Geoscience Frontiers 8 (2017) 1247–1252

HOSTED BY

Contents lists available at ScienceDirect

China University of Geosciences (Beijing)

Geoscience Frontiers

journal homepage: www.elsevier.com/locate/gsf

Research Paper

Visualising data distributions with kernel density estimation and reduced chi-squared statistic

C.J. Spencer [a,*], C. Yakymchuk [b], M. Ghaznavi [c]

[a] Earth Dynamics Research Group, The Institute of Geoscience Research, Department of Applied Geology, Curtin University, Perth, Australia
[b] Department of Earth and Environmental Sciences, University of Waterloo, Waterloo, Canada
[c] David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Canada

ARTICLE INFO

Article history:
Received 17 January 2017
Received in revised form
11 May 2017
Accepted 21 May 2017
Available online 3 June 2017
Handling Editor: Nick M.W. Roberts

Keywords:
Data visualisation
Kernel density estimation
Reduced chi-squared statistic
Mean square weighted deviation
Geostatistics

ABSTRACT

The application of frequency distribution statistics to data provides objective means to assess the nature of the data distribution and viability of numerical models that are used to visualize and interpret data. Two commonly used tools are the kernel density estimation and reduced chi-squared statistic used in combination with a weighted mean. Due to the wide applicability of these tools, we present a Java-based computer application called KD**X** to facilitate the visualization of data and the utilization of these numerical tools.

1. Introduction

The calculation of weighted means along with its accompanied reduced chi-squared statistic and visualisation of univariate data density provide important insight into the nature and usability of said data. Many numerical tools have been developed for this purpose, but few are available across multiple operating systems, and are generally restricted to antiquated programming languages. We provide a new cross-platform application called KD**X** (kay-dee-kai; a combined initialism of kernel density estimation and reduced chi-squared). This application is designed to perform the aforementioned functions in the robust Java platform. In addition to providing the ability to export the visualisations in a variety of file formats, it provides an extensive customization for production of publication-quality figures not requiring additional editing in vector image editing software. This application and source code is available free of charge from the author's websites.

2. Weighted mean and reduced chi-squared

The weighted mean simply calculates a mean value based on a particular weighting (in this case weighted by the uncertainty of each datum). It is calculated using:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i/\sigma_i^2}{\sum_{i=1}^{n} \sigma_i^2} \text{ or } \overline{x} = \frac{\left(x_1/\sigma_1 + x_2/\sigma_2 + \cdots + x_n/\sigma_n\right)^2}{(\sigma_1 + \sigma_2 + \cdots + \sigma_n)^2} \quad (1)$$

where x_i and σ_i are the analyses and uncertainties, respectively. The weighted uncertainty is calculated with:

$$\sigma_{\overline{x}} = \sqrt{\frac{1}{\sum_{i=1}^{n} 1/\sigma_i^2}} \text{ or } \sigma_{\overline{x}} = \sqrt{\frac{1}{1/(\sigma_1 + \sigma_2 + \cdots + \sigma_n)^2}} \quad (2)$$

The reduced chi-squared statistic (X_ν^2) is used extensively as a goodness of fit test between a model and set of data. It is often referred to as the mean square weighted deviation (MSWD) and is defined by the chi-squared per degree of freedom (Bevington, 1969; Wendt and Carl, 1991). In this case the model that is being tested is the weighted mean described above and the reduced chi-squared statistic provides a goodness of fit assessment. It is calculated using the following equation:

* Corresponding author.
E-mail addresses: cspencer@curtin.edu.au, spenchristoph@gmail.com (C.J. Spencer).
Peer-review under responsibility of China University of Geosciences (Beijing).

http://dx.doi.org/10.1016/j.gsf.2017.05.002

$$X_\nu^2 = \frac{1}{n-1} \times \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma_{x_i}^2} \quad (3)$$

where n is the number of analyses, x_i is the individual analyses, $\bar{x}$ is the weighted mean, and σ_{x_i} is the corresponding uncertainty of x_i.

