Estimating missing data in hierarchical space-time series with a short temporal extent

Martin Charlton
Chris Brunsdon
Conor Cahalane
Lars Pforte

National Centre for Geocomputation,
National University of Ireland Maynooth,
Maynooth,
IRELAND

martin.charlton@nuim.ie

Abstract

A challenging problem exists in the estimation of missing space-time data where the time series are relatively short, and the space series belong to a spatial hierarchy. An example is provided by the population estimates for regions belonging to the NUTS hierarchy which are available from the EUROSTAT data portal. The table *demo_r_gind3* provides estimates of the population of NUTS0/1/2/3 regions at the 1st January 2000…2012 inclusive. Inspection of the table reveals that estimates are missing for 2000-2003 for two of the five NUTS3 regions in the NUTS2 region of Liège. There are other instances of missing data at NUTS3 where there are data for the corresponding higher level NUTS regions. The EUROSTAT table *demo_r_d2jan* provides estimates of the population on the 1st January for a longer time period, 1990…2012 inclusive, but these are only to NUTS2. Again, there is missing data. The question then arises as to whether it is possible to estimate the missing series. The NUTS2 values act as a constraint on the NUTS3 values – the total population of the NUTS3 regions should equal those of the corresponding NUTS2 regions. However, the relative shortness of the available series is a challenge if conventional methods of time series analysis are adopted. Furthermore, the imposition of the spatial constraints is both a check as well as a challenge.

For the purposes of this exercise, a time series is a set of observations of some characteristic, taken at regular time intervals. Annual estimates of population would be one example of a series; monthly counts of unemployment would be another. One convention in representation is as an ordered vector of observations relative to the current time period *t*: $x_t, x_{t-1}, x_{t-2}, \ldots, x_{t-n}$.

One of the goals of time series analysis is to make forecasts – either into the future, or the past as a backcast. This usually requires a model which is fitted to the existing series to be calibrated through the estimation of some parameters. Reliable forecasts need reliable estimate of parameters, in turn needing a plentiful supply of data. There are a range of potential models which might be used.

Exponential smoothing models are a generalisation of the simple moving average: $s_t = \alpha x_t + (1 - \alpha)s_{t-1}$, where α is the smoothing factor which lies between 0 and 1 (Holt, 1957). Variations include double exponential smoothing if the data exhibit trend, and triple exponential smoothing if the data exhibit seasonal variation (this last source of variation can be a characteristic of monthly employment series). A common form of these models is Holt-Winters smoothing. Box and Jenkins (1970) popularised a form of model with autoregressive and moving average components; in the autoregressive component,

the terms of the series are modelled as a function of the previous terms: . $x_t = c + \sum_{i=1}^{p} \vartheta_i x_{t-i} + \varepsilon_t$, and the moving average component models the serial relationship in the errors: . $x_t = \mu + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$.

The models are usually applied to a stationary version of the series which may be obtained through differencing successive terms: $d' = d_t - d_{t-1}$. Such models are generally referred to as ARIMA(*p*,*d*,*q*) models: autoregressive integrated moving average models with *p* autoregressive parameters, *d* degrees of differencing and *q* moving average parameters, where the values for *p*, *d* and *q* are typically in the range 0…2. In Box and Jenkins' description of model fitting, the process involves a deal of manual intervention, although more recently, automated fitting functions have been developed which find suitable values of *p*, *d*, and *q* to minimise some function such as the AIC (Hyndman and Khandakar, 2008).

The drawbacks to taking these approaches lie in both the insufficiency of the data which is available (leading to large standard errors on the parameter estimates) and the necessity to treat the individual series as independent. Furthermore, if the missing data lie in the middle of a series, then the data insufficiency problem is doubled. One might forecast forwards from the early section of the series, backforecast from the later section, and perhaps average the forecasts, but this has an element of the ad hoc about it which is unsatisfactory and has no theoretical basis.

A more consistent approach is to consider the non-missing data as the evidence that we have already, and seek methods which provide a more coherent approach to the problem of estimating the missing data. To this end, we consider a Bayesian approach to the problem.

Given a set of data *D* and model parameters θ, the data can be modelled by the probability distribution: $P(D|\theta)$. *D* might be a 2 x n matrix of x and y pairs for a regression model, and θ would be the triplet of slope, intercept and error variance from the model. Using Bayes theorem we can make a probabilistic statement about θ given *D* thus:

$$P(\theta \mid D) = P(\theta)\frac{P(D \mid \theta)}{\int_{\theta} P(D \mid \theta) d\theta}$$

where $P(\theta)$ is the Bayesian prior distribution, representing the analyst's belief in the value of θ prior to the analysis. $P(\theta \mid D)$ is the posterior distribution for the parameters θ, given the data *D*. The denominator in the expression is often not analytically tractable and in these cases Markov Chain Monte Carlo methods may be helpful. This approach simulates random variables drawn from $P(\theta \mid D)$ rather than estimating the distribution itself. The simulated values are then used to investigate the characteristics of $P(\theta \mid D)$.

The MCMC approach can be used to estimate missing data within the same consistent framework. We treat the missing data as unobserved variables, D^*, and derive the expression for the posterior predictive distribution of the missing data:

$$P(D^* \mid D) = \int P(D^* \mid \theta) P(\theta \mid D) d\theta$$

As an example, we estimate missing data in a relatively short series with MCMC. The example time series has 25 observations, from which 5, near the middle, have been removed. The remaining data are then used to estimate the missing values.

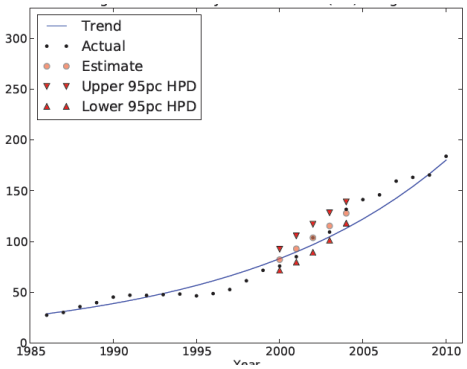

Figure 1

The results, shown in Figure 1, exhibit a reassuringly close match to the original values. The technique also allows us to provide a credible interval around each estimated value: in a Bayesian framework the credible interval corresponds to the confidence interval in classical statistical inference.

The approach is flexible enough to allow the inclusion of hierarchical constraints required to maintain the consistency of the values through the various levels of the NUTS hierarchy when the data are treated in a cross-sectional manner.

In this paper we show the MCMC approach in use to provide examples of missing data estimation for the EUROSTAT population series.

References

BOX, GEP, AND JENKINS GJ, 1970, *Time series analysis: forecasting and control*, San Francisco: Holden-Day

HOLT, CE, 1957, *Forecasting Trends and Seasonal by Exponentially Weighted Averages*, ONR Memorandum No. 52, Pittsburgh: Carnegie Institute of Technology

HYNDMAN RJ, AND KHANDAKAR Y, 2008, Automatic time series forecasting: the *forecast* package for R, *Journal of Statistical Software*, 27(3),