Mineralogical Magazine, December 2012, Vol. 76(8), pp. 3355–3364

Data analysis toolkit for long-term, large-scale experiments

D. P. Bennett[1,*], R. J. Cuss[2], P. J. Vardon[1,3], J. F. Harrington[2], R. N. Philp[1] and H. R. Thomas[1]

[1] Geoenvironmental Research Centre, Cardiff School of Engineering, Cardiff University, Queen's Buildings, The Parade, Cardiff CF24 3AA, UK
[2] British Geological Survey, Keyworth, Nottingham NG12 5GG, UK
[3] Geo-Engineering Section, Department of Geoscience and Engineering, Delft University of Technology, PO Box 5048, 2600 GA Delft, The Netherlands

[*Received 18 December 2011; Accepted 2 May 2012; Associate Editor: Nicholas Evans*]

ABSTRACT

A new data analysis toolkit which is suitable for the analysis of large-scale, long-term datasets and the phenomenon/anomalies they represent is described. The toolkit aims to expose and quantify scientific information in a number of forms contained within a time-series based dataset in a quantitative and rigorous manner, reducing the subjectivity of observations made, thereby supporting the scientific observer. The features contained within the toolkit include the ability to handle non-uniform datasets, time-series component determination, frequency component determination, feature/event detection and characterization/parameterization of local behaviours. An application is presented of a case study dataset arising from the 'Lasgit' experiment.

Keywords: Lasgit, large-scale experiment, large dataset, non-uniform, time-series analysis.

Introduction

The development of the safety case for deep geological repositories has prompted a series of experimental activities ranging from laboratory scale to field scale. A proportion of such experiments have a significant duration owing to the longevity of the problems being studied, e.g. the (re)saturation of clay buffers (SKB, 2007; Dixon *et al*., 2002). The large-scale, long-term nature of these experiments can yield vast datasets that would be impractical to examine entirely by hand.

Time-series analysis and signal processing techniques applied computationally can help to reduce subjectivity in observations made (e.g. Chatfield, 1989), therefore supporting the scientific observer. This is achieved by providing a uniform and automated procedure for making observations that, by its nature, performs the task in a quantitative way. This approach improves such observations by turning them into measurements.

Such computational analysis may also be instrumental in uncovering a wealth of information contained within the dataset (e.g. Box and Jenkins, 1976; Chatfield, 1989), and addressing the practicalities of analysing a large-scale dataset. In addition to any primary observations or data processing, an analysis focussing on smaller scale features can be performed to expose this information and can potentially yield extra value from the dataset. Such an analysis is termed a 'second order' analysis in this work.

As an additional consequence of the longevity of such experimental undertakings, non-uniform datasets commonly arise due to circumstances outside of the experimental control such as hardware malfunction/failure. The probability of experiencing an event causing non-uniformity increases with the length of the experiment (Halpern, 1978; O'Connor, 1995). Typically automated computational analysis algorithms take uniform input data, and as such can be

* E-mail: bennettdp@cardiff.ac.uk
DOI: 10.1180/minmag.2012.076.8.48

hampered by the non-uniformities present in such datasets (e.g. Box and Jenkins, 1976).

This paper describes the development at the Geoenvironmental Research Centre (GRC) of a series of data analysis tools designed around long-term, large-scale datasets capable of performing time-series and signal processing procedures on non-uniform datasets. These approaches, based on statistics, Fourier analysis and component analysis are described within the paper as a 'data analysis toolkit'. Each data processing and analysis component has been chosen based on its potential to expose useful scientific information and implemented in such a way as to accommodate the input of such a dataset. An initial analysis is then presented of a case study dataset that aims to computationally identify and quantify smaller scale and otherwise difficult to observe behaviours and features.

Case study

An example of a long-term, large-scale experiment with a large non-uniform dataset is the large scale gas injection test or 'Lasgit', a field-scale experiment located at the Äspö Hard Rock Laboratory (HRL) in Sweden. The project is a full-scale demonstration test based on the KBS-3V high-level radioactive waste disposal concept (SKB, 2006) and is designed to examine the impact of gas (generated primarily by corrosion of metal) in compact bentonite within a deposition hole.

Lasgit has been in continuous operation since February 2005 (Cuss *et al*., 2010) and has undergone more than 90,000 logging cycles (i.e. the recording of a single datum point at all installed sensors at a specific time) leading to an acquired dataset in excess of 14.7 million datum points owing to its highly instrumented nature.