Where the reduced chi-squared statistic equals 1, it indicates the observed values conform to a statistically univariate normal distribution, and the corresponding weighted average and uncertainty are an appropriate representation of those data. If the reduced chi-squared statistic is greater than one, the observed scatter of the data exceeds that predicted by the datapoint uncertainties. This 'overdispersion' is either due to an underestimation of the uncertainties or the presence of 'natural' or 'geologic' scatter. If the reduced chi-squared statistic is less than one, the data display 'underdispersion' and the analytical uncertainties are overestimated (Horstwood, 2008; Horstwood et al., 2016).

'Real-world' examples rarely provide a reduced chi-squared statistic of exactly 1. However, one can assess the statistical probability that the reduced chi-squared represent a single population using the maximum of reduced chi-squared frequency distributions (Wendt and Carl, 1991; Horstwood, 2008; Spencer et al., 2016). The distribution functions for both values less than and greater than one approach one asymptotically with increasing n and the reduced chi-squared distribution approaches a normal distribution (Wendt and Carl, 1991). Therefore the acceptable reduced chi-squared statistics can be calculated within $\pm 2\sigma$ using $1 + 2\sqrt{2/n-1}$. If the reduced chi-squared falls within the 2σ envelope there is a >95% probability the data form a single population and the weighted average is appropriate. On the other hand, if the reduced chi-squared does not fall within this envelope then the data do not conform to a single population and the reduced chi-squared should only be used with this caveat, i.e. there is less than 5% probability that the data form a statistical single population.

3. Kernel density estimation

As discussed at length by Vermeesch (2012), the kernel density estimation (KDE) (Silverman, 1986) provides a more robust alternative to the commonly used 'Probability Density Plot' (PDP) when visualizing frequency data. The kernel density estimation estimates data frequency by summing a set of Gaussian distributions, but in contrast to the 'Probability Density Plot', does not take into account the analytical uncertainty. This is particularly useful in looking for a cluster of analyses in spectra of data. It is calculated using:

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

where $K(\bullet)$ is the kernel—a non-negative function that integrates to one and has a mean of 0 and h (always > 0) is a smoothing parameter called the bandwidth. A kernel with subscript h is called the *scaled kernel* and defined as:

$$K_h(x) = 1/h\, K(x/h) \quad (5)$$

Various kernels have been devised including:

$$\text{Epanechnickov}: \quad K(t) = \max\left[0, \frac{3}{4}\left(1 - \frac{1}{5}t^2\right)\Big/\sqrt{5}\right] \quad (6)$$

$$\text{Gaussian}: \quad K(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \quad (7)$$

