Systematic Analysis of Time Series - CReMIT

Zoltán Pödör[1], Márton Edelényi[2], László Jereb[2]

***Abstract*—One of the problems frequently arising in connection with the study of time series is the most thorough examination of any correlations between them. In the case of time series with periodicity, study of the effects of periods with time shifting, delayed and varying length can be also examined. We present a data transformation procedure that allows more intensive studies by the systematic expansion of the basic data set due to the applied special window technique (CReMIT). To apply it in the practice, the method has been integrated in a uniform analyzing process that includes operations from the preparation of data lines for analysis, via their systematic transformation to the implementation of the analyses. The modular structure of the system provides high flexibility due to which the process can be suitable for the study of relations between time series of any type in practice by using the adequate expansions.**

***Index terms*—time series, systematic extension, CReMIT, data transformation, analyzing process**

I. Introduction

Searching for relations between the time series is a major area of statistics and data mining. A vast number of techniques are available and among them correlation and regression analysis [15], [20] defining connections between one or more independent and dependent variables are the most frequently used ones. At the same time the success and completeness of the studies can be significantly affected by the sphere of the involved dependent and independent parameters and in addition, the applied methods of their use in the analysis process. For example, the Principal Component Analysis (PCA) [1] or cluster analysis [11] can be suitable for the reduction of dimensions by combining the relevant variables.

In certain steps of the analysis generally only a specific slice specified by a window is used instead of the full range of the available time series. In the case of data lines of proper length, the temporal change of the connections can be examined using the forward and backward evolution techniques and moving intervals [3]. The evolution technique means, in fact, that the width of the applied window increases by one in each iteration without changing its starting point. In the case of moving intervals, the length of the examined interval is fixed in a suitable way and the starting point is moved forward by one cycle in every iteration step.

Manuscript received November 23, 2013, revised March 20, 2014.

[1] Institute of Mathematics, Faculty of Forestry, University of West Hungary, 4 Bajcsy-Zsilinszky street, Sopron 9400, Hungary, e-mail: podzol@emk.nyme.hu

[2] Institute of Informatics and Economy, Simonyi Károly Faculty of Engineering, Wood Sciences and Applied Arts, University of West Hungary, 9 Bajcsy-Zsilinszky street, 9400 Sopron, Hungary, e-mail: edelenyim@inf.nyme.hu, jereb@inf.nyme.hu

The window technique is used also in many other statistical procedures such as simple moving averaging, cover-up of events in time series [19] or segmentation procedures [8], [12] used also in the mining of time series. In all mentioned cases, an actual transformation or observation function is taken into account over the defined window. These procedures are characterized by the fact that only the width of the examined window or the starting point of the window changes in the individual iterations. The fuzzy time-series includes special window based techniques. These time series can be defined for example by differential equations. To predict the value at certain point $t + P$ in the future we want to use known values of the time series up to the point t in time. The standard method for this type of prediction is to create a mapping from D sample data points, sampled every α units in time $\big(x(t-(D-1)\alpha); \dots; x(t-\alpha); x(t)\big)$ to a predicted future value $x(t+P)$. The D, P and α values are usually defined by the user.

The window techniques can be used not only during the breakdown of the whole examined data line into intervals but also in a more specific meaning in the case of periodic time series. Periods can be often inherently assigned to time series used in studies relating the observations of the natural environment. Let's think of the temperature, rainfall, atmospheric pressure, tree growth, reproduction etc. data for which e.g. annual cycles can be defined. The relevant studies often require the creation of windows covering more than one previous cycles and having a varying width within a certain period and containing given aggregated data sets so that we can study their effects.

In our researches, basically we examined problems relating time series of forestry nature, with a special regard to the typical climate parameters on the forestry variables (tree growth, healthy conditions, lighttrap data). In the literature there are a great number of window techniques used. Let's take into consideration the analysis of the effects of the monthly rainfall and temperature data on the annual tree growth [7], [9], [10], [13], [14]. Moving window [5], [18], [22] and evolution techniques [2], [3] are often used to show the varying effects of the climatic components, and specially developed periodic data are also used [4], [17], [21]. However, they are created generally on experimental observations, and are not intended to global use.

