Contributions to Methodology and Statistics
A. Ferligoj and A. Kramberger (Editors)
Metodološki zvezki, 10, Ljubljana: FDV 1995

Searching for Patterns in Sequences of Data — Visual Approach

Andrej Blejec[1]

Abstract

Recognition of patterns in sequences of data is often needed in many fields of research.

Among the practical problems, where the search for patterns in data is common, one can find studies of cycle length in moulting animals, finding sequences of lithologic states in sedimentary succession, determination of patterns in DNA sequences, recognition of absorption lines in stellar spectra and many others.

A tool for analysis of non-numerical (categorical) sequences of data, using dynamic computer graphics, is presented.

As a measure of concordance, different measures of association may be used, composing the auto- and cross- associative functions for lagged sets of compared data sequences.

To maintain clear connection between cross-associative function and original data, linked plots are used for visual presentation. On such combined plots one can slide one original data sequence across the other. At the same time one can observe relative position of data and compare it with plotted cross-associative function.

Visual presentation can ease understanding of association between lagged data sequences. On screen measurement of lag are built for measurement of the duration of cycles in sequences of data. Since connection with original data is emphasised, artefacts in association measuring can be clearly identified.

Keywords: Statistical graphics; Time series; Statistics.

1 Introduction

Many phenomena are observed as sequences of events in time or space. To understand or describe such phenomena, reappearance of sub sequences of events, their periodicity and similar pattern related features are important. There are many methods available to analyse such phenomena. Collected data sequences can be

[1]Institute of Biology, University of Ljubljana, Karlovška 19, P.O.Box 141, 61001 Ljubljana, Slovenia; E-mail: andrej.blejec@uni-lj.si

recorded on a numerical or non-numerical (categorical) scale. The arsenal of analytic methods is richer for the analysis of numerical data. Many methods, ranging from cross correlation to Fourier analysis are used for the analysis of time series and other types of numerical sequences of data. In the case of non-numerical (categorical) data, methods are mostly based on the comparison of lagged data sequences. In both cases, results of analyses are somehow "disconnected" from the data. The relationship of the measure of concordance with the corresponding actual situation in lagged data is not always clear. The actual situation is to be imagined in most cases if plots are not available. The interpretation of results is not always easy or require certain level of skill and experience that can not be taken for granted for all users of statistics.

The purpose of this paper is to present a dynamic graphical method, which can help students, or consultees, to understand some aspects of interpretation of measures of similarity. Since high similarity means pattern reappearance, we can use it for pattern searching.

2 Method description

The motivation and inspiration to develop computer supported visual method for pattern searching was a problem to determine the duration of the moulting cycle for small woodlices of the species *Ligia italica* (Štrus et al., 1992). Animals can be classified into different stages, according to the state of their cuticle which they shed periodically. Series of recordings of stages were available and the question was to determine the duration between successive moults.

According to the periodic nature of moulting, a method of cross-association (or auto-association) was proposed. The difference of successive high association values (peaks) in associative function should be an estimate of cycle duration. Many peaks were present in a typical associative function and we decided to identify individual peak values by data from recorded sequences. It is inconvenient to perform such comparisons on paper and we decided to compose all relevant graphs on computer screen. Using the method described below, we were able to collect a series of moult cycle durations, measured as a distance between successive high values of associative function. Not all high values were meaningful, which was clearly shown on the plots of actual data sequences.

The central part of computer screen printout, labelled as A on Figure 2,[1] represent the graph of the associative function.[2] Associative function can be any measure of association, according to the problem considered: percentage of matches, χ^2 values, match of selected states are some of the possible measures. On the upper right half of the screen (B), compared data sequences are shown. It is convenient to indicate different states by black dots, which merge into thick black horizontal lines when

[1]Most figures are composed from actual computer screen printouts. Some texts, which are not relevant for understanding of examples, are left untranslated and appear in Slovene language.

