Earth and Planetary Science Letters 274 (2008) 179–188

Contents lists available at ScienceDirect

Earth and Planetary Science Letters

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

Finding peaks in geochemical distributions: A re-examination of the helium-continental crust correlation

John F. Rudge *

Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York 10964, USA

ARTICLE INFO

Article history:
Received 8 December 2007
Received in revised form 7 July 2008
Accepted 13 July 2008
Available online 23 July 2008

Editor: R.D. van der Hilst

Keywords:
ocean island basalt
continental crust
helium isotopes
zircon age
density estimation

ABSTRACT

Parman [Nature (2007) 446, 900–903] has recently suggested that a correlation exists between peaks in the ocean island basalt (OIB) $^4He/^3He$ distribution and peaks in crustal zircon ages. This correlation is based on matching peaks seen in smooth kernel density estimates. Kernel density estimation is a very useful technique, but care is required when choosing the smoothing bandwidth as spurious peaks can be produced if the bandwidth is too small. Here I provide an introduction to a general statistical technique for determining whether peaks in density estimates are significant, know as SiZer, focusing on its application to the $^4He/^3He$ data. SiZer identifies only two statistically significant peaks in the OIB $^4He/^3He$ distribution, compared with the eight peaks identified by Parman. The helium-continental crust correlation does not seem to be supported by the current data.

1. Introduction

Parman (2007) has recently shown a correlation between peaks in ocean island basalt (OIB) $^4He/^3He$ distributions and peaks in the age distributions of crustal zircons (Condie, 1998; Kemp et al., 2006). Such a correlation has intriguing geochemical consequences (Porcelli, 2007) — in particular, it links a record of mantle depletion ($^4He/^3He$) with a record of crustal production (zircons), and thus provides a key constraint on the chemical evolution of the Earth. It suggests that the continents have grown through distinct episodes of mantle melting over the Earth's history.

Parman's correlation raises some important statistical questions: How do we identify peaks in distributions? How do we know if a peak we observe in a histogram or a density estimate is really there? Can we distinguish between real peaks and the spurious peaks that can arise as artifacts of the sampling process? In fact, statistical methods for answering these questions have been developed, and the aim of this paper is to provide an accessible introduction to some of them. In particular, I review kernel density estimation (Silverman, 1986), a recently developed method for identifying significant peaks known as SiZer (Chaudhuri and Marron, 1999), and Gaussian mixture modelling (McLachlan and Peel, 2000). I also examine the problems and pitfalls of attaching significance to spurious peaks. While the techniques described apply generally to any data that can plotted in a histogram, I focus here on the helium isotopic data. For a more rigorous and detailed exposition of these ideas, the reader is referred to the statistics literature. Formal mathematical definitions of the techniques can be found in the appendices.

2. Kernel density estimation

The main focus of Parman's analysis are the probability density functions (PDFs) of $^4He/^3He$ for different groups of basalts, shown in figs. 1 and 2 of Parman (2007). These PDFs were generated by a statistical technique known as kernel density estimation (Silverman, 1986), which can be thought of as refinement over histograms. Kernel density estimates have two main advantages over histograms: they are smooth, and they do not require the choice of end points of bins. However, there is still one key parameter in kernel density estimation that must be chosen by the user, known as the bandwidth, which is analogous to the choice of bin size in a histogram. One must also choose the shape of the kernel function (typically a Gaussian, as assumed here), but this choice is generally less important than the bandwidth. To form the kernel density estimate, each data point in the sample is represented by a Gaussian centred on the data point, with standard deviation given by the bandwidth. The smooth density estimate curve is simply the sum of these individual Gaussians. Different curves result from different choices of bandwidth.

Fig. 1 illustrates the problem of bandwidth selection. 1340 random samples (the same number of samples as Parman's OIB data set) were drawn from a specified bimodal distribution with PDF shown by the dashed curves. The goal is to estimate this true underlying PDF from the random samples. If the chosen bandwidth is too large, only a single peak is

* Tel.: +1 845 365 8676; fax: +1 845 365 8150.
E-mail addresses: jrudge@ldeo.columbia.edu, rudge@esc.cam.ac.uk.

0012-821X/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.epsl.2008.07.021

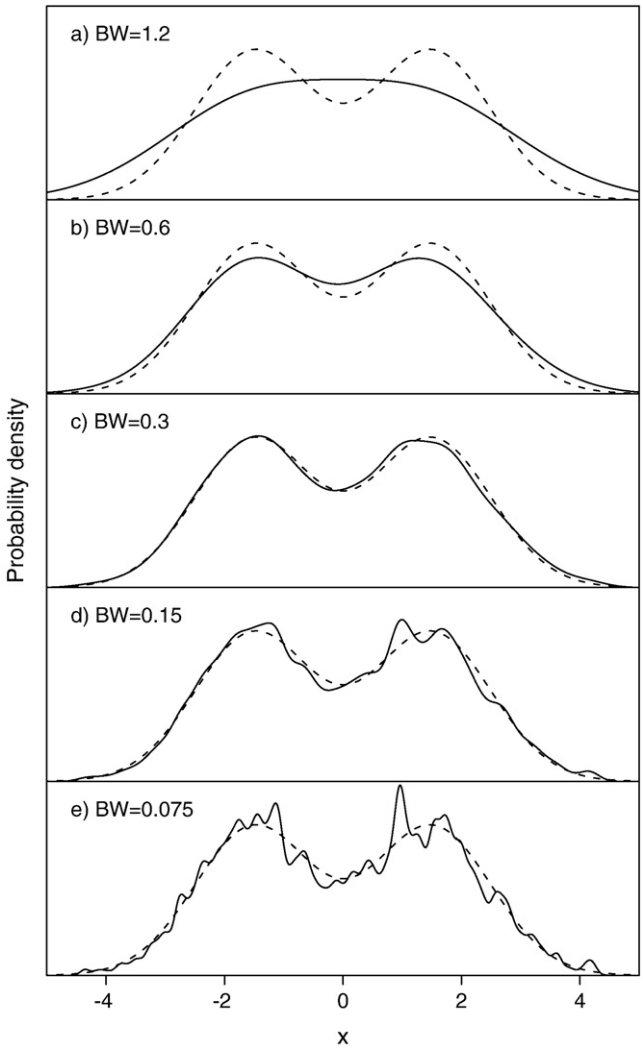

Fig. 1. Kernel density estimates (solid lines) of 1340 random samples drawn from a distribution with true density shown by the dashed line. Five different choices of bandwidth are shown. a) is certainly over smoothed, as the estimate has only a single peak. c) shows a bandwidth choice that is considered optimal by the method of Sheather and Jones (1991). e) is certainly under smoothed, as the estimate shows numerous spurious peaks that are not features of the true distribution.

found, and the estimate is said to be oversmoothed: we have missed important features of the underlying distribution by this choice. On the other hand, if the chosen bandwidth is too small, we undersmooth: the density estimate has far too many peaks, and the many peaks that are observed do not reflect any feature of the true underlying distribution, but are instead a spurious artifact of the sampling. The same effect can be seen in histograms by varying the bin size.

The important question is then, how to choose the bandwidth? In fact, there are a number of techniques that automatically choose a good bandwidth (Jones et al., 1996), and all software packages that implement kernel density estimation come with a default method. These automatic choices of bandwidth typically try to minimise the mean integrated squared error between the density estimate and the unknown true density, based on various assumptions and approximations. For example, the estimate shown in Fig. 1c is close to the bandwidth that is automatically selected by the method of Sheather and Jones (1991) (BW = 0.31), which matches the true density rather well. Silverman's rule of thumb (Silverman, 1986) for a good bandwidth gives a similar estimate (BW = 0.38). Silverman's rule of thumb only works well for near-Gaussian densities, whereas the Sheather and Jones method is more flexible and gives good results for a wider range of densities (see Appendix A and Jones et al. (1996)). While there is still some debate over the best way to automatically choose a good bandwidth, an automatic choice is generally preferable to a manual choice.

