Adjusting Time Series of Possible Unequal Lengths

Ilaria L. Amerise and Agostino Tarsitano

Abstract Previous researches dealing with combined time-series cross-section (TSCS) data usually assume that complete time-series exist for all the variables and all the cross-section units under consideration. In many contexts, however, this may be a very restrictive assumption. Our paper is concerned with problems of distance computation when every cross-section unit has contiguous data, but the time frames over which indicators are measured may differ for starting and/or ending dates.

Key words: TSCS data, Padding

1 Introduction

TSCS data are usually organized into a three-way array of data referring to a set of m units, indexed by $r = 1, 2, \ldots, m$, in which each unit is described by a set of p different time series or variable, indexed by $j = 1, 2, \ldots, p$. Both the set of units and the set of variables remain unchanged during the period of observation. In most cases, the lengths of time series are short, *e.g.* consisting of only $5 - 10$ time steps and seldom more than 20. In this case, it is essential to make the most efficient use of the few observations that are available. While there are a variety of problems related to TSCS data, in this paper we focus on the equalization of time series of unequal lengths in situations where the time series are fairly short for one reason or another. Our goal is to facilitate the computation of distance functions (*e.g.* the Minkowski

Ilaria L. Amerise
Dipartimento di Economia e Statistica, Università della Calabria, Cubo 0C, 87036 Rende (CS), Italy, e-mail: iamerise@unical.it

Agostino Tarsitano
Dipartimento di Economia e Statistica, Università della Calabria, Cubo 1C, 87036 Rende (CS), Italy,e-mail: agotar@unical.it

metrics) for clustering algorithms, which cannot be applied to time series with unequal length.

Let us consider the time series $\mathbf{x}_{r,j}$ observed on the unit r for the variable j for the period $[a_{r,j},\ldots,b_{r,j}]$ and suppose that there are data not reported at the start and/or at the end of a time series with respect to the longest time interval for which the j-variable of the TSCS data set is observed: $\alpha_j = \min_{1\le r\le n} a_{r,j}$ and $\beta_j = \max_{1\le r\le n} b_{r,j} \quad j = 1,2,\ldots,m$. Data falling in $[\alpha_j, a_{r,j}-1]$ or $[b_{r,j}+1,\beta_j,]$ are considered missing. It is important to distinguish between missing values due to a shorter time interval and values that are omitted for other reasons (*e.g.* experimental inaccuracies, irregularities in data collection). Only the former type is considered in our paper. Also, we will not discuss of missing values that appear after the beginning or before the end of the time series, because they require techniques which are outside the scope of this paper but have been clearly established elsewhere.

Lack of values at the extremes of the time series raise problems for the computation of distance functions requiring continuous data points. It is therefore customary to remove the leading and trailing "not available" observations and confining the computation to the shortest overlapping time period. Truncation, however, forces the longer time series to shrink (downsampling) to the length of the shorter which implies a waste of potentially useful information. Similar problems arise if one concentrates attention to the cross-section units for which a complete time series is available (thus forming a balanced sub-panel). Such a procedure appears to be unreliable due to the limited number of time points suffered by both time series and cross-sectional analysis.

2 Review of the Proposed Approaches

When two time series have a different length, it is possible to increase the number of data points by padding, upsampling, or interpolating the time-series to obtain sequences of the same length (see [2]). For reasons of space, we shall not pursue the first class of techniques beyond observing that their overall results are less satisfactory than those obtained by upsampling or interpolation.

Uniform scaling. [2] suggested a simple technique for transforming a time series $\mathbf{x}_{r,j}$ of length $\eta_{r,j}$ to produce a new time series of length η_j. The formula is: $x^{+}_{r,j,t} = x_{r,j,s}$ for $s = \lceil tf \rceil\, t = 1,2,\ldots,\eta_j$ where $f = \eta_{r,j}/\eta_j$ is the scaling factor and $\lceil . \rceil$ denotes the smallest integer not less than the argument. Uniform scaling implies the duplication of a certain number of values of the shorter time series using multiple copies of a value until its length equals that of the longest time series (with respect of the j-th variables of the panel). The values to be duplicated are selected uniformly (upsampling) in the range of the time series being stretched.

Asymmetrical filters. The completion of a sequence is similar in a sense to the calculation of the asymmetrical moving average filters used at the end-points of a time series to estimate the trend-cycle curve or the seasonal factor curve. See, for example, [3]. At each time point, the estimate $\hat{x}_{r,j,t}$ can be obtained by fitting a local

polynomial just to the first $(2h_{r,j}+1)$ or the last $(2k_{r,j}+1)$ observations. This is equivalent to taking linear combinations with weights that depends only upon the degree of the polynomial and the number of points to fit.

$$\hat{x}_{r,j,t} = \frac{\sum_{u \in U_{r,j}} w_{t,u} x_{r,j,u}}{\sum_{u \in U_{r,j}} w_{t,u}},\ t \in T_{r,j} \tag{1}$$

where $U = [1,\ldots,2h_{r,j}+1]$, $T_{r,j} = [\alpha_j,\ldots,a_{r,j}-1]$ for the missing values on the left and $U = [2k_{r,j},\ldots,b_{r,j}]$, $T_{r,j} = [b_{r,j}+1,\ldots,\beta_j]$ for the missing values on the right.
Box-Jenkins models. The simplest encountered model is the AR(1) model

$$x_{r,j,t} = \phi_{0,r,j} + \phi_{1,r,j} x_{r,j,t-1} + \varepsilon_t \quad t = \alpha_j,\ldots,\beta_j, \quad \left|\phi_{1,r,j}\right| < 1 \tag{2}$$

where $\phi_{0,r,j}/\left(1-\phi_{1,r,j}\right)$ is the expected value and $\phi_{1,r,j}$ determines the linear dependency in $\mathbf{x}_{r,j}$. The ε_t are independent errors with mean zero and finite variance. Estimation of the parameters and the associated standard errors can be done using maximum likelihood after skipping the missing observations. The time series are then enlarged with the forecasted values (forward or backward). See [4].