[2]Compare also Figure 6 showing more straightforward example of cross-correlation of numerical data.

succeeding states are the same (see oval inserts C, D and E).[3] Thin slanted lines, connecting the dots, merely indicate the succession of states and have no other meaning. Small slanted arrowhead in the centre of the screen printout indicate the mouse pointer position and points to the vertical marker line. This marker line can be moved (by mouse) to any position of interest. Thus one can point to peaks or some other points on the associative function. At the same time, relative position of compared data sequences is shown in the upper right part of the screen (B). In oval inserts (C,D,E) some possible positions of data sequences are shown. They are connected with arrowed lines to the relevant points on the associative function. The points that match are hollowed, turning to white circles or bars on screen. The hollow series in the upper right part (B) of screen printout on Figure 2 correspond to zero lag (the sequence is compared to itself), showing the perfect match. It would be more impressive to see the data sequences sliding one across another, comparing the matched pattern to the value of association, but the paper is not the proper media for dynamical presentations. Some relatively high peaks can be identified as artefacts resulting from the small number of overlapped values or some other reasons which are obvious from the data sequence plot.

If one wants to measure the difference between the positions of peaks, or some other positions of interest, one simply points to the first point and selects it by a mouse click. While the mouse is moved away to the second point, the bi-arrowed horizontal line is spanned between the first and next point. While mouse is moving, the lags and lag difference are reported in the third line of the screen. By clicking the mouse on the second position, lag difference is recorded in a small database. At the end of a measurement phase, simple statistics (number of measurements, mean value, standard deviation, standard error of estimate for the mean) is displayed.

Original step values are often not in a natural order because they may be determined by technical reasons. In order to put them to a natural and more meaningful scale they may be redefined (= "recoded") and amalgamated to fewer number

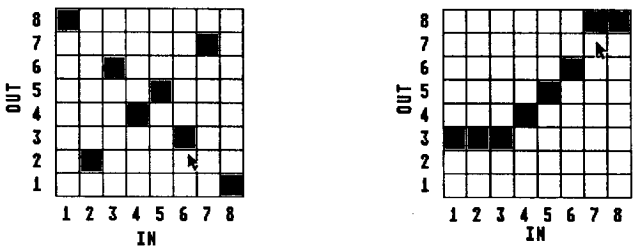

Figure 1: Recoding "checkerboard". IN: before recoding, OUT: after recoding. Left: one-to-one recoding (1↔8 ; 3↔6). Right: many-to-one recoding: (1,2,3→3 ; 7,8→8).

[3]The ovals are not parts of the actual screen. They are presented here as three typical examples of the compared sequences, appearing dynamically at the position B on the screen.

of classes than the number of original steps. For convenient redefinition of data, a "checkerboard" for redefinition of "IN" step values (before recoding) to "OUT" scale (after recoding) is displayed on the screen (Figure 1). This redefined scale is used in all following computations. Black squares, indicating the recoding of values, change their position with a mouse click.

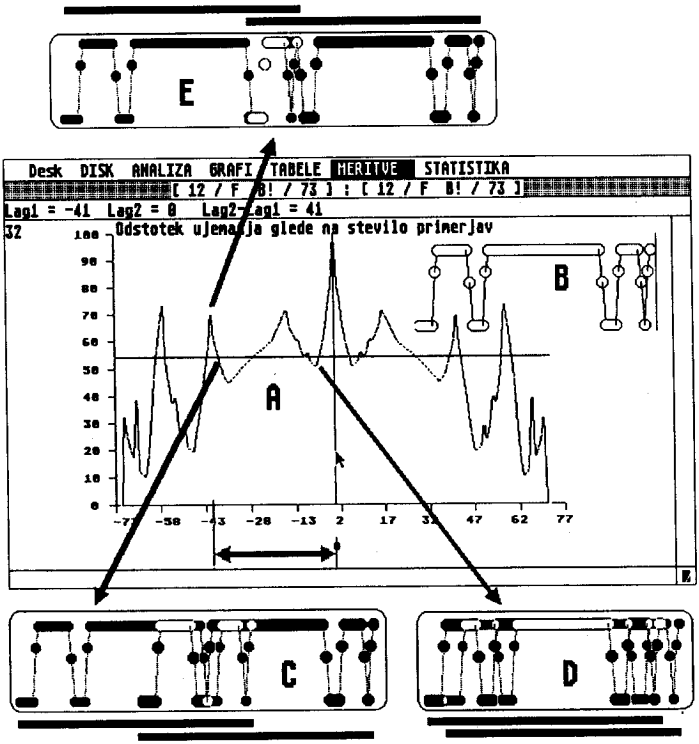

Figure 2: Moult cycle analysis of *Ligia italica*. A Box representing actual screen snapshot with associative function. B Actual plot of lagged data sequences; C,D,E Actual lagged data sequences corresponding to locations on the associative function (indicated by arrowed lines). They are dynamically shown on the screen at position B. Lines outside of the ovals indicate the span and lag of the compared sequences.

