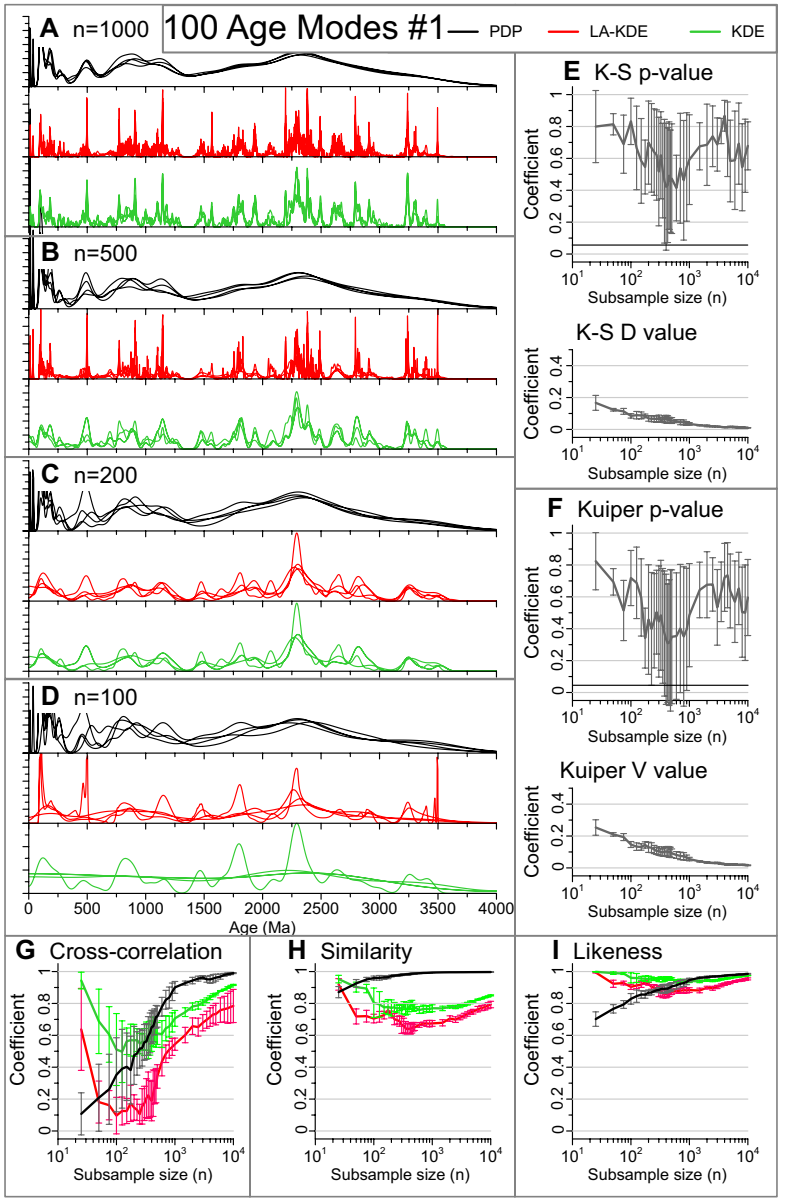


Figure 7. Summary of data and comparison for the synthetic population with 100 age modes. Probability density plots (PDPs in black), adaptive Kernel density estimates (KDEs, in green) and locally adaptive (LA) KDEs (in red) are shown for subsamples. (A) For n = 1000. (B) For n = 500. (C) For n = 200. (D) For n = 100. X-axis for density distributions is as the lowest sample. The tendency of the KDE algorithm to oversmooth the distribution is particularly evident in D. (E) Mean and standard deviation for Kolmogorov-Smirnov (K-S) test p-values and D values. (F) Mean and standard deviation for Kuiper test p-values and V values. (G) Coefficients of PDPs, KDEs, and LA-KDEs calculated using Cross-correlation. (H) Calculated using Similarity. (I) Calculated using Likeness. 1σ error bars calculated from repeated subsampling the four samples. K-S and Kuiper tests are calculated independently of either PDPs or KDEs.

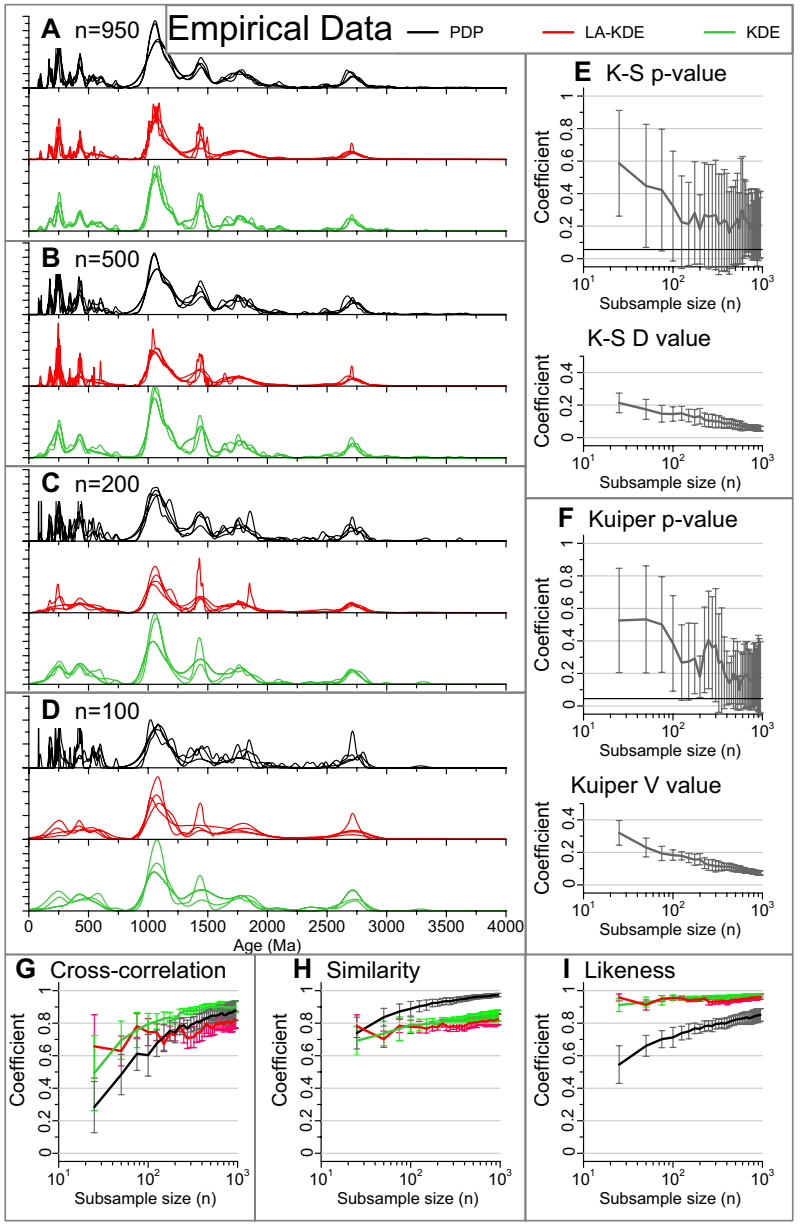


Figure 8. Summary of data and comparison for the four empirical data sets (Pullen et al., 2014). Probability density plots (PDPs in black), adaptive Kernel density estimates (KDEs, in green) and locally adaptive (LA) KDEs (in red) are shown for subsamples. (A) For n = 950. (B) For n = 500. (C) For n = 200. (D) For n = 100. (E) Mean and standard deviation for Kolmogorov-Smirnov (K-S) test p-values and D values. (F) Mean and standard deviation for Kuiper test p-values and V values. (G) Coefficients of PDPs, KDEs, and LA-KDEs calculated using Cross-correlation. (H) Calculated using Similarity. (I) Calculated using Likeness. 1σ error bars calculated from repeated subsampling the four samples. K-S and Kuiper tests are calculated independently of either PDPs or KDEs.