Kernel density estimates for Parman's OIB dataset are shown in Fig. 2. In Parman's plots the bandwidth was manually chosen to be around 1500 (compare Fig. 2d of this paper to figs. 1 and 2 of Parman (2007)). A bandwidth chosen by the method of Sheather and Jones (1991) is around 3000 (Fig. 2c), and by Silverman's rule of thumb around 5000, which suggests Parman's density estimates may be undersmoothed and suffer from spurious peaks. There can be good reasons for manually choosing a smaller bandwidth: for example, if one is interested in small scale features of the density function, or if the density is thought to have well separated peaks. However, there is always the danger that many of the peaks found with a small bandwidth are artifacts of the sampling and do not reflect the true distribution. Even with an automatic choice of bandwidth, a few peaks may be seen that do not reflect the true distribution.

3. Feature significance

Since Parman's analysis is based on attaching physical significance to peaks in the density estimates, it is crucial to determine which peaks are statistically significant. Which peaks are really there? This is

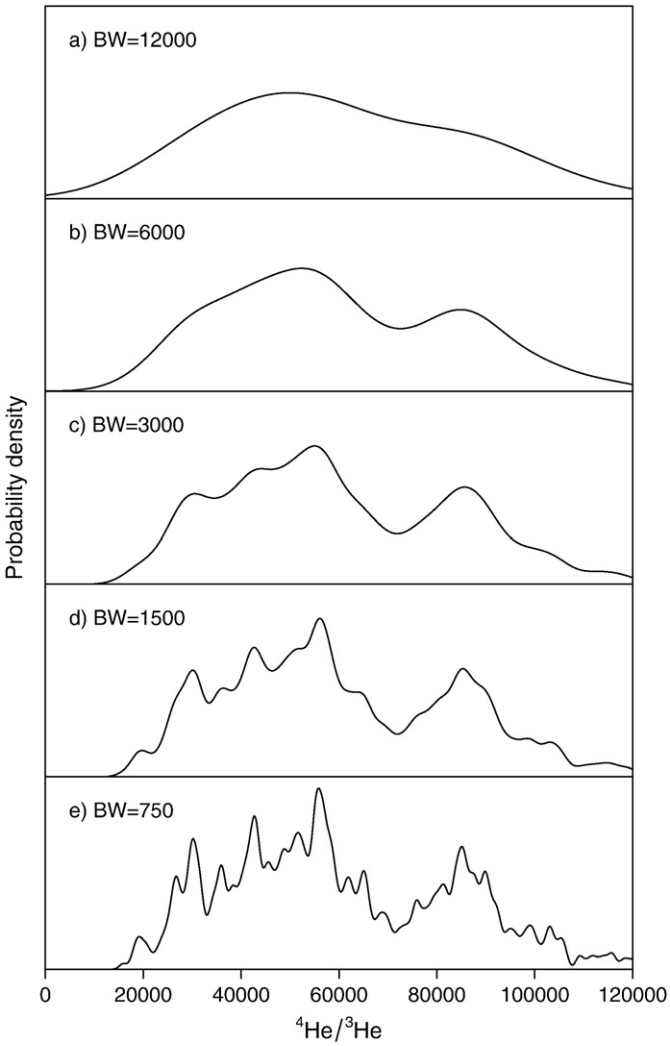

Fig. 2. Kernel density estimates of $^4He/^3He$ OIB data for five different choices of bandwidth. There are 1340 observations in the dataset. a) BW = 120,000 is certainly over smoothed. c) BW = 3000 is the bandwidth that would be automatically chosen by the method of Sheather and Jones (1991). d) BW = 1500 is closest to Parman's choice of bandwidth (see Figs. 1 and 2 of Parman, 2007). e) BW = 750 is certainly under smoothed.

a question that has received a lot of attention in recent years (Chaudhuri and Marron, 1999; Duong et al., in press; Godtliebsen et al., 2002; Hannig and Marron, 2006), and is known in the statistics literature as the problem of feature significance.

One technique for feature significance is the SiZer (Significant Zero crossings of the derivative) method of Chaudhuri and Marron (1999), which is straight-forward to perform in modern software (see Appendix D for a discussion of available software). There are two main ideas to SiZer: First is the notion of scale space — that instead of trying to find the one bandwidth that provides the closest match to the unknown true density, we instead look at the whole range of bandwidths, and explore the different features that occur on different scales. Secondly, peaks and troughs are identified by finding the regions of significant gradient (zero crossings of the derivative). The information is presented in a simple visual way by the SiZer map.

The SiZer map shows location (what is being measured, e.g. ^{4}He/^{3}He) and scale (the bandwidth). The map displays the results of multiple hypothesis tests, which determine whether the gradient at a particular location viewed at a particular bandwidth is significant. As with all hypothesis testing, a significance level must be chosen for the test, and a level of 5% is used here throughout. There are three outcomes of the test: either there is significant positive gradient, a significant negative gradient, or there is simply not enough evidence to say either way (the gradient could be zero, a null hypothesis). In the SiZer map, regions of significant positive gradients are coloured blue; significant negative gradients, red; and no significant gradient, purple. There is also a fourth colour, grey, corresponding to those regions where there is simply not enough data to perform the test. At a given bandwidth, a significant peak can be identified when a region of significant positive gradient is followed by a region of significant negative gradient (i.e. blue–red), and a significant trough by the reverse (red–blue).

A SiZer map for the example bimodal distribution is shown in Fig. 3. In this example, the bimodal structure is clearly brought out by the SiZer map, as the two peaks and the trough can be identified by the left to right blue–red–blue–red pattern on the map that occurs for a range of bandwidths (0.29 to 0.74). Note that for very large bandwidths (>0.74) there is just one significant peak (just blue–red), as would be expected for a large amount of smoothing (see Fig. 1a). For very small bandwidths (<0.18) there is simply not enough information to say anything statistically significant (all purple), and this reflects the fact that large amounts of data are required to resolve small scales.

Fig. 4 shows a SiZer map for Parman's OIB dataset. Only two peaks can be identified: one around 83,000–86,000 (close to the main MORB peak) and one less well located around 46,000–56,000. There is a significant trough around 71,000–74,000. There are additional peaks visible in the Sheather and Jones (1991) density estimate, but these are not statistically significant peaks. Parman's analysis is based on attaching significance to more peaks than just the two found by SiZer.

The story is similar for Parman's other data sets: there are only one or two statistically significant peaks in each set. Fig. 5 shows the corresponding SiZer maps for MORB, Hawaii, and Iceland, alongside kernel density estimates with an automatically chosen bandwidth. The well-known sharp peak in MORB around 88,000–91,000 is clearly identified as a significant peak by the SiZer map. A second peak in MORB is seen in the Sheather and Jones density estimate, but is not a significant SiZer peak. In fact, an additional red region does start to appear at slightly lower significance levels (~10%), so there is possibly a second peak in MORB. In Hawaii there are two peaks: a sharp one around 29,000–31,000, and a less well constrained one around 79,000–86,000. Iceland has a single statistically significant peak around 43,000–53,000. This is at odds with Parman's analysis, in which it is claimed that Hawaii and Iceland have eight statistically significant peaks in their ^{4}He/^{3}He distributions rather than the one or two peaks picked up by the SiZer analysis here.

