Review of Income and Wealth
Series 53, Number 4, December 2007

# USING OVERLAPPING AND INCOMPLETE TIME SERIES FOR THE ESTIMATION OF COST OF LIVING INDICES

BY P. M. PÉREZ-CASTROVIEJO AND F. TUSELL*

*University of the Basque Country, Spain*

We address the estimation of cost of living indices from time series which are incomplete, in a way that exploits all available information, while also giving an indication of the uncertainty associated with the estimation. The method used allows for multiple sources of prices for a single item, extending over the same or partially overlapping time ranges. We describe summarily the methodology and demonstrate its use in the estimation of a cost of living index for Biscaye (North of Spain), for the period 1862–1940.

## 1. INTRODUCTION

Let $\boldsymbol{p}^t = \left(p_1^t, \ldots, p_n^t\right)^{\mathrm{T}}$ be the vector of prices of $n$ items at time $t$. Similarly, let $\boldsymbol{q}^t$ be the vector of quantities (or weights) of said items, also at time $t$. Two popular index numbers are those named after Laspeyres and Paasche. Those indices at time $t = 1$ with base at time $t = 0$ are given by

$$I_{\mathrm{LASP}} \stackrel{\mathrm{def}}{=} \frac{\sum_{i=1}^{n} p_i^1 q_i^0}{\sum_{i=1}^{n} p_i^0 q_i^0} = \frac{\boldsymbol{p}^1 \boldsymbol{q}^0}{\boldsymbol{p}^0 \boldsymbol{q}^0} \tag{1}$$

$$I_{\mathrm{PAAS}} \stackrel{\mathrm{def}}{=} \frac{\sum_{i=1}^{n} p_i^1 q_i^1}{\sum_{i=1}^{n} p_i^0 q_i^1} = \frac{\boldsymbol{p}^1 \boldsymbol{q}^1}{\boldsymbol{p}^0 \boldsymbol{q}^1}. \tag{2}$$

The numerators in formulae (1)–(2) are the values of a basket made of items $i = 1, \ldots, n$ at the prices prevailing at time $t = 1$. The denominator is the value of the same basket at the prices prevailing at time $t = 0$. The difference between the two indices is in the basket object of valuation, which is made of quantities

*Note*: We thank, for constructive comments and ideas, Esther Ruiz and Juan Romo, two anonymous referees, the editor and participants at seminars in Bilbao, Pamplona and Madrid. P. M. Pérez-Castroviejo was supported by MCyT (grant BEC2002-03927) and UPV/EHU (grant UPV-12.321-H-14860). F. Tusell was supported by (grant 9/UPV 00038.321-13631/2001) and MCyT (grant BEC2003-02273).

*Correspondence to: Fernando Tusell, Departamento de Econometría y Estadística, Facultad de CC.EE. y Empresariales, Avenida Lendakari Aguírre, 83, E-48015 Bilbao, Spain (fernando.tusell@ehu.es).

$q_1^0, \ldots, q_n^0$ for $I_{\text{LASP}}$ and quantities $q_1^1, \ldots, q_n^1$ for $I_{\text{PAAS}}$. There are proposals theoretically superior to $I_{\text{LASP}}$ and $I_{\text{PAAS}}$, like the so-called "ideal index number of Fisher" (see Vogt and Barta, 1996, p. 14),

$$I_{\text{FISHER}} \stackrel{\text{def}}{=} \sqrt{I_{\text{PAAS}} \cdot I_{\text{LASP}}};$$

but the differences between $I_{\text{FISHER}}$ and either of $I_{\text{LASP}}$ or $I_{\text{PAAS}}$ are usually small (cf. Feinstein, 1998, p. 639).