As of the end of 2011 three gas injection tests have been undertaken. Tortuous gas flow paths have been detected and a number of externally caused events have been observed within the dataset (Cuss *et al*., 2010, 2011). However a large quantity of uninterpreted information may remain within the dataset making it a candidate for analysis of the kind described in this paper.

The Lasgit dataset possesses a number of phenomena that require consideration during computational analysis. Primarily the logging (time) interval is not consistent across the dataset. This non-uniformity is due, in part, to increases of sample rates during periods of critical interest such as gas injection phases and also to hardware limitations and unavoidable breakdown causing interruptions and loss of data. Some data streams also occasionally 'spike', characterized by a single datum point, disparate from the consensus of surrounding data and therefore can be considered to not be a representative measurement at that time, or of the trend of the process of interest at the observed scale.

Data analysis toolkit

To address the above problems a toolkit capable of performing a 'second order' analysis (as defined above) on long-term, large-scale datasets with non-uniformities has been developed.

Options regarding reformatting of a dataset

Computational time-series analysis and algorithmic procedures typically require uniform datasets. To overcome the difficulty associated with the computational analysis of a non-uniform dataset the option exists to reprocess the data into a uniform form. One method of achieving this is by down-sampling the dataset to a uniform time step (i.e. taking the points that correspond to the lowest common sampling rate within the dataset) at the cost of a loss of data resolution.

There may exist, however, as shown in the case-study dataset, the situation where there is no global timing grid that a base sample rate can be affixed to, i.e. data logged at hourly intervals that correspond with the top of an hour in one section of the dataset do not necessarily correspond to the top of an hour in other sections of the dataset that are sampled at an hour. In the case study dataset this is primarily due to logging hardware requiring restarts at various points in the experimental history, leading to an arbitrary interval between two logging cycles.

An alternative to this approach is to define an appropriate assigned and arbitrary global sample rate for the dataset and to interpolate the defined points using the original non-uniform data. However a high percentage of the original data could be abandoned in favour of the 'created' data using this method. There is also the risk that detail present in the original dataset will be lost in the created dataset, particularly local maxima and minima if they do no coincide closely with an interpolated point.

Neither of these approaches was considered entirely satisfactory therefore an adaption of

https://doi.org/10.1180/minmag.2012.076.8.48 Published online by Cambridge University Press

relevant algorithms to a non-uniform applicability was undertaken.

Generalization of algorithms to non-uniform input

Time-series analysis can be considered as the measurement or exploitation of the fact that proximate datum points in a time-series are interconnected to (or non-independent of) each other (Box and Jenkins, 1976). Approaching time-series analysis with this in mind helps ensure that any modification/generalization of existing algorithms towards non-uniform application preserve the intent of the original algorithms and the mathematics underpinning them.

Processes that involve moving windows across a time-series, i.e. processes that allow the analyst to specify the scale over which the non-independence of points is assumed to occur, can be utilized. These can be implemented by either collating a fixed number of points around the centre of the window or by collating all the points that fall within a fixed time span around the centre of the window.

When a time-series has been sampled uniformly the two approaches are equivalent, however when applied to a time-series that is non-uniform, the method using a window fixed in time maintains the scale over which the observation is made. Figure 1 depicts a comparison of the two 'windowing' methods when applied to a time-series at the point the sample rate changes. It can be seen that the approach using a fixed number of points places a forward bias on the actual time indices of the time-series.

Basic time windowing applications and implications

Applying time windowing to a weighted moving average algorithm requires a continuous user defined weighting function in place of pre-defined fixed weighting constants used in the fixed number of points windowing method. This allows the modified algorithm to adapt to non-uniformities in the input time-series by deriving weight as a function of each point's time relative to the centre of the window. Information such as the local mean, standard deviation and characterising values of the window can be calculated.

Spike identification can be achieved by comparing the absolute deviation of a point from its local mean (excluding the point in question) to a threshold defined as a multiple of the local standard deviation around the point (again excluding the point itself). Exceeding that threshold classes the point as a spike, either suggesting rapid process evolution or possible measurement error at that point.

The time windowing implementation of point gathering to determine the local parameters that are required to make such a comparison ensure that the observed density of spikes with time will be unaffected by a change in sample rate, assuming the frequency or likelihood of spike occurrence is not a function of the sample rate and that the data is sampled at a greater rate than the underlying process evolution of interest.