3 Experimental Results

To illustrate the two procedures, we choose the "Synthetic Control Chart Time Series" data set which contains 600 examples of time series each with 60values generated by the process introduced by [1]. There are 6 different classes with 100 representative examples from each class. We divide the 60 observations of each series into m consecutive subsequences of equal length $\eta_m = 60/m, m = 3,4,5$ which acts as the variables of the TSCS data set. For a number of cluster $c \in \{2,\ldots,6\}$ we select $\{31,21,16,13,11\}$ cases (cross-section units) for each cluster from the classes $1,2,\ldots,c$. When $c \neq 6$, the classes are chosen at random (without replacement) from $1,\ldots,6$ in each repetition of the experiment. To simulate time spans of different length, we randomly omit $0,1,\ldots,4$ contiguous entries at the beginning and end respectively of each subsequence. For all pairs of the m units we compute the city-block distance between the adjusted time series $d(\mathbf{x}_{r,j},\mathbf{x}_{s,j})$ which form the distance matrix $\mathbf{D}_j$ relative to the j-th variable for $j = 1,2,\ldots,m$ so that a comparison of two cross-section units becomes a comparison of two set of matrices. Finally, the units are classified using the Partitioning around medoids (PAM) algorithm on the weighted distance matrix $\mathbf{D} = \sum_{j=1}^{p} (w_j/\sum_{i=1}^{m} w_j)\mathbf{D}_j$ with $w_j = \max_{1\le r,s\le m} d(\mathbf{x}_{r,j},\mathbf{x}_{s,j})$. The adjusted Rand index (ARI) is chosen as validation measure. To compare the stability of the results, the data generation is repeated for 1000 times for each k and for each m. In all the experiments, we set k to the number of classes in the data set. The numerical summaries are given in Table 1.1.
It can be easily seen that, for all the methods, the mean values and the standard deviation of the ARI criterion, decrease as the number of cluster increases. This can

Table 1 Adjusted Rand index for interpolation methods

	No imputation			Filtering			AR(1) model			Upsampling		
c	3	4	5	3	4	5	3	4	5	3	4	5
Mean												
2	0.716	0.756	0.779	0.765	0.785	0.809	0.779	0.811	0.816	0.747	0.773	0.798
3	0.653	0.683	0.686	0.695	0.699	0.715	0.709	0.721	0.729	0.678	0.689	0.704
4	0.607	0.633	0.653	0.647	0.646	0.674	0.651	0.661	0.680	0.631	0.643	0.669
5	0.568	0.591	0.604	0.605	0.609	0.616	0.607	0.618	0.622	0.595	0.605	0.614
6	0.540	0.561	0.577	0.569	0.575	0.585	0.569	0.580	0.593	0.555	0.572	0.582
St. Dev.												
2	0.339	0.329	0.314	0.324	0.327	0.311	0.322	0.317	0.313	0.325	0.331	0.315
3	0.172	0.178	0.176	0.174	0.176	0.178	0.174	0.182	0.181	0.172	0.178	0.172
4	0.110	0.108	0.110	0.114	0.107	0.114	0.114	0.111	0.113	0.110	0.110	0.110
5	0.073	0.075	0.071	0.070	0.068	0.066	0.068	0.072	0.070	0.074	0.070	0.071
6	0.057	0.058	0.057	0.060	0.057	0.062	0.058	0.056	0.060	0.059	0.059	0.050

be considered an indication of the diminishing ability of the clustering algorithm PAM/City block to determine the appropriate number of clusters. On the other hand, the ARI is stable across the number of segments in which the original time series are divided, save perhaps an almost imperceptible increase for $m = 5$. The effectiveness of the padding procedures discussed in the previous section, can be determined by comparing their results with those obtained with no imputation. The improvement due to padding is evident, although not very pronounced. As a whole, the performance of the three proposed methods is of the same order of magnitude, although the impression is that there is a difference in favor of the AR(1) modeling.

There is no single best procedure to adjust time series of unequal length and a variety of techniques are available to make up the required length. Our modest experiment, however, should act as a deterrent to those who would quickly compute distances only for overlapping intervals.

References

1. Alcock, R.J. and Manolopoulos, Y.: Time-series similarity queries employing a feature-based approach. 7th Hellenic Conference on Informatics. August 27-29. Ioannina, Greece (1999). http://kdd.ics.uci.edu/databases/synthetic_control/synthetic_control.html
2. Keogh, E.: Efficiently finding arbitrarily scaled patterns in massive time series databases. In: Lavrac, N. and Gamberger, D. and Todorovski, L. and Blockeel, H. (eds.) Knowledge Discovery in Databases, PKDD, pp. 253–265. Springer, Heidelberg (2003)
3. Mills, T.C.: A note on trend decomposition: the classical approach revisited with an application to surface temperature trends. Journal of Applied Statistics, **34**, 963–972 (2007).
4. Velicer, W. F., Colby, S. M.: A comparison of missing-data procedures for arima time-series analysis. Educational and Psychological Measurement, **65**, 596–615 (2005)