For Hawaii and Iceland the sample sizes are both less than 500, and it is highly unlikely that eight peaks so close together can be resolved from such a small sample size. An example demonstrating this is shown in Fig. 6, where random samples have been taken from a chosen density that has eight peaks. With only 500 random samples, the eight peaks cannot be resolved, and this is clearly shown in the SiZer map and the density estimate. However, with 5000 random samples, the eight peak structure is clear in both SiZer map and density estimate. For this example, a sample size of around 4000 or higher seems to be required before the eight peaks can be resolved. In general, the number of peaks that can be resolved depends not only on the sample size, but also on the underlying distribution. While eight statistically significant peaks cannot be identified in the helium data sets, it should be noted that the helium–continental crust correlation is based on four main peaks.

4. Gaussian mixture modelling

Kernel density estimation and SiZer are both non-parametric methods: that is to say, they make no assumptions about the underlying functional form of the true density. In some circumstances the underlying functional form may be known, or we may have good reasons for assuming a certain functional form.

In such cases, parametric methods are preferred. Mixture modelling is one such parametric method (e.g. McLachlan and Peel, 2000; Fraley and Raftery, 2002), and typically assumes that the true density is a sum of small number of Gaussian distributions. Given a sample from the distribution, the goal of mixture modelling is to identify the number of Gaussian components, along with their means and standard deviations. Mixture modelling provides an alternative approach to density estimation, and is commonly used in geochronology (Sambridge and Compston, 1994; Jasra et al., 2006). Mixture modelling was also used to suggest a link between the zircon ages and osmium isotopic measurements of mantle samples in a recent study by Pearson et al. (2007).

Fig. 7 shows best fitting Gaussian mixture models for two example distributions and the OIB, MORB, Hawaii and Iceland data sets. The first example shows a mixture model calculated from the same 1340 samples of the bimodal distribution used in Fig. 1. In this case, the true underlying density is a sum of two Gaussians with means at −1.5 and 1.5, both with unit variance, and added together in equal proportion. Unsurprisingly, the mixture model calculated from the random samples does a very good job at recovering the two components.

The second example shows a mixture model calculated from 1340 samples from a uniform distribution (used again later in Fig. 9). In this case, the true density cannot be expressed in the form of a Gaussian mixture model, and thus applying a mixture model to the samples will lead to spurious results. In this example, seven spurious components are identified. This example serves as a warning — one can always calculate a Gaussian mixture model from a set of data, but the results are only meaningful if the underlying assumption can be justified.

If the assumptions are true, then mixture modelling is a much more powerful technique than the non-parametric techniques described earlier. For example, to identify the eight peaks of the example distribution shown in Fig. 6 typically takes at least 4000 samples with SiZer. However, if a mixture model is assumed, the eight components can be identified with around 1500 samples or more. If it is additionally assumed that each of the components have equal variance, then 500 samples are enough.

Mixture modelling identifies three components in the OIB dataset, at 36,000, 55,000, and 86,000, and these three components have reasonable overlap and occur in roughly equal proportion (Fig. 7). The components at 55,000 and 86,000 are close to the two SiZer peaks, but the component at 36,000 does not correspond to a SiZer peak. For MORB two components are identified, the dominant one at 90,000 (74% of the total) matching the SiZer peak, and a smaller component

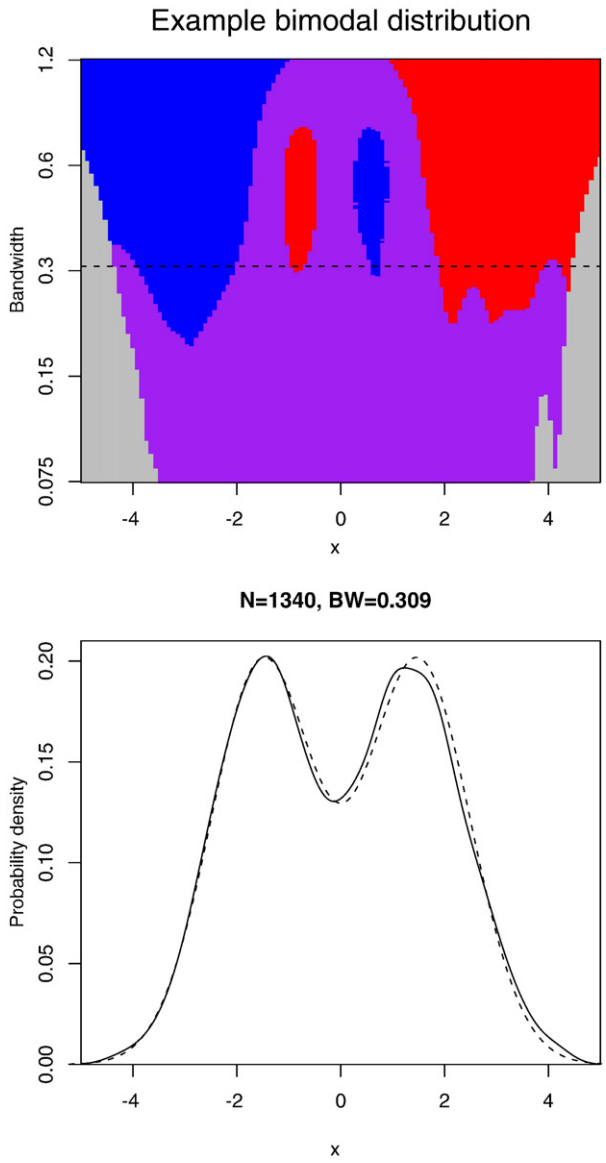

Fig. 3. SiZer map (top) and kernel density estimate (bottom) for the 1340 random samples used in Fig. 1. In the SiZer map, blue shows regions of significant positive gradient; red, region of significant negative gradient; purple, regions where no significant gradient can be identified: the slope could be zero, positive, or negative. Grey regions have insufficient data to make inferences. The bandwidth is on a logarithmic scale. Two peaks and a trough are clearly identified from the blue–red–blue–red pattern, showing that we are able to resolve the bimodal structure of the true density from the random samples. The dashed horizontal line on the SiZer map shows the bandwidth used in the kernel density estimate below, which is as in Fig. 1c with the Sheather and Jones (1991) choice of bandwidth. Notice that the peaks and troughs in the density estimate lie in the purple regions between red and blue on the SiZer map. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

around 62,000. Hawaii splits into three overlapping components with roughly equal proportions, two of which correspond to the SiZer peaks (29,000 and 82,000) and one which does not (51,000). Iceland has two components, a dominant one at 47,000 (73% of the total) corresponding to the SiZer peak, and a broader smaller second component at 82,000 which does not cause a peak in the overall density estimate.

On the whole, the density estimates produced for the helium data by the mixture models look reasonably similar to the kernel density estimates with the Sheather and Jones choice of bandwidth (Figs. 4 and 5). There are some small differences, e.g. Hawaii lacks the deep trough around 74,000 that is seen in the kernel density estimate. In each case mixture modelling has identified one more component than significant peaks identified by SiZer, which is unsurprising since non-parametric techniques such as SiZer will tend to be more conservative than parametric techniques. Despite the extra components, the number of peaks in the mixture model densities is still far fewer than identified by Parman.