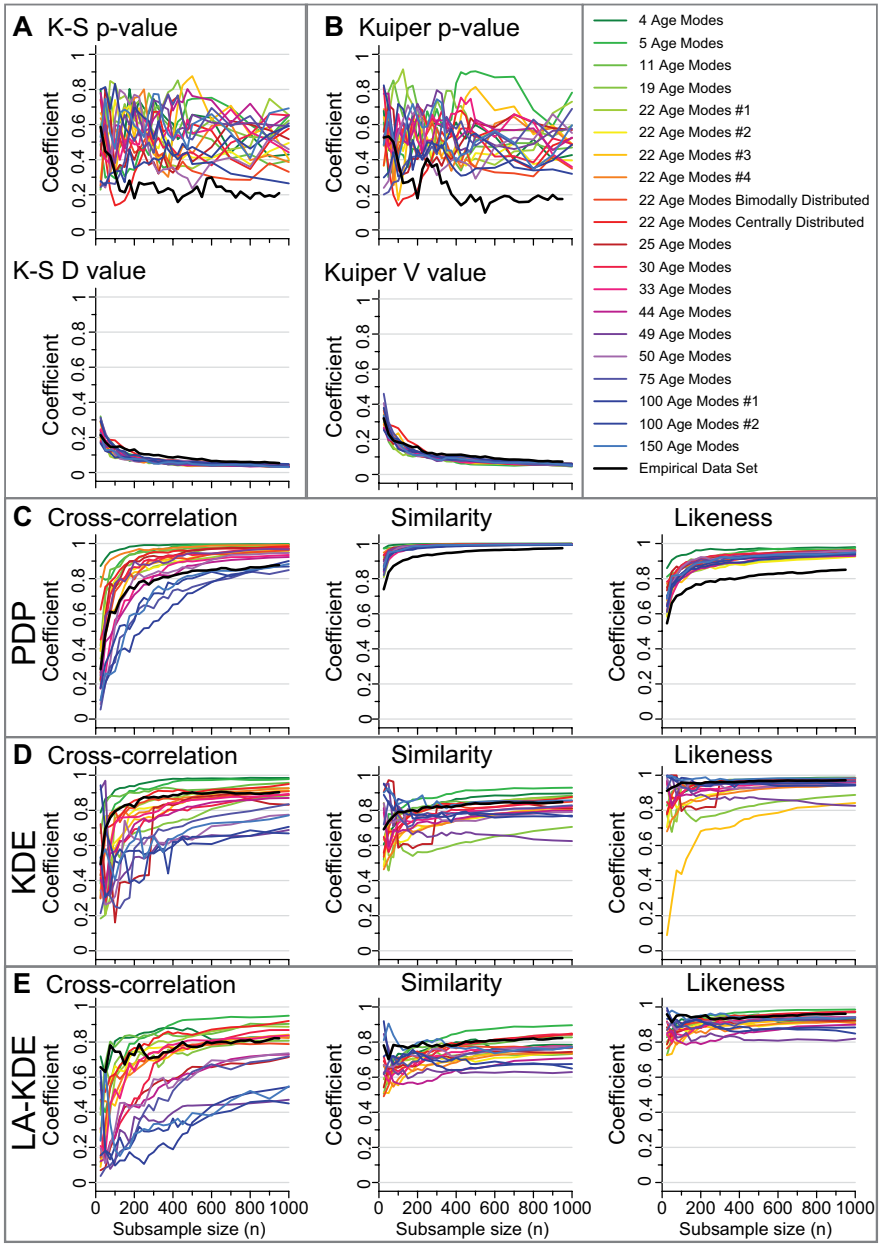


Figure 9. Summary of metrics for all synthetic populations and the empirical data set. (A) Kuiper test p-value and D value. (B) Kuiper test p-value and V values. (C) Cross-correlation, Similarity, Likeness coefficients from comparison of subsample probability density plots (PDPs). (D) Cross-correlation, Similarity, Likeness coefficients from comparison of subsample Kernel density estimates (KDEs). (E) Cross-correlation, Similarity, Likeness coefficients from comparison of subsample locally adaptive (LA) KDEs. X-axis for all plots is as the lowest row. Kolmogorov-Smirnov (K-S) and Kuiper tests are calculated independently of either PDPs or KDEs.