Based on the above methods, a systematic window concept can be created by using their beneficial properties. The solution combines the essence of moving intervals and evolution techniques, the systematic movement of the windows and the systematic creation of windows with different widths for the individual steps. The procedure ensures the combination of the

time shifting and width values of the windows in a single process, significantly widening the sphere of the analyzing possibilities. The method has been basically developed for the solution of forestry related problems but it can be applied to any other periodic time series as well.

The procedure is integrated in an analyzing process that includes preparation of the data, data transformation and analysis procedural modules. In view of the possibilities, these modules are independent and can be adjusted thereby the spectrum of analyses can be increased and widened. The transformation module (CReMIT: Cyclic Reverse Moving Intervals Techniques) widens the sphere of variables that can be included in the study by using systematic transformation of the basic data on the basis of the above window technique.

II. Data Transformation – CReMIT

Let there be a time series and its natural period is marked by P. The elements of the time series are stored in vector x. Let its first element be the chronologically latest element, and natural numbers will be assigned to the data accordingly.

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \tag{1}$$

Let K $(1 \leq K \leq P)$ mark the starting point of the currently applied study, this is the Kth element of the vector. The time shifting (i) and width (j) values of the window applied on the time series are defined on the basis of this index. The minimal value of time shifting can be 0 $(i = 0)$, and the window width can be 1 $(j = 0)$. Using the periodicity of the time series the window defined in the above way will be periodically repeated, and the maximum cycle number (MCN) depending on the parameters and the length of the data line can be created. This MCN value defines the number of windows and the dimension of the transformed vector.

The starting and end point indexes of the windows created with the actual K, i and j values can be defined as $[K + i + l * P\,; K + i + j + l * P]$, $0 \leq l \leq MCN - 1$. Two temporal vectors are defined for the storage of the index values determining the limits of the windows using these parameters. Let us denote by:

$$index_b = \begin{pmatrix} K + i + 0 * P \\ K + i + 1 * P \\ \vdots \\ K + i + (MCN - 1) * P \end{pmatrix} \tag{2}$$

$$index_e = \begin{pmatrix} K + i + j + 0 * P \\ K + i + j + 1 * P \\ \vdots \\ K + i + j + (MCN - 1) * P \end{pmatrix} \tag{3}$$

By using the above indexes pre-defined transformation function TR can be applied on the elements of the individual windows.

$$tr\_x_{K,i,j} = \begin{pmatrix} TR(index_b[1]; index_e[1]) \\ TR(index_b[2]; index_e[2]) \\ \dots \\ TR(index_b[MCN]; index_e[MCN]) \end{pmatrix} \tag{4}$$

Based on the starting point K $(1 \leq K \leq P)$, the maximum time shifting value I $(0 \leq i \leq I)$ pre-defined on the basis of the task, the maximum window width J $(0 \leq j \leq J)$ and all the potential $tr\_x_{K,i,j}$ transformed vectors a systematic analysis procedure can be generated.

The value of the maximum cycle number, i.e. the dimension of the transformed vector is of significance in the individual iteration steps. In relation to vector x, this is determined by (MCN_x), the last (chronologically the oldest) element of the first window currently examined in a given iteration, $(K + i + j)$th element and the value of P period:

$$MCN_x = \left[\frac{n-(K+i+j)}{P}\right]+1, \tag{5}$$

where [] is the entire function.

The above transformation procedure generalizes the window techniques presented in the introductory chapter. Methodologically, it combines their beneficial properties, therefore it majors them. On the one hand it is able to change systematically the window width used in the evolution technique and to change systematically the starting point determining the moving intervals on the other hand. The advantage of CReMIT against these methods is that when the two principles are combined it can simultaneously and perfectly handle the windows with combined starting points and widths.

If the time shifting i is considered fixed and the window width is changed $(0 \leq j \leq J)$ the principle of the evolution technique is used: the window width is increased by one in each iteration step with a given starting point. If the window width is considered fixed with a given value j and the time shifting i is changed $(0 \leq i \leq I)$, then the method of moving intervals is implemented: the window width is fixed, and the window is moved forward by one in each iteration. With the choice $i = 0$ and $j = 0$ parameters, the relation between the actual $K, K + P, K + 2P$ etc. dependent and the independent variables of the same indexes are investigated. If $P = 1$, then with the given time shifting and window width values the original evolution and moving interval techniques are implemented for non-periodical time series.