Comparatively, using a fixed number of points will result in variation in the threshold determination with changes in sample rate. Figure 2 depicts

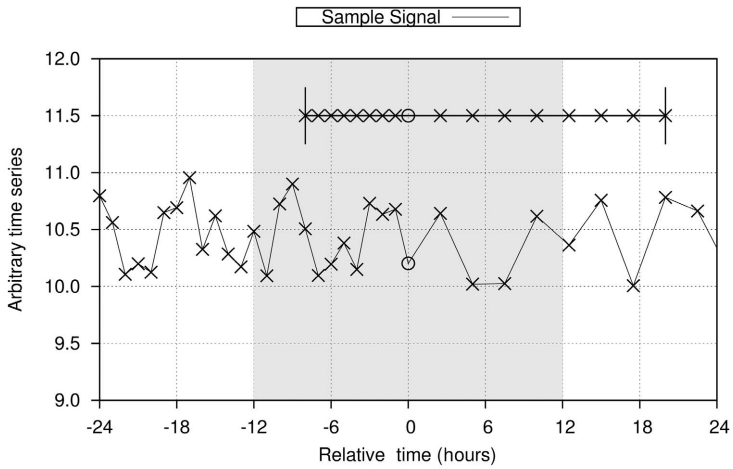

FIG. 1. Comparison of time-series windowing methodologies.

https://doi.org/10.1180/minmag.2012.076.8.48 Published online by Cambridge University Press

spike threshold evolution along a time-series with varying local environments as determined by time windowing.

Information quantification or parameterization

Quantification or parameterization of local aspects of a dataset with respect to time can potentially expose information about the system it represents. Examples include quantification of local standard deviation of the signal, as discussed in the previous section, to characterize the noise and/or rate of change within the time-series with time. This method uses the time windowing method for point gathering discussed in the previous section.

Isolated peaks in the local standard deviation or changes in the inherent amount of noise within the system can indicate events within the dataset that are candidates for scientific investigation, with their relative prominence suggesting possible relevance. Figure 3 depicts an arbitrary data series, with the local standard deviation with time. The large peak in standard deviation corresponds to the large ramp in time-series magnitude. The local standard deviation peak, highlighted as an 'anomalous noise peak', corresponds to a small scale event in the time-series which is shown in more detail in Fig. 3*b*. The effect identified by the local peak is unlikely to appear significant when observed at the macro scale of the dataset as a whole, however, could indicate a second order process occurring.

Aggregation of the derived local standard deviation series across sensor types can be performed allowing a visual indicator of synchronization within a group of sensors. The result of this process is shown in Fig. 5.

Frequency domain analysis

Cyclic information over a domain, i.e. frequency components along with their amplitudes and phases, can be quantified with application of a Fourier transform (FT) analysis. Application to a time-series requires a discrete Fourier transform (DFT). The standard DFT process converts a time-series into a power series consisting of a discrete set of frequency locations defined by the time-series length and sample rate.

The mathematical representation of a DFT is shown in equation 1 (Bagchi and Mitra, 1999), where P(k) is the power series of the time-series F with points at n or t_n (represented by a complex number) at frequencies that are n multiples of k/N, N is the number of points in the time-series and $0 \leqslant k \leqslant N-1$.

To overcome the requirement for a uniform sample rate in the time-series a modification to equation 1 can be applied, resulting in an implementation of a non-uniform discrete Fourier transform (NDFT). This modification is shown in equation 2.

$$P(k) = \sum_{n=0}^{N-1} F_n \cdot e^{-i2\pi\frac{k}{N}n} \qquad (1)$$

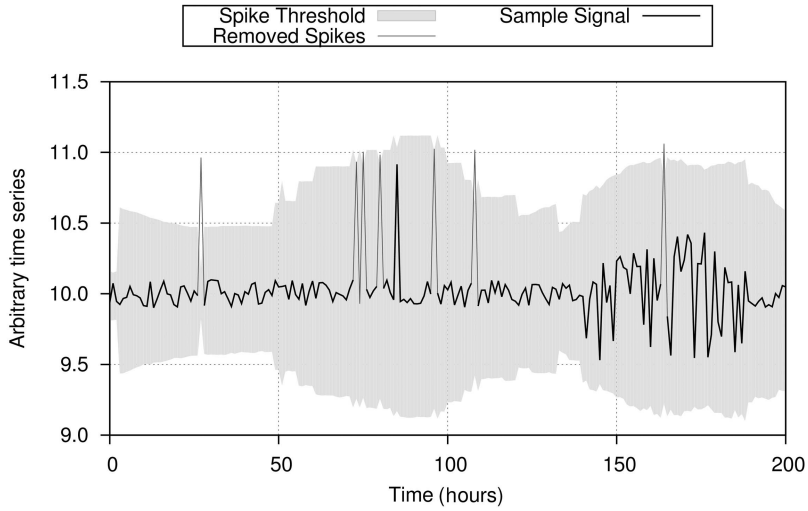

FIG. 2. Evolution of spike threshold with different levels of spike frequency and signal noise.