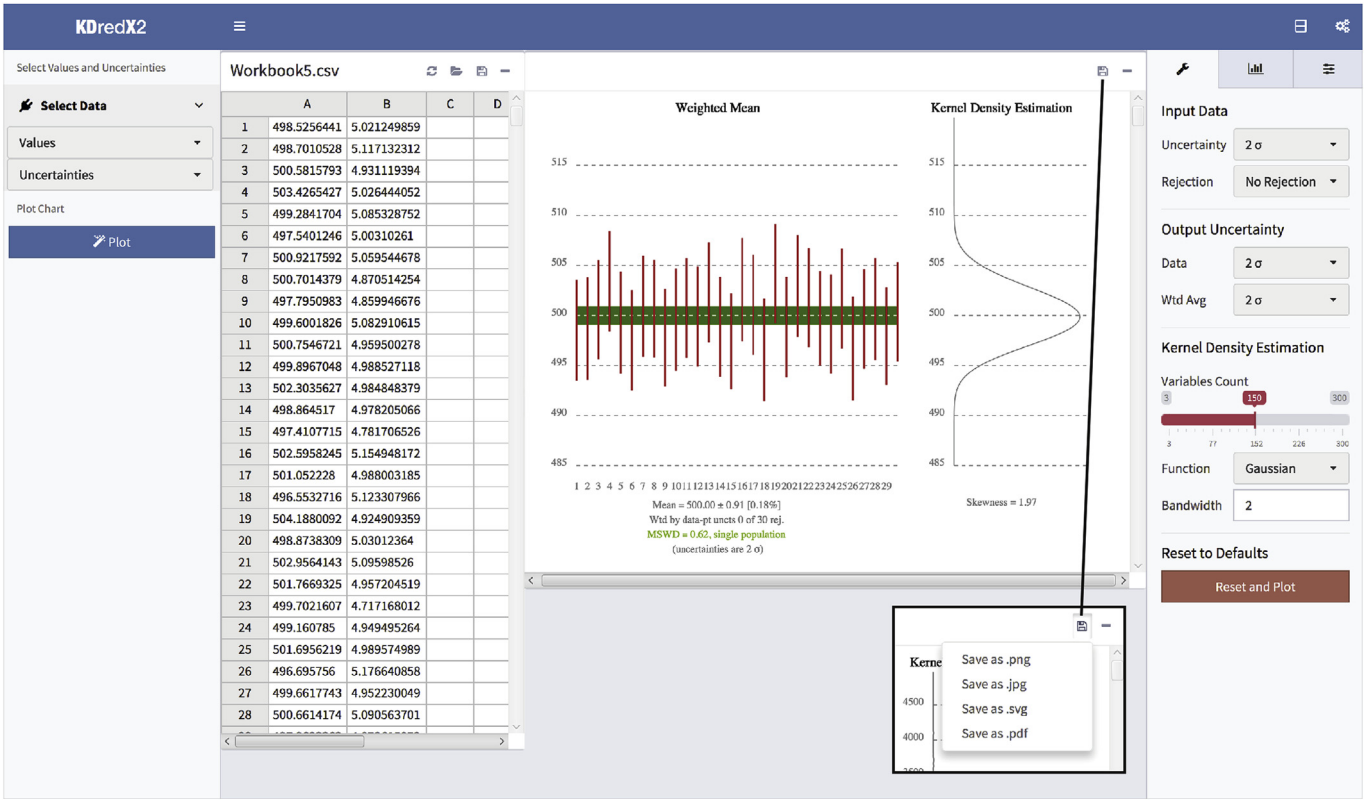

Figure 1. KD**X** application, inset showing supporting file save options (png, jpg, svg, pdf).

Vermeesch (2012) noted that the choice of the kernel only determines the smoothness characteristics of the density estimation and does not broadly affect the resulting KDE. However, the choice of bandwidth will dramatically change the KDE, as a bandwidth that is too high or low will result in over- or undersmoothing, respectively. We suggest the choice of bandwidth should be assigned based the fundamental limitations of the instrument producing the data. Otherwise, a bandwidth significantly smaller than typical uncertainties expected by a particular instrument is likely to produce unrealistic structure in the density estimation.

Lastly, in addition to passing the reduced chi-squared test, data that conforms to a single population should also display a skewness approaching 1. As discussed by Spencer et al. (2016), it is possible for a set of data to pass the chi-squared test, but displays a negatively skewed tail. Defining the skewness can provide further support for or against the reduced chi-squared test.

4. The application KDX

The application presented is called KD**X** (kay-dee-kai), which combines kernel density estimation with the reduced chi-squared statistic (also known as the mean square weighted deviation or MSWD). The KD**X** interface is designed using HTML, CSS, and Javascript and its core is developed using Javascript and Java.

KD**X** consists of the *data* and *chart* modules (Fig. 1). The data module provides the functionality of the spreadsheet supporting the comma separated values (CSV) format. The data loaded in the spreadsheet is used to plot the charts. The chart module provides the plotting functionality. This module allows a user to modify the chart appearance using the plotting preferences and save the plot in 4 different file formats (png, jpg, svg, pdf).

4.1. Data module

Users can manually enter or import the data from a CSV file into the program's spreadsheet. The spreadsheet provides the functionality to copy, cut, paste, edit, add, and remove cells, columns, and rows. Up to 15 columns and 10,000 rows of data can be accommodated. This module can also export the spreadsheet data into a CSV file. Two columns of input are required in the data module: (1) the values of the data and (2) the corresponding uncertainty (at the 1σ or 2σ level) for each value in the same row.