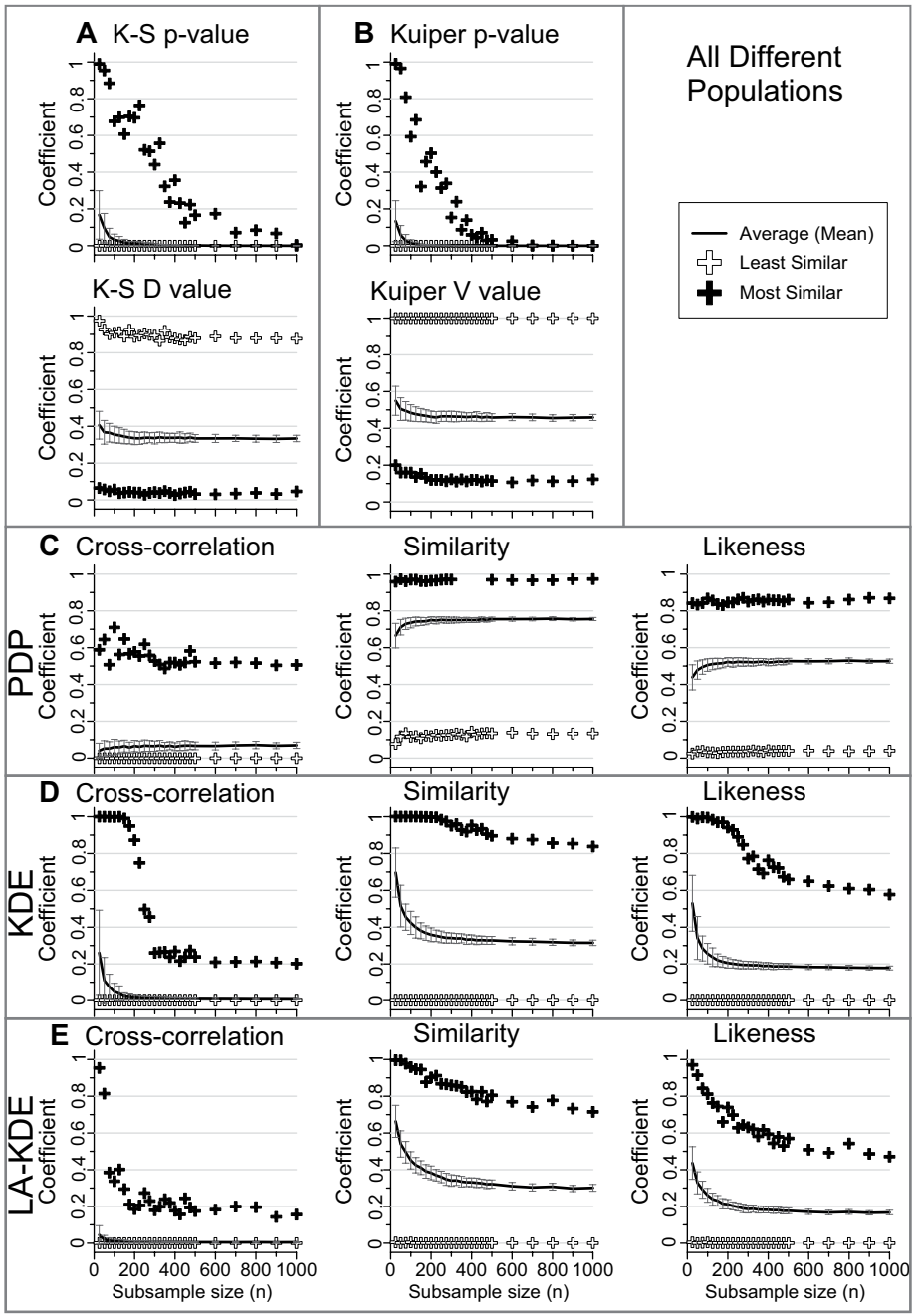


Figure 10. Summary of results of comparison of all 20 different synthetic populations (190 comparison pairs for each subsample size). The solid black line is the average of all 190 comparison pairs for each subsample size. Error bars show the 1σ based on the 190 comparisons. White crosses indicate the least similar of the 190 sample pairs for each subsample size. Black crosses indicate the most similar sample pair for each subsample size. X-axis for all plots is as the lowest row. K-S—Kolmogorov-Smirnov; PDP—probability density plot; KDE—Kernel density estimate; LA—locally adaptive.

DISCUSSION

In the following discussion, we test each metric's ability to quantify subsample similarity. We then evaluate each metric as a hypothesis test to discriminate subsamples from the same versus different populations. Finally, we suggest interpretive guidelines for metrics that are strictly a measure of similarity lacking a clear hypothesis test criterion.

Metrics as Descriptors of Sample Similarity

Criteria for Evaluation

We use the following three criteria to evaluate each of the metrics as quantitative descriptors of sample similarity (Table 1). (1) For subsamples of the same population, the comparison metric should show a monotonic increase in measured subsample similarity with increasing subsample size. (2) Comparison metrics should maximize the range of possible coefficient values (0–1). (3) The comparison metric should minimize nonsystematic artifacts either due to subsample size or the complexity of the sample (number of age modes).

As the size of randomly selected subsamples of a population increases, they will be increasingly skewed toward modes that contribute greater proportions to the population (Andersen, 2005). A perfect match between subsamples and a parent population is therefore unlikely. However, the probability of nearly matching, for example within one standard deviation, increases with increasing data set size (Andersen, 2005). In order to take advantage of the increasing likelihood of nearly matching despite the decreasing likelihood of an exact match with larger data set sizes, we incorporate the standard deviation of the statistical metrics in the following discussion. Standard deviation is calculated based on the 6 pairings of 4 samples drawn from each population in the case of comparison of identical populations, or based on intercomparison of all 20 different populations when analyzing different populations. Because this variability is inherent in sampling a multimodal population with varying modal proportions, we calculate the mean and standard deviation of the comparison metric for all sample pairs (Figs. 4–8).

K-S and Kuiper Tests

Use of the K-S or Kuiper test p-values for quantitative similarity analysis of detrital geochronological data sets is likely to lead to incorrect conclusions (Satkoski et al., 2013; Vermeesch, 2013). The K-S and Kuiper tests yield highly variable, nonsystematic p-values when applied to subsamples drawn from the same synthetic population, i.e., there is no convergence in p-values for larger subsamples sizes (Figs. 9A, 9B). Taken as a measure of similarity (which the p-value is not, though it is sometimes treated as one), this would suggest that samples do not become more similar with increasing data set size. Interpreted in the same way, the decrease in p-values with increasing subsample

TABLE 1. EVALUATION OF ALL COMPARISON METRICS CONSIDERED IN THIS STUDY USING THE FOUR CRITERIA OUTLINED IN THE TEXT

Goal	Quantitative assessment of similarity			Hypothesis test
Criterion	1. Systematic increase in measured similarity with subsample size	2. Utilize the entire range of possible coefficients	3. Minimize artifacts due to sample size or population complexity	Consistently discriminate same versus different populations*
K-S p-value	*No*	**Yes**	*No*	Mean: Yes Individual: No
K-S D value	**Yes**	*No*	**Yes**	n > 950
Kuiper p-value	*No*	**Yes**	*No*	Mean: Yes Individual: No
Kuiper V value	**Yes**	*No*	**Yes**	n > 500
PDP Cross-correlation	**Yes**	**Yes**	**Yes**	n > 300
PDP Similarity	**Yes**	*No*	**Yes**	*No*
PDP Likeness	**Yes**	*No*	**Yes**	*No*
KDE Cross-correlation	*No*	**Yes**	*No*	n > 375
KDE Similarity	*No*	*No*	*No*	*No*
KDE Likeness	*No*	*No*	*No*	n > 350
LA-KDE Cross-correlation	*No*	**Yes**	*No*	n > 500
LA-KDE Similarity	*No*	*No*	*No*	*No*
LA-KDE Likeness	*No*	*No*	*No*	n > 400

Note: K-S—Kolmogorov-Smirnov; PDP—probability density plot; KDE—kernel density estimate; LA—locally adaptive (see text).
*Maximum similarity metric from different populations < minimum similarity metric from the same population and empirical data set consistently within the region of the same population.

size of the empirical data set would suggest a decrease in sample similarity with larger data sets when applying the K-S or Kuiper tests (Figs. 8E, 8F). As a measure of similarity, p-values fail the first and third criteria: they do not reflect increasing subsample similarity and they incorporate artifacts that are a function of subsample size rather than subsample similarity. The K-S and Kuiper test p-values are also sensitive to the number and distribution of population age modes when applied to the synthetic data sets (Fig. 9). The high standard deviation further suggests that they are unreliable indicators of sample similarity because the p-values vary dramatically from repeated selection of samples of the same size from the same population. Because greater p-values do not correlate with greater sample similarity (Figs. 9A, 9B), they cannot be used to assess degrees of sample similarity or mixing proportions. We therefore turn to the alternative metrics: the K-S D values, Kuiper V values, and the coefficients for Similarity, Likeness, and Cross-correlation.

As noted by Vermeesch (2013), the D or V values provide more robust assessment of the dissimilarity between samples than do p-values. However, both D and V values also fail the second criterion: they typically vary by <0.4 with increasing subsample size from 25 to 10,000 (Figs. 9A, 9B), suggesting that they are relatively insensitive to differences between samples. The D value is most sensitive to variation in CDFs near the median. In addition, D values and, to a lesser extent, V values are most sensitive to the relative proportions of age modes rather than the mean age of age modes. They will return low D or V values for samples whose modal ages do not overlap as long as the relative proportion of the age modes are similar and age modes are approximately equidistant (Fig. 11C). In the example shown in Figure 11, comparison of distributions A and D, which share no age modes, returns a lower D value (0.2) than comparison of B and D, which share 3 of 5 age modes (D = 0.4). V values of comparison of both A-D and B-D yield V values of 0.4 (Table 2). Thus, they do not reflect the inversion of expectations observed in D values, nor do they reflect the clear correlation between sample age modes.

Cross-Correlation

PDP Cross-correlation fulfills all of the criteria laid out here. It shows a systematic increase in subsample similarity with increasing subsample size, while utilizing almost the full range of possible values (Fig. 9C). It does not reflect artifacts either due to subsample size or the complexity of the sample (number of age modes). However, as with most of the other metrics considered, Cross-correlation is a poor indicator of sample similarity when applied to either the KDE or LA-KDE (Figs. 9D, 9E). When applied to the KDE or LA-KDE of complex samples, Cross-correlation values decrease over the range from n = 25 to n = 300 (Figs. 4G, 5G, 6G, 7G, 9D, 9E). Because there is no way to know, a priori, whether a sample is complex or not, this requires analysis of at least 300 ages before cross-correlation can be confidently applied to assess sample similarity using either the KDE or LA-KDE. We attribute this decrease in mean Cross-correlation coefficients to occasional oversmoothing, and hence low Cross-correlation coefficients between subsamples either subject to oversmoothing or not, at low to moderate subsample sizes (Figs. 6D and 7D).

For applications requiring a measure of sample dissimilarity such as multidimensional scaling (MDS; Vermeesch, 2013), the compliment of PDP Cross-correlation (the coefficient of nondetermination, $1 - R^2$) may provide a suitable alternative to D or V values (Fig. 11). In this limited experiment, PDP Cross-correlation, Similarity, and Likeness reproduce the degree of overlap of the age modes better than either D or V values (Figs. 11A, 11C). Of these three, PDP Cross-correlation allowed for three-dimensional (3-D) MDS, providing greater resolution of intersample distance than did two-dimensional (2-D) (Similarity or Likeness) MDS (Fig. 11B).

Similarity and Likeness Coefficients

We consider Similarity and Likeness together because they show many similar results. When applied to subsample PDPs the Similarity and Likeness coefficients systematically increase with increasing subsample size, fulfilling criteria one and three (Fig. 9C). However, like the D and V values, PDP Similarity and Likeness coefficients show minimal variation, and so fail criterion two. Minimum Similarity coefficients are typically >0.6 (for n = 25) and increase rapidly to ~1. Minimum Likeness coefficients are typically >0.5 and rapidly increase to between 0.8 and 1.

When applied to KDEs or LA-KDEs of subsamples of the same population the Similarity coefficient decreases over subsample sizes as large as n = 1000, depending on the population complexity (Figs. 9D, 9E), failing criteria one and three. As with the PDP Similarity and Likeness coefficients, these show minimal (<0.4) variation with increasing subsample size.

Metrics as Hypothesis Tests

Our criterion for evaluating the suitability of a metrics as a hypothesis test is simply that it should be able to consistently discriminate subsamples drawn from the same population from those drawn from different populations. We also determine the minimum sample size at which the least similar result for samples drawn from the same parent population do not overlap the most similar result for comparison of different data sets (see Fig. 12; summary in Table 1). These provide conservative data set sizes for application of their respective statistical metrics to discriminate identical from different source populations.

K-S and Kuiper P-Values

Use of individual K-S or Kuiper test p-values as hypothesis tests to evaluate sourcing of sediments from a common parent is problematic. Comparing subsamples of the synthetic populations with n > 125, at least 1 of the 6 pairs yield K-S p-values < 0.05 in as much as 73% of the subsamples (mean of 14%

