Math Geol (2007) 39: 409–417
DOI 10.1007/s11004-007-9103-y

Statistical Segmentation of Geophysical Log Data

Danilo R. Velis

Received: 2 June 2006 / Accepted: 18 December 2006 / Published online: 14 August 2007

Abstract Stationary segments in well log sequences can be automatically detected by searching for change points in the data. These change points, which correspond to abrupt changes in the statistical nature of the underlying process, can be identified by analysing the probability density functions of two adjacent sub-samples as they move along the data sequence. A statistical test is used to set a significance level of the probability that the two distributions are the same, thus providing a means to decide how many segments comprise the data by keeping those change points that yield low probabilities. Data from the Ocean Drilling Program were analysed, where a high correlation between the available core-log lithology interpretation and the statistical segmentation was observed. Results show that the proposed algorithm can be used as an auxiliary tool in the analysis and interpretation of geophysical log data for the identification of lithology units and sequences.

Keywords Data mining · Segmentation · Zonation · Change point · Probability density function

Introduction

Segmentation is an important data mining process. One important application is the identification of locally stationary intervals or, equivalently, the location of change points. In this context, segmentation (also known as zonation) is the dividing of a sequence into relatively homogeneous and stationary intervals such that each segment

D.R. Velis (✉)
Facultad de Ciencias Astronómicas y Geofísicas, Universidad Nacional de La Plata, La Plata, Argentina
e-mail: velis@fcaglp.unlp.edu.ar

D.R. Velis
CONICET, Paseo del Bosque s/n, B1900FWA La Plata, Argentina

is distinctive from the adjacent ones. Well logs can be subdivided into relatively uniform segments that represent zones of similar lithologic character (stratigraphic units and formations). Segment boundaries correspond to abrupt changes in the layering and conform the limits of relatively stable periods or geologically meaningful zones. These elementary units of similar properties can then be used as the basis for inferring correlations between wells. A different approach consists of blocking or filtering the data to get a simpler approximation (e.g. piecewise constant segments). This segmentation problem will not be considered here, and the reader is referred to, for example, Kaaresen and Taxt (1998) and the references therein for details. In this work, the focus is on the identification of statistically distinct intervals in the log sequences.

There are various strategies for addressing this segmentation problem. Classical approaches include the detection of abrupt changes in the mean (Webster 1973) or in the variance (Gill 1970; Hawkins and Merriam 1973). A general description of these techniques is given in Davis (1986). Recent studies include zonation by means of cluster analysis (Gill et al. 1993), spectral analysis for identifying stationary intervals (Ligges et al. 2002), etc. The method presented here takes into account both the mean and the variance, and also higher-order robust statistics such as certain non conventional skewness and kurtosis measures (Velis 2003) to identify change points. Essentially, a split window is moved along the sequence and the probability density functions (pdf) of the two adjacent half-windows are compared. When a significant difference is detected, a change point is identified. Smooth pdfs are estimated using the maximum entropy method as described in Velis (2003), which guarantees robustness when dealing with short data sequences. Finally, a criterion for deciding which is the number of segments that comprise the data is proposed.

The effectiveness of this strategy is supported by the analysis of various examples using simulated and real data sequences derived from well-logs. The real data comprises various borehole measurements which are part of the Ocean Drilling Program, Leg 197, Site 1203 (Tarduno et al. 2002). At this site, the lithology interpretation based on extensive core samples and logging data analysis was available, so it was possible to make a comparison between this interpretation and the statistical segmentation. Results show that there is a high correlation between the published core-log lithology and the segmentation generated by the proposed statistical procedure.