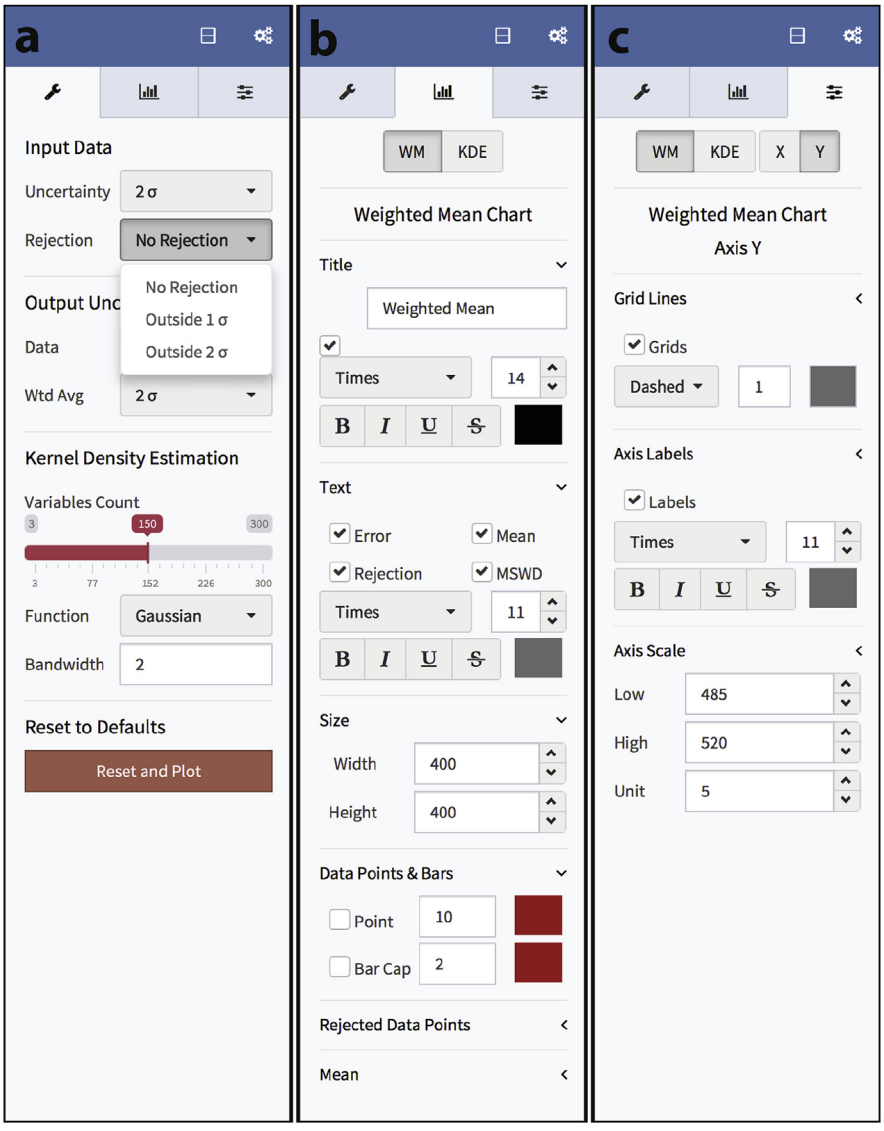

Figure 2. (a) Data settings, (b) general settings, (c) axis settings.

4.2. Chart module

The chart module plots values from the data module. KD**X** plots the weighted mean and kernel density estimation charts based on the columns that a user selects as values and uncertainties. This module provides extensive settings to customize the charts. These settings are arranged into three categories: the data settings, the general chart settings, and the axis settings.

4.3. Data settings

The *Data settings* tab includes Input Data, Output Uncertainty, Kernel Density Estimation, and Reset to Defaults (Fig. 2a). The same as for the equations above, σ represents the standard deviation, $\overline{x}$ is the weighted mean, and $\sigma_{\overline{x}}$ is weighted uncertainty.

Input Data settings allow users to specify that if the uncertainties for the values are at the 1σ or 2σ level. There is an option to choose if values should be rejected from the calculations. The options for rejection are '*No Rejection*', '*Outside* 1σ', and '*Outside* 2σ' meaning that no value will be rejected, and values are rejected if they resides outside of $\overline{x} \pm \sigma_{\overline{x}}$ or $\overline{x} \pm 2\sigma_{\overline{x}}$, respectively.

The *Output Uncertainty* settings control how the values and weighted mean are plotted in the *Chart Module*. The '*Data*' setting controls if uncertainties of values are plotted at σ or 2σ. The '*Wtd Avg*' setting controls if the weighted average is calculated and plotted at $\sigma_{\overline{x}}$ or $2\sigma_{\overline{x}}$, respectively.