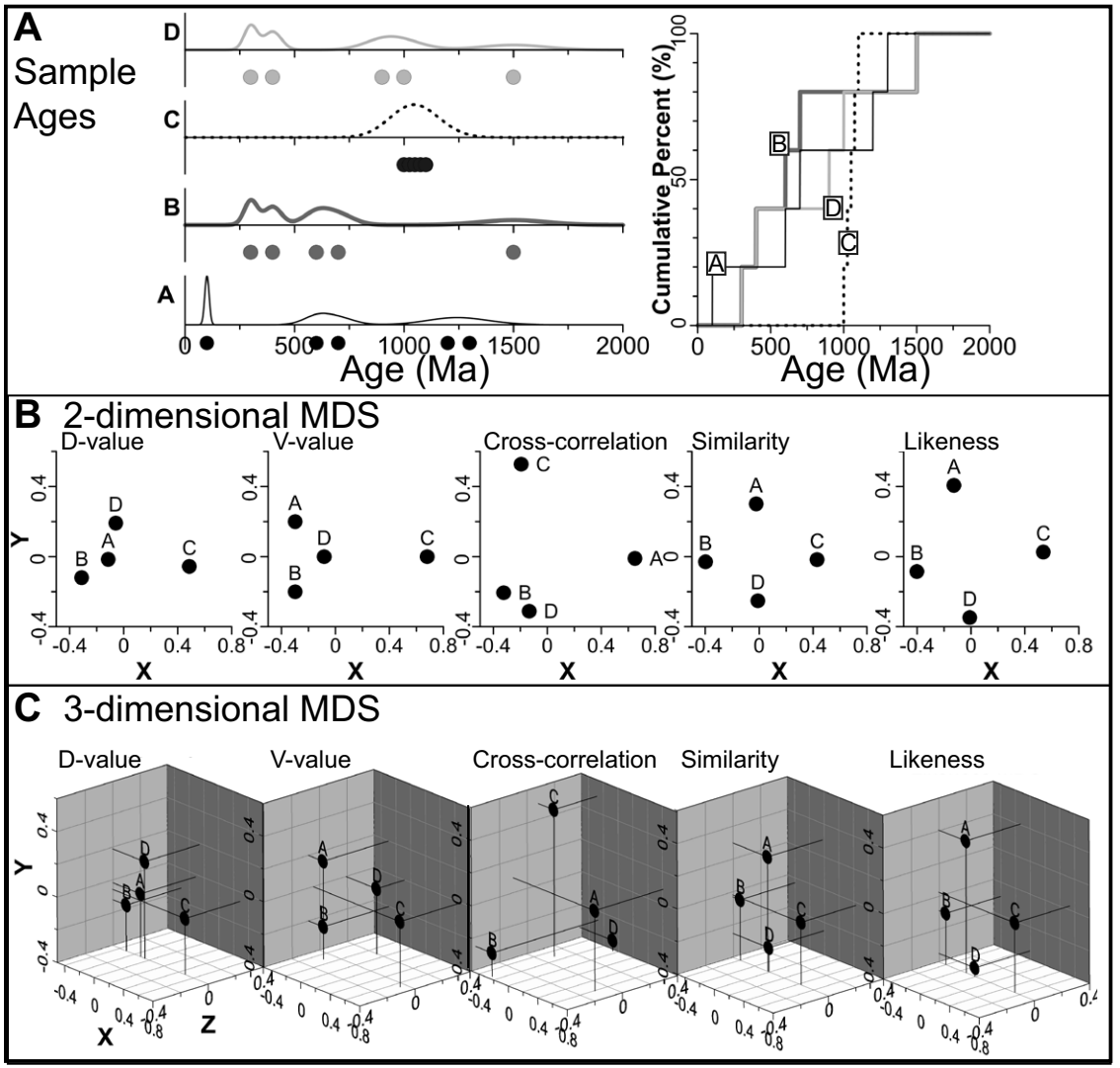


Figure 11. (A) Probability density plots and cumulative distribution plots of hypothetical samples A–D. Each sample has 5 modal ages, with an associated uncertainty of 10%. (B) Two-dimensional (2-D) nonmetric multidimensional scaling (MDS) of the distributions shown in A using the various metrics evaluated in this paper. (C) Three-dimensional (3-D) nonmetric MDS. Only 3-D MDS based on V values and Cross-correlation have a z-axis component. Because the z-axis for D values, Likeness, and Similarity are 0, they are effectively 2-D.

for the 20 synthetic data sets; Figs. S1–S16 in Supplemental File 2). Similarly, the Kuiper test yields p-values <0.05 in as much as 58% of the subsamples (mean of 51% for the 20 synthetic data sets). This suggests that individual K-S and Kuiper p-values are not reliable indicators as hypothesis tests, because too often they reject the null hypothesis at a higher rate than predicted for the selected p-value. We attribute this to the distortion introduced into the distribution of small sample sizes by random sampling of a larger population. Mean K-S and Kuiper p-values calculated from repeated subsampling (a minimum of four trials), however, behave as expected: not rejecting the hypothesis of a common source for subsamples of the same population (Figs. 9A, 9B). The p-values consistently reject the null hypothesis when applied to sufficiently large subsamples of different populations (Figs. 10A, 10B). We conclude that use of K-S or Kuiper p-values as binary hypothesis tests requires random subsampling without replacement of geochronological data sets and calculation

TABLE 2. RETURNED VALUES FOR EACH METRIC EXPLORED IN THIS STUDY FOR THE SAMPLES SHOWN IN FIGURE 11

D value	A	B	C	D
A	0.00	0.20	0.60	0.20
B	0.20	0.00	0.80	0.40
C	0.60	0.80	0.00	0.60
D	0.20	0.40	0.60	0.00
V value	**A**	**B**	**C**	**D**
A	0.00	0.40	1.00	0.40
B	0.40	0.00	1.00	0.40
C	1.00	1.00	0.00	0.80
D	0.40	0.40	0.80	0.00
PDP Cross-correlation	**A**	**B**	**C**	**D**
A	1.00	0.01	0.00	0.08
B	0.01	1.00	0.12	0.21
C	0.00	0.12	1.00	0.08
D	0.08	0.21	0.08	1.00
PDP Similarity	**A**	**B**	**C**	**D**
A	1.00	0.61	0.51	0.46
B	0.61	1.00	0.17	0.73
C	0.51	0.17	1.00	0.62
D	0.46	0.73	0.62	1.00
PDP Likeness	**A**	**B**	**C**	**D**
A	1.00	0.50	0.26	0.24
B	0.50	1.00	0.06	0.65
C	0.26	0.06	1.00	0.39
D	0.24	0.65	0.39	1.00

Note: PDP—probability density plot (see text).

of mean p-values from multiple applications of these tests to the randomly generated subsamples. Robust application of mean K-S p-values to discriminate identical versus different sources requires sample sizes >1000 because maximum p-values remain above 0.05 for subsamples of different populations with n < 1000. Maximum Kuiper p-values >0.05 for n < 475 indicate that sample sizes ≥475 are required for application of mean Kuiper p-values to discriminate identical versus different sources.

D Values and V Values

The conclusion that D and V values are relatively insensitive to sample dissimilarity noted here is borne out when they are used to discriminate subsamples drawn from different populations. The minimum D values for subsamples of different parent samples are less than maximum values for subsamples from the same parent sample for all n < 950 (Fig. 12A). At the same time, D values of the empirical data set overlap those from subsamples of different synthetic populations for all n < 600 (Fig. 12A). They are within 1σ of the minimum coefficient from different populations for all n < 1000. For Kuiper test V values, the crossover in these metrics for identical versus different populations occurs at n = 600. Similarly, V values from the empirical data set overlap V values from subsamples of different synthetic populations for n < 300 and are within 1σ for n < 600. We interpret this as indicating that the D and V values are insufficiently sensitive to consistently discriminate comparison of subsamples from different or the same population for n < 300–600 and may continue to overlap values expected from different populations for significantly larger sample sizes.

Cross-Correlation

Cross-correlation is able to consistently discriminate samples from identical versus different populations for sample sizes that can reasonably be produced on a regular basis. PDP, KDE, and LA-KDE Cross-correlations of subsamples of the empirical data set are well within the field of coefficients of subsamples of the same population (Fig. 12). Of the three, the maximum coefficients for Cross-correlation of two PDPs of unrelated subsamples diverge from the minimum coefficients for subsamples drawn from the same parent sample at the smallest sample sizes, n > 300 (Fig. 12C). In contrast, we attribute maximum KDE Cross-correlation coefficients of 1 from samples of different populations at n < 175 to oversmoothing of the KDE, and hence minimization of the differences between samples, at these low to moderate sample sizes (e.g., Figs. 6D, 7D, and 10D). We note that the mean KDE Cross-correlation coefficient remains low over these sample sizes (Fig. 9D), consistent with only occasional oversmoothing by the KDE bandwidth optimization algorithm. For sample sizes > 375 the KDE Cross-correlation coefficient is able to consistently distinguish samples drawn from the same parent (coefficients >~0.5) from those drawn from different parents (coefficients <~0.2) (Fig. 12F). For sample sizes >475 the LA-KDE Cross-correlation coefficient is able to consistently distinguish samples drawn from the same population versus those from different populations. However, the difference in coefficients between these two sets is less than for the PDP or KDE Cross-correlation coefficients (Fig. 12I).