The Problem

Let $\vec{x} = (x_1, x_2, \ldots, x_N)$ be the sequence of well log data. The objective of the segmentation process is to subdivide this series into smaller segments so that each interval is relatively locally stationary. That is, we look for a sequence of change points

$$\vec{t} = (t_1, t_2, \ldots, t_M) \tag{1}$$

which satisfy

$$1 = t_1 < t_2 < \cdots < t_{M-1} < t_M = N. \tag{2}$$

These indexes determine a set of $M - 1$ segments of length

$$T_j = t_{j+1} - t_j. \tag{3}$$

In practice the algorithm proceeds iteratively by searching successive change points $\{t_j\}$ based on the assumption that two adjacent intervals are distinct when the pdfs of the data on each side of t_j are significantly different. For this purpose, a split window of length $2L$ is centered at location t_j, and the corresponding pdfs are estimated and compared appropriately.

Here, L should be short enough to allow for the identification of short stationary intervals. Thus, a robust pdf estimation method that works well even for short data sequences is required. The maximum entropy (MaxEnt) method with moment constraints described in Velis (2003) produces smooth non-parametric pdfs which are consistent with the data. The approach utilizes robust statistics computed directly from the data to constrain the maximization of the pdf entropy. These statistics (called S-measures) involve the non-conventional skewness and kurtosis indices that measure shape and proved to be appropriate to identify the main features of the distribution of primary reflection coefficients (reflectivity). In exploration seismology, the reflection coefficient is the ratio of the amplitude of the displacement of the reflected wave to that of the incident wave. The reflectivity is one of the components, together with the seismic wavelet, of the so-called convolutional model of the seismic trace, especially valid for layered geological models (Yilmaz 2001).

The strategy to carry out the segmentation is based on the sliding window approach, which consists of moving the analysing window along the whole sequence and assigning a change point when a significant difference between the pdfs is observed. To avoid the assigning of change points which are too close, we found it more appropriate to look for a single change point at a time. Starting with $j = 2$ (recall that $t_1 = 1$), we look for optimum change points until the next change point that is added does not yield a significant difference between the adjacent pdfs. These optimum change points correspond to the smallest probabilities along the whole sequence for the current iteration.

The Algorithm

Let $\hat{t}$ be the current estimate of the jth change point. Let $\vec{u} = (x_{\hat{t}-L}, x_{\hat{t}-L+1}, \ldots, x_{\hat{t}})$ and $\vec{v} = (x_{\hat{t}}, x_{\hat{t}+1}, \ldots, x_{\hat{t}+L})$ be the two subsets of $\vec{x}$ spanned by the split window, and let $\hat{p}_u(\vec{u})$ and $\hat{p}_v(\vec{v})$ be the corresponding estimated pdfs, which are to be compared. Rather than measuring the difference between $\hat{p}_u$ and $\hat{p}_v$, we measure the difference between their respective cumulative distribution functions (cdf), $\hat{P}_u$ and $\hat{P}_v$, using the Kuiper test (the cdfs are calculated numerically by integrating the pdf estimates). The Kuiper test, a variant of the well-known Kolmogorov–Smirnov test (Press et al. 1992), quantifies the difference between two cdfs. The Kuiper statistic is given by

$$V = \max_{a \le x \le b} \left(\hat{P}_u - \hat{P}_v\right) + \max_{a \le x \le b} \left(\hat{P}_v - \hat{P}_u\right), \tag{4}$$

where a and b define the region of support of the cdf (usually the minimum and maximum values in the data set). It turns out that the distribution in the case of the null hypothesis that the two data segments come from the same distribution can be

calculated asymptotically, giving rise to a formula that allows one to compute the significance level (Press et al. 1992)

$$\text{Probability}(V > \text{observed}) = 2\sum_{i=1}^{\infty}\left(4i^2\lambda^2 - 1\right)e^{-2i^2\lambda^2}, \tag{5}$$

where

$$\lambda = \left(\sqrt{\frac{L}{2}} + 0.155 + 0.24\sqrt{\frac{2}{L}}\right)V. \tag{6}$$

The segmentation algorithm is a three stage process. In the first stage, the probability (5) is calculated for every possible change point location throughout the whole sequence in the range $(L, N-L)$. In the second stage, change points candidates are added according to the following strategy: at the beginning, the point with the smallest probability is selected as a candidate for the first change point, yielding t_2 and the new segmentation (t_1, t_2, t_3), which is comprised of two segments of lengths T_1 and T_2, respectively. Then, a new change point is added by selecting the smallest probability within the current longest segment (largest T_j), giving rise to a new partition (t_1, t_2, t_3, t_4). This process is repeated and new change points are added (within the longest segments obtained so far) until all segments are shorter than a given minimum length, $T_{\min}$.