The *Kernel Density Estimation* settings provide three controls. KD**X** divides the range of *Y*-axis in the weighted mean chart into a number of analyses that the '*Variables Count*' option specifies. Results are a set of x_i values that are used to compute the kernel density estimation. The kernel function, either *Gaussian* (Eq. 6) or *Epanechnikov* (Eq. 7), and the bandwidth are selected using '*Function*' and '*Bandwidth*' controls. Finally, using the button '*Reset and Plot*' in the '*Reset to Defaults*' section, all settings are reset and the charts are replotted.

4.4. General settings

The *General Settings* tab (Fig. 2b) is used to adjust various visual aspects of the *Weighted Mean Chart* and the *KDE* chart. The title can be assigned (the default is 'Weighted Mean') as well as the font, point size of the title and other text. Various information can be included or excluded from the plots, including: the errors, rejected points, mean and the MSWD.

The height of both charts is adjusted using the height control of the weighted mean chart. Various visual aspects of the charts can be modified, including colors, point size of lines and the nature of data points, error bars and caps. For the KDE chart, the line section controls the style of the KDE line.

The axis settings (Fig. 2c) controls the appearance of *X* and *Y* axes for both charts. 'Grid Lines', 'Axis Labels', and 'Axis Scales' sections control the style of gridline, labels, and the scale of the selected axis in the selected chart. The scales of *Y* axes in both charts are controlled using the 'Axis Scales' section of the weighted mean chart.

5. Comparison with other software

Within the Earth Science community, there are a number of software programs that are used for statistical treatment and visualization of frequency data. For the calculation of weighted averages and plotting of probability density plots, the most popular and widely used is the Visual Basic for Applications (VBA) Microsoft Excel-Add-in Isoplot/Ex (Ludwig, 2001). Although Isoplot has proven itself a staple in isotope chemistry and geochronology, it is only compatible with Microsoft Excel with VBA (only on MS Windows). Furthermore, Isoplot is no longer being updated by the author and is unlikely to remain viable in its current state. Another program called *Geodate* (Eglington and Harmer, 1999) has similar functionality to Isoplot, but is only available for computers running MS Windows.