Similarity and Likeness

For PDP Similarity maximum coefficients for different populations and minimum coefficients from the same population overlap for n < 200, though the two curves are virtually indistinguishable for larger n. For PDP Likeness, the crossover occurs at n = 525. Unexpectedly, when applied to subsamples of different populations the mean PDP Similarity and Likeness coefficients show a minor increase for very small subsample size but show no change for subsamples >200 (Fig. 10C). However, KDE and LA-KDE Similarity and Likeness

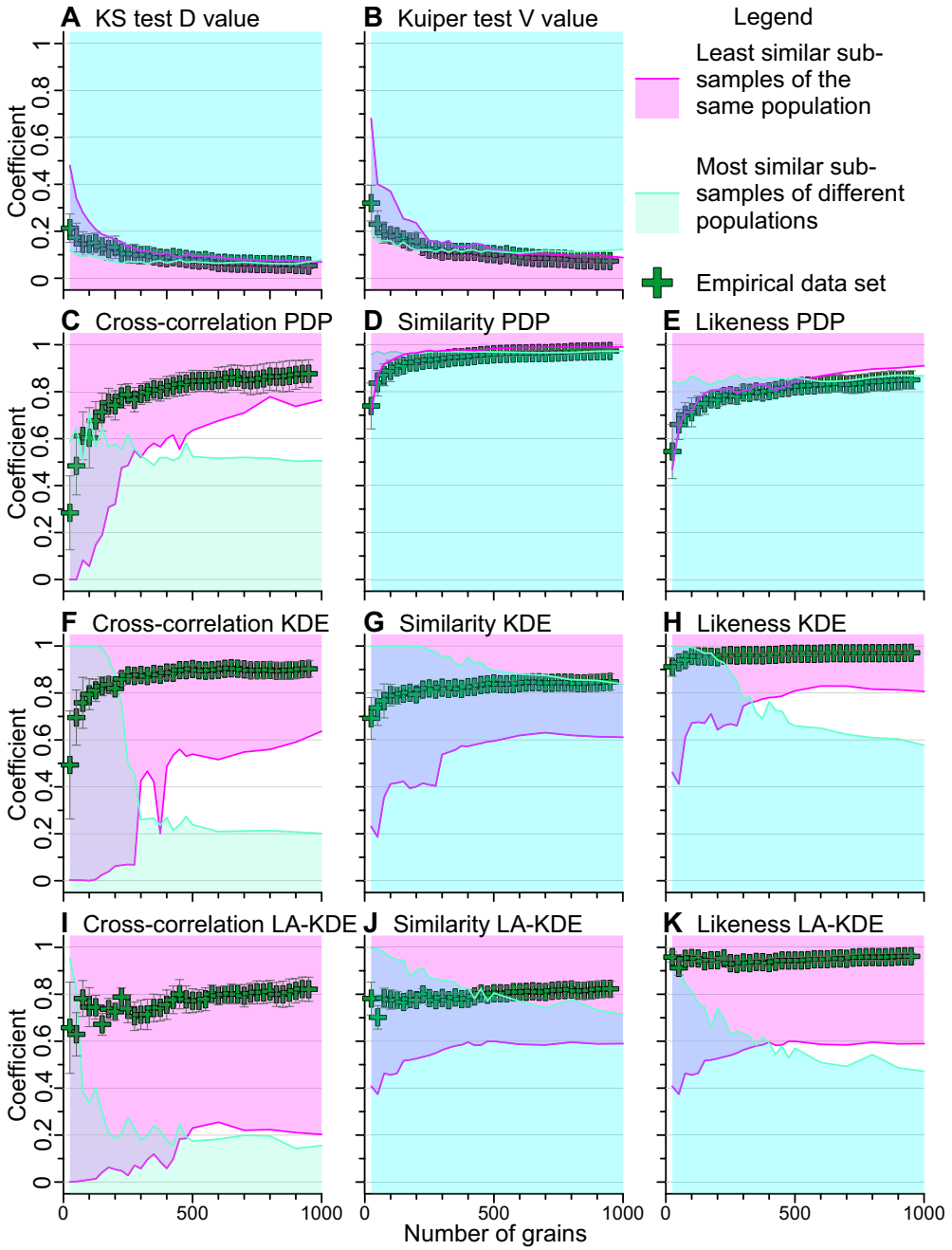


Figure 12. Comparison of the most similar subsamples of different populations and the least similar subsamples from identical populations indicates that there is significant overlap in these coefficients for low to moderate sample sizes. The blue field is the region occupied by coefficients from intercomparison of all 20 different populations. The purple field indicates the region populated by comparison of the 4 samples of each of the 20 populations for each subsample size. X-axis for all plots is as the lowest row. KS—Kolmogorov-Smirnov; PDP—probability density plot; KDE—Kernel density estimate; LA—locally adaptive.

show a much larger decrease in mean coefficients over sample sizes of as much as 300 (Figs. 10D, 10E).

Similarity yields overlapping coefficients when comparing KDEs or LA-KDEs samples of the same population or different populations for all n < 1000, suggesting that it is insufficiently sensitive to discriminate identical versus different populations. However, KDE and LA-KDE Likeness coefficients for identical versus different populations diverge at n >350 or 400, respectively. For these tests, the empirical data set yields coefficients well within the range of values expected from samples of the same population. We conclude that Likeness may be better used in conjunction with KDEs or LA-KDEs rather than PDPs. However, KDE Likeness yields coefficients >0.9 for samples of different populations to n = 200, reflecting the tendency of the KDE to oversmooth at low to moderate sample sizes, and suggesting that larger sample sizes are needed for robust application of the Likeness coefficient.

Interpretive Guidelines

The discussion herein provides a method for defining conservative interpretative guidelines for coefficient metrics, including those that lack a clear hypothesis test. Rather than a determining a single critical coefficient that defines whether samples are drawn from the same population, Figure 9 indicates that all the metrics except p-values vary as a function of sample size. Hence, the guidelines are expressed as a function of sample size such that for a given sample size, if the coefficient is within the field occupied by subsamples of the same population in Figure 12, we cannot reject the hypothesis that the samples were drawn entirely from the same parent population. For example, for a sample of n = 500 grains, if the PDP Cross-correlation coefficient is >0.64, we cannot reject the hypothesis that these samples were drawn from the same source (Fig. 12C). Similarly, this approach for a given metric identifies the maximum value that may be expected from two unrelated sources. For example, in the case of PDP Cross-correlation with values <0.52 derived from a data set with n = 500, we cannot reject the hypothesis that two samples were drawn entirely from unrelated populations (Fig. 12C). Where the two fields overlap we cannot discriminate identical from different populations. For example, for a PDP Cross-correlation coefficient of 0.4 from 2 samples of n = 100, we cannot rule out derivation either from the same or different populations. Coefficients that do not plot in either of the fields in Figure 12 are consistent with mixing of different sources. For example, for a PDP cross-correlation coefficient of 0.6 from 2 samples of n = 600, we can rule out total derivation from identical sources and total derivation from different sources. We emphasize that Figure 12 only provides preliminary guidelines for interpretation of analysis results, as these cutoff values are based on synthetically produced data sets. All of these metrics are best applied in a relative rather than absolute sense, and application of a cutoff value without consideration of the geologic context is likely to result in inaccurate interpretations.