The third stage of the algorithm consists of discarding those change points whose associated probabilities are larger than a predefined threshold. Also, the change points with largest probabilities in excess of a predefined number of change points are deleted. Note that a large probability is indicative of a high degree of confidence on the null hypothesis that the two distributions are the same, so low values of probability are desired to obtain a high confidence on the hypothesis that the two distributions are different. To avoid too fine segmentations (i.e. two change points separated by a few samples), a minimum separation Δ between two consecutive change points is forced by adjusting the search range accordingly.

Step by step, the algorithm is as follows.

1. Set $j = 1$.
2. For every $\hat{t}$ in the initial search range $(L, N-L)$:
   (a) Set $\vec{u} = (x_{\hat{t}-L}, x_{\hat{t}-L+1}, \ldots, x_{\hat{t}})$ and $\vec{v} = (x_{\hat{t}}, x_{\hat{t}+1}, \ldots, x_{\hat{t}+L})$.
   (b) Estimate $\hat{p}_u$ and $\hat{p}_v$ using the MaxEnt method.
   (c) Estimate $\hat{P}_u$ and $\hat{P}_v$ by numerical integration.
   (d) Compute V and evaluate the probability (5).
3. Set $j = j + 1$.
4. Find the smallest probability within the current search range to get a new optimum change point t_j.
5. Sort, in ascending order, the current set of change points and update the segmentation $(t_1, t_2, \ldots, t_j, t_M)$.
6. Compute segment lengths according to (3).
7. Update the search range: $(t_{j\max} + \Delta, t_{j\max+1} - \Delta)$, where $t_{j\max}$ is the beginning of the longest segment, $T_{\max}$.
8. If $T_{\max} > T_{\min}$, go to Step 3.

9. Delete all change points whose probabilities are larger than a predefined significance level.
10. Delete all change points in excess of a predefined maximum number of change points whose probabilities are largest.

Test Results

To check the consistency of the segmentation algorithm, we applied it to the simulated sequence (8400 random values) shown in Fig. 1. The sequence was generated by concatenating samples drawn from eight different non-parametric distributions. These distributions were selected so as to simulate a realistic reflectivity sequence (Velis 2003) and are shown in Fig. 2.

In the segmentation process we set $L = 250$ and $\Delta = 200$, and change points were added until no segment was larger than $T_{\min} = 200$ samples. At the end of the process, the change points with the associated probability larger than 0.01 were discarded. This significance level was chosen based on the inspection of Fig. 3, where the probability (5) was plotted in ascending order for all the identified change points. For values larger than about 0.01, the probability of the null hypothesis that the two distributions are the same increases rapidly. The estimated change points are shown in Fig. 1 and in Table 1, along with the correct change points. All eight segments were identified correctly.

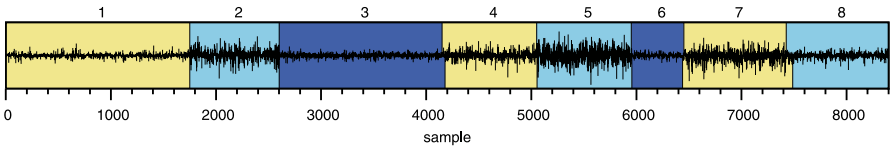

Fig. 1 Simulated random sequence comprised of eight statistical independent segments. The segmentation is indicated by *vertical lines*: true (*top*) and estimated (*bottom*). Table 1 shows the exact location of the change points

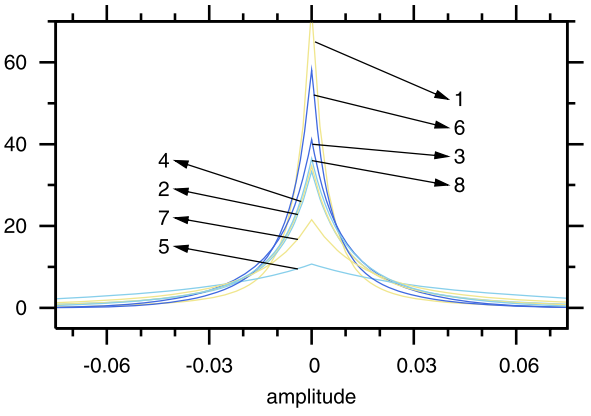

Fig. 2 Probability density functions used to generate the reflectivity sequence shown in Fig. 1

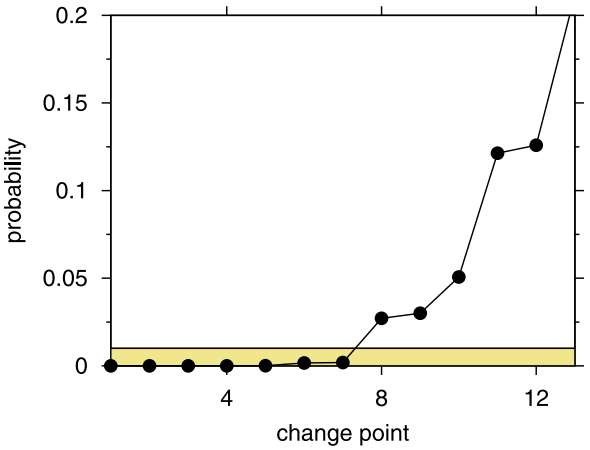

Fig. 3 Probability of the null hypothesis that the two distributions are the same. The plot reveals an abrupt change at about 0.01, a value which is selected as a threshold to discard change points with high probabilities in the third stage of the segmentation process

Table 1 The eight segments used to build the sequence shown in Fig. 1 and their corresponding change points (true and estimated), Kuiper statistics and associated probability

pdf	t_j	$\hat{t}_j$	V	Prob
1	1	–	–	–
2	1751	1751	0.360	0.00000
3	2601	2601	0.198	0.00162
4	4151	4179	0.196	0.00189
5	5051	5055	0.318	0.00000
6	5951	5957	0.469	0.00000
7	6451	6439	0.355	0.00000
8	7426	7487	0.230	0.00006

The next example shows the results of the segmentation process when applied to various geophysical logging data sequences. The data, which are part of the Ocean Drilling Program (Leg 197, Site 1203), were collected to characterize the southward motion of the Hawaiian Hotspot in the Emperor Seamount trend (Tarduno et al. 2002). The drilling achieved moderate basement penetration and high recovery allowing for a detailed lithostratigraphy analysis. Downhole measurements, which are of very good quality in the basement sections, included various standard and non-standard tool strings and passes.

Figure 4 displays all data sequences used in the segmentation procedure. In particular, we selected total natural gamma ray, electrical resistivity, bulk density, porosity and S-wave velocity. The sampling interval is 0.1524 m, and each data sequence contains about 3300 samples in the interval considered. The last column of the figure, labeled "mean standardized log data", was built so as to take into account all logging data into a single sequence. For this purpose, the five previous sequences were standardized and averaged with equal weight and appropriate sign into a single sequence (the logarithm of the electrical resistivity was used in this sum. The polarity of both the total natural gamma ray and porosity was changed before the sum). The resulting "mean" sequence was the data that we actually used to carry out the statis-

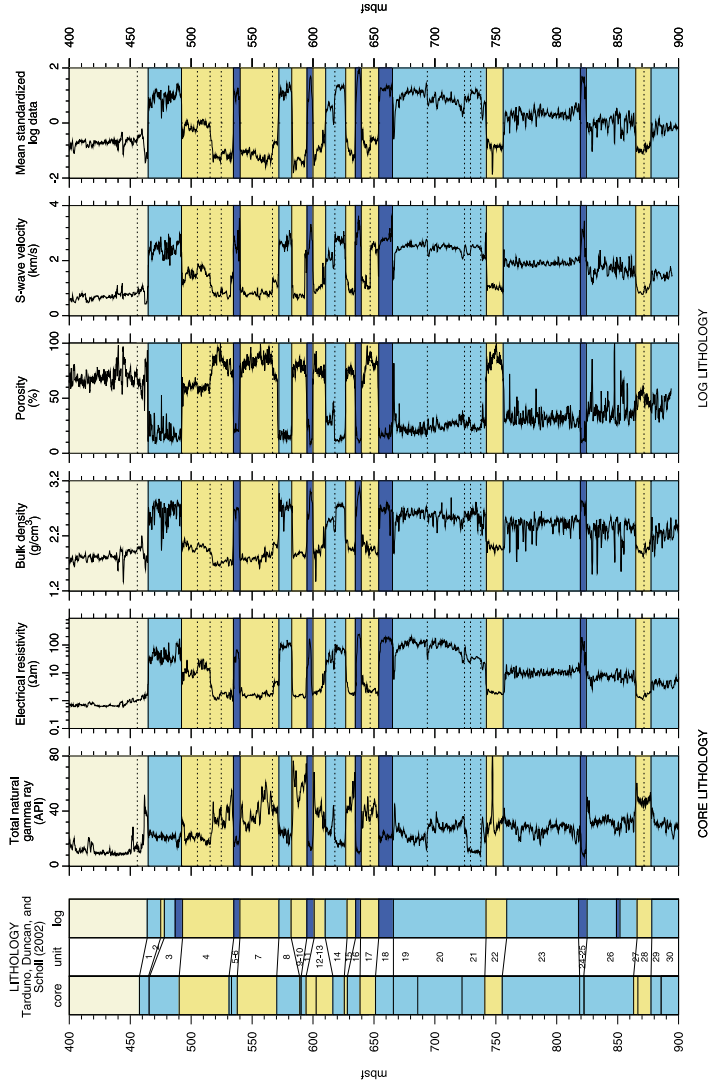

Fig. 4 Segmentation of the Ocean Drilling Program Downhole logs (Leg 197, Site 1203). The *left panel* shows the core and log lithology columns after Tarduno et al. (2002), together with the unit number identification. Total natural gamma ray, electrical resistivity, bulk density, porosity and S-wave velocity were combined into a single sequence named "mean standardized log data" (see the text for details). The statistical segmentation algorithm detected the 32 change points indicated by *horizontal lines* (*dashed lines* correspond to change points that do not correlate with the available log lithology column). The same color code was used in all cases in order to facilitate the visual comparison between the available core-log interpretation and the results of the statistical segmentation

tical segmentation. Note that this sequence exhibits features of the five previous data sequences, allowing for a full multivariate segmentation of the whole data set.

The results of the segmentation are shown in the same figure along with the available core-log lithology interpretation. The available interpretation in the analysed interval (400–900 meters below sea-floor, mbsf) comprises 30 units of alternating sediments (ooze and volcaniclastic) and basalts. The basement starts at about 460 mbsf. For details, see Tarduno et al. (2002). As for the statistical segmentation process, we set $L = 31$, $\Delta = 31$ and $T_{\min} = 50$ (sample units) and we kept the 32 change points with the lowest probabilities. In general, the correlation between the statistical segmentation and the core-log interpretation is from good to excellent. All the major units were correctly identified. In particular, all the units identified in the available log interpretation were automatically detected by the statistical segmentation algorithm, except for the thin layers at about 475 mbsf (unit 2), 490 mbsf and 850 mbsf. In any case, these thin beds are not so clear by inspecting the considered log data. Actually, except for unit 2, the other two beds were not identified in the core lithology analysis. On the other hand, there is a total of 12 change points (denoted by dashed lines in the figure) which are not identified in the available log lithology interpretation. Some of these change points may be associated to units clearly identified in the core lithology. For example, units 27 and 28 are identified as a single unit in the available log lithology, but as two units in the statistical segmentation, in accordance to the core lithology analysis. The same can be said for units 19 and 20. Moreover, units 12, 13 and 14 are identified as two units in the available log interpretation, but correctly as three distinct units in the statistical segmentation. Finally, some of the remaining detected change points that do not correlate with the available core-log interpretation may be associated to a fine lithology layering (e.g. division of units into sub-units, etc.). In effect, the three change points detected in unit 4 by the statistical segmentation at 505, 515 and 524 mbsf, for example, correlate very well with core-lithology subunits 4i, 4k and 4m at 508, 516 and 523 mbsf, respectively. These subunits are not indicated in the core lithology column, and the interested reader can find detailed information in Tarduno, Duncan, and Scholl (2002). A deeper analysis regarding subunits is beyond the scope of this work.