https://doi.org/10.1180/minmag.2012.076.8.48 Published online by Cambridge University Press

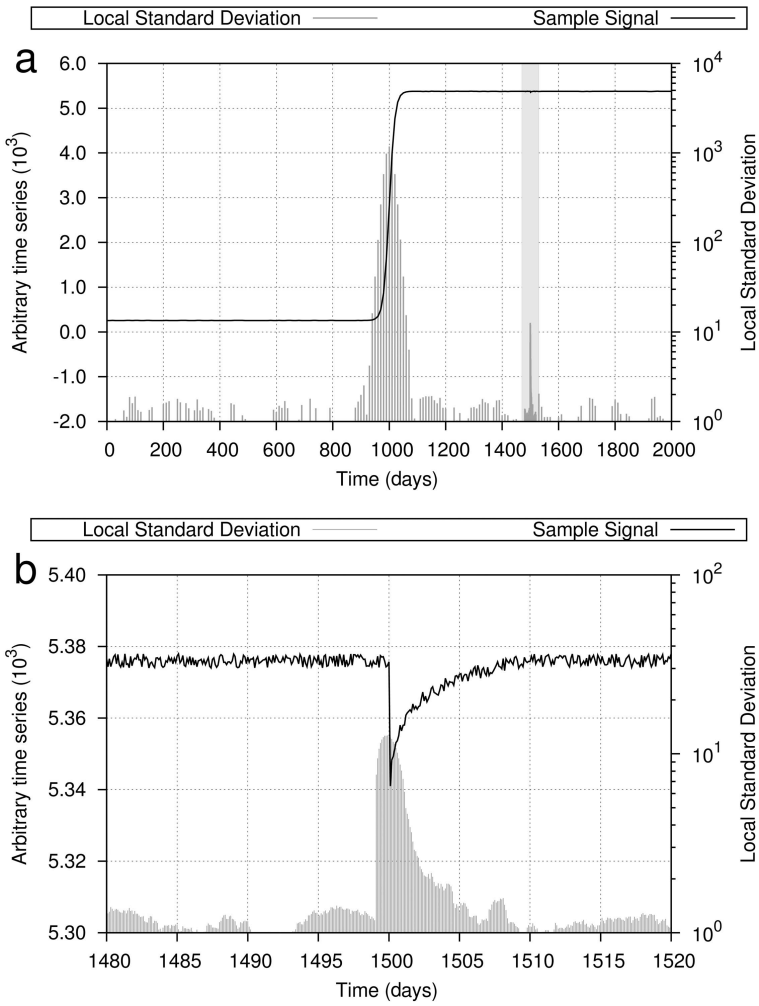

FIG. 3. (*a*) Local standard deviations of a time-series. An isolated peak is highlighted as a small scale event candidate. (*b*) Detailed view of time-series at section highlighted in (*a*).

$$P(\omega) = \sum_{n=0}^{N-1} F(t_n) \cdot e^{-i\omega t_n} \qquad (2)$$

The modification replaces the terms $2\pi k/N$ with the term ω ($=2\pi h$) allowing the power series to be specified arbitrarily over a range of frequencies (h) and references each point's relevant time index (t_n). The algorithm implementing the NDFT (equation 2) allows for calculation of the zero frequency component, P(0), and subsequent normalization of the time-series to zero average magnitude. Normalization of the signal in such way may help reduce distortion (aliasing) of the power series.

Local peak detection algorithms can be used in an iterative manner to highlight local maxima in the resulting power series with the frequency, amplitude and phase information associated with these identified peaks easily obtainable.

The NDFTs can however present a positive bias in amplitude calculations due to aliasing effects (a type of distortion present in NDFTs). Additionally, frequency domain filtering (normally achieved through excluding unwanted frequency content and performing an inverse Fourier transform) is hampered by the arbitrary nature of the frequency domain investigated. Fourier transform procedures in general also benefit from input data with no overall trend

https://doi.org/10.1180/minmag.2012.076.8.48 Published online by Cambridge University Press

