Adjusting time series of possible unequal lengths

Ilaria Amerise - Agostino Tarsitano

Dipartimento di economia e statistica
Università della Calabria
RENDE (CS), Italy
Outline

Adjusting Time Series of Possible Unequal Lengths

- Time series cross-section data
- Missing data at the extremes
- Techniques of padding
- Experimental results
- Limitations and future research
TSCS data

TSCS data usually organized into a three-way array of data referring to a set of \( m \) entities, in which each entity is described by a set of \( p \) different time series.

\[
X_j = \begin{bmatrix}
  x_{1,j,a_{1,j}} & x_{1,j,a_{1,j}+1} & x_{1,j,a_{1,j}+2} & \cdots & x_{1,j,b_{1,j}} \\
  x_{2,j,a_{2,j}} & x_{2,j,a_{2,j}+1} & x_{2,j,a_{2,j}+2} & \cdots & x_{2,j,b_{2,j}} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  x_{r,j,a_{r,j}} & x_{r,j,a_{r,j}+1} & x_{r,j,a_{r,j}+2} & \cdots & x_{r,j,b_{r,j}} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  x_{m,j,a_{m,j}} & x_{m,j,a_{m,j}+1} & x_{m,j,a_{m,j}+2} & \cdots & x_{m,j,b_{m,j}}
\end{bmatrix}
\]

\( j=1,2,\ldots,p \)

Our paper is concerned with problems of classification of short TSCS data where the time frames over which indicators are measured may differ for starting and/or ending dates.
In some cases, time series are short (20 time points or fewer). There are several reasons.

1) The process of gathering data is very expensive or endangers the health of the subjects.

2) Delivering data to users takes long time.

3) Data are acquired by an infrequent survey due to experimental factors or logistical constraints.

4) Methodological changes in data collection.

5) Interruption of survey execution or data access.

6) Limited comparability between different sources.

These datasets present unique challenges.
Missing data at the extremes

For greater generality, we have admitted the possibility that the time spans can have different starting and ending points for both different indicators and different entities.

The time series $x_{r,j}$ concerning the entity $r$ is observed for the variable $j$ for the period $[a_{r,j}, \ldots, b_{r,j}]$. Let

$$
\alpha_j = \min_{1 \leq r \leq n} \{a_{r,j}\}; \quad \beta_j = \max_{1 \leq r \leq n} \{b_{r,j}\}
$$

Data falling in the following intervals are considered missing values

$$
[\alpha_j, a_{r,j} - 1]; \quad [b_{r,j} + 1, \beta_j]
$$

<table>
<thead>
<tr>
<th>Type</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Na</td>
<td>Na</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Na</td>
<td>Na</td>
<td>Na</td>
<td>Na</td>
<td>Na</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Na</td>
<td>Na</td>
<td>1</td>
<td>Na</td>
</tr>
<tr>
<td>3</td>
<td>Na</td>
<td>Na</td>
<td>Na</td>
<td>Na</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Na</td>
<td>Na</td>
<td>Na</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Na</td>
<td>Na</td>
<td>Na</td>
</tr>
<tr>
<td>5</td>
<td>Na</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Na</td>
<td>Na</td>
</tr>
<tr>
<td>6</td>
<td>Na</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

1=Valid datum
Na=Not available

j-th variable
Empty data/Missing data

It is important to distinguish between missing values due to a shorter time interval (empty data) and values that are omitted for other reasons (e.g. experimental inaccuracies or values not released by government data agencies).

Only the former type is considered in our paper.

“empty data” are values that are unknown and do not have a real value. For instance, because the time series data were not collect or because data observed up to a certain time point, but not beyond.

Also, we will not discuss about of missing values appeared 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 somewhere else.
Distance between short time series

A simple distance function between two time series is the city-block metric

\[ d_{j,r,s} = \sum_{t=t_{1,r,s}}^{t_{n,r,s}} |x_{j,r,t} - x_{j,s,t}| \]

Lack of values at the extremes raise problems for the computation of distances.

It is customary to remove the leading/trailing “Na” and confining the computation to the shortest overlapping time period \( t_{1,r,s} - t_{n,r,s} \).
Truncation

Truncation, however, forces the longer time series to shrink to the length of the shorter, which implies a waste of potentially useful information.

Similar problems arise if one concentrates attention to the cross-sections for which a complete time series is available. Such a procedure appears to be unreliable due to the limited number of time points suffered by both time series and cross-sectional analysis.

The ability of any measurement of distance to distinguish between objects decreases with the reduction of the number of comparisons.

When two time series have a different length, it is possible to increase, albeit artificially, the number of points.
Valid data in between two sequences of missing data can be interpreted as an effect of an intervention on some response variable.

To extend the effect, consecutive Na’s at the start and/or at the end are replaced by the arithmetic average (or the median) across time.

In practice, constants are added to the distance reflecting the number of non-overlapping periods.

<table>
<thead>
<tr>
<th>entity</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>m1</td>
<td>m1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>m1</td>
<td>m1</td>
<td>m1</td>
<td>m1</td>
<td>m1</td>
<td>m1</td>
<td>m1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>m2</td>
<td>m2</td>
</tr>
<tr>
<td>3</td>
<td>m3</td>
<td>m3</td>
<td>m3</td>
<td>m3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>m4</td>
<td>m4</td>
<td>m4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>m4</td>
<td>m4</td>
<td>m4</td>
<td>m4</td>
</tr>
<tr>
<td>5</td>
<td>m5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>m5</td>
</tr>
<tr>
<td>6</td>
<td>m6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

This approach is fast and simple, but introduces distortions: the constants might be different for each comparison, thus potentially altering the Euclidean property of a distance function.
Separate paddings

Separate paddings at the start (end) using the mean or the median of the first (last) few values before (after) the gap are not more effective.

A similar problematic result occurs with assigning to all consecutive missing points the value at the closest extreme

<table>
<thead>
<tr>
<th>entity</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>m1a</td>
<td>m1a</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>m1b</td>
<td>m1b</td>
<td>m1b</td>
<td>m1b</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>m2</td>
<td>m2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>m3</td>
<td>m3</td>
<td>m3</td>
<td>m3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>m4a</td>
<td>m4a</td>
<td>m4a</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>m4a</td>
<td>m4a</td>
<td>m4a</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>m5a</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>m5b</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>m6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Here is a basic lesson where time series analysis is about comparing pre- and postintervention trends. You cannot know what a trend looks like if you do not have enough data either before or after the intervention.

However, with increasing number of missing points and/or reducing number of valid observations, doubts increase about the reliability of the interpolated estimates.
Keog (2003) suggested a simple technique for transforming a time series $x_{r,j}$ of length $\eta_{r,j}$ to produce a new time series of length $\eta_j$

$$x_{r,j,f} = x_{r,j,f}, \quad t = 1,2,\ldots,\eta_j$$

$$f = \left\lfloor t\left(\frac{\eta_{r,j}}{\eta_j}\right)\right\rfloor$$  Scaling factor

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 to the longest time series (with respect of the j-th variables).

The values to be duplicated are selected uniformly (upsampling) in the range of the time series being stretched.
Asymmetric filters

The completion of a time series 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.

At each time point, the estimate can be obtained by fitting a local polynomial just to the last \((2h_{r,j} + 1)\) or the first \((2k_{r,j} + 1)\) observations.

This is equivalent to taking linear combinations of the observations with weights that depend only on the degree of the polynomial and the number of points to fit.

\[
\hat{x}_{r,j,t+n|t} = \sum_{u=-h_{r,j}}^{h_{r,j}} \frac{w_{u,n}x_{r,j,u}}{\sum_{u=-h_{r,j}}^{h_{r,j}} w_{u,n}};
\]

\[
\hat{x}_{r,j,t-n|t} = \sum_{u=-k_{r,j}}^{k_{r,j}} \frac{w_{u,n}x_{r,j,u}}{\sum_{u=-k_{r,j}}^{k_{r,j}} w_{u,n}}
\]

We obtained the end-point fitting a cubic polynomial to the nearest 5, 7, 9, valid points. The number of points increases according to the number of empty values at the extremes.
Cubic fitting

Note that the number of points needed to estimate the first terms and might be different from the number of points used to estimate the last terms.

We note further that either of $h_{r,j}$ or $k_{r,j}$ may be zero.

The time series is reversed for backforecasting.
Linear Gaussian state space models

The general linear Gaussian state space model for the n-dimensional observation sequence $x_t$ can be given by

$$
\begin{cases}
    x_t = Z_t w_t + \varepsilon_t, & \varepsilon_t \sim NID(0,H_t) \\
    w_{t+1} = T_t w_t + R_t \eta_t, & \eta_t \sim NID(0,Q_t)
\end{cases}, \quad t = 1,2, \ldots
$$

Where $w_t$ is the state vector, $\varepsilon_t$ and $\eta_t$ are disturbance vectors and the system matrices $Z_t$, $T_t$, $R_t$, $H_t$ and $Q_t$ are fixed and known but a selection of elements may depend on an unknown parameter vector.

By appropriate choices of the elements, a wide range of different time series models can be derived. For example, the local linear trend model is given by

$$
\begin{cases}
    \text{Observation equation:} & x_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \sim NID(0,\sigma^2_{\varepsilon}) \\
    \text{State equation, level:} & \mu_{t+1} = \mu_t + v_t + \eta_t, \quad \eta_t \sim NID(0,\sigma^2_{\eta}) \\
    \text{State equation, slope:} & v_{t+1} = v_t + \zeta_t, \quad \varepsilon_t \sim NID(0,\sigma^2_{\zeta})
\end{cases}, \quad t = 1,2, \ldots,n
$$

where the $\varepsilon_t$s, $\eta_t$s and $\zeta_t$s are all mutually independent.
Local linear trend model

![Graph showing local linear trend model](image)

Although it is simple, the local linear trend model is not an artificial model and it provides the basis for the treatment of important series in practice.
Short interrupted time series

An ITS is a special kind of time series in which we know the specific point in the series where an intervention occurred on some response variable.