Another point worth mentioning is the accuracy of the detected change points. The statistical segmentation procedure automatically detects very accurately the presence of a change point. Based on the information provided by the core interpretation, the thickness of unit 22, for example, is about 14 m. The same value is obtained after the statistical segmentation. On the contrary, the available log interpretation estimates a thickness of about 16.5 m.

Conclusions

The detection of stationary segments in geophysical log data sequences can be carried out in a quasi-unsupervised mode by searching for change points in the data. The MaxEnt method using robust non-conventional statistics that measure shape provides an appropriate technique to estimate the distributions that are to be compared. After estimating the distributions of the two halves of a moving window, abrupt changes

are easily identified based on the analysis of the probability of the null hypothesis that the two distributions are the same. The Kuiper test proved to be a useful criterion to decide which change points lead to significant differences between adjacent distributions. This provides a means of choosing the appropriate number of locally stationary segments that the data sequence can be subdivided into.

The statistical segmentation algorithm presented in the work is viewed as an auxiliary tool that may contribute useful information in the identification of the main lithology units and sequences as derived from measured geophysical log data. The zonation of a borehole environment is an essential step in the correlation of subsurface layers between wells, with application in oil exploration and reservoir evaluation.

Acknowledgements This work was partially supported by Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina. The data were obtained from the ODP Well Log Database at the Borehole Research Group of the Lamont Doherty Earth Observatory.

References

Davis J (1986) Statistics and data analysis in geology. Wiley, New York, p 646

Gill D (1970) Application of a statistical zonation method to reservoir evaluation and digitized-log analysis. Am Assoc Pet Geol Bull 54(5):719–729

Gill D, Shomrony A, Fligelman H (1993) Numerical zonation of log suites and logfacies recognition by multivariate clustering. Am Assoc Pet Geol Bull 77(10):1781–1791

Hawkins DM, Merriam DF (1973) Optimal zonation of digitized sequential data. Math Geol 5(4):389–395

Kaaresen KF, Taxt T (1998) Multichannel blind deconvolution of seismic signals. Geophysics 63(6):2093–2107

Ligges U, Weihs C, Hasse-Becker P (2002) Detection of locally stationary segments in time series. In: Härdle W, Rönz B (eds) Proc. of the 15th conference on computational statistics. Physika, Heidelberg, pp 285–290

Press WH, Teukolsky S, Vetterling W, Flannery B (1992) Numerical recipes in FORTRAN: the art of scientific computing, 2nd edn. Cambridge University Press, New York

Tarduno JA, Duncan RA, Scholl DW (2002) Motion of the Hawaiian hotspot: a paleomagnetic test. In: Proc. ODP, initial reports, vol 197, available from Word Wide Web: http://www-odp.tamu.edu/publications/197_IR

Velis DR (2003) Estimating the distribution of primary reflection coefficients. Geophysics 68(4):1417–1422

Webster R (1973) Automatic soil-boundary location from transect data. Math Geol 5(1):27–37

Yilmaz Ö (2001) Seismic data analysis: processing, inversion, and interpretation of seismic data. Society of Exploration Geophysicists, Tulsa