The fuzzy time series techniques mentioned in the introductory chapter resembles to the CReMIT method, but they are able to implement only a part of CReMIT. These techniques use a starting point t, a time shifting α, but use only single elements, do not create windows with different width to define transformed time series.

III. The Analysing process

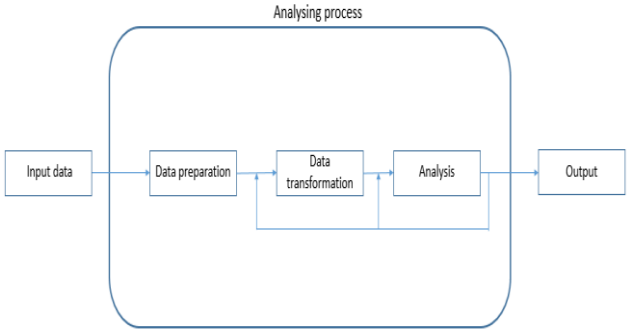

Fig. 1. The analysing process

The CReMIT transformation procedure has been established for the systematic study of more complex relations between periodic time series. The analyzing process built around the transformation procedure allows the use of the complete preparation-transformation-analysis process as a single unit, while the modular structure provides high flexibility in relation to the variables involved in the study, their transformed derivatives and the various analysis methods too.

The first main module is the data preparation that makes the raw data suitable for the integration into the transformation module. This involves not only the traditional data preparation (data cleaning, handling missing data) but also the combination of the raw data by time or by other aspects, or its breakdown, for example conversion of event based detections into identical time units. This data preparation allows us to organize the data into relevant periods as well.

The second module CReMIT is the transformation module that includes the essence and novelty of the procedure. This block is responsible for the systematic preparation of secondary data sets (time series) based on the transformation technique presented in the previous section.

The third module of the process receives data lines prepared during the transformation and carries out the pre-defined analysis procedure. The analysis performs basically the testing of connections between the time series including most frequently the correlation and the regression analysis.

The modules of the process are superimposed on each other but on the other hand, they are still independent to a certain extent and this feature provides significant flexibility and the possibility of combining and expanding the preparation, the transformation and the analysis methods. Next examples illustrate the possibilities of the use and the expansion of the process.

A. The Application of CReMIT

The basic method offers a number of development and expansion possibilities either in view of transformation functions used in the analysis module or of expansion to several variables. Transformation function TR used in the module can be a simple elemental conversion, for example average, sum, minimum, maximum. However, more complex other possibilities than the elementary functions can be also defined when for example some weight functions (which is exponential so that the effects of the chronologically older elements are exponentially reduced, or even other weighs considering special aspects can be defined) or non-linear functions are applied on the elements of the windows. The use of binary weight factors is a special application case that defines non-continuous windows. In addition, the method offers also the possibility when the transformation function is used not on all elements of the designated window but only on its elements meeting one or more given conditions.

The operation of the transformation module was presented in connection with vector x in general. Therefore it is a natural possibility of use when the transformation procedure is applied on the single independent parameter with the static dependent variable. This allows us to study the effects of the data segments of the examined independent variable having different lengths and time shifting on the dependent parameter.

Moreover, the procedure can be simply expanded to the dependent or more independent variables too. Applying it to the dependent variable we can analyze the effects of the independent parameter(s) on the various periods of the dependent parameter. Furthermore, the expansion to several independent variables makes the system also suitable for the implementation of analysis with multiple variables. If the periodicity of the different vectors is different or if we want to define various I and J values for them then all the applied vectors must be taken into consideration for defining MCN in the given iteration step (if the applied analysing method requests identical vector dimensions):

$$MCN = min\{MCN_{x1}; MCN_{x2}; \dots; MCN_{xw}\}, \qquad (6)$$

where w indicates the number of variables involved in the study, and MCN_{xm} $(1 \le m \le w)$ defines the maximum cycle number for vector x_m $(1 \le m \le w)$. In addition, the temporal vectors (w vector pairs) storing the starting and end indexes of the windows must be separately defined for each variable, and the transformation function TR can be also defined differently by any vectors.

