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 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

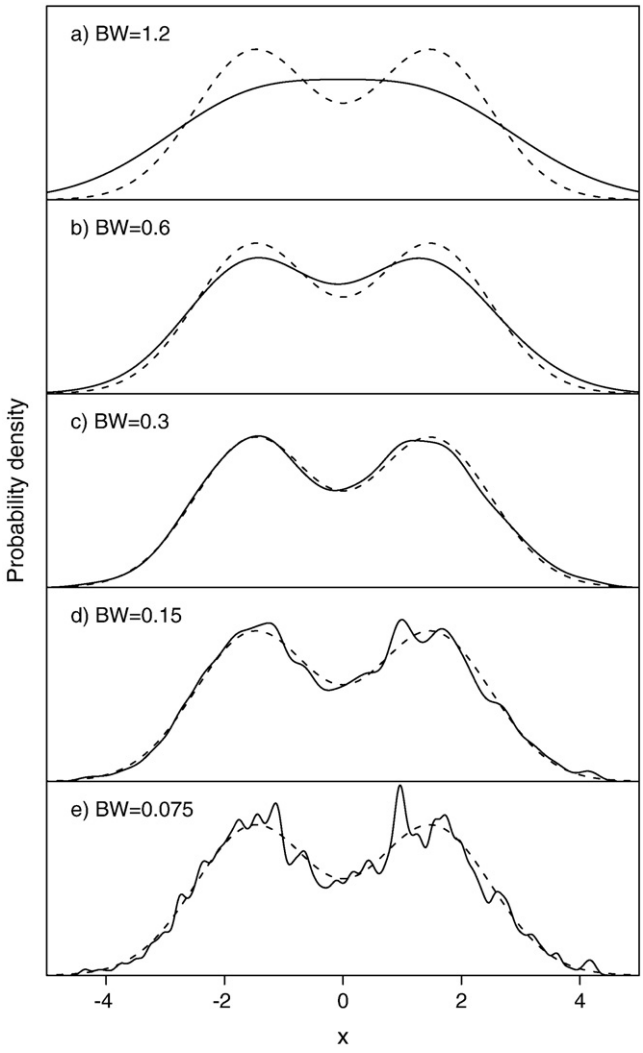


**Fig. 1.** Kernel density estimates (solid lines) of 1340 random samples drawn from a distribution with true density shown by the dashed line. Five different choices of bandwidth are shown. a) is certainly over smoothed, as the estimate has only a single peak. c) shows a bandwidth choice that is considered 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

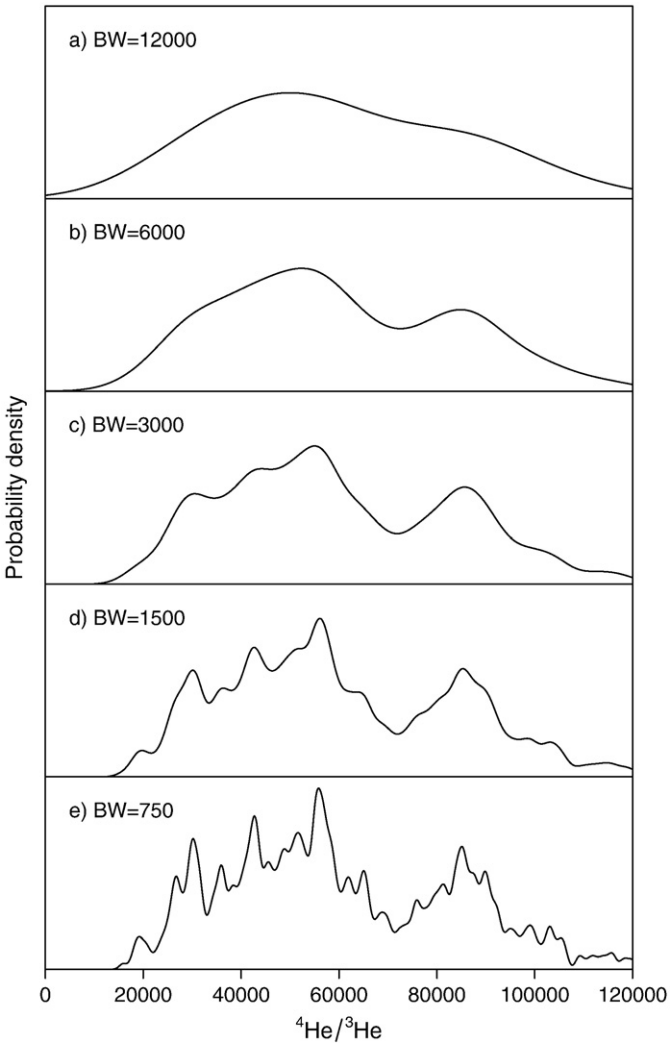


**Fig. 2.** Kernel density estimates of $^4He/^3He$ OIB data for five different choices of bandwidth. There are 1340 observations in the dataset. a) BW = 120,000 is certainly over smoothed. c) BW = 3000 is the bandwidth that would be automatically chosen by the method of Sheather and Jones (1991). d) 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.