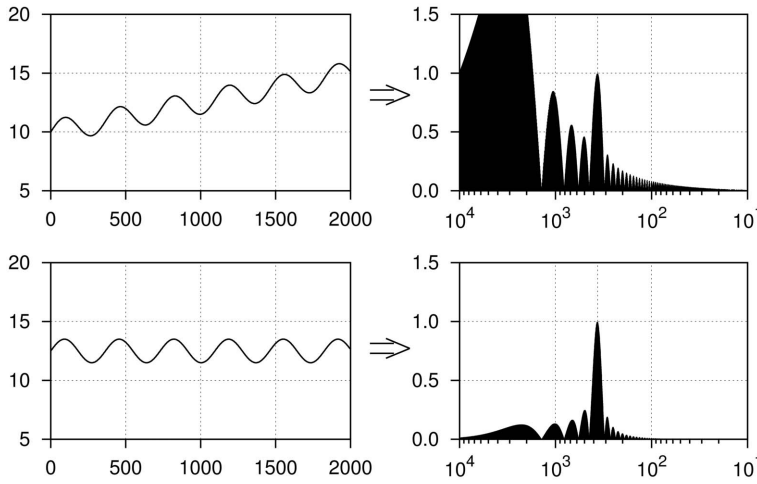

FIG. 4. Diagrammatic depiction of NDFT process on trending (upper) and non-trending (lower) time-series.

associated with them as detailed diagrammatically in Fig. 4. The peak in the lower transform is more defined due to an absence of low frequency ramping present in the conversion resulting from the trending time series. The frequency peak is distinguishable in both power series in Fig. 4. However, as the trending of the input time series increases, the low frequency ramping in the power series will obscure greater portions of the frequency domain.

Non parametric data inspection techniques

Non-parametric data inspection of a time-series consists of analysing the form of a time-series without categorizing it with a mathematical

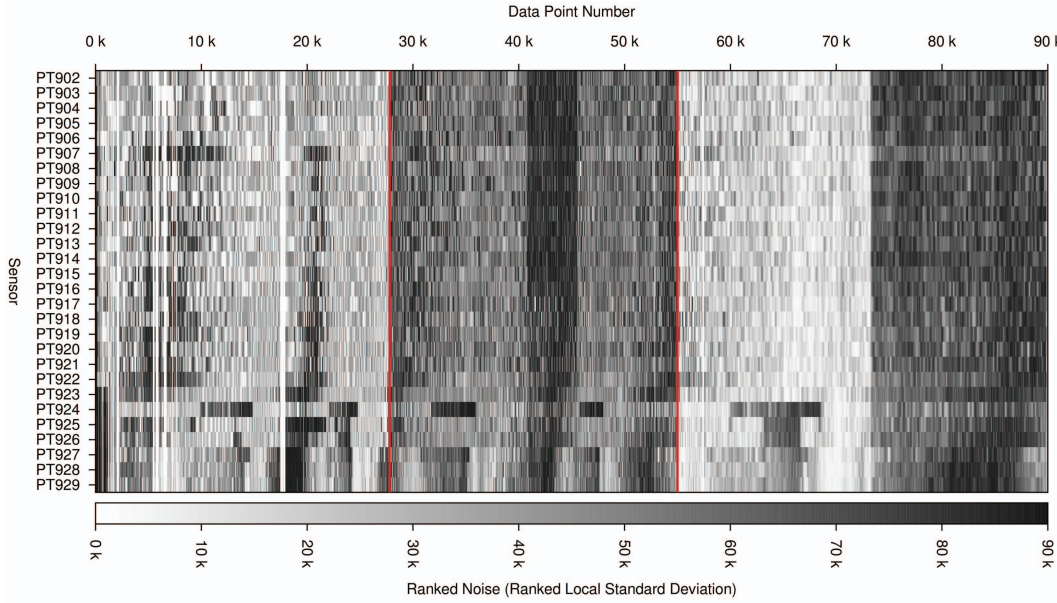

FIG. 5. Comparison of local standard deviations with time across multiple temperature sensors installed in the Lasgit experiment. Vertical lineation indicates synchronicity in events or system behaviour changes.

https://doi.org/10.1180/minmag.2012.076.8.48 Published online by Cambridge University Press

relationship, i.e. not defining a continuous function to model it.