Whether we use $I_{\text{LASP}}$, $I_{\text{PAAS}}$, $I_{\text{FISHER}}$ or, indeed, most anything else which has been put forward as an index number, we need the full vector of prices for both $t = 1$ and $t = 0$, if we are going to compare the index values at those time points. This poses a problem, for that information is usually unavailable, either because the sources are incomplete—as is frequently the case with historical registers—or because the price series are intrinsically unobservable—as is the case of agricultural prices, only observed seasonally.

A solution, admittedly *ad hoc* but frequently used, is to set the missing prices at plausible values. Sometimes, resort is made to simple linear interpolation (Reher and Ballesteros, 1993, note 31). Another solution is to restrict the computation in both numerator and denominator of (1)–(2) to items for which we have prices available. This is easy to implement but not exempt from problems: it makes the index numbers inhomogeneous over time. In extreme cases, the value of the index at time points with scarce information may reflect the prices of only a tiny minority of the items.

Our proposal proceeds differently: we impute unobserved prices, so expressions (1)–(2) are always computed for the same basket of items, with partly observed and partly imputed prices. This preserves the conceptual homogeneity of the index, which is always the ratio of values of *the same* basket of items at two different points in time.

Clearly, this is not problem free: for one thing, we cannot consider on an equal footing prices really observed and imputed prices. The imputation process creates the fiction that all the prices have been observed, when in fact some of them have been "fabricated." We have to account for the uncertainty introduced in the imputation process.

We present in the following a methodology to deal with incomplete and overlapping time series in the computation of price indices. In order to make the description concrete, we consider a particular historical data set, whose peculiarities epitomize well the problems and features the historian or economist is likely to find in the construction of indices. It will become apparent in the sequel that the methodology proposed is quite general and not limited to the example shown.

The rest of this paper is structured as follows. Section 2 describes the sources of several data sets with prices from the province of Biscaye (North of Spain) in the period 1862–1940. Section 3 introduces the model and the technique used. Section 4 presents and comments on the results obtained. Finally, Section 5 provides some context of our work and mentions refinements, applications and extensions.

TABLE 1
SUMMARY OF DATA SOURCES

| Source | Period | Series |
|---|---|---|
| Boletín Oficial de Vizcaya (Valmaseda) | 1862–1890 | 10 |
| Boletín Oficial de Vizcaya (Bilbao) | 1862–1890 | 10 |
| Hospital Civil de Basurto | 1879–1935 | 25 |
| Santa Casa de Misericordia de Bilbao | 1881–1924 | 13 |
| Ayuntamiento de Barakaldo | 1891–1899 | 18 |
| Cooperativa Altos Hornos de Vizcaya | 1893–1903 | 22 |
| Ayuntamiento de Barakaldo | 1906–1927 | 45 |
| Boletín Estadística Ayto. de Bilbao | 1913–1940 | 36 |

## 2. DATA

Some modelling decisions have been made in the light of the data, and require an understanding of it. We therefore revert to the common practice of proposing a model, then illustrating its use, and describe first the problem to better motivate the model later.

### 2.1. *Prices*

The information used has been described at length in Pérez-Castroviejo (2006). It consists of price time series from different sources in Biscaye. They extend over different intervals in the range 1862–1940. Table 1 and Figure 1 give a synopsis of the number of series available from each source and the years they cover. We comment briefly on the sources and features of the data.

The time series from the Cooperativa de Altos Hornos de Vizcaya[1] give prices charged to workers for basic food staples and other items. Presumably, these prices were lower than those charged in retail shops to the public at large.

Another source of data is the Ayuntamiento[2] de Barakaldo,[3] which compiled statistics of average annual retail prices for a variety of consumption items, stretching over the periods 1891–99 and 1906–27. In the second period, the statistics covers more items. The Ayuntamiento de Bilbao also published series of average annual retail prices for 1913–40.

The time series from the Hospital Civil de Basurto[4] come from the account books of that institution. These are prices that can be assumed below street prices, given the large quantities purchased by the hospital.

The Boletín Oficial de Vizcaya[5] also provides time series with average annual retail prices at two locations, Bilbao and nearby Valmaseda, for the period 1862–90.

[1]Altos Hornos de Vizcaya was established in 1902, the outcome of a merger of several preexisting iron and steel producing companies. It has since been the flagship of the steel industry in Biscaye, up until the end of the 20th century. The Cooperativa offered workers consumption goods at preferential prices. Prices predating 1902 come from similar cooperatives in existence at the merged companies.

[2]Local Council.

[3]Barakaldo is an industrial town, close to Bilbao. It hosts in particular a large part of the operations of Altos Hornos de Vizcaya.

[4]The Hospital Civil de Basurto opened its doors as a privately promoted charity in 1908. The archives contain information from the former Hospital de Bilbao (Achuri) prior to 1908.

[5]A publication of the Diputación de Vizcaya, the provincial authority. It carried statistical information among other things.

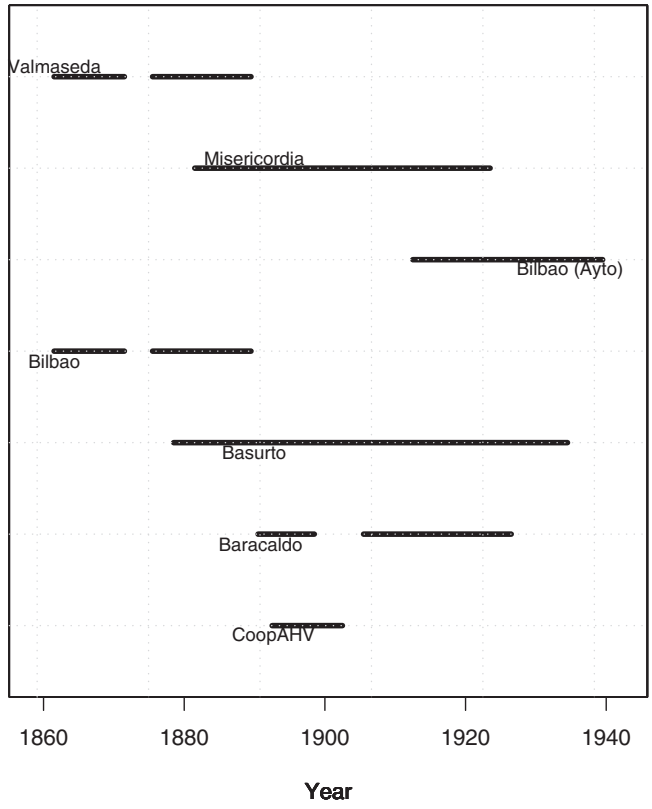


Figure 1. Time Span of the Time Series Provided by Each Source

Finally, the account books of the Santa Casa de Misericordia de Bilbao[6] also provide information on prices. As with the Hospital Civil de Basurto, these are neither retail nor wholesale prices, but can be assumed lower than retail prices for equal qualities.[7]

Pérez-Castroviejo and Martínez-Mardones (1996) provide some background for several of the sources mentioned. The paper also contains information on the evolution of consumption patterns in the period 1840–1940.

Some of the series have been discarded, either because the information they provide is extremely sparse or because they are manifestly incoherent with the rest of the prices for the same item. The number of series finally used is thus considerably reduced from that presented in Table 1. The upper panel of Figure 2 shows

[6]The Santa Casa de Misericordia de Bilbao, a charity, was established in its present location in 1872, but the institution was in existence before. It served first as orphanage, then as a house for elderly persons.

[7]Aside from their buying power, institutions such as the Hospital de Basurto and Santa Casa de Misericordia likely made their purchases through medium or long term contracts. As Ballesteros (1997) points out, such practice makes prices less responsive to short term fluctuations. Since data consist for the most part of average annual prices, we do not think this is a problem, as it would have been had we tried to compute, say, a monthly price index.

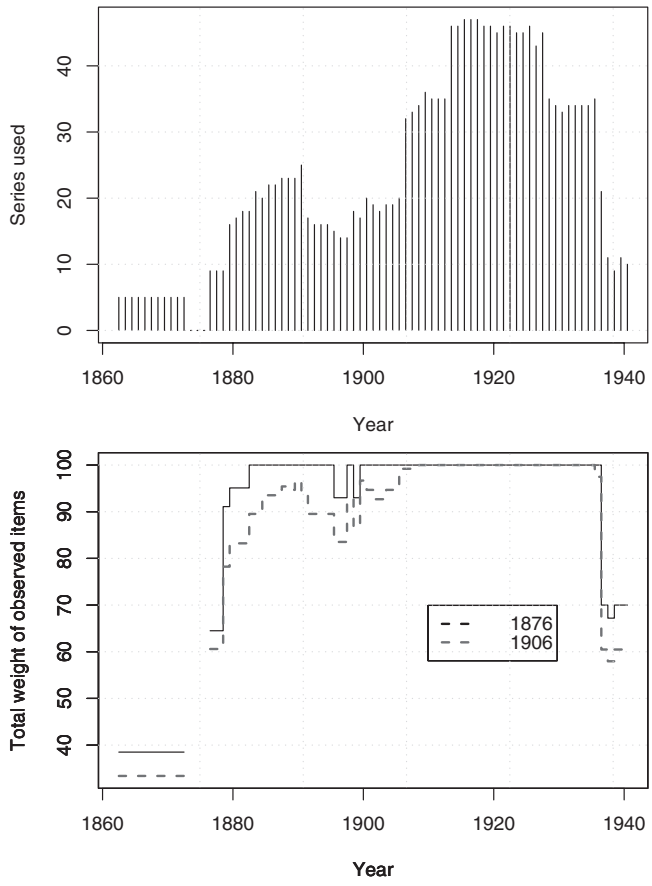


Figure 2. Number of Time Series Available by Year (Top Panel) and Total Percentage Weight of Items Observed from At Least One Source, for Each of the Two Weightings Used (Bottom Panel)

the number of series used for each year of the period 1862–1940. The lower panel shows the total weight over time of the items entering the indices which have been observed from at least one source. (We have computed two price indices, with weights aimed at approximating the consumption structure circa 1876 and 1906 respectively.)

It is apparent that statistical information is very sparse for the years from 1862 to 1872, and completely non-existent from 1873 to 1875 (Carlist war). On the other hand, from 1920 onwards virtually all prices have been observed from at least one source (and usually more than one).

It is worth pausing for a moment to consider the peculiarities of the time series just described, as they will condition our choice of method later. First, all of the time series extend over only part of the period 1862–1940, sometimes overlapping (except for 1873–1875, as mentioned).

Second, even for items nominally the same—beef, wine, coffee—different series provide in general different prices. On the one hand, we are using a mixture

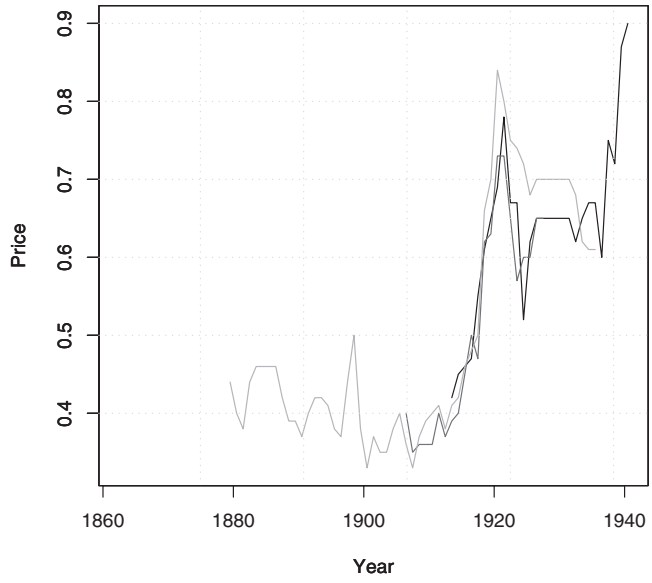


Figure 3. Wheat Bread Prices from Different Sources

of sources, giving retail, cooperative and institutional prices. On the other hand, it is possible (and likely) that there are differences in the qualities surveyed by the different sources. Sometimes the differences are apparent from the sources (e.g. "beef with bones" and "beef with no bones"); sometimes they can only be suspected.

Thus, about *the only thing* we can expect for time series referring to the same item but coming from different sources is that they display a similar profile, even if at a different level. For instance, if it were the case that a certain item was offered by the Cooperativa de Altos Hornos de Vizcaya to its workers at a price $p_t^{\mathrm{AHV}}$ about 15 percent less than retail prices $p_t$, we would expect $p_t^{\mathrm{AHV}} \approx 0.85\, p_t$ over the range where $p_t$ and $p_t^{\mathrm{AHV}}$ were both observed. As we shall see, the statistical model used for merging information from different sources allows (and automatically performs) the adjustments needed.

As an example, Figure 3 shows the price time series available for wheat bread. It can be seen that there is good agreement among the profiles of all series, even if some of them appear shifted with respect to the others.

A final comment refers to the fact that prices are either yearly averages, presumably computed from a sample, or prices obtained from the records of institutions which, however large, account for only a tiny fraction of the total volume negotiated in the market. In either case it makes sense to view those prices as composed of "signal plus noise." In other words, we take $p_{it} = s_{it} + \varepsilon_{it}$, where $p_{it}$, the observed price at time $t$ of item $i$, is the "true" average price $s_{it}$ (the "signal," which we might compute, at least in principle, had we exhaustive knowledge of all transactions carried in the market) plus a random variable $\varepsilon_{it}$ accounting for the observation error.

TABLE 2

Weights for the Cost of Living Indices in Biscaye. Two Different Sets of Weights are Given, for the Periods 1876–1905 and 1906–36

| | 1876–1905 Weight in | | 1906–1936 Weight in | |
|---|---|---|---|---|
| Group/Item | Index | Group | Index | Group |
| *Food* | 70.00 | 100.00 | 63.00 | 100.00 |
| Bread | 26.60 | 38.00 | 17.64 | 28.00 |
| Meat | 13.30 | 19.00 | 13.23 | 21.00 |
| Wine | 7.70 | 11.00 | 5.67 | 9.00 |
| Oil | 4.20 | 6.00 | 5.04 | 8.00 |
| Vegetables | 10.50 | 15.00 | 6.93 | 11.00 |
| Rice | 2.80 | 4.00 | 2.52 | 4.00 |
| Potatoes | 4.90 | 7.00 | 5.04 | 8.00 |
| Sugar | | | 1.26 | 2.00 |
| Fresh fish | | | 2.52 | 4.00 |
| Milk | | | 1.89 | 3.00 |
| Eggs | | | 1.26 | 2.00 |
| *Housing* | 13.00 | 100.00 | 14.00 | 100.00 |
| Rents | 13.00 | 100.00 | 14.00 | 100.00 |
| *Clothing* | 7.00 | 100.00 | 10.00 | 100.00 |
| Cotton cloth | 7.00 | 100.00 | 6.00 | 60.00 |
| Linen cloth | | | 2.00 | 20.00 |
| Wool cloth | | | 2.00 | 20.00 |
| *Toiletries* | 4.00 | 100.00 | 5.00 | 100.00 |
| Soap | 4.00 | 100.00 | 5.00 | 100.00 |
| *Fuel and energy* | 6.00 | 100.00 | 8.00 | 100.00 |
| Charcoal and firewood | 6.00 | 100.00 | 3.60 | 45.00 |
| Coal | | | 3.60 | 45.00 |
| Electricity | | | 0.80 | 10.00 |

*Source*: Table 1 in Pérez-Castroviejo (2006).

## 2.2. *Expenditure Structure*

The computation of indices such as those in expressions (1)–(2) requires the specification of a vector of quantities or coefficients $\boldsymbol{q}^t$ giving a measure of the relative importance of each item in family expenditure. Pérez-Castroviejo (2006) provides two different weightings, for the periods starting in 1876 and 1906. They have been obtained through consideration of over 20 dietaries of working class families and the budgets and accounts of the Santa Casa de Misericordia. The weights arrived at are shown in Table 2.

The reason we have opted for two different sets of weights is that, starting in the first decade of the 20th century, new items became commonplace in the budgets of working class families. Somewhat later, the same items could be observed in the documentation from the institutions, which, albeit with a lag, reproduced external expenditure patterns (Pérez-Castroviejo and Martínez-Mardones, 1996; Pérez-Castroviejo, 2006).

Food accounted for the largest share of expenditure in workers' families. The basic diet included only a few staples (bread, meat, vegetables, wine, rice, oil and potatoes). In the first decade of the 20th century, this diet was enlarged and diversified to include fresh fish, milk, eggs and sugar. Bread, meat and vegetables remained the basis of the diet in both periods, though. Expenditure on food

declined from about 70 percent in the last quarter of the 19th century to about 63 percent in the first third of the 20th century.

Even though food was all important, the indices constructed also weight expenditure in housing (rents), toiletries (soap) and energy (coal, charcoal and firewood in the first period, with electricity entering in the second period). Expenditure on these items increased seven percentage points from the first period to the second, to make up for the decline of the same magnitude in food expenditure.

## 3. The Statistical Model For Imputation

In principle, any model giving a good fit can be used for the imputation of missing data, with the requirement that it keeps the number of parameters to be estimated moderate relative to the number of effective observations. We have found particularly useful local level and local linear trend models (Harvey, 1989; Harvey *et al.*, 2004), which can be cast as state–space models. The estimation by maximum likelihood of such models is simplified by the use of the Kalman filter (Harvey, 1989; Durbin and Koopman, 2001). We describe in the following several alternative models which seem plausible, and the alterations we have made to have them suit our needs. While these models are adequate for the problem at hand, it should be remarked that further elaboration is possible. For instance, seasonal effects could be handled if we had monthly data. See Harvey and Chung (2000) for an application which demands a more elaborate model than ours.

### *Local Level Model*

Let $y_{it}$ be an observation at time $t$ of the price of item $i$, $1 \leq i \leq p$. We can consider such observation generated as:

$$y_{it} = \alpha_{it} + \varepsilon_{it} \tag{3}$$

(see last paragraph of Section 2.1), where $\alpha_{it}$ is the "state," in principle unobservable, of the average price of item $i$ at time $t$; $\varepsilon_{it}$ is the observation error. Stacking equations (3) for $1 \leq i \leq p$ in a single matrix expression we have the *observation equation*

$$\boldsymbol{y}_t = \boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t = \boldsymbol{Z}_t\boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t \tag{4}$$

with $\boldsymbol{Z}_t = \boldsymbol{I}$.

Let us assume $\varepsilon_t \sim N(\mathbf{0}, \boldsymbol{H}_t)$. We can choose the covariance matrix $\boldsymbol{H}_t$ so that the observation errors are correlated or uncorrelated. (We can also take $\boldsymbol{H}_t = \mathbf{0}$, if we consider that $y_{it}$ are observations of $\alpha_{it}$ with no error whatsoever.) In our application, it seems natural to consider observation errors uncorrelated and hence diagonal $\boldsymbol{H}_t$.

Our goal will be to obtain estimates $\tilde{\boldsymbol{\alpha}}_t$ of $\boldsymbol{\alpha}_t$, the *state vector*, for all $t$. One possibility is to assume the state vector evolves with a particular simple dynamics, given by the *state equation*

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t + \boldsymbol{\eta}_t; \tag{5}$$

$\boldsymbol{\eta}_t$ is a random vector driving the price vector from time $t$ to $t + 1$. We shall assume $\boldsymbol{\eta}_t \sim N(\mathbf{0}, \boldsymbol{Q}_t)$.

The choice of the covariance matrix $\boldsymbol{Q}_t$ controls, among other things, the degree of correlation between components of the state vector $\alpha_{it}$, $\alpha_{jt}$, $i \neq j$. We can opt for independent (diagonal $\boldsymbol{Q}_t$) or non-independent (general $\boldsymbol{Q}_t$) random walks. In our application, it seems natural to consider non-diagonal $\boldsymbol{Q}_t$: we expect average prices of the different items to evolve in a similar fashion, hence showing positive correlation.

The model made up of equations (4)–(5) with the assumptions mentioned is simple, flexible and intuitively appealing. In general, though, the covariance matrices $\boldsymbol{H}_t$ and $\boldsymbol{Q}_t$ will be unknown and hence require estimation; and this will almost invariably force some compromises.

First, we will consider only time invariant covariance matrices: $\boldsymbol{H}_t = \boldsymbol{H}$ and $\boldsymbol{Q}_t = \boldsymbol{Q}$ for all $t$, and, as stated previously, we will take $\boldsymbol{H}$ diagonal.

Second, even if we set $\boldsymbol{Q}_t = \boldsymbol{Q}$ invariant over time, the number of parameters to estimate may be quite substantial. A simplifying assumption which drastically reduces the number of parameters is to prescribe identical correlation between components $\alpha_{it}$, $\alpha_{jt}$, for all $i, j$, $i \neq j$. For a state vector $\boldsymbol{\alpha}_t$ of dimension $p$ this takes down the number of parameters in $\boldsymbol{Q}$ from $p(p + 1)/2$ to $p + 1$. Assuming identical correlations is consistent with the idea that prices move, on average, in the same direction.[8] We will call the local level model with this assumption the *equicorrelated local level model.*

If there were prices evolving in a markedly different way, we could consider equicorrelation by groups of products, with a less dramatic, but still substantial, drop in the number of parameters to be estimated. Matrix $\boldsymbol{Q}_t = \boldsymbol{Q}$ would then have $p$ variances plus one correlation to estimate for each homogeneously correlated set of prices; we will refer to this model as the *group equicorrelated local level model.* This is pursued in Section 4.

*Local Linear Trend Model*

We can choose a different dynamics instead of the simple random walk of equation (5). We might consider replacing (5) with

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t + \boldsymbol{\beta}_t + \boldsymbol{\eta}_t \tag{6}$$

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + \boldsymbol{\delta}_t. \tag{7}$$

If $\boldsymbol{\beta}_t$ were fixed, we would have a random walk with drift for $\boldsymbol{\alpha}_t$. If, however, $\boldsymbol{\beta}_t$ evolves according to a random walk, we have a local linear trend; $\boldsymbol{\beta}_t$ plays the role of a local slope at time $t$. Now equation (4) is replaced with

[8]Aside from being a reasonable assumption, it makes it easy to check that $\boldsymbol{Q}$ is positive semidefinite (Abadir and Magnus, 2005, Problem 8.74). This is helpful when we use an iterative algorithm to estimate $\boldsymbol{Q}$ and want to check at the end of each iteration that we are within the feasible set.

$$\boldsymbol{y}_t = [\boldsymbol{I} \quad \boldsymbol{0}]\begin{bmatrix}\boldsymbol{\alpha}_t \\ \boldsymbol{\beta}_t\end{bmatrix} + \boldsymbol{\varepsilon}_t = \boldsymbol{Z}_t^*\boldsymbol{\alpha}_t^* + \boldsymbol{\varepsilon}_t, \tag{