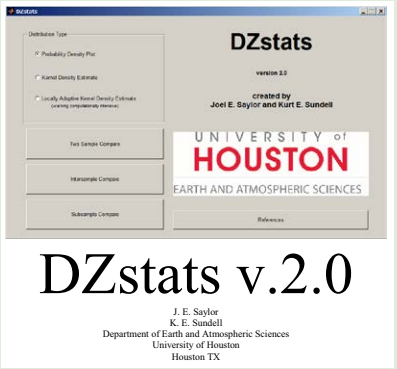
DZstats APPLICATION

In order to facilitate application of quantitative metrics to multiple large detrital geochronology data sets, we developed a MATLAB-based application, DZstats, that implements all of the metrics discussed in this paper through a graphical user interface (Supplemental File 3[3]). Although originally envisioned to be applied to detrital zircon U-Pb data sets, DZstats can be applied to any discrete data with associated uncertainties. DZstats has three main modules. The first module, Two Sample Compare, allows quick comparison of two samples by calculating any individual or all of the statistical metrics mentioned here, as well as plotting their PDPs, KDEs, LA-KDEs, and CDFs. The second module, Intersample Compare, is used for comparison of multiple data sets in their entirety. It outputs all of the metrics mentioned here for each sample pair as well as their PDPs, KDEs, LA-KDEs, and CDFs and the bandwidths used to calculate the KDEs. The third module, Subsample Compare, calculates the minimum, maximum, mean, and standard deviation of each statistical metric for each input sample pair by creating a specified number of random subsamples of each data set and applying all of the statistical metrics to each subsample pair. Because subsampling is done without replacement, the specified subsample size must be less than or equal to the smallest input data set. For modules two and three, the number of data sets and data set size are limited only by the computer's memory. For all of these modules the user can select KDEs calculated based on the LA-KDE model of Shimazaki and Shinomoto (2010) or the adaptive KDE model of Botev et al. (2010) in addition to standard PDPs (in the former two cases the code used in DZstats was written by the original authors; see Supplemental Figure 4[4]).

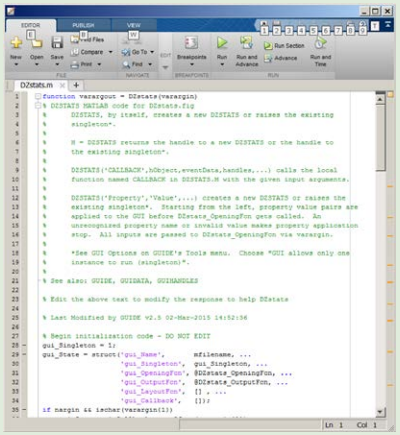
CONCLUSIONS

We applied three criteria to evaluate quantitative assessment of sample similarity using the K-S test p-values and D values, Kuiper test p-values and V values, and Similarity, Likeness, and Cross-correlation coefficients. K-S or Kuiper test p-values, currently the statistical metrics most widely used to describe the degree of similarity between detrital geochronological data sets, fail two of the criteria. The D and V values do not utilize the full range of possible coefficients and so fail one of the criteria. Cross-correlation, Similarity, and Likeness coefficients fail at least two of the criteria when applied to KDEs or LA-KDEs. Cross-correlation fulfills all three criteria when applied to PDPs.

We caution that the application of individual K-S or Kuiper test p-values to detrital geochronology data sets as a hypothesis test of a common sedimentary source is likely to lead to incorrect conclusions. Individual K-S and Kuiper p-values are too sensitive to variation between samples of the same population by rejecting the null hypothesis at a significance level of 95% for more the 5% of the samples drawn from identical populations. Instead, we recommend calculation of mean p-values based on multiple subsampling of geochronological data sets without replacement and assessment of mean p-values as



[3]Supplemental File 3. Executable DZstats file and User Manual. Please visit http://dx.doi.org/10.1130/GES01237.S3 or the full-text article on www.gsapubs.org to view Supplemental File 3.



[4]Supplemental File 4. Annotated MATLAB files for DZstats. Please visit http://dx.doi.org/10.1130/GES01237.S4 or the full-text article on www.gsapubs.org to view Supplemental File 4.

a binary hypothesis test. Maximum K-S and Kuiper p-values for comparison of subsamples of different populations remain above 0.05 even for large data set sizes. Due to this very high sample size requirement, we recommend that p-values be used in conjunction with the other metrics explored here for quantitative evaluation.

For metrics lacking a clear hypothesis test criterion, we propose interpretive guidelines whereby we cannot reject the hypothesis that two samples are drawn from the same parent population if their coefficients are greater than the minimum coefficient from similar-sized subsamples of any of the 20 synthetic populations presented here. Similarly, we cannot reject the hypothesis that 2 samples were drawn entirely from different populations if they yield coefficients less than the maximum from intercomparison of subsamples of the 20 synthetic populations. As with other hypothesis tests, this guideline cannot be used to determine which samples are drawn from the same parent population, but only to identify those samples that are unlikely to be drawn from the same parent population. Although this provides guidelines for data interpretation, we emphasize that metrics lacking a clear hypothesis test are best utilized to determine the degree of similarity of samples rather than to establish an absolute pass-fail coefficient.

Of metrics lacking a clear hypothesis test, PDP Cross-correlation requires the smallest data set for discrimination of samples drawn from the same parent population verses those drawn from different parent populations. Coefficients from data sets drawn from the same parent population and data sets drawn from different parent populations may overlap for small ($n < 300$) data sets. We conclude that large data sets are necessary to quantitatively describe similarity or difference between samples. Because a perfect match between samples and a source population is increasingly unlikely with increasing sample size, we advocate incorporation and reporting of uncertainties of comparison metrics through repeated random sampling of geochronological data sets and application of the statistical metrics to each sampled pair.

ACKNOWLEDGMENTS

We thank G.E. Gehrels, C. Peters, and four anonymous reviewers for helpful comments on earlier versions of this manuscript and the recommendation that we include kernel density estimates in this study and in our software. T. Andersen also provided constructive insights on this paper. We thank Editor R. Russo for handling this paper.

REFERENCES CITED

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, doi:10.1111/j.1365-2117.2005.00279.x.

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, p. 244–260, doi:10.1016/j.epsl.2005.03.019.

Andersen, T., 2005, Detrital zircons as tracers of sedimentary provenance: Limiting conditions from statistics and numerical simulation: Chemical Geology, v. 216, p. 249–270, doi:10.1016/j.chemgeo.2004.11.013.

Behboodian, J., 1970, On a mixture of normal distributions: Biometrika, v. 57, p. 215–217, doi:10.1093/biomet/57.1.215.

Botev, Z.I., Grotowski, J.F., and Kroese, D.P., 2010, Kernel density estimation via diffusion: Annals of Statistics, v. 38, p. 2916–2957, doi:10.1214/10-AOS799.

Brandon, M.T., 1996, Probability density plot for fission-track grain-age samples: Radiation Measurements, v. 26, p. 663–676, doi:10.1016/S1350-4487(97)82880-6.

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, doi:10.1198/jasa.2009.tm07543.