Scientific investigation of a time-series may require separation of a number of underlying contributions to the time-series. Singular spectrum analysis (SSA) is an analysis method capable of decomposing a time-series into a collection of mathematically independent component time-series that sum to the original input. The process (as described in Golyandina *et al*., 2001) involves mapping a time-series $F = (f_0, \cdots, f_{N-1})$ of length N into "a sequence of [K (= N − L + 1)] multidimensional lagged vectors" of length L, such that when collected into a matrix, **X**, they yield:

$$\mathbf{X} = \begin{bmatrix} f_0 & f_1 & f_2 & \cdots & f_{K-1} \\ f_1 & f_2 & f_3 & \cdots & f_K \\ f_2 & f_3 & f_4 & \cdots & f_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{L-1} & f_L & f_{L+1} & \cdots & f_{N-1} \end{bmatrix} \quad (3)$$

Subsequently decomposing **X** into a series of components such that $\mathbf{X} = \mathbf{X}_1 + \cdots + \mathbf{X}_L$. The decomposition is achieved by performing a singular value decomposition (SVD) (e.g. Golyandina *et al*., 2001). Each component $\mathbf{X}_i$ is associated with an eigenvalue (λ) and eigenvector (U) of the matrix $\mathbf{X}\mathbf{X}^T$ such that $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^T$ and $V_i = \mathbf{X}^T U_i / \sqrt{\lambda_i}$.

Computationally the eigenspace solution that forms the majority of the work can be achieved using the Jacobi eigenvalue algorithm (Golub and van der Vorst, 2000). To produce a derived time-series component, F_i, from each component matrix, $\mathbf{X}_i$, the average of the elements of the matrix along the diagonals defined by indices i + j = *constant*, i.e. the off diagonal, is taken. The form of the derived time-series components is determined by the SSA process with no user defined specification other than number of decomposition components. Components of the input time-series typically have decreasing magnitude as the magnitude of the eigenvalue they are associated with decreases. Dependent upon form, they are typically characterized as either trends, oscillatory components or noise. Measurements of the independence of each component from the others can be performed, as can statistical tests on residual components, to confirm only noise remains. This can aid understanding of the meaning of each component or set of components.

The trends determined by SSA can also potentially be used to pre-process the original

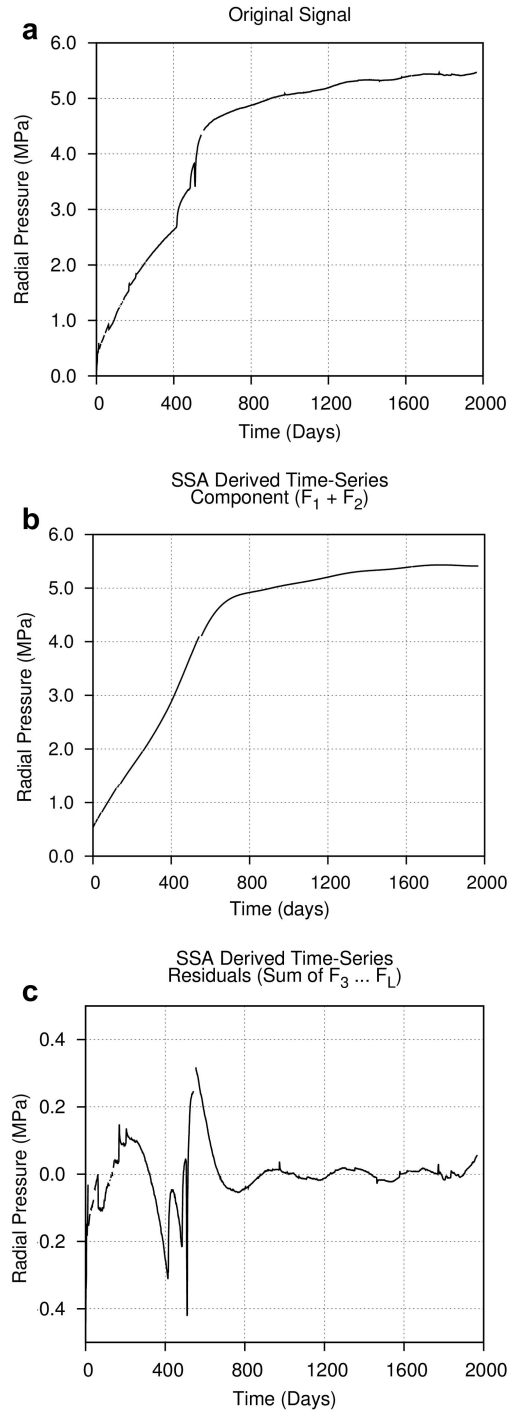

FIG. 6. Comparison of original time-series with derived component time-series and residual derived components.

https://doi.org/10.1180/minmag.2012.076.8.48 Published online by Cambridge University Press

signal (e.g. be subtracted from the original signal) in order to improve frequency domain analysis or can be used to examine physical processes.