3 Examples

To demonstrate different possible applications of the approach, data from some other fields of research; ranging from behaviour of bees, DNA sequence analyses, stratigraphic sequences in geology to absorption spectrum analyses; were entered. Actually, any sequence of qualitative data that has to be compared with another sequence can be analysed with such approach.

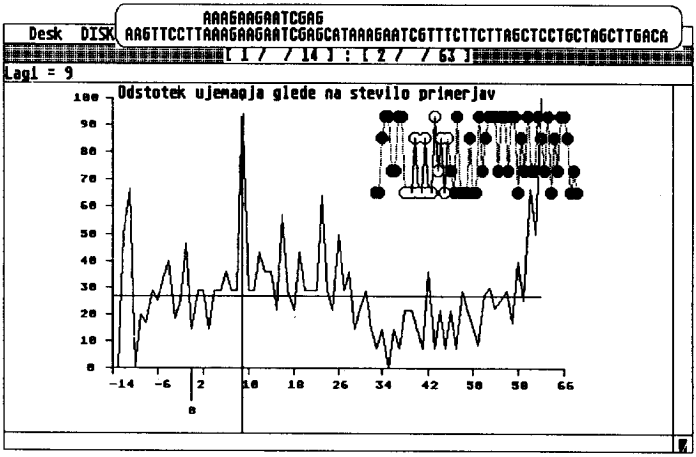

Figure 3: Finding the original DNA pattern

On Figures 3 and 4 an example of DNA sequence analysis is shown. One has to identify the position of a subsequence (**AAAGAAGAATCGAG**) in a longer sequence shown in oval insert on Figures 3 and 4. This sequence should also be matched to the complementary sequence, TTTCTTCTTAGCTC, since base A correspond to T and G correspond to C. The complementary sequence can be easily constructed by the recoding "checkerboard". Short data sequences, up to few hundred states, can be analyzed. For analysis of longer sequences different techniques should be used, see Church and Helfman (1993).

Another example, shown on Figure 5, shows identification of spectral lines. In the example shown, spectrum of hydrogen (represented by spikes with black dots) is perfectly matched with the corresponding lines in solar spectrum. High values at the beginning of the associative function are due to single spectrum line matches, which is clearly shown on the dynamically changing screen.