If the treatment has a causal impact, the post-intervention series will have a different level or slope than the pre-intervention series.

Filling in consecutive missing data at the very beginning or end of a time-series is similar to the idea of ITS.

Causal hypothesis is that observations during treatment will have a different level or slope from those after and before Intervention.
Statistical analysis of ITS are intended to reveal the impact caused on the dependent variable by the intervention. Box-Jenkins modeling is frequently suitable for this purpose.

The simplest and most commonly encountered model sciences is

$$x_{r,j,t} = \phi_{0,r,j} + \phi_{1,r,j}x_{r,j,t-1} + \epsilon_t; \quad |\phi_{1,r,j}| \leq 1, \quad E(x_{r,j,t}) = \frac{\phi_{0,r,j}}{1 - \phi_{1,r,j}}$$

Where the $\epsilon_t$s 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 at the extremes.

The time series are then enlarged with the forecasted values (forward or backward).
Synthetic Control Chart Time Series data set contains 600 examples, each with 60 time series values and there are 6 different classes with 100 representative examples from each class.

We divide the 60 observations of each time series into \( m \) consecutive subsequences of equal length \( p=60/m, \, m=3,4,5 \) which acts as the variables of a TSCS data set.

For a number of cluster 2,3,…,6 we select (31,21,16,13,11) cases for each cluster from the classes 1,2,…c. When \( c\neq 6 \), the classes are chosen randomly (without replacement) from 1,…,6 in each repetition.
To simulate time spans of different length, we randomly omit 0, 1, 2, 3, 4 contiguous entries at the beginning and at the end respectively of each subsequence.

For all pairs of the entities we compute the city-block distance between the adjusted time series, which form the distance matrix $D_j$ relative to the $j$-th variable of the TSCS for $j=1,2,\ldots,m$.

Finally, the units are classified using the PAM algorithm on the weighted distance matrix

$$D = \sum_{j=1}^{m} \left( \frac{w_j}{m} \sum_{j=1}^{w_j} D_j \right)$$

$$w_j = \max_{1 \leq r \leq p} \{d_{j,r,s}\}$$

To compare the stability of the results, the data generation is repeated for 500 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 adjusted Rand index (ARI) is chosen as validation measure.
Experimental results

<table>
<thead>
<tr>
<th>Adjusted Rand index for interpolation methods</th>
<th>Keog</th>
<th>Cubic</th>
<th>L.L.T.</th>
<th>Ar(1)</th>
<th>Null</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.75</td>
<td>0.77</td>
<td>0.80</td>
<td>0.77</td>
<td>0.79</td>
</tr>
<tr>
<td>3</td>
<td>0.68</td>
<td>0.69</td>
<td>0.70</td>
<td>0.70</td>
<td>0.70</td>
</tr>
<tr>
<td>4</td>
<td>0.63</td>
<td>0.64</td>
<td>0.67</td>
<td>0.65</td>
<td>0.65</td>
</tr>
<tr>
<td>5</td>
<td>0.60</td>
<td>0.61</td>
<td>0.61</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td>6</td>
<td>0.56</td>
<td>0.57</td>
<td>0.58</td>
<td>0.57</td>
<td>0.69</td>
</tr>
<tr>
<td>St.Dev.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.33</td>
<td>0.33</td>
<td>0.32</td>
<td>0.32</td>
<td>0.33</td>
</tr>
<tr>
<td>3</td>
<td>0.17</td>
<td>0.18</td>
<td>0.17</td>
<td>0.17</td>
<td>0.18</td>
</tr>
<tr>
<td>4</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>5</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>6</td>
<td>0.06</td>
<td>0.06</td>
<td>0.05</td>
<td>0.06</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Mean values and standard deviations of ARI, decrease as k increases. This can be considered an indication of the diminishing ability of the PAM/City block to confirm the appropriate number of clusters.

The mean values of ARI increase and the st.devs decrease, with increasing m, the number of time series in which the original sequences are divided.

The improvement due to padding is evident, although not very impressive.
Limitations and Future Research

TSCS data are intrinsically multivariate; the analysis of many sequences of each cross section entity is necessary to adequately capture interspecific relationships.

A multivariate version of the autoregressive model or of the state space model should be evaluated, where it is possible with appropriate trials.

There are a lot of procedures to adjust time series of unequal length. We have performed some preliminary work on simulated data with various techniques and feel that this area requires further research. At present, the impression is that there is a small difference in favor of the ar(1) modeling.

However, our modest experiment, however, acts as a deterrent to those who would quickly compute distances only for overlapping intervals.

The End