Application of data analysis toolkit

The developed toolkit has been applied to the case study (Lasgit) dataset as described earlier and an initial analysis of the dataset and assessment of the toolkit was performed.

Exposed/highlighted and quantified phenomena

Application of the spike identification algorithm to the Lasgit dataset was performed utilizing a 48 hour window and a threshold of $3\sigma_{local}$. The $3\sigma_{local}$ threshold was chosen to coincide with the upper limit of the empirical 68–95−99.7 rule for normally distributed data. This limits false positive detection of spikes to 0.3% assuming the noise is normally distributed. The 48 hour window was chosen to ensure a large enough population from which to calculate the standard deviation at lower sample rates that occur within the Lasgit dataset. The algorithm identified 117,165 spikes (approximately 0.8% of the recorded datum points), with a notable concentration of those occurring in the temperature records, and within them concentrated during the annual minimum temperature regions.

Aggregation of the local standard deviations into time-series (Fig. 7) identifies second order events as indicated in Fig. 3*a*,*b*, with second order features in the time-series being highlighted. Further aggregation of local standard deviation series into groups (defined by sensor type or proximity for example) allows for a visual assessment of the independence of each time-series. Figure 5 shows the ranked standard deviation with time of each temperature sensor installed at the Lasgit experiment. Vertical lineation indicates synchronized behaviour. The abrupt changes in standard deviation level indicated by the red lines in Fig. 5 coincide with entries in the experimental log describing 'over-pressurizations' of the experiments fluid injection system. Additional behaviours and events can be identified and associated in other sensor sets using this method, e.g. the reciprocating nature of the injection pumps and the closure of pressure relief holes close to the experiment.

Frequency domain analysis performed on the temperature sensors confirmed an annual temperature cycle in all down-hole positions. Phase (offset from air temperature cycle) and amplitude of the temperature cycles with depth is shown in Fig. 8. The only time-series in which a daily frequency component was detected was the air temperature record of the tunnel.

Signal component identification using SSA was performed on a reduced resolution version of the dataset to reduce the computational burden during the initial investigation phase. Representative daily values were calculated and a vector length of 365 days was applied (corresponding to an annual scale).

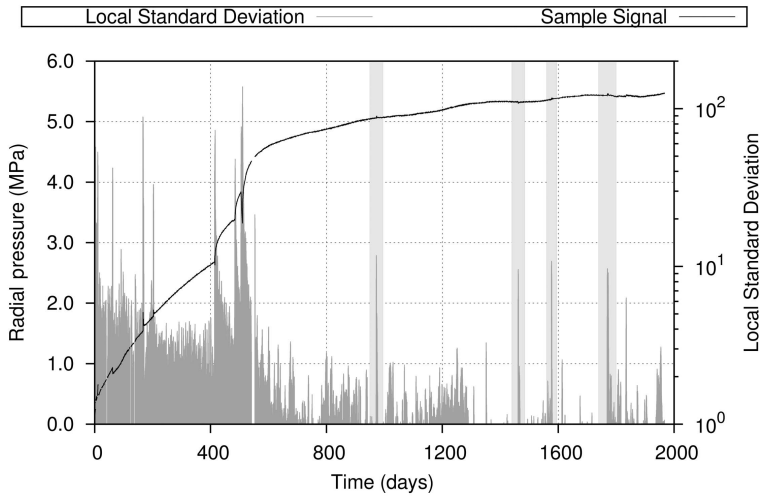

FIG. 7. Second order events identified in a radial pressure record from the Lasgit dataset.

https://doi.org/10.1180/minmag.2012.076.8.48 Published online by Cambridge University Press

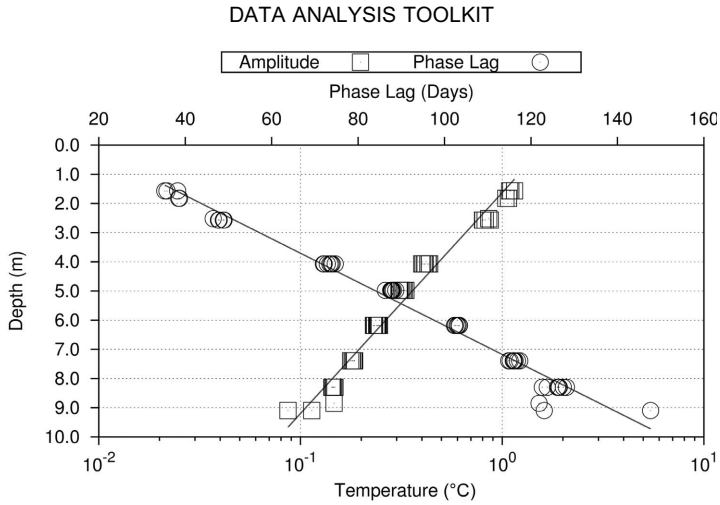

FIG. 8. Annual temperature cycle phase and amplitude information with respect to depth.

