Computers & Geosciences 95 (2016) 99–104

Contents lists available at ScienceDirect

Computers & Geosciences

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

Case study

Spectral analysis of time series of categorical variables in earth sciences

Eulogio Pardo-Igúzquiza [a,*], Francisco J. Rodríguez-Tovar [b], Javier Dorador [b]

[a] Instituto Geológico y Minero de España (IGME), Ríos Rosas 23, 28003 Madrid, Spain
[b] Departamento de Estratigrafía y Paleontología, Universidad de Granada, Campus Fuentenueva, s/n, 18071 Granada, Spain

ARTICLE INFO

Article history:
Received 29 June 2015
Received in revised form
24 May 2016
Accepted 8 July 2016
Available online 9 July 2016

Keywords:
Categories
Spectral envelope
Indicator variable
Ichnofabric

ABSTRACT

Time series of categorical variables often appear in Earth Science disciplines and there is considerable interest in studying their cyclic behavior. This is true, for example, when the type of facies, petrofabric features, ichnofabrics, fossil assemblages or mineral compositions are measured continuously over a core or throughout a stratigraphic succession. Here we deal with the problem of applying spectral analysis to such sequences. A full indicator approach is proposed to complement the spectral envelope often used in other disciplines. Additionally, a stand-alone computer program is provided for calculating the spectral envelope, in this case implementing the permutation test to assess the statistical significance of the spectral peaks. We studied simulated sequences as well as real data in order to illustrate the methodology.

1. Introduction

A categorical variable is one whose values are categories, that is, qualitative attributes. For this reason they may also be referred to as qualitative or attribute variables. In the most general cases there is no intrinsic or logical ordering of the categories, and one has the so-called nominal variable. An intrinsic order is also possible, e.g. when the categorical variable is obtained by discretization (thresholding a real variable), hence we speak of an ordinal variable. The simplest example of a categorical variable is when the number of categories is only two, and the variable is called a binary, dichotomous or indicator variable. There are many examples in the Earth Sciences: facies, petrofabric features, ichnofabrics, fossil assemblages, or mineral compositions, among others. When the variable is measured over time (or space), for example when measured along a borehole or core, or over a stratigraphic sequence, one obtains a time series of a categorical variable.

Binary or indicator variables are often found in geostatistics (Goovaerts, 1997) for modeling spatial categories. Many properties of these indicator functions are described in Journel and Posa (1990) and Dowd et al. (2003). Particularly in the case of a binary indicator, a categorical variable with two categories $\{A, B\}$, the spectral information of both categories is equivalent. Such studies usually focus on the plane or three-dimensional space, however, the main interest is in spatial interpolation and spatial simulations, not their spectral analysis or the study of their cyclic behavior. For this reason such studies make no reference to time series or 1D sequences.

The earliest attempt to define a mathematical geological cycle of categorical variables dates back to Schwarzacher (1969), where it is shown how the eigenvalue of the Markov matrix provides a measure of cyclicity. Although the concept of a Markov chain for describing cyclic patterns is pursued by others (Hattori, 1976), the proper tool for searching for cycles in sequences of categorical data is the power spectrum, whose estimation is a problem solved by spectral analysis.

The first rigorous development of a theory for the spectral analysis of categorical time series was presented by Stoffer et al. (1993), motivated by the analysis of DNA sequences in biology. Two aspects were resolved by Stoffer et al. (1993): the first is scaling, finding numerical values for the categories in an optimal way, such that the selected values help emphasize any periodic feature that may exist in the categorical sequence; and secondly, finding a frequency function, the so-called spectral envelope, that provides a kind of maximum attainable power spectrum for each frequency. With a minimum loss of information, it provides as much information as possible on the cyclic behavior of the sequence.

Thus a second-order stationary categorical time series (that is, without trend or where the trend has been removed) can be represented as a sequence $\{X_1, X_2,\ldots,X_n\}$ where X_t is a categorical variable measured at time t. The categorical variable takes values from a finite set of k categories $\{c_1, c_2,\ldots,c_k\}$. We are interested in the spectral analysis of the sequence $\{X_1, X_2,\ldots,X_n\}$, that is, we

* Corresponding author.
E-mail addresses: e.pardo@igme.es (E. Pardo-Igúzquiza), fjrtovar@ugr.es (F.J. Rodríguez-Tovar).

http://dx.doi.org/10.1016/j.cageo.2016.07.005
0098-3004/

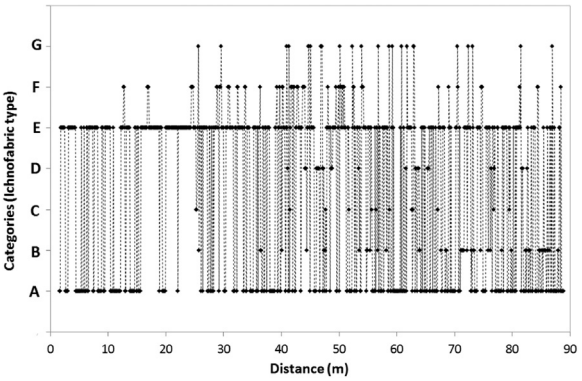

Fig. 1. Example of a sequence of a categorical variable along a core. The attributes are seven types of ichnofabric. It can be seen as a time series as distance can be transformed to time using the sedimentation rate.