The question remains as to whether mixture models are appropriate for studying the helium isotope distributions: Are the components we find spurious, as they are in uniform example of Fig. 7? Is there a good reason for believing that the helium isotope distributions result from a sum of a small number of individual Gaussians? For example, a more appropriate model for the Iceland data might be a single skewed non-Gaussian distribution rather than a sum of two Gaussians (Fig. 7). There are generalisations of mixture modelling that consider sums of non-Gaussian distributions (e.g. Jasra et al. (2006)), but for the helium isotopes it is not clear what the appropriate parametric model should be. Without a clear reason for believing a particular functional form for the underlying density, non-

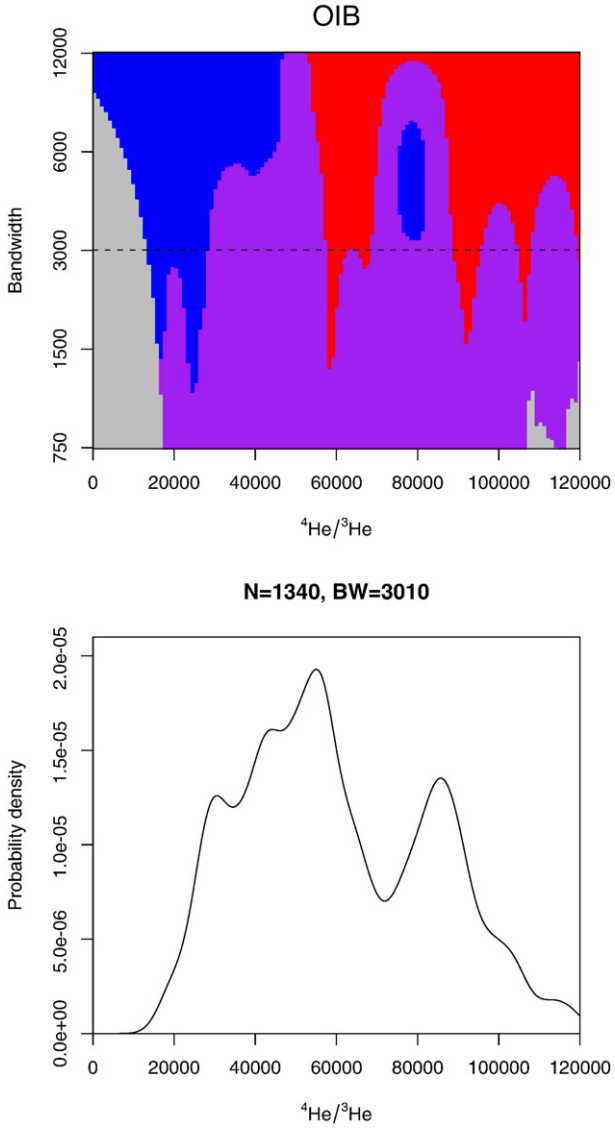

Fig. 4. SiZer map (top) and kernel density estimate (bottom) of Parman's OIB dataset as used in Fig. 2, plotted in the same style as Fig. 3. Kernel density estimate is as in Fig. 2c, with the Sheather and Jones (1991) choice of bandwidth. From the SiZer map it seems there are only two statistically significant peaks that can be resolved, despite the greater number of peaks that are seen in the density estimate.

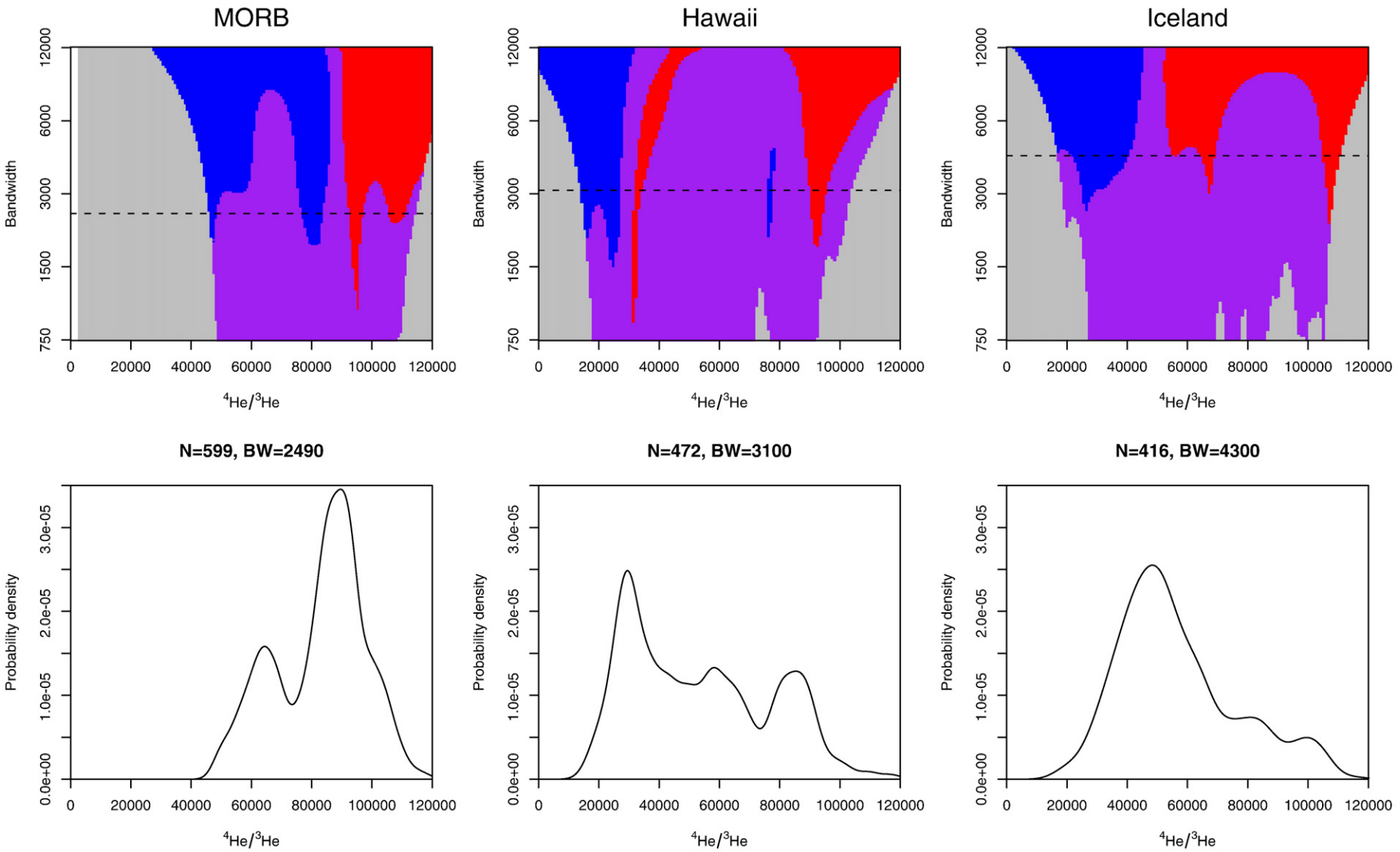

Fig. 5. SiZer maps and kernel density estimates for Parman's MORB, Hawaii and Iceland datasets. Bandwidths for the kernel density estimates are chosen using the method of Sheather and Jones (1991). MORB and Iceland only seem to have one significant peak, Hawaii has two. The dashed horizontal lines on the SiZer maps show the bandwidth corresponding to the kernel density estimates. Notice that the peaks in the kernel density estimates do not match up between Hawaii and Iceland.

parametric methods such as kernel density estimation and SiZer are preferred.

5. From spurious peaks to spurious correlations

From the analyses above, it seems that many of the peaks identified by Parman are not statistically significant. However, the question remains as to why Parman's major peaks correlate so well with the zircon age peaks. Shown in Fig. 8 are kernel density estimates for OIB, along with the four major peaks (69,000, 56,000, 43,000 and 30,000) that Parman associated with the four main zircon age peaks (1.2, 1.9, 2.7, and 3.3 Gyr). Notice that these four ages are almost evenly spaced. As such the four ages will correlate well with any four numbers that are reasonably evenly spaced, e.g. the correlation coefficient of 1, 2, 3, 4 with these ages is 0.9986.

By undersmoothing a density estimate, spurious peaks are produced. Moreover these peaks tend to be reasonably evenly spaced (although not precisely evenly spaced). The spacing of the spurious peaks is controlled by the bandwidth rather than any inherent feature of the underlying distribution. An example is shown in Fig. 9. 1340 random samples were taken from a uniform distribution, and an undersmoothed kernel density estimate calculated. The peaks seen in Fig. 9 are completely spurious, as they do not reflect any feature of the underlying distribution, and yet the first four peaks correlate with the four main zircon ages with a correlation coefficient of 0.9992. When undersmoothing a uniform distribution, the spacing between the spurious peaks is on average around five times the bandwidth, although there is a spread around this average value. Given enough spurious peaks (enough undersmoothing), one can always find a set of four that are very nearly evenly spaced and can be well correlated with the zircon ages. Some of Parman's peaks are likely to be similarly spurious, and their correlation with the zircon ages may not reflect a common underlying physical cause, but may be simply an artifact of the statistical technique.

Another argument Parman puts forth to justify the significance of peaks is their recurrence across different islands. However, if undersmoothed density estimates for different populations use the same bandwidth (as they do in Parman's fig. 1), then artificial correlations

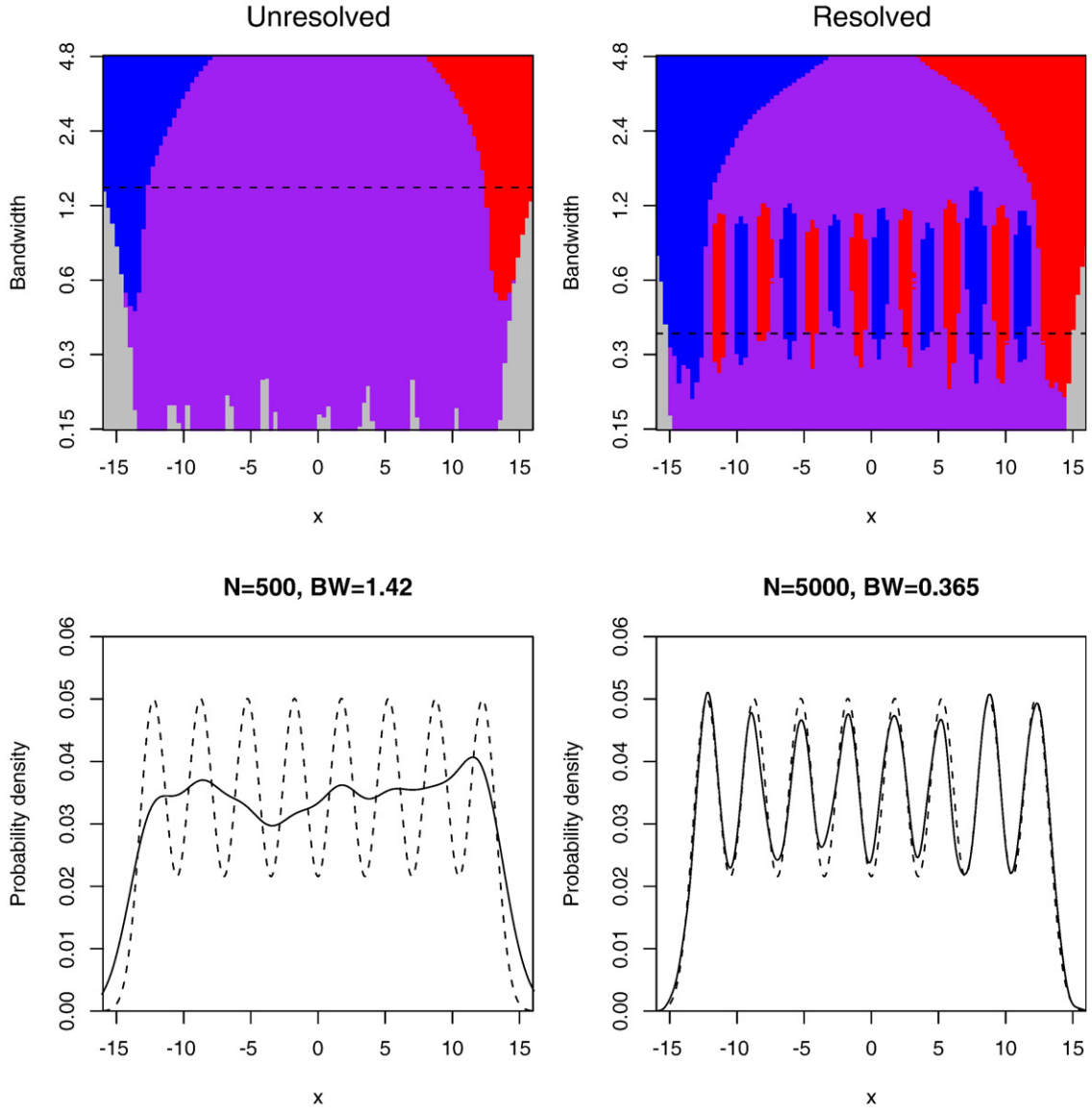

Fig. 6. An example of increasing resolution with increasing sample size, showing SiZer maps and kernel density estimates. Both cases have the same underlying true density, which has eight peaks and is shown by the dashed line in the density plots. The left plots were generated from 500 random samples from the distribution, which is clearly not enough to resolve the peaks. The right plots have a larger sample size (5000), and can clearly resolve the eight peaks.