The transformation procedure offers a number of uses and expansions in relation to either the individual handling of variables involved in the study or the applied transformation functions TR.

B. The Analysis Module

The transformed vectors prepared by the CReMIT module form the input for the analysis module. The available analysis procedures are preliminarily implemented by the user, and it is important to mention that the analysis module only uses the results of the transformation, and therefore, it is totally independent from the operation of the transformation module.

Let us suppose that a single variable correlation and regression analyses are carried out with transformations applied to the independent variable. In this case the transformation step prepares a 2 dimension matrix output depending on the value of parameters applied to the

independent variable in the analyzing module, where the lines indicates the current time shifting and the columns indicate the window widths. The individual cells contain the results obtained with the given time shifting i and window width j. Each cell contains a simple correlation coefficient or a more complex data structure depending on the nature of the examination. In order to provide more useful information for the users the test of the correlation significance is performed. Based on new statistical approaches [16] the effect size value is a nice measure for quantifying the difference between two groups on a common scale. But it measures only the effect size, not the sign of the effect. In this paper we used the correlation test based on t-test, which depends on the length of examples, because we have to know the sign of the effects too

Table 1 shows the results of the single variable correlation analysis. In the cells of the table the correlation coefficients can be seen only in those cases when the significance test was positive for the given set of data.

TABLE 1
OUTPUT MATRIX (EXAMPLE)

		J					
		0	***1***	***2***	***3***	***4***	***5***
I	***0***	0.43					
	1				0.65		
	2		0.54				0.5
	3					0.71	
	4	0.76			0.32		
	5			0.68			0.56

The applied analysis method in the analysis module can be simple linear correlation analysis, but in this module other more complex nonlinear methods can be applied too.

The software frame is implemented in the open source R (R 2.15.2. version) environment and the implemented elementary TR functions are the mean, sum, minimum and maximum ones. The default analysing procedure is the linear correlation analysis. The values of the applied parameters (K, I, J) depend on the actual problem. If the transformation procedure is expanded also to the dependent or to several independent variables the dimension of the output matrix will increase accordingly.

IV. CASE STUDY

The Forestry Light Trap Network has been operating in Hungary since the beginning of the 1960s which is a long period even internationally. The catching data of the various insects and butterflies are yearly summarized. The monthly precipitation sum and the average temperature are the studied meteorological parameters. This is explained by the fact that these two components are generally available in that breakdown, and the obtained results can be expanded also to other areas.

The purpose of the studies was to examine the dependence of the population dynamics of the selected species on the climate parameters. The time series of 23 catch places and 9 insect populations with different time length were examined. The analysis comprised the weather effects of the subject year and of the previous year by applying the CReMIT transformation procedure to monthly precipitation sum and average temperature data from April of the previous year to August of the actual year in a maximum length period of 6 months ($K = 5$, the index of month August, $I = 16$ and $J = 5$). The notation of previous year is *p* and the actual year is *a*. The resulting statistical correlations were analyzed in view of many aspects.

First the insect species and catch places were examined separately. Therefore we got 9*23=207 separate tables, with statistically significant correlation values ($\alpha = 0.05$). The relationships – concerning the given species and catch places – can be examined one by one based on these tables.

TABLE 2
SIMPLE RESULT TABLE OF ONE SPECIES AND CATCH PLACE

K=5		J					
		0	***1***	***2***	***3***	***4***	***5***
I	***0***		0.428	0.519	0.576	0.423	0.416
	1		0.505	0.605			
	2	0.482	0.557				
	3	0.503					
	4						
	5						
	6						
	7		-0.487	-0.432			
	8	-0.459					
	9			-0.523			
	10						
	11						
	12						
	13						
	14						
	15						
	16						

In order to obtain a more overall view of the individual species the significant correlation values of all catch locations can be arranged in a single table. With the purpose of getting the most useful general relations we can define a threshold value *k* that defines the minimum number of catch places where the correlation is significant in the given time period, generated by the CReMIT procedure.