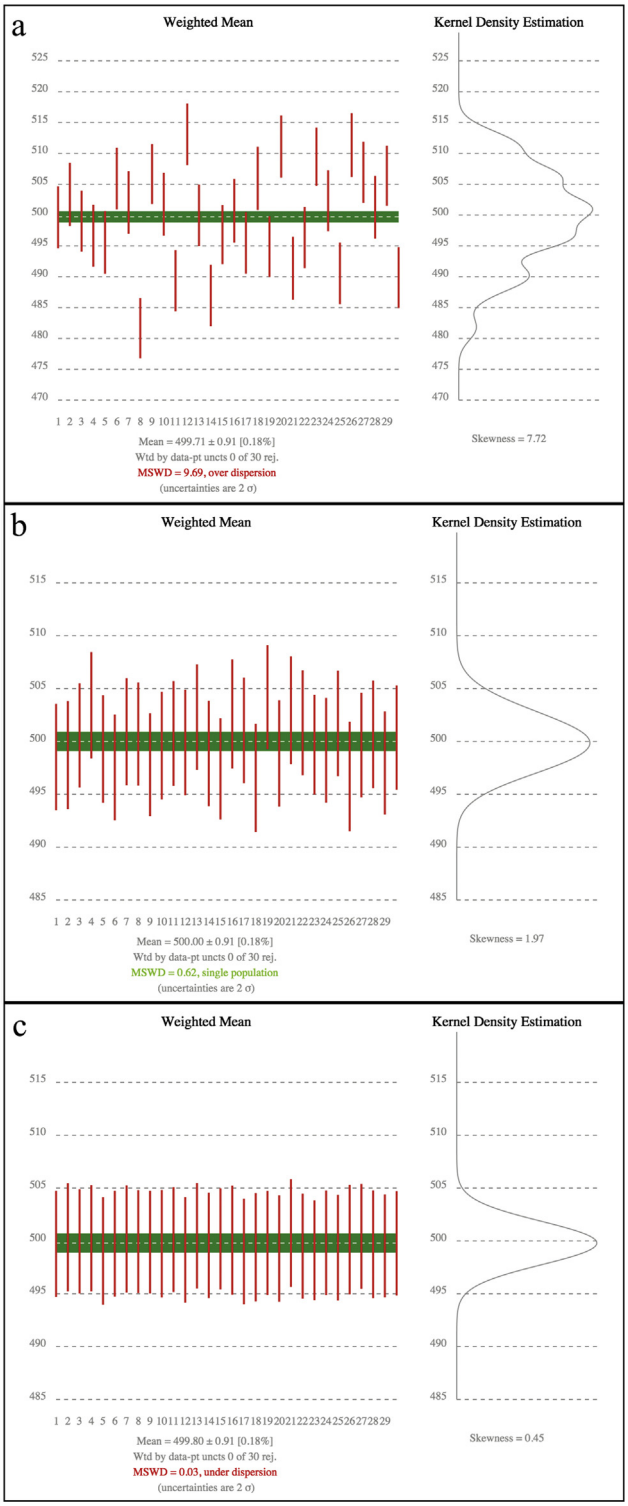

Figure 3. Three examples of data displaying over-dispersion (a), a single population (b), and underdispersion (c), respectively.

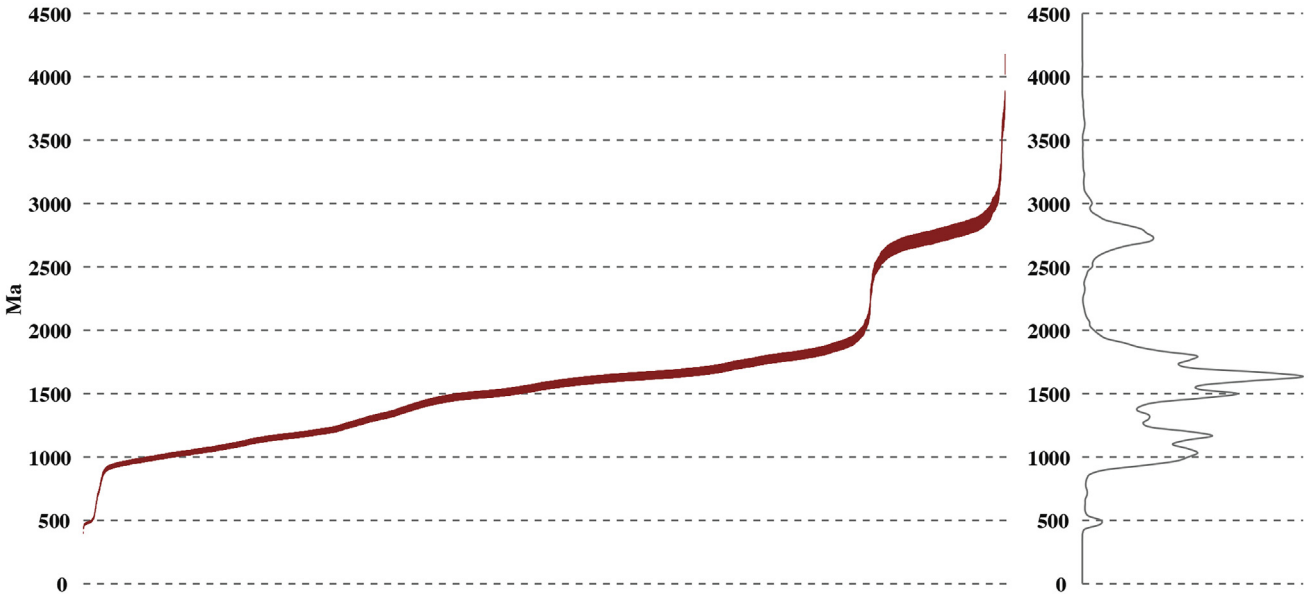

Figure 4. A large n example ($n = 4000$) of detrital zircon U-Pb ages.

For plotting KDE and PDP diagrams, densityplotter (Vermeesch, 2012) is a useful program, but does not assess the weighted average, reduced chi-squared, or skewness.

6. System requirements and application distribution

KD**X** is cross-platform. This application with full capabilities runs on any Apple computer under Mac OS X El Capitan (OS version 10.11) or later and runs with limited capabilities (no copy and paste) on Mac OS X 10.3 to 10.10. It will not run on earlier versions of the Mac OS. The application will also run on MS Windows running the Windows Vista operating system and later. The application is freeware and can be obtained as gzipped tar archive following the links from http://KD**X**.travelinggeologist.com.