Figure 6 compares a radial pressure record from Lasgit with the sum of the first two components and presents the residual of the original series without the first two components. The presence of two trending components may be an indicator that two processes are driving the time-series in question.

Frequency content is observable in the residual along with anomalies that coincide with closure of the pressure relief holes near the experimental set-up. The presence of this anomaly is more pronounced in other sensor types. Analysing the dataset post pressure relief hole closure may produce more informative components due to the SSA process not attempting to account for a significant imposed change in system behaviour.

Conclusions

A toolkit has been developed to analyse long-term, large-scale datasets, particularly of the nature arising from geomechanical and geoenvironmental experiments. The toolkit has been designed to support the scientific observer, reducing the subjectivity of observations through qualitative processes.

The toolkit is capable of performing a range of analyses on non-uniform datasets including: event candidate detection; quantification and parameterization/characterization of time-series data; frequency domain analysis; non-parametric trend and component derivation; and system synchronization/association visualizations.

The toolkit was applied to the Lasgit dataset as a case study and an initial analysis was undertaken. The toolkit highlighted possible events in individual time-series records and synchronizations across sensor types at specific points in time. A correlation of temperature cycles within the system (characterizing both amplitude and phase) with depth was also achieved.

Acknowledgements

The research leading to these results has received funding from the European Atomic Energy Community's Seventh Framework Programme (FP7/2007-2011) under Grant Agreement No. 230357, the FORGE project.

References

Bagchi, S. and Mitra, S.K. (1999) *The Nonuniform Discrete Fourier Transform and its Applications in Signal Processing*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Box, G.E.P. and Jenkins, G.M. (1976) *Time Series Analysis: Forecasting and Control*. Revised Edition. Holden-Day, San Francisco, California, USA.

Chatfield, C. (1989) *The Analysis of Time Series: An Introduction*. Fourth Edition. Chapman & Hall, London.

Cuss, R.J., Harrington, J.F. and Noy, D.J. (2010) *Large Scale Gas Injection Test (Lasgit) Performed at the Äspö Hard Rock Laboratory: Summary Report 2008*. SKB Technical report TR-10-38. Swedish Nuclear Fuel and Waste Management Company, Stockholm.

Cuss, R.J., Harrington, J.F. and Noy, D.J., Wikman, A. and Selin, P. (2011) Large scale gas injection test (Lasgit): results from two gas injection tests. *Physics*

https://doi.org/10.1180/minmag.2012.076.8.48 Published online by Cambridge University Press

and Chemistry of the Earth, **36**, 1729–1742.

Dixon, D.A., Chandler, N., Graham, J. and Gray, M.N. (2002) Two large-scale sealing tests conducted at Atomic Energy of Canada's underground research laboratory: the buffer-container experiment and the isothermal test. *Canadian Geotechnical Journal*, **39**, 503–518.

Golub, G.H. and van der Vorst, H.A. (2000) Eigenvalue computation in the 20th century. *Journal of Computational and Applied Mathematics*, **123**, 35–65.

Golyandina, N., Nekrutkin, V. and Zhigljavsky, A. (2001) *Analysis of Time Series Structure: SSA and Related Techniques*. Chapman & Hall/CRC, Boca Raton, Florida, USA.

Halpern, S. (1978) *The Assurance Science: An Introduction to Quality Control and Reliability*. Prentice-Hall, Englewood Cliffs, New Jersey, USA.

O'Connor, P.D.T. (1995) *Practical Reliability Engineering*. Third Edition Revised. John Wiley & Sons, Chichester, UK.

SKB (2006) *Long-term Safety for KBS-3 Repositories at Forsmark and Laxemar – A First Evaluation: Main Report of the SR-Can Project*. SKB Technical Report TR-06-09. Swedish Nuclear Fuel and Waste Management Company, Stockholm.

SKB (2007) *RD&D Programme 2007: Programme for Research, Development and Demonstration of Methods for the Management and Disposal of Nuclear Waste*. SKB Technical Report TR-07-12. Swedish Nuclear Fuel and Waste Management Company, Stockholm.

https://doi.org/10.1180/minmag.2012.076.8.48 Published online by Cambridge University Press