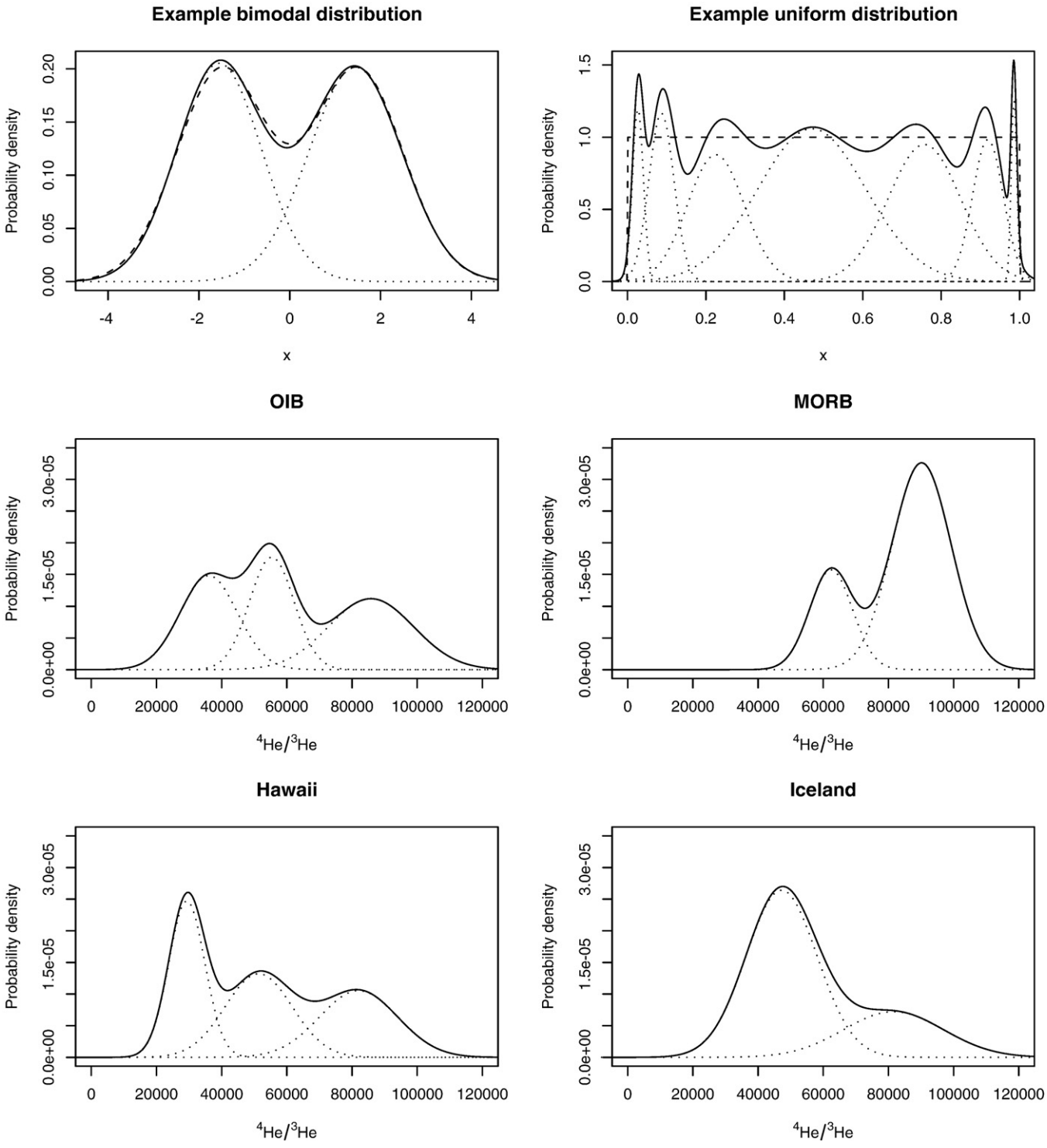

Fig. 7. Best fitting Gaussian mixture models using the method of Fraley and Raftery (2008) (see Appendix C for more details of the fitting procedure). Shown are the bimodal example of Figs. 1 and 3, a uniform example (calculated from 1340 random samples from a uniform distribution on the interval 0 to 1, also used in Fig. 9), and the OIB, MORB, Hawaii and Iceland data sets. Solid lines show the density estimate generated by the mixture model. Dotted lines show the individual Gaussian components which are summed to form the density estimate. Dashed lines show the true densities for the two example distributions.

can arise between the peaks of the different density estimates due to this common choice of bandwidth. An example of this effect is shown for random data in Fig. 10. Two sets of samples were taken from two different uniform distributions, but the bandwidth of their density estimates was chosen to be the same. The spurious peaks caused by undersmoothing line up in some places, just as some of Parman's peaks in Hawaii and Iceland line up. This matching phenemona arises whenever one compares two random signals that have the same average frequency content, where the signals are occasionally seen to be going approximately in phase. Here the two kernel density estimates have very similar average frequency, since their frequency content is controlled by the bandwidth. The matching of Parman's peaks between different islands is likely to be a consequence of the common bandwidth rather than a common physical cause.

There is no recurrence between the different islands using only the statistically significant peaks. The SiZer analysis suggests Hawaii has

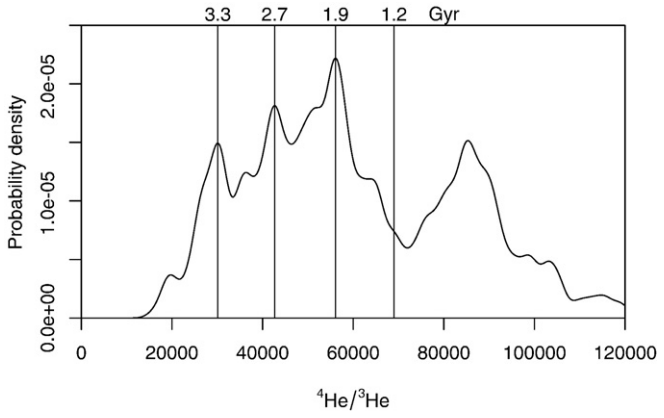

Fig. 8. Kernel density estimate of OIB data with Parman's choice of bandwidth = 1500. Solid vertical lines show the four main peaks corresponding to the four main zircon ages in Parman's correlation (zircon ages shown at top of plot). In fact, the fourth line doesn't intersect a OIB peak, but instead intersects a MORB peak with Parman's choice of bandwidth (see Fig. 1 of Parman (2007)). Notice that the four lines are nearly equally spaced.

peaks at around 29000 and 82000, and Iceland has a peak at around 47000. The only other individual island data set that might be well characterised is Reunion (with 76 samples, see Parman (2007)) which has a peak at around 56000. All other islands have fewer than 70 samples, and are not yet sufficiently well sampled to start identifying clear peaks. Mixture modelling identifies only one further peak, at around 51000 in Hawaii, and that too does not match with any other island. MORB has peaks at around 90000 and maybe also 62000, but these do not match any of the above peaks either.

6. Discussion

It would be extremely useful to perform a SiZer analysis on the other data set used in Parman's correlation, namely the crustal zircon age data (Condie, 1998). Unfortunately, unlike the helium isotopic data (Abedini et al., 2006), there is not a publicly available compilation of zircon ages, and so the analysis has not been done. The zircon age histogram appears much more strongly peaked than the ^{4}He/^{3}He distributions, and I suspect that a SiZer analysis will find several statistically significant zircon age peaks. I hope that future zircon age compilations will be more accessible (e.g. Voice et al., 2007), and that the statistical significance of peaks will be thoroughly tested.

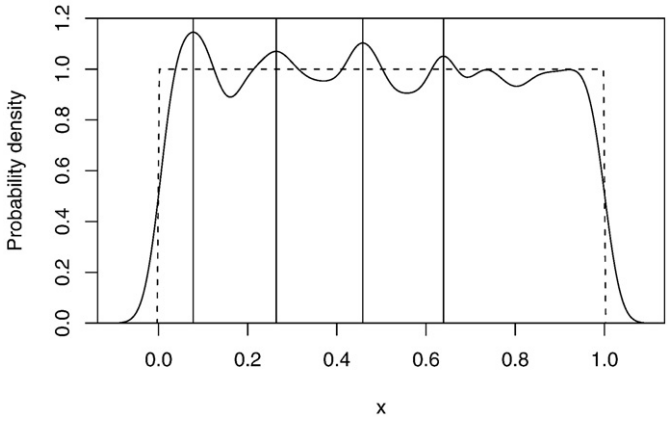

Fig. 9. An example of nearly evenly spaced spurious peaks. 1340 random samples have been taken from a uniform distribution (with true density shown by the dashed line), and a kernel density estimate calculated with a bandwidth of 0.03. Four evenly spaced spurious peaks have been marked with vertical lines, and correlate with the four main zircon ages (1.2, 1.9, 2.7, 3.3 Gyr) with a correlation coefficient of 0.9992.

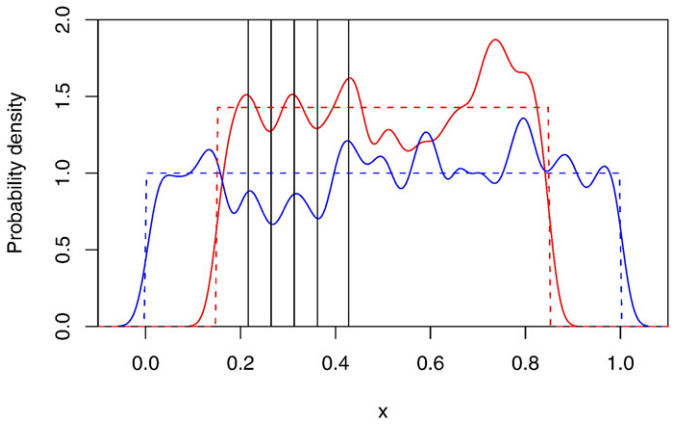

Fig. 10. An example of matching of spurious peaks across different distributions. Shown are undersmoothed kernel density estimates for two separate samples. The first (red) is 500 samples from a uniform distribution on the interval 0 to 1, the second (blue) is 500 samples from a uniform distribution on interval 0.15 to 0.85. Dashed lines show true PDFs. Both kernel density estimates have a bandwidth of 0.02. Vertical lines illustrate matching between some of the spurious peaks. The average spacing between spurious peaks is around five times the chosen bandwidth (0.10).