7. Applications

The main application of the reduced chi-squared statistic in the geosciences is evaluating the robustness of ages obtained through geochronology. The statistic can be applied to isochron calculations in Rb-Sr, Sm-Nd, Lu-Hf, and Re-Os geochronology as well as discordia calculations of U-Pb geochronology (see Rink and Thompson, 2015). Unlike current software available for evaluating geochronology results, with KD**X**, the user can evaluate the reduced chi-squared statistic (also known as MSWD) as well as the skewness of the distribution, which is essential for a robust interpretation (Fig. 3). For example, it is possible for a set of U-Pb age data to pass the chi-squared test, but displays a negatively skewed tail caused by lead loss (e.g. Spencer et al., 2016).

In addition to geochronology, the reduced chi-squared statistic has been applied to other geoscience disciplines, including: neotectonics and seismology (Aloisi et al., 2013), seawater geochemistry (e.g. Stoll et al., 1999) and metamorphic thermobarometry (e.g. Applegate and Hodges, 1994). Other scientific fields also use this statistic, including chemistry (e.g. Enyedy and Kovach, 2004), astronomy (e.g. Xu et al., 2013; Kay et al., 2015) and medicine (Al-Issa et al., 2015).

In the geosciences, kernel density estimation is used to assess the different populations in provenance studies using datable detrital minerals (e.g. Vermeesch, 2012; Gehrels, 2014) as well as to evaluate the global record of supercontinents and large igneous provinces (e.g. Condie et al., 2015) and seismic risk analysis (Danese et al., 2008). KD**X** is able to handle large datasets ($n < 10{,}000$) allowing for unprecedented evaluation and visualisation of large amounts of data (Fig. 4).

Outside of the geosciences, KDE is used in ecological studies (e.g. Worton, 1989; Fleming and Calabrese, 2016), astronomy (Helmi and De Zeeuw, 2000; Ferdosi et al., 2011; Trapero, 2016), economics (Bolancé et al., 2003; Ruppert, 2011), Archaeology (e.g. Baxter et al., 1997), medicine (Rossiter, 1991) and chemistry (Unke and Meuwly, 2015).

Acknowledgements

This manuscript benefited greatly from the reviews by Pieter Vermeesch and an anonymous reviewer and the editorial handling of Associate Editor Dr. Nick Roberts.

References

Al-Issa, Y., Njagi, J., Schuckers, S.C., Suni, I.I., 2015. Amperometric bioelectronic tongue for glucose determination. Sensing and Bio-Sensing Research 3, 31–37.
Aloisi, M., Bruno, V., Cannavò, F., Ferranti, L., Mattia, M., Monaco, C., Palano, M., 2013. Are the source models of the M 7.1 1908 Messina Straits earthquake reliable? Insights from a novel inversion and a sensitivity analysis of levelling data. Geophysical Journal International 192, 1025–1041.
Applegate, J.D.R., Hodges, K.V., 1994. Empirical evaluation of solution models for pelitic minerals and their application to thermobarometry. Contributions to Mineralogy and Petrology 117, 56–65.
Baxter, M.J., Beardah, C.C., Wright, R.V., 1997. Some archaeological applications of kernel density estimates. Journal of Archaeological Science 24, 347–354.
Bevington, P.R., 1969. Data Reduction and Error Analysis for the Physical Sciences. McGraw Hill Book Co, New York.
Bolancé, C., Guillen, M., Nielsen, J.P., 2003. Kernel density estimation of actuarial loss functions. Insurance: Mathematics and Economics 32, 19–36.
Condie, K.C., Davaille, A., Aster, R.C., Arndt, N., 2015. Upstairs-downstairs: supercontinents and large igneous provinces, are they related? International Geology Review 57, 1341–1348.
Danese, M., Lazzari, M., Murgante, B., 2008. Kernel density estimation methods for a geostatistical approach in seismic risk analysis: the case study of potenza hilltop town (Southern Italy). In: International Conference on Computational Science and its Applications. Springer, pp. 415–429.
Eglington, B.M., Harmer, R.E., 1999. GEODATE for Windows Version 1: Isotope Regression and Modeling Software. Counc. Geosci. Open File Rep, pp. 1–51.