DeGraaff-Surpless, K., Mahoney, J.B., Wooden, J.L., and McWilliams, M.O., 2003, Lithofacies control in detrital zircon provenance studies: Insights from the Cretaceous Methow basin, southern Canadian Cordillera: Geological Society of America Bulletin, v. 115, p. 899–915, doi:10.1130/B25267.1.

Delaigle, A., and Meister, A., 2008, Density estimation with heteroscedastic error: Bernoulli, v. 14, p. 562–579, doi:10.3150/08-BEJ121.

Dickinson, W.R., and Gehrels, G.E., 2008, Sediment delivery to the Cordilleran foreland basin: Insights from U-Pb ages of detrital zircons in Upper Jurassic and Cretaceous strata of the Colorado Plateau: American Journal of Science, v. 308, p. 1041–1082, doi:10.2475/10.2008.01.

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 [London], v. 145, p. 977–983, doi:10.1144/gsjgs.145.6.0977.

Everitt, B., and Hand, D.J., 1981, Finite Mixture Distributions: New York, Chapman and Hall, 143 p.

Everitt, B.S., 2005, Finite mixture distributions, *in* Everitt, B., and Howell, D., eds., Encyclopedia of Statistics in Behavioral Science: Chichester, John Wiley & Sons, Ltd., doi:10.1002/9781118445112.stat06216.

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: Reviews in Mineralogy and Geochemistry Volume 53, p. 277–303, doi:10.2113/0530277.

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: Geological Society of America Special Paper 347, p. 1–17, doi:10.1130/0-8137-2347-7.1.

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: Chichester, John Wiley & Sons, p. 47–62, doi:10.1002/9781444347166.ch2.

Gehrels, G.E., Valencia, V.A., and Ruiz, J., 2008, Enhanced precision, accuracy, efficiency, and spatial resolution of U-Pb ages by laser ablation–multicollector–inductively coupled plasma–mass spectrometry: Geochemistry, Geophysics, Geosystems, v. 9, Q03017, doi:10.1029/2007GC001805.

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, p. 269–277, doi:10.1017/S0016756800029162.

Jessberger, E.K., Dominik, B., Staudacher, T., and Herzog, G.F., 1980, ^{40}Ar-^{39}Ar ages of Allende: Icarus, v. 42, p. 380–405, doi:10.1016/0019-1035(80)90103-7.

Kuiper, N.H., 1960, Tests concerning random points on a circle: Koninklijke Nederlandse Akademie van Wetenschappen, v. 63, p. 38–47.

Lawrence, R.L., Cox, R., Mapes, R.W., and Coleman, D.S., 2011, Hydrodynamic fractionation of zircon age populations: Geological Society of America Bulletin, v. 123, p. 295–305, doi:10.1130/B30151.1.

L'Ecuyer, P., and Simard, R., 2007, TestU01: A C library for empirical testing of random number generators: ACM Transactions on Mathematical Software, v. 33, no. 4, p. 22, doi:10.1145/1268776.1268777.

Lo, Y.T., Mendell, N.R., and Rubin, D.B., 2001, Testing the number of components in a normal mixture: Biometrika, v. 88, p. 767–778, doi:10.1093/biomet/88.3.767.

Marsaglia, G., and Tsang, W.W., 1984, A fast, easily implemented method for sampling from decreasing or symmetric unimodal density functions: SIAM Journal on Scientific and Statistical Computing, v. 5, p. 349–359, doi:10.1137/0905026.

Martin, B.R., 2012, Statistics for Physical Sciences: An Introduction: Waltham, Academic Press, 301 p.

Matsumoto, M., and Nishimura, T., 1998, Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator: ACM Transactions on Modeling and Computer Simulation, v. 8, no. 1, p. 3–30.

McIntyre, J., and Stefanski, L., 2011, Density estimation with replicate heteroscedastic measurements: Annals of the Institute of Statistical Mathematics, v. 63, p. 81–99, doi:10.1007/s10463-009-0220-x.

Mélard, G., 2014, On the accuracy of statistical procedures in Microsoft Excel 2010: Computational Statistics, v. 29, p. 1095–1128, doi:10.1007/s00180-014-0482-5.

Miller, M.B., 2014, Mathematics and Statistics for Financial Risk Management (second edition): Hoboken, New Jersey, John Wiley & Sons, Inc., 336 p.

Pearson, R.K., 2011, Exploring Data in Engineering, the Sciences, and Medicine: New York, Oxford University Press, 792 p.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P., 2007, Numerical Recipes: The Art of Scientific Computing (third edition): New York, Cambridge University Press, 1256 p.

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, p. 971–980, doi:10.1039/c4ja00024b.

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, p. 1783–1799, doi:10.1130/B30888.1.

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, p. 762–779, doi:10.1130/B30534.1.

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, p. 17–33, doi:10.1086/668683.

Scott, D.W., 1992, Multivariate Density Estimation: Theory, Practice, and Visualization: New York, John Wiley & Sons, Inc., 317 p.

Shaulis, B., Lapen, T.J., and Toms, A., 2010, Signal linearity of an extended range pulse counting detector: Applications to accurate and precise U-Pb dating of zircon by laser ablation quadrupole ICP-MS: Geochemistry, Geophysics, Geosystems, v. 11, Q0AA11, doi:10.1029/2010GC003198.

Shimazaki, H., and Shinomoto, S., 2010, Kernel bandwidth optimization in spike rate estimation: Journal of Computational Neuroscience, v. 29, p. 171–182, doi:10.1007/s10827-009-0180-4.

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, p. 21–31, doi:10.1016/j.cageo.2003.09.006.

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, doi:10.1038/191688a0.

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, doi:10.1198/016214508000000328.

Stephens, M.A., 1970, Use of the Kolmogorov-Smirnov, Cramer-Von Mises and related statistics without extensive tables: Royal Statistical Society Journal, ser. B, v. 32, p. 115–122.

Titterington, 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., 2004, How many grains are needed for a provenance study?: Earth and Planetary Science Letters, v. 224, p. 441–451.

Vermeesch, P., 2012, On the visualisation of detrital age distributions: Chemical Geology, v. 312, p. 190–194, doi:10.1016/j.chemgeo.2012.04.021.

Vermeesch, P., 2013, Multi-sample comparison of detrital age distributions: Chemical Geology, v. 341, p. 140–146, doi:10.1016/j.chemgeo.2013.01.010.

Vermeesch, P., and Garzanti, E., 2015, Making geological sense of 'Big Data' in sedimentary provenance analysis: Chemical Geology, v. 409, p. 20–27, doi:10.1016/j.chemgeo.2015.05.004.

Weislogel, A.L., Graham, S.A., Chang, E.Z., Wooden, J.L., and Gehrels, G.E., 2010, Detrital zircon provenance from three turbidite depocenters of the Middle-Upper Triassic Songpan-Ganzi complex, central China: Record of collisional tectonics, erosional exhumation, and sediment production: Geological Society of America Bulletin, v. 122, p. 2041–2062, doi:10.1130/B26606.1.

Wilk, M.B., and Gnanadesikan, R., 1968, Probability plotting methods for the analysis of data: Biometrika, v. 55, p. 1–17, doi:10.2307/2334448.