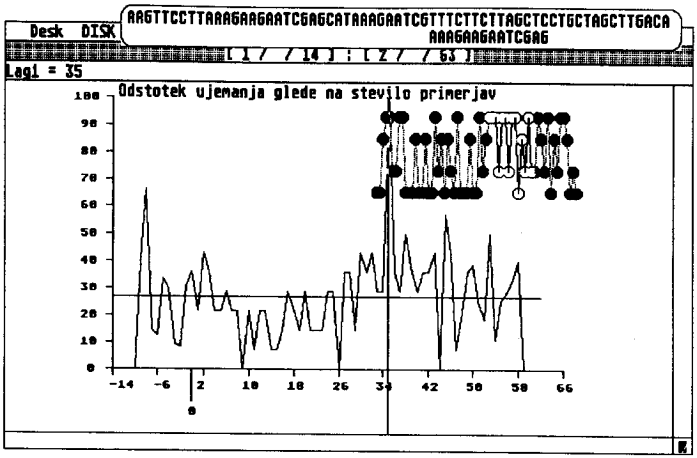

Figure 4: Finding location of the pattern complementary to the original pattern

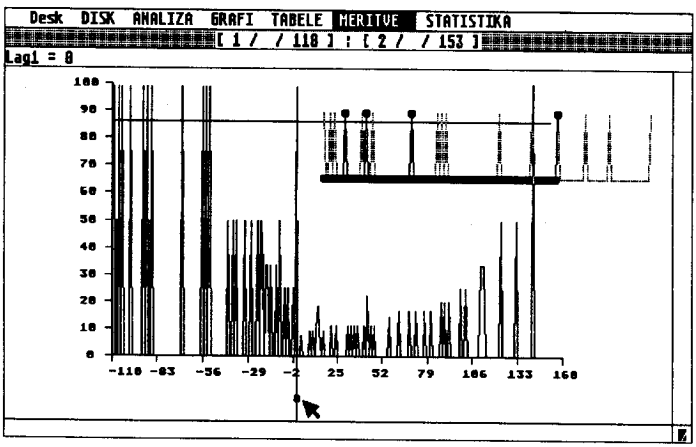

Figure 5: Identification of hydrogen spectrum in solar spectrum

4 Correlation

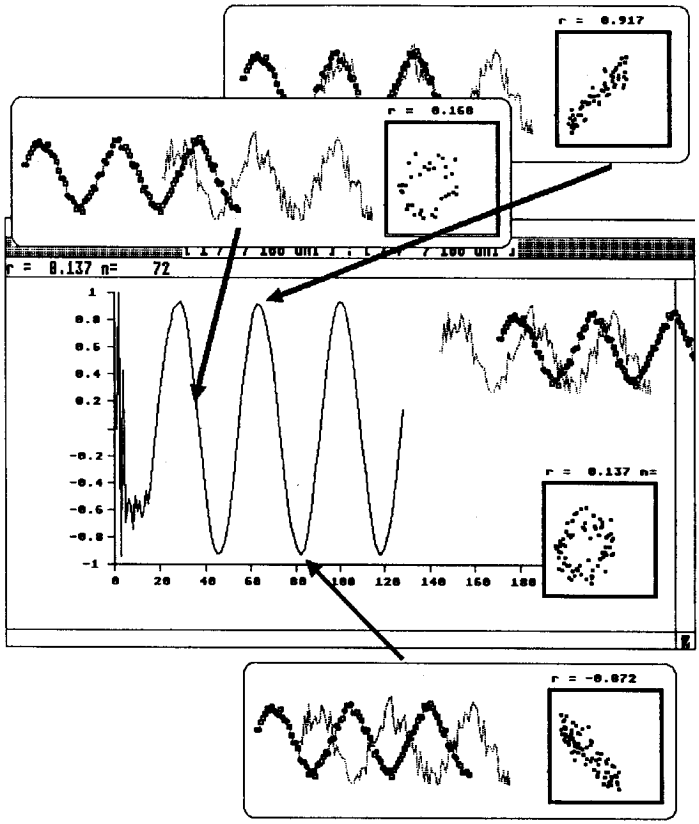

Figure 6: Crosscorelation analysis

With minor adaptations, analysis of numerical data sequences was incorporated (Figure 6). Matching of corresponding values was replaced with correlation. Cross- or auto-correlograms are plotted while data sequences slide one across another. In small squares in the lower-right portion of the screen, scattergrams of lagged data sequences are plotted. Dynamic positioning and measurements are possible but data recoding is disabled.

5 Discussion

Though the idea of searching for patterns which reveal sort of correlative relations in a series of data by means of visual display of the original series and the series with some lag is not new, we developed a tool that enables such searches in dynamical way. Our program effectively combines plots of some kind of association (correlation) function with plots of lagged data sequences so that one can study the formal tool and the visual tool in combination. This combination enables the end users, familiar with the subject under study, to interpret specific values of associative functions according to the display of actual data sequences. It is especially useful for analysis of non-numerical data where special care should be given to possible artefacts in formal measurement of correlation.

Described approach, with analytic and data description graphs presented simultaneously and dynamically, has the advantage to obviously present some interpretations of association and correlation when dealing with data sequences. It is useful for demonstrative and teaching purposes as well as for research.

6 Software

The program can be used on ATARI ST computers and is written in GFA Basic. The program is available from the author upon request. Since ATARI computers are not widely available, the program will be in future rewritten for use on more common computers, probably in Xlisp-Stat.

References

[1] Štrus, J., A. Blejec and P. Ličar (1992): Ultrastructure and Formation of the Cuticule in *Ligia italica* During the Moult Cycle. *First European Crustacean conference, Paris.*

[2] Church, K.W. and J.I. Helfman (1993): Dotplot: A Program for Exploring Self-Similarity in Millions of Lines of Text and Code. *Journal of Computational and Graphical Statistics*, **2**, 153-174.