The techniques discussed here take no account of measurement error: it is assumed we are sampling from the true underlying density without error. Deconvolution kernel density estimation (Stefanski and Carroll, 1990; Delaigle and Gijbels, 2004) is one technique that takes account of measurement error, but it is more involved and not yet available in standard software. Techniques that take account of measurement error may prove very useful.

Feature significance is an area of active research (Duong et al., in press; Hannig and Marron, 2006), and its future development is likely to benefit geochemistry greatly. Feature significance extends beyond the 1D density estimation problem described here. The multidimensional generalisations of SiZer (Duong et al., in press; Godtliebsen et al., 2002) could be useful in finding the interesting peaks in multi-dimensional isotope ratio space (e.g. including ^{143}Nd/^{144}Nd, ^{207}Pb/^{204}Pb, etc.) Moreover, SiZer can also be used for examining scatterplot smooths (Chaudhuri and Marron, 1999, 2000), which could potentially be useful for identifying significant features in the spatial pattern of isotope ratios (e.g. isotopic data sets along the ridges, Agranier et al., 2005).

As Parman (2007) has remarked, it should be emphasised that there are further complications in attaching meaning to peaks, even if they have been shown to be statistically significant. Perhaps the most important issue is that in practice we are not simply drawing random samples from a population: geochemists' sampling is anything but random. Some areas are sampled more frequently than others, simply because they are more accessible, and this may well result in statistically significant peaks that have nothing to do with mantle evolution. Even statistically significant peaks require care before attaching any physical meaning to them.

7. Conclusions

Kernel density estimation is a powerful technique for estimating PDFs, and deserves to be used routinely. However, one must always be conscious of bandwidth effects, and in particular it is best not to change from the default choice without good reason. Moreover, one must be careful not to attach too much significance to every small bump found in such density estimates. Feature significance techniques such as SiZer are one way of deciding which peaks are significant, and SiZer is particularly straightforward to perform in modern software.

SiZer analysis suggests that there are only two statistically significant peaks in the OIB ^{4}He/^{3}He distribution. The zircon age data has not been analysed, but probably contains more peaks. The

more measurements that are made, the better we will be able resolve small scale features in the distributions. Maybe we will start seeing additional peaks in the helium data that can be related to the crustal ages, and a helium–continental crust correlation will be established. But for now, the correlation is not supported by the data.

Acknowledgements

I am very grateful to Steve Parman for providing me with the helium isotopic data used in Parman (2007), which is a filtered version of the compilation in Abedini et al. (2006) that includes only values of $^4He/^3He<120000$ ($^3He/^4He>6$ R_A). I thank Marc Spiegelman for all his advice and support, and thank Steve Parman and Francis Albarède for their constructive reviews. This work was inspired by discussions at the "FUN" geochemistry reading group at the Lamont–Doherty Earth Observatory, and I thank Cornelia Class, Rajdeep Dasgupta, Al Hofmann, John Maclennan, Steve Marron and Terry Plank for their feedback. This work was supported by NSF grant OCE-0452457.

Appendix A. Kernel density estimation

Given a sample of n data points $X_1, X_2, \ldots, X_n$ independently drawn from a distribution with probability density $f(x)$, the kernel density estimator $\hat{f}_h(x)$ is defined by

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i), \tag{A.1}$$

where $K_h(y)=(1/h)K(y/h)$ and h is the bandwidth. $K(y)$ is the chosen kernel function, which is assumed throughout this paper to be a Gaussian,

$$K(y) = \frac{1}{\sqrt{2\pi}}\exp(-y^2/2). \tag{A.2}$$

The problem of bandwidth selection is to choose a value of h such that the estimator $\hat{f}_h(x)$ provides a good estimate of the true probability density $f(x)$. One measure of the quality of the estimator is the mean integrated square error (MISE), defined by

$$\mathrm{MISE} = E\int_{-\infty}^{\infty}\left(f(x)-\hat{f}_h(x)\right)^2 dx, \tag{A.3}$$

where E denotes expectation. Ideally, we would like to minimise the MISE by an appropriate choice of bandwidth. However, we cannot calculate the MISE because we don't know $f(x)$. Indeed, if we knew $f(x)$ there would be no point in doing density estimation! One solution to this problem is to minimise an approximation to the MISE rather than the MISE itself. When the number of data points is large ($n\to\infty$) the MISE is well approximated by the asymptotic mean integrated squared error (AMISE) given by

$$\mathrm{AMISE} = \frac{R(K)}{nh} + \frac{h^4 R(f'')S(K)^2}{4}, \tag{A.4}$$

where

$$R(\phi) = \int_{-\infty}^{\infty}\phi(x)^2 dx, \tag{A.5}$$

$$S(\phi) = \int_{-\infty}^{\infty}x^2\phi(x)dx. \tag{A.6}$$

The first term on the right hand side of Eq. (A.4), $R(K)/nh$, is the integrated variance of the estimator, and the second term is the integrated squared bias of the estimator. The two terms behave in opposite ways as the bandwidth h is varied. For large values of the bandwidth h the estimator has low variance but high bias, whereas for small values of h it has low bias but high variance. The AMISE is minimised when $h=h_{\mathrm{AMISE}}$ where

$$h_{\mathrm{AMISE}} = \left(\frac{R(K)}{nR(f'')S(K)^2}\right)^{1/5}. \tag{A.7}$$

Importantly, from this expression we see that the optimal bandwidth decreases with increasing n, scaling as $h_{\mathrm{AMISE}}\sim n^{-1/5}$. This reflects the fact that given more data we can resolve smaller scale features of the density.

For a Gaussian kernel, $R(K)=1/(2\sqrt{\pi})$ and $S(K)=1$. Unfortunately, h_{AMISE} still depends on the unknown true density $f(x)$ through $R(f'')$. The final step is to replace $R(f'')$ by an approximation based on the data (known as the "plug-in" approach), and it is in this step that the different bandwidth selection methods used in this paper differ.

A.1 Silverman's rule of thumb

Silverman's rule of thumb (Silverman, 1986) is based on estimating $R(f'')$ by assuming f is a Gaussian distribution with variance σ^2, namely

$$R(f'')\approx\frac{3}{8\sqrt{\pi}\sigma^5}, \tag{A.8}$$

leading to

$$h_{\mathrm{Gaussian}} = \left(\frac{4\sigma^5}{3n}\right)^{1/5} = 1.06\sigma n^{-1/5}. \tag{A.9}$$

The standard deviation σ can then be estimated using the sample standard deviation. Silverman's rule of thumb is actually a slight modification of the above, where

$$h_{\mathrm{Silverman}} = 0.9\sigma n^{-1/5}, \tag{A.10}$$

where σ is the minimum of the sample standard deviation and the interquartile range divided by 1.34 (which makes the estimator a bit more robust against outliers). This choice of bandwidth tends to be good if the true distribution is near Gaussian (as would be expected, given the assumption above), but may perform poorly on other distributions, particularly multi-modal ones.

A.2 Sheather and Jones

The method of Sheather and Jones (1991) uses a more sophisticated technique to estimate $R(f'')$, and is based on solving the fixed point equation

$$h = \left(\frac{R(K)}{nR\left(\hat{f}''_{g(h)}\right)S(K)^2}\right)^{1/5}. \tag{A.11}$$

where a kernel density estimator has been used in place of f''. The bandwidth $g(h)$ for the estimator of $R(f'')$ differs from h because bandwidths that are good for curve estimation differ from those appropriate for estimating $R(f'')$. How this better bandwidth $g(h)$ is chosen is an important part of the method, and the interested reader is referred to Sheather and Jones (1991) for the technical details. The Sheather and Jones method provides a much better approximation to h_{AMISE} than Silverman's rule of thumb but is computationally more expensive.