would like to find if there are statistically significant cyclic behavior in the occurrence of the categories. In order to be able to perform the inference of the spectral content of the sequence, it is necessary to have a quantitative sequence. This can be done by using indicator variables or by assigning a number to each category (the scaling problem). The indicator sequences are binary sequences that express the presence (1) or absence (0) of a given category of a set of categories. The scaling problem is to find a set of k real numbers $\{\beta_1, \beta_2,\ldots,\beta_k\}$ so that there is a one-to-one correspondence between each category in the set of categories and each real number in the set that defines the scaling. The concept of scaling is closely related to the spectral envelope because the spectral envelope $\lambda(\omega)$ is the maximum attainable standardized spectrum of any scaled process: $f(\omega;\beta) \le f(\omega)$ (Stoffer et al., 1993). Many applications have been found for the spectral envelope as described in Stoffer et al. (2000). In this paper we introduce the spectral analysis of a categorical time series in geosciences by using the full indicator analysis of the sequence as complement to the well-known method of the spectral envelope. It should be noted that in the general case of a nominal variable the only information at each time (t) is the presence or absence of each category and that the indicator variable for any set of categories is already a scaling of the categories.

In addressing the problem, we introduce in Fig. 1a real time series with seven categories of ichnofabrics identified along a core: {green mottled, *Planolites*, *Taenidium* and *Planolites*, *Thalassinoides*-like and *Palaeophycus*, *Planolites* and *Thalassinoides*/*Thalassinoides* like, *Zoophycos*, and *Chondrites*}. The main question to be answered is, "What is the spectral content of this sequence?" In the next section we introduce a methodology for answering such a question.

2. Methodology

As mentioned above, a categorical time series can be represented as $\{X_1, X_2,\ldots,X_n\}$ where X_t is a categorical variable measured at time t and where the categorical variable takes values from a finite set of k categories $\{c_1, c_2,\ldots,c_k\}$. For each category it is possible to define a presence/absence indicator variable:

$$Y_t = \begin{cases} 1 & \text{if} \quad X_t = c_k \\ 0 & \text{if} \quad X_t \neq c_k \end{cases} \tag{1}$$

Thus, $\mathbf{Y}_t$ is the $k \times 1$ vector of indicator variables, one for each category. The scaled (i.e. quantitative time series) would be defined by:

$$\mathbf{X}_t = \beta^{\mathrm{T}}\mathbf{Y}_t \tag{2}$$

Where the superscript T denotes transposed vector or transposed matrix.

As may be found in Stoffer et al. (1993), the spectral envelope for each frequency $\lambda(\omega)$ is defined by an optimality criterion that is calculated as the largest eigenvalue of the determinant equation

$$\left|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{f}^{R}(\omega)\boldsymbol{A} - \lambda(\omega)\boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{A}\right| = 0, \tag{3}$$

where $\boldsymbol{V}$ is the $k \times k$ variance-covariance matrix between the indicator variables, $\boldsymbol{f}^{R}(\omega)$ is the real part of the $k \times k$ spectrum-cross-spectrum between the indicators and $\boldsymbol{A}$ is a $k \times k$ matrix $\boldsymbol{A} = \left[\mathbf{I}_{k-1}|0\right]^{\mathrm{T}}$, with $\mathbf{I}_{k-1}$ a $(k-1)\times(k-1)$ identity matrix and $\mathbf{0}$ is a $(k-1)\times 1$ vector of zeros. Thus $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{f}^{R}(\omega)\boldsymbol{A}$ and $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{A}$ are the upper $(k-1)\times(k-1)$ blocks of $\boldsymbol{f}^{R}(\omega)$ and $\boldsymbol{V}$, respectively (Stoffer et al., 2000). This selection corresponds to scaling as zero the last category, i.e., β_k=0, while the rest of the optimal scaling factors are a function of the eigenvector corresponding to the largest eigenvalue of Eq. (3). The procedure has been implemented in R statistical language by Shumway and Stoffer (2011). The authors of this paper implemented a FORTRAN stand-alone program that is public domain and will be provided upon request to anyone interested.

The cyclic content of a sequence of categories can be exhaustively analysed if all the possible combinations of categories are considered. For instance, 7 categories give 7 different combinations of one category (or the complementary six categories), 21 combinations of two categories (or the complementary five categories) and 35 combinations of three categories (or the complementary four categories). All the possibilities are therefore covered with 63 indicator variables. Hence, the number of indicator variables is:

$$\binom{k}{1} + \binom{k}{2} + \ldots + \binom{k}{\ell}, \tag{4}$$

with $\ell = \lceil k/2 \rceil$, where $\lceil k/2 \rceil$ is the integer part of $k/2$.

If we consider the following sequence of 7 categories (from A to G):

E	B	A	A	C	B	D	D	F	G	D	A	C	D	D	B	E	B	D	A	C

and one takes the indicator as 1 if the category is in the subset {C, E, F}, or 0 otherwise (if the category belongs to the complementary set), then in the indicator sequence there emerges a perfectly cyclic behavior, one cycle every four points.

1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1

The mean of the indicator is equal to m_I, the number of ones divided by the total number of data, and the variance is $\sigma_I^2 = m_I(1-m_I)$.

The shortest cycle or highest frequency cycle would consist of one cycle every two points:

1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

This cycle has the same number of ones and zeroes, so the mean is 0.5 and the variance 0.25. Longer cycles may have smaller variance (like the one cycle every four points represented above) or they can have the same variance if they are more similar to a square wave: