Multiple imputation of time series with an application to the construction of historical price indices

Fernando TUSELL[*]

Abstract

Time series in many areas of application, and notably in the social sciences, are frequently incomplete. This is particularly annoying when we need to have complete data, for instance to compute indices as a weighted average of values from a number of time series; whenever a single datum is absent, the index cannot be computed. This paper proposes to deal with such situations by creating multiple completed trajectories, drawing on state space modelling of time series, the simulation smoother and multiple imputation ideas.

Keywords: Multiple imputation; Time series; Missing data; Kalman filter; Index computation

[*] Departamento de Estadística y Econometría. Facultad de CC.EE. y Empresariales, Avda. Lehendakari Aguirre, 83, 48015 BILBAO. E-mail: `fernando.tusell@ehu.es`.

1 Introduction

Missing data is a pervasive problem, afflicting not only the social sciences but also the physical and medical sciences.

Missing data in cross-sectional data is handled in a variety of ways, many admittedly *ad hoc* and making sense only in particular cases: cold deck and hot deck imputation, using only complete records, replacing the mean for missing data, using all available (possible incomplete) records, etc.

It has long been recognized that a sounder, more principled approach is desirable, and considerable effort has been expended in this direction. Much of it stems from the seminal work in Little and Rubin (1987) (an enlarged second edition appeared in 2002). The monograph Schafer (1997) describes in considerable detail a methodology to deal with missing data in cross sectional data, and Rubin (1996) provides a useful overview of the ideas on multiple imputation and its impact on statistical practice.

The literature dealing with missing data in (multiple) time series is nonetheless sparse. Missing data in time series is considered in Little and Rubin (2002); conceptually, the problem can be handled in the same way as in cross sectional data. However, the problem is both harder and more important. Harder, because an additional level of complexity exists when dealing with multivariate time series: both contemporaneous and lagged relationships between components need to be considered when imputing a missing data point. More pressing, because strategies like using only complete data records are no longer feasible. With cross sectional data, discarding records with data missing completely at random (MCAR) has no other effect than reducing the available sample. In a time series, each record is unique: dropping it would leave us with a series with holes, unusable for many purposes.

In the last fifteen years, state space modelling of time series has seen wider acceptance and is now a well established tool in the kit of the applied statistician. A number of theoretical breakthroughs like the simulation smoother (cf. Harvey et al. (2004), de Jong (1995), Durbin and Koopman (2002)) and Markov Chain Monte Carlo (see for instance Gamerman (1997)), along with ever increasing computing power at the desktop, have improved our chances of dealing with missing data in multivariate time series. We describe in this paper a possible approach.

The rest of the paper is organized as follows. Section 2 reviews some of the basic theory on state space models and Kalman filtering and smoothing that we will be using. Section 3 discusses some of the models we have found useful for the purposes of this paper. Section 4 shows an example and discusses some further applications and the issues that they raise.

2 State space models

Let $\{\boldsymbol{y_t}\}$ be a p-variate time series, observed (perhaps partially) at times $t = 1, \ldots, n$. We are concerned with the imputation of the missing values.

We will assume $\{\boldsymbol{y_t}\}$ is generated by an state space model (see, for instance, Anderson and Moore (1979) or Durbin and Koopman (2001)):

$$\begin{aligned} \boldsymbol{y_t} &= \boldsymbol{Z_t \alpha_t} + \boldsymbol{\epsilon_t} && (1) \\ \boldsymbol{\alpha_{t+1}} &= \boldsymbol{T_t \alpha_t} + \boldsymbol{\eta_t} && (2) \end{aligned}$$

where $\boldsymbol{\epsilon_t} \sim N(\boldsymbol{0}, \boldsymbol{H_t})$ and $\boldsymbol{\eta_t} \sim N(\boldsymbol{0}, \boldsymbol{Q_t})$. Matrices $\boldsymbol{Z_t}, \boldsymbol{T_t}, \boldsymbol{Q_t}$ and $\boldsymbol{H_t}$ are in general time-varying, and may depend on parameters that need to be estimated.

Let $\boldsymbol{\mathcal{Y}_s} \overset{\text{def}}{=} \{\boldsymbol{y_1}, \ldots, \boldsymbol{y_s}\}$, i.e. the section of the time series up to and including time s. Given $\boldsymbol{Z_t}, \boldsymbol{T_t}, \boldsymbol{Q_t}$, $\boldsymbol{H_t}$ and $\boldsymbol{\mathcal{Y}_t}$, the Kalman filter (see for instance Anderson and Moore (1979), Durbin and Koopman (2001)) gives the conditional mean value and covariance matrix of state vector $\boldsymbol{\alpha_t}$ at each point in time, $\boldsymbol{a_{t|t-1}} = E[\boldsymbol{\alpha_t}|\boldsymbol{\mathcal{Y}_{t-1}}]$ and $\boldsymbol{P_{t|t-1}} = \text{Cov}(\boldsymbol{\alpha_t}|\boldsymbol{\mathcal{Y}_{t-1}})$. Defining,

$$\begin{aligned} \boldsymbol{F_t} &= \left(\boldsymbol{Z_t P_{t|t-1} Z_t}^{\mathrm{T}} + \boldsymbol{H_t}\right) && (3) \\ \boldsymbol{K_t} &= \boldsymbol{T_t P_{t|t-1} Z_t}^{\mathrm{T}} \boldsymbol{F_t}^{-1} && (4) \\ \boldsymbol{L_t} &= \boldsymbol{T_t} - \boldsymbol{K_t Z_t} && (5) \\ \boldsymbol{v_t} &= \boldsymbol{y_t} - \boldsymbol{Z_t a_{t|t-1}} && (6) \\ \boldsymbol{M_t} &= \boldsymbol{P_{t|t-1} Z_t}^{\mathrm{T}} && (7) \end{aligned}$$

the formulae for one-step-ahead updating are:

$$\begin{aligned} \boldsymbol{a_{t|t-1}} &= \boldsymbol{T_{t-1} a_{t-1|t-2}} + \boldsymbol{K_{t-1} v_{t-1}} && (8) \\ \boldsymbol{P_{t|t-1}} &= \boldsymbol{T_{t-1} P_{t-1|t-2} L_{t-1}}^{\mathrm{T}} + \boldsymbol{Q_{t-1}}. && (9) \end{aligned}$$

In order to start the iteration, either it is assumed that $\boldsymbol{\alpha_0} \sim N(\boldsymbol{a_{0|-1}}, \boldsymbol{P_{0|-1}})$ or a diffuse prior is used (see Durbin and Koopman (2001), § 5.2).

Similar algorithms, collectively known as *smoothers*, give $\boldsymbol{\hat{\alpha}_t} \overset{\text{def}}{=} E[\boldsymbol{\alpha_t}|\boldsymbol{\mathcal{Y}_n}]$ and $\boldsymbol{V_t} \overset{\text{def}}{=} \text{Cov}(\boldsymbol{\alpha_t}|\boldsymbol{\mathcal{Y}_n})$, i.e. conditional on the full length of time series. For instance, defining

$$\begin{aligned} \boldsymbol{r_{t-1}} &\overset{\text{def}}{=} \boldsymbol{Z_{t-1}}^{\mathrm{T}} \boldsymbol{F_{t-1}}^{-1} \boldsymbol{v_t} + \boldsymbol{L_t}^{\mathrm{T}} \boldsymbol{r_t} && (10) \\ \boldsymbol{N_{t-1}} &\overset{\text{def}}{=} \boldsymbol{Z_t}^{\mathrm{T}} \boldsymbol{F_t}^{-1} \boldsymbol{Z_t} + \boldsymbol{L_t}^{\mathrm{T}} \boldsymbol{N_t L_t}, && (11) \end{aligned}$$

we have

$$\begin{aligned} \boldsymbol{\hat{\alpha}_t} &= \boldsymbol{a_{t|t-1}} + \boldsymbol{P_{t|t-1} r_{t-1}} && (12) \\ \boldsymbol{V_t} &= \boldsymbol{P_{t|t-1}} - \boldsymbol{P_{t|t-1} N_{t-1} P_{t|t-1}}. && (13) \end{aligned}$$

The iteration is initialized with $\boldsymbol{N_n} = 0$ and $\boldsymbol{r_n} = 0$ (see Durbin and Koopman (2001), § 4.3.3).

Assume that the system matrices $\boldsymbol{Z_t}, \boldsymbol{T_t}, \boldsymbol{Q_t}$ and $\boldsymbol{H_t}$, possibly depending on a parameter vector $\boldsymbol{\theta}$, are known. Algorithms referred to as *simulation smoothers* afford easy generation of trajectories of $\boldsymbol{\alpha_t}$, $\boldsymbol{\epsilon_t}$ and $\boldsymbol{\eta_t}$ conditional on both $\boldsymbol{\mathcal{Y}_n}$ and $\boldsymbol{\theta}$, by drawing from the distributions $p(\boldsymbol{\alpha_t}|\boldsymbol{\mathcal{Y}_n}, \boldsymbol{\theta})$, $p(\boldsymbol{\epsilon_t}|\boldsymbol{\mathcal{Y}_n}, \boldsymbol{\theta})$, $p(\boldsymbol{\eta_t}|\boldsymbol{\mathcal{Y}_n}, \boldsymbol{\theta})$ (see de Jong (1995) and a simpler algorithm in Durbin and Koopman (2002)).

In practice, $\boldsymbol{Z_t}, \boldsymbol{T_t}, \boldsymbol{Q_t}$ and $\boldsymbol{H_t}$ are seldom known and need to be estimated, at least in part. If this is the case, the uncertainty introduced by the use of estimated parameters in place of $\boldsymbol{\theta}$ has to be accounted for. Very little work seems to have dealt with this issue: Watanabe (1985) gives some asymptotic results. Both Hamilton (1986) and Pfefferman and Tiller (2005) describe techniques to account for the influence of estimated parameters on the variance of $\hat{\boldsymbol{\alpha}}_{\boldsymbol{t}}$ and prediction mean square error.

3 Models and strategy for imputation

There are no limitations in the choice of imputation models, other than the requirement to keep the number of estimated parameters down to a manageable size. Although, in principle, any model that fits the data reasonably well can be used, we have found simple structural models (see Harvey (1989),Harvey et al. (2004) and Durbin and Koopman (2001) for instance) well suited for the task of imputation.

Local level full dimension multivariate model. Taking $\boldsymbol{T_t} = \boldsymbol{Z_t} = \boldsymbol{I_p}$ in equations (1)-(2) we have what may be the simplest multivariate model for $\boldsymbol{y_t}$: each component y_{it} $(1 \leq i \leq p)$ fluctuates about a component α_{it} of $\boldsymbol{\alpha_t}$.

The choice of the covariance matrix $\boldsymbol{Q_t}$ governs the degree of correlation among components α_{it}, α_{jt}, $i \neq j$. We can choose to have independent random walks (diagonal $\boldsymbol{Q_t}$), non-independent random walks ($\boldsymbol{Q_t}$ with non null off-diagonal terms) or even a reduced rank model ($\boldsymbol{Q_t}$ not of full rank; except for the possible influence of the prior distribution on $\boldsymbol{\alpha_1}$ this would be equivalent to a reduced dimension state vector).

Regarding $\boldsymbol{H_t}$, we can choose independent or correlated observation disturbances. We can also take $\boldsymbol{H_t} = \mathbf{0}$, effectively saying that the components α_{it} of the state can be observed without error, whenever the corresponding y_{it} is observed. In either case, interest normally centers in the generation of the vector $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{t}}$ of simulated trajectories.

Local level reduced rank multivariate model. An alternative to the full rank model consists in keeping the random walk dynamics for th vector state $\boldsymbol{\alpha_t}$ while taking $p = \dim(\boldsymbol{y_t}) > \dim(\boldsymbol{\alpha_t}) = s$. In that case, the observed time series $\{\boldsymbol{y_t}\}$ evolve as linear combinations of a small number s of unobserved components. In this case, matrix $\boldsymbol{Z_t}$ will typically contain regression coefficients of the $\boldsymbol{y_t}$ on the $\boldsymbol{\alpha_t}$. The model can be seen as a dynamic factor analysis model.

Generation of imputed trajectories. A random sample of m trajectories from $(\boldsymbol{\alpha_t}|\mathcal{Y}_n)$ can then be obtained as sketched in Algorithm 1.

Each time it is run, the loop in steps 2 to 5 in the Algorithm 1 generates a random $\hat{\boldsymbol{\theta}}$ from the posterior distribution of the parameters by using rejection sampling (see for instance Gamerman (1997), p. 85).

Algorithm 1 – Simulation of the state with prior information on $\boldsymbol{\theta}$

```
Require: $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$, $p(\boldsymbol{\theta})$, $m$
1: for $i = 1$ to $m$ do
2:   repeat
3:     Draw $\hat{\boldsymbol{\theta}}$ from the prior distribution $p(\boldsymbol{\theta})$.
4:     Draw $U$ from the uniform distribution on $[0, 1]$
5:   until $\ell(\hat{\boldsymbol{\theta}})/\ell(\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}) > U$
6:   $\hat{\boldsymbol{\theta}}^{(\boldsymbol{i})} \leftarrow \hat{\boldsymbol{\theta}}$
7:   Draw the $i$-th random trajectory $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{t}}^{(i)}$ from $p(\boldsymbol{\alpha_t}|\mathcal{Y}_n, \hat{\boldsymbol{\theta}}^{(\boldsymbol{i})})$.
8: end for
```

The likelihood for each $\hat{\boldsymbol{\theta}}$ needed at step 5 can be computed by running the Kalman filter, generating the set of innovations $\boldsymbol{v_t}$ and their covariance matrices $\boldsymbol{F_t}$ from (6) and (3) above and setting

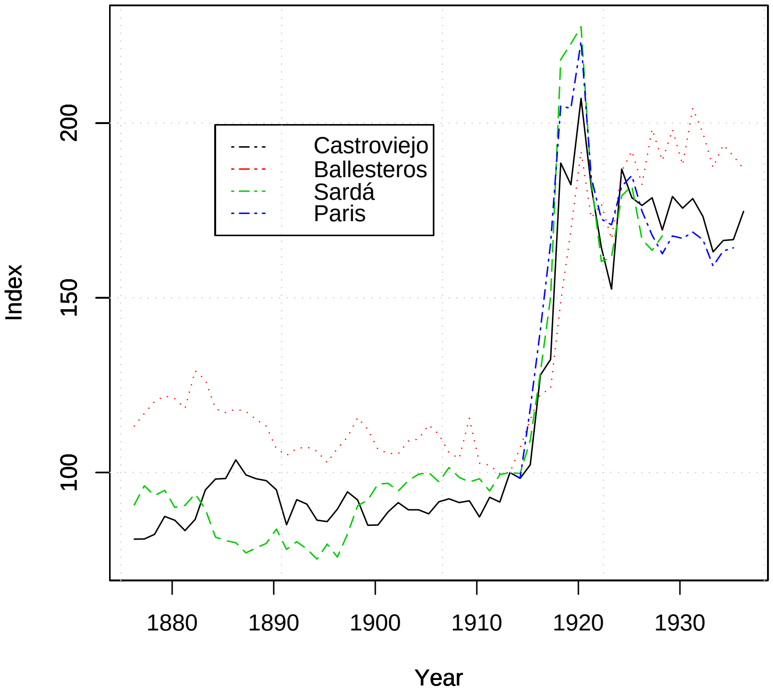
$$\log \ell(\hat{\boldsymbol{\theta}}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_t \left(\log|\boldsymbol{F_t}| + \boldsymbol{v_t}^T \boldsymbol{F_t}^{-1}\boldsymbol{v_t}\right);$$

the inversion of $\boldsymbol{F_t}$ at each step may be avoided altogether by treating the time series $\boldsymbol{y_t}$ as univariate (Durbin and Koopman (2001), § 6.4, Anderson and Moore (1979), § 6.4), so the computation of the likelihood is reasonably fast.

Step 7 is handled with the simulation smoother (see Durbin and Koopman (2001), § 5.3).

There are particular cases that can be handled faster. For certain prior distributions $p(\boldsymbol{\theta})$ we may be able to exploit conjugacy and sample directly from the posterior.

Figure 1: Cost of living indices computed for the whole or part of Spain by four historians

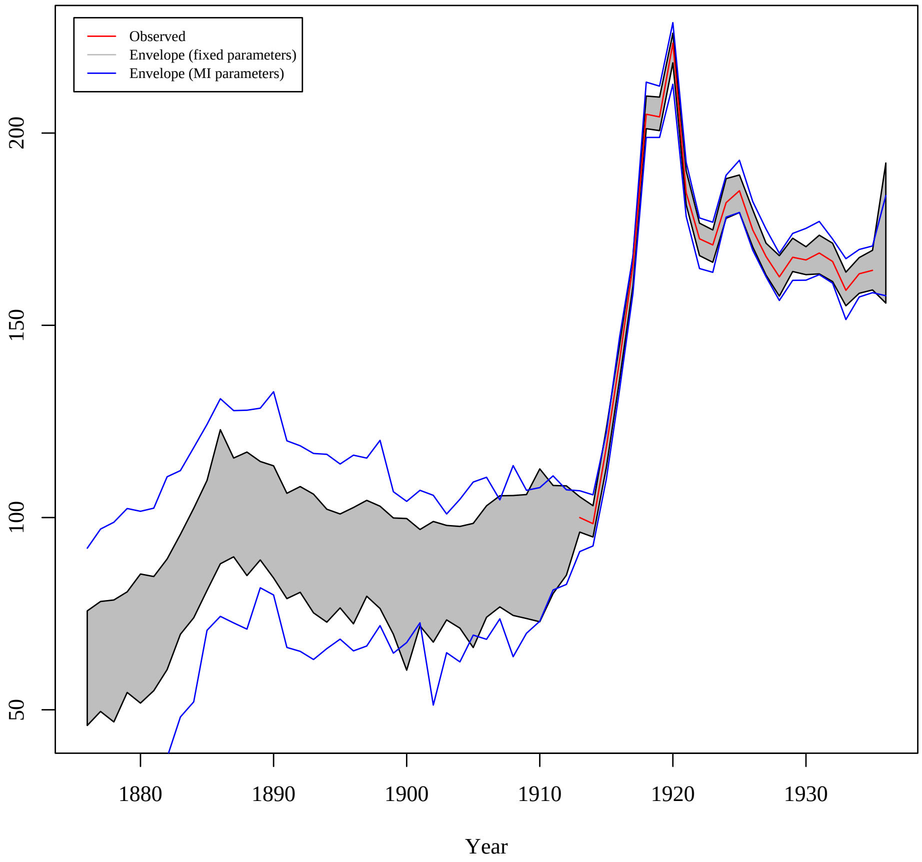


On the other hand, if we have no prior information whatsoever, we may choose to sample $\hat{\boldsymbol{\theta}}^{(i)}$ from the asymptotic distribution $\mathrm{N}(\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}, \mathcal{I}(\hat{\boldsymbol{\theta}}_{\mathrm{MLE}})^{-1})$, where $\mathcal{I}(\hat{\boldsymbol{\theta}}_{\mathrm{MLE}})$ is the information matrix. (A similar approach in another context has been proposed by Hamilton (1986).) Thus, steps 2 to 7 in Algorithm 1 are replaced by a single draw from the asymptotic distribution of the estimates.

4 An illustration

Figure 1 shows four cost of living indices, computed by four historians. They refer to the period 1876–1936 and different regions or the whole of Spain. (For a description of the indices, see Pérez Castroviejo (submitted for publication) and references therein.) As could be expected, the four indices show a similar pattern: no long term trend before World War I, a phenomenal inflation during the war, then a drop of prices which nonetheless failed to return to

Figure 2: Observed index and confidences bands for imputation of the price index by Paris. Grey band is the envelope of 100 trajectories from the conditional distribution $p(\boldsymbol{\alpha_t}|\mathcal{Y}_n, \hat{\boldsymbol{\theta}}_{\mathrm{MLE}})$. The outer band is the envelope of 100 trajectories from $p(\boldsymbol{\alpha_t}|\mathcal{Y}_n, \hat{\boldsymbol{\theta}}^{\boldsymbol{(i)}})$ with $\hat{\boldsymbol{\theta}}^{\boldsymbol{(i)}}$ varying.



the pre-war levels.

Not all four indices are available for the whole period, the one computed by Paris being the shortest.

Given the similar patterns displayed by the indices over the period where all four are observed, we can attempt to impute the missing years of an index using the past and future observations of itself and the others. To do so, we have set up the full dimension local level model described in Section 3 above.

Thus, we consider:

$$\boldsymbol{\alpha_{t+1}} = \boldsymbol{\alpha_t} + \boldsymbol{\eta_t} \tag{14}$$

$$\boldsymbol{y_t} = \boldsymbol{Z_t}\boldsymbol{\alpha_t} + \boldsymbol{\epsilon_t} \tag{15}$$

where $\{\boldsymbol{y_t}\}$ is the four dimensional time series and $\boldsymbol{Z_t}$ is a matrix whose rows are a subset of the rows of the unit $\boldsymbol{I_4}$ matrix: those rows are taken that correspond to observed components of $\boldsymbol{y_t}$.

The covariance matrices are assumed time invariant; $\boldsymbol{H_t} = \boldsymbol{H}$ is chosen diagonal while $\boldsymbol{Q_t} = \boldsymbol{Q}$ is a full general covariance matrix, with no other restriction than being symmetric non-negative definite. Thus, we are assuming that what is observed are the "true" underlying indices plus observation error, and the observation errors are unrelated for each of the four historians. On the other hand, the disturbances driving the state vector are correlated, as seems natural in this case.

With the model thus specified, two sets of one hundred trajectories of the state $\boldsymbol{\alpha_t}$ have been generated with the simulation smoother. In one case, the parameters were kept fixed at the values estimated by maximum likelihood, while in the other each trajectory was generated with a vector of parameters $\hat{\boldsymbol{\theta}}^{(\boldsymbol{i})}$ sample from $\mathrm{N}(\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}, \mathcal{I}(\hat{\boldsymbol{\theta}}_{\mathrm{MLE}})^{-1})$. The envelopes for each set of trajectories are represented in Figure 2 and can be interpreted as approximate (simultaneous) confidence bands for the state. It can be noticed that taking into account the variability of the parameters increases substantially the width of the band, which nearly doubles in the regions where no observations were available and the Sardá index had to be imputed with information from the other three.

All computations were programmed in R (see R Development Core Team (2004)). Software is available from the author.

5 Conclusion

An approach has been proposed for the imputation of multivariate time series, and its use illustrated imputing a price index with unavailable information. The approach is general enough to cope with the general situation where several, partially overlapping sources of information are available and we need to construct an index.

We also may notice that the Kalman filter and smoother can deal with series with disparate observation periods, i.e., some series could be observed monthly and others quarterly. The use of the Kalman filter in such situations is demonstrated in Harvey and Pierse (1984), where the emphasis is in prediction or benchmarking while in our case is imputation.

Finally, we would like to point out that the purpose of multiple imputation in the example shown is to account for the uncertainty in the estimation of parameters, an hence be able to produce “honest” confidence intervals or bands for the estimands of interest. This issue is all too often neglected in time series analysis: Chatfield (2001) is one of the rare monographs to discuss this issue in his Chapter 8. Perhaps in many applications the uncertainty due to imprecise estimation of the parameters is likely to be negligible, given large enough sample sizes. That this is not always so is well epitomized by the example above.

Acknowledgments. Support from UPV/EHU (Grupo Consolidado 9/UPV 00038.321-13631/2001) and MCyT (grant BEC2003-02273) is gratefully acknowledged.

References

Anderson, B. and Moore, J. (1979), *Optimal Filtering* (Prentice-Hall).

Chatfield, C. (2001), *Time-Series Forecasting* (Chapman and Hall/CRC).

de Jong, P. (1995), The simulation smoother for time series models, *Biometrika*, **82**, 339–350.

Durbin, J. and Koopman, S. (2001), *Time Series Analysis by State Space Methods* (Oxford Univ. Press).

Durbin, J. and Koopman, S. (2002), A simple and efficient simulation smoother for state space time series analysis, *Biometrika*, **89**, 603–615.

Gamerman, D. (1997), *Markov chain Monte Carlo* (Chapman and Hall).

Hamilton, J. (1986), A standard error for the estimated state vector of a state-space model, *Journal of Econometrics*, **33**, 387–397.

Harvey, A. (1989), *Forecasting, structural time series models and the Kalman filter* (Cambridge Univ. Press).

Harvey, A., Koopman, S. and Shephard, N., editors (2004), *State Space and Unobserved Components Models* (Cambridge Univ. Press).

Harvey, A. and Pierse, R. (1984), Estimating missing observations in economic time series, *Journal of the American Statistical Association*, **79**, 125–131.

Little, R. and Rubin, D. (1987), *Statistical Analysis with Missing Data.* (John Wiley and Sons, New York), second edition appeared 2002.

Little, R. and Rubin, D. (2002), *Statistical Analysis with Missing Data* (Wiley), second edition.

Pfefferman, D. and Tiller, R. (2005), Bootstrap approximation to prediction MSE for state-space models with estimated parameters, *Journal of Time Series Analysis*, **26**, 893–916.

Pérez Castroviejo, P. (submitted for publication), Poder adquisitivo y calidad de vida de los trabajadores vizcainos, 1876–1936, *Revista de Historia Industrial*.

R Development Core Team (2004), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org`, ISBN 3-900051-07-0.

Rubin, D. (1996), Multiple imputation after 18+ years (with discussion), *Journal of the American Statistical Association*, **91**, 473–489.

Schafer, J. (1997), *Analysis of Incomplete Multivariate Data* (Chapman and Hall, London).

Watanabe, N. (1985), Note on the Kalman filter with estimated parameters, *Journal of Time Series Analysis*, **6**, 269–278.