Appendix B. Feature significance

The basis of the SiZer method is the notion of scale space. Instead of trying to find a single bandwidth h which produces an estimator $\hat{f}_h(x)$ that best matches the true underlying curve $f(x)$, the whole range of bandwidths is considered important to constrain features on different scales. Moreover, instead of basing the statistical inference on a comparison between $\hat{f}_h$ and f, the comparison is made between $\hat{f}_h$ and f_h where $f_h = f * K_h$ is a convolution of the true density with the kernel (i.e. we are comparing a kernel density estimate with an appropriately blurred version of the true density).

In constructing the SiZer map, at each point (x,h) a confidence interval is constructed for the derivative of the kernel density estimate. The confidence interval takes the form

$$\hat{f}'_h(x) \pm q \cdot \hat{SD}(\hat{f}'_h(x)), \tag{B.1}$$

where $\hat{SD}$ denotes the estimate of standard deviation and q is an appropriate quantile. If 0 lies outside this interval the gradient is said to be significant, and it appears as red (negative) or blue (positive) on the SiZer map. If 0 is within the interval, the gradient is not significant, and it appears as purple on the SiZer map. q depends on the chosen significance level α, and its calculation is rather subtle due to the simultaneous nature of the hypothesis test (see Chaudhuri and Marron (1999) and Duong et al. (in press) for further details).

Appendix C. Gaussian mixture modelling

In an m-component Gaussian mixture model it is assumed that the true probability density takes the form

$$f(x) = \sum_{i=1}^{m} w_i p(x;\mu_i,\sigma_i), \tag{C.1}$$

where

$$p(x;\mu_i,\sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\left(\frac{x-\mu_i}{\sigma_i}\right)^2\right), \tag{C.2}$$

and

$$\sum_{i=1}^{m} w_i = 1, \quad w_i > 0. \tag{C.3}$$

The $\{\mu_i\}$ and $\{\sigma_i\}$ are the means and standard deviations of the m individual Gaussian components, and $\{w_i\}$ are the weights, which specify the proportion each component contributes to the overall population.

Given a sample of n data points $X_1, X_2, \ldots X_n$ from the distribution with density $f(x)$, the goal in mixture modelling is to estimate the parameters, m, $\{\mu_i\}$, $\{\sigma_i\}$, and $\{w_i\}$ from the data. Some parameters may be fixed in advance and not determined from the data: for example, a certain number of components might be assumed, or it might be assumed that all the components have the same variance. However, for the best fitting mixture models shown in Fig. 7, no parameters were fixed in advance. The best fit means, variances, and proportions were estimated from the data by a maximum likelihood approach, using the Expectation–Maximization (EM) algorithm. The best-fit number of components was determined by the Bayesian Information Criterion (BIC), which penalizes models that are overly complex (have too many free parameters, see Fraley and Raftery (2008) for further details).

Appendix D. Software availability

All data analysis was performed in the freely available "R" statistical software (R Development Core Team, 2008), version 2.6.2 (http://www.r-project. org). Kernel density estimation was performed with the built-in function "density", SiZer maps were calculated using the "feature" package, version 1.1–12 (Duong et al., in press), and Gaussian mixture models were calculated using the "mclust" package, version 3.1–3 (Fraley and Raftery, 2008). In MATLAB, kernel density estimation can be performed with "ksdensity" in the Statistics toolbox, and routines for performing SiZer analysis are available at http://www.stat.unc.edu/faculty/marron/marron_software.html. There are also online web-based applets available for these techniques: kernel density estimation at http://www.wessa.net, and SiZer at http://www.wagner.com.

References

Abedini, A.A., Hurwitz, S., Evans, W.C., 2006. USGS-NoGaDat — a global dataset of noble gas concentrations and their isotopic ratios in volcanic systems. US Geological Survey Digital Data Series 202. http://pubs.usgs. gov/ds/2006/202.

Agranier, A., Blichert-Toft, J., Graham, D., Debaille, V., Schiano, P., Albarède, F., 2005. The spectra of isotopic heterogeneities along the mid-Atlantic Ridge. Earth Planet. Sci. Lett. 238, 96–109.

Chaudhuri, P., Marron, J.S., 1999. SiZer for exploration of structures in curves. J. Am. Stat. Assoc. 94, 807–823.

Chaudhuri, P., Marron, J.S., 2000. Scale space view of curve estimation. Ann. Stat. 28, 408–428.

Condie, K.C., 1998. Episodic continental growth and supercontinents: a mantle avalanche connection? Earth Planet. Sci. Lett. 163, 97–108. (doi:10.1016/S0012-821X(98)00178-2.).

Delaigle, A., Gijbels, I., 2004. Practical bandwidth selection in deconvolution kernel density estimation. Comput. Stat. Data Anal. 45, 249–267. (doi:10.1016/S0167-9473(02)00329-8.).

Duong, T., Cowling, A., Koch, I., Wand, M.P., 2008. Feature significance for multivariate kernel density estimation. Comp. Stat. Data. Anal. 52, 4225–4242. (doi:10.1016/j.csda.2008.02.035).

Fraley, C., Raftery, A., 2008. mclust: model-based clustering/normal mixture modeling. R package version 3, pp. 1–3. URL http://www.stat.washington.edu/fraley/mclust.

Fraley, C., Raftery, A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. 97, 611–631.

Godtliebsen, F., Marron, J.S., Chaudhuri, P., 2002. Significance in scale space for bivariate density estimation. J. Comput. Graph. Stat. 11, 1–21.

Hannig, J., Marron, J.S., 2006. Advanced distribution theory for SiZer. J. Am. Stat. Assoc. 101, 484–499 (doi:10.1198/016214505000001294.).

Jasra, A., Stephens, D.A., Gallagher, K., Holmes, C.C., 2006. Full Bayesian mixture modelling in geochronology via Markov Chain Monte Carlo. Math. Geol. 38, 269–300 (doi:10.1007/s11004-005-9019-3.).

Jones, M.C., Marron, J.S., Sheather, S.J., 1996. A brief survey of bandwidth selection for density estimation. J. Am. Stat. Assoc. 91, 401–407.

Kemp, A.I.S., Hawkesworth, C.J., Paterson, B.A., Kinny, P.D., 2006. Episodic growth of the Gondwana supercontinent from hafnium and oxygen isotopes in zircon. Nature 439, 580–583 (doi:10.1038/nature04505.).

McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley.

Parman, S.W., 2007. Helium isotopic evidence for episodic mantle melting and crustal growth. Nature 446, 900–903 (doi:10.1038/nature05691.).

Pearson, D.G., Parman, S.W., Nowell, G.M., 2007. A link between large mantle melting events and continent growth seen in osmium isotopes. Nature 449, 202–205 (doi:10.1038/nature06122.).

Porcelli, D., 2007. When crust is bred. Nature 446, 863–864 (doi:10.1038/446863a.).

R Development Core Team, 2008. : A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Sambridge, M.S., Compston, W., 1994. Mixture modeling of multi-component data sets with application to ion-probe zircon ages. Earth Planet. Sci. Lett. 128, 373–390 (doi:10.1016/0012-821X(94)90157-0.).

Sheather, S.J., Jones, M.C., 1991. A reliable data-based bandwidth selection method for kernel density estimation. J. R. Stat. Soc., Ser. B Stat. Methodol. 53, 683–690.

Silverman, B.W., 1986. Density Estimation. Chapman and Hall, London.

Stefanski, L.A., Carroll, R.J., 1990. Deconvoluting kernel density estimators. Statistics 21, 169–184.

Voice, P.J., Eriksson, K., Kowaleski, M., 2007. Global episodic growth of continental crust recorded by detrital zircons. GSA Denver Annual Meeting.