Enyedy, E.J., Kovach, I.M., 2004. Proton inventory studies of α-thrombin-catalyzed reactions of substrates with selected P and P′ sites. Journal of the American Chemical Society 126, 6017–6024.
Ferdosi, B., Buddelmeijer, H., Trager, S., Wilkinson, M., Roerdink, J., 2011. Comparison of density estimation methods for astronomical datasets. Astronomy & Astrophysics 531, A114.
Fleming, C.H., Calabrese, J.M., 2016. A new kernel-density estimator for accurate home-range and species-range area estimation. Methods in Ecology and Evolution. http://dx.doi.org/10.1111/2041-210X.12673.
Gehrels, G., 2014. Detrital zircon U-Pb geochronology applied to tectonics. Annual Review of Earth and Planetary Sciences 42, 127–149.
Helmi, A., De Zeeuw, P.T., 2000. Mapping the substructure in the Galactic halo with the next generation of astrometric satellites. Monthly Notices of the Royal Astronomical Society 319, 657–665.
Horstwood, M.S.A., 2008. Data reduction strategies, uncertainty assessment and resolution of LA-(MC-)ICP-MS isotope data. Mineralogical Association of Canada Short Course 40, 283–303.
Horstwood, M.S.A., Košler, J., Gehrels, G., Jackson, S.E., McLean, N.M., Paton, C., Pearson, N.J., Sircombe, K., Sylvester, P., Vermeesch, P., Bowring, J.F., Condon, D.J., Schoene, B., 2016. Community-derived standards for LA-ICP-MS U-Th-Pb geochronology – uncertainty propagation, age interpretation and data reporting. Geostandards and Geoanalytical Research. http://dx.doi.org/10.1111/j.1751-908X.2016.00379.x n/a–n/a.
Kay, C., Dos Santos, L., Opher, M., 2015. Constraining the masses and the non-radial drag coefficient of a solar coronal mass ejection. The Astrophysical Journal Letters 801, L21.
Ludwig, K.R., 2001. Isoplot 3.0—A Geochronological Toolkit for Microsoft Excel: Special Publication No. 4. Berkeley Geochronol. Center, Berkeley, Calif.
Rink, W.J., Thompson, J.W., 2015. Encyclopedia of Scientific Dating Methods. Springer.
Rossiter, J., 1991. Calculating centile curves using kernel density estimation methods with application to infant kidney lengths. Statistics in Medicine 10, 1693–1701.
Ruppert, D., 2011. Statistics and Data Analysis for Financial Engineering. Springer.
Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. CRC press.
Spencer, C.J., Kirkland, C.L., Taylor, R.J.M., 2016. Strategies towards statistically robust interpretations of in situ U-Pb zircon geochronology. Geoscience Frontiers 7, 581–589. http://dx.doi.org/10.1016/j.gsf.2015.11.006.
Stoll, H.M., Schrag, D.P., Clemens, S.C., 1999. Are seawater Sr/Ca variations preserved in Quaternary foraminifera? Geochimica et Cosmochimica Acta 63, 3535–3547.
Trapero, J.R., 2016. Calculation of solar irradiation prediction intervals combining volatility and kernel density estimates. Energy 114, 266–274.
Unke, O.T., Meuwly, M., 2015. Kernel density estimation-based solution of the nuclear Schrödinger equation. Chemical Physics Letters 639, 52–56.
Vermeesch, P., 2012. On the visualisation of detrital age distributions. Chemical Geology 312–313, 190–194. http://dx.doi.org/10.1016/j.chemgeo.2012.04.021.
Wendt, I., Carl, C., 1991. The statistical distribution of the mean squared weighted deviation. Chemical Geology Isotope Geoscience Section 86, 275–285. http://dx.doi.org/10.1016/0168-9622(91)90010-T.
Worton, B.J., 1989. Kernel methods for estimating the utilization distribution in home-range studies. Ecology 70, 164–168.
Xu, S., Jura, M., Klein, B., Koester, D., Zuckerman, B., 2013. Two beyond-primitive extrasolar planetesimals. The Astrophysical Journal 766, 132.