TABLE 3
ONE SPECIES AND ALL CATCH PLACES (*K*=3)

time intervals	**Loc.1**	**Loc.2**	**Loc.3**	**Loc.4**	**Loc.5**	**Loc.6**
p8-a2		-0.723	0.564		0.515	
p10	0.497			0.421	0.626	
p12	-0.580	-0.581		-0.459		
p12-a1	-0.580	-0.581		-0.487		
p12-a4		-0.732			0.469	-0.522
a1-a4		-0.747			0.502	-0.525
a5	0.519			0.503	0.485	
a5-a6	0.519			0.557	0.508	
a5-a7	0.570			0.605	0.649	
a5-a8	0.535			0.576	0.549	0.508
a6-a7	0.490			0.505	0.628	
a6-a8	0.500			0.519	0.530	0.560

In Table 3 the significant correlation values for one of the species in all possible locations and with threshold k=3 are depicted. These tables and results provide many additional examination opportunities, for example it is possible to identify similar catch places based on the correlation values and to analyze the reason of these similarities (maybe using some other parameters for the catch places). Using the significant time periods determined by CReMIT – multivariable analysis can be also performed and models for the catch data based on the climate parameters can be also derived.

V. Summary/Conclusions

The study of the time series and the efficiency of the search for connections between them can be affected by the applied analysis methods and the sphere and use of variables involved in the study. The paper presented a transformation procedure CReMIT that supports the analysis of complex relations in a systematic way. Based on the systematic expansion of the width and time shifting of the windows applied to the variables, this will allow us to carry out more extensive studies than the previous, typically specific concepts. The method is implemented in an analyzing process, and it has been already used in practice on several data lines in forestry.

The independent parameters are often connected to climate or meteorology (precipitation, temperature, air moisture, blast, soil moisture, etc.) while the typical dependent variables are different measures of forestry, for example tree growth, healthy conditions, mortality, lighttrap data of pests and moths. Generally, the data lines show periodic properties and the identified correlations can provide information for experts to make scientifically supported decisions. Although our aim was to develop a method and apply for the given problems and not to derive professional forestry conclusions the applicability of the method to such problems has been shown by the investigations performed so far [6].

Acknowledgement

This work was supported by the TAMOP-4.2.2.C–11/1/KONV-2012-0015 (Earth-system) project sponsored by the EU and European Social Foundation.

Zoltán Pödör received the M.Sc. degree in Mathematics and Computer Science from the University of Szeged in 1999. In 2006 he joined to the Institute of Informatics and Economics at the University of West Hungary as a Ph.D. student under the supervision of Prof. László Jereb. His research interests cover the time series analysis and data mining techniques in practice. Currently he is working at University of West Hungary, Faculty of Forestry, Institute of Mathematics.

Márton Edelényi received the M.Sc. degree Business Information Systems from the University of West Hungary in 2008. After that he joined the Institute of Informatics and Economics at the University of West Hungary as a Ph.D. student under the supervision of Prof. László Jereb and Prof. László Szabó. His primary research interest is application possibilities of data mining in different areas.

László Jereb graduated in electrical engineering from the Technical University of Budapest (TUB) in 1971. He received the candidate of science degree and the Doctor of the Hungarian Academy of Sciences title in 1986 and 2004, respectively. Since 1971 he is with the Department of Communications of TUB, and since 2002 with the Institute of Informatics and Economics, University of West Hungary (UWH). With TUB his main teaching and research interests are the planning and performability analysis of infocommunications systems, while with UWH he deals with setting up information technology support for research in wood sciences and forestry.

References

[1] Abdi, H., Williams,. L. J. (July/August 2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, Volume 2, Issue 4, pp. 433–459.

[2] Biondi, F. (1997). Evolutionary and moving response functions in dendroclimatology. *Dendrochronologia* 15 (1997) pp. 139-150.

[3] Biondi F, Waikul K. (2004). DENDROCLIM2002: A C++ program for statistical calibration of climate signals in tree-ring chronologies. *Comp Geosci* 30, pp. 303–311.

[4] Briffa, K. R., Osborn, T. J., Schweingruber, F. H., Jones, P. D., Shiyatov, S. G., Vaganov, E. A. (2002). Tree-ring width and density data around the Northern Hemisphere: Part 1, local and regional climate signals. *The Holocene*. 12(6), pp. 737–757.

[5] Büntgen, U., Frank D. C., Schmidhalter, M., Neuwirth, B., Seifert, M., Esper, J. (2006). Growth/climate response shift in a long subalpine spruce chronology. *Trees*. 20(1), pp. 99–110.

[6] Csóka,Gy., Pödör, Z., Hirka, A., Führer, E. and Szőcs, L. (10–14 September 2012). Influence of weather conditions on population fluctuations of the oak processionary moth (Thaumetopoea processionea L.) in Hungary. *Joint IUFRO 7.03.10 – "Methodology of forest insect and disease survey" and IUFRO WP 7.03.06 – "Integrated management of forest defoliating insects" Working Party* Meeting, Palanga

[7] Dittmar, C., Zech, W., Elling, W. (2003). Growth variations of common beach (Fagus Sylvatica L.) under different climatic and environmental conditions in Europe – a dendroecological study. *Forest Ecology and Management* 173, pp. 63-78.

[8] Fu, T. (2010). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24, pp. 164-181.

[9] Führer, E. (2010). Tree growth and the climate (in Hungarian). „*KLÍMA-21" Füzetek* 61, pp. 98-107.

[10] Führer, E.; Horváth, L.; Jagodics, A.; Machon, A. and Szabados, I. (2011). Application of a new aridity index in Hungarian forestry practice. *Időjárás* 115, pp. 103-118.

[11] Han, J. and Kamber, M.*: Data Mining: Concepts and Techniques*, 2nd ed.. Morgan Kaufmann Publishers, 2006.

[12] Keogh, E., Chu, S., Hart, D., Pazzani, M. (1993). Segmenting Time Series: A Survey and Novel Approach. In an Edited Volume, *Data mining in Time Series Databases*. Published by World Scientific

[13] Manninger, M. (2004). Annual and periodic growth pattern of forest trees and its relation to ecological factors (in Hungarian). In: Mátyás, Cs., Vig, P. (Eds.), *Proceedings of the 4th Forests and Climate Conference*, University of West Hungary, Sopron, pp. 151-162.

[14] Maxime, C., Hendrik, D. (2011). Effects of climate on diameter growth of co-occurring Fagus sylvatica and Abies alba along an altitudinal gradient. *Trees* 25, pp.265–276.

[15] Miles, J., Shevlin, M.: *Applying Regression and Correlation: A Guide for Students and Researchers*. Sage publications Ltd., 2001.

[16] Nakagawa S, Cuthill IC. 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biological Reviews 82:591-605.

[17] Novák, J., Slodicák, M., Kacálek, D., Dusek, D. (2010). The effect of different stand density on diameter growth response in Scots pine stands in relation to climate situations. *Journal Of Forest Science*. 56(10), pp. 461–473.

[18] Oberhuber, W., Kofler, W., Pfeifer, K., Seeber, A., Gruber, A., Wieser, G. (2008). Long-term changes in treering– climate relationships at Mt. Patscherkofel (Tyrol, Austria) since the mid-1980s. *Trees*. 22(1) pp. 31–40.

[19] Pełech-Pilichowski, T., Duda,T. J. (2010). A two-level algorithm of time series change detection based on a unique changes similarity method. Proceedings of the International *Multiconference on Computer Science and Information Technology* pp. 259–263.

[20] Raymond H. Myers: *Classical and modern regression with applications* (second edition). Virginia Polytechnic Institute and State University, Duxbury, Thomson Learning, 1990.

[21] Scharnweber, T., Manthey, M., Criegee, C., Bauwe, A., Schroder, C., Wilmking, M. (2011). Drought matters – Declining precipitation influences growth of Fagus Sylvatica L. and Quercus robur L. in north-eastern Germany, *Forest Ecology and Management* 262(6), pp. 947-961.

[22] Wilczynski, S., Podlaski, R. (2007). The effect of climate on radial growth of horse chestnut (Aesculus hippocastanum L.) in the Swietokrzyski National Park in central Poland. *Journal of Forest Research*. 12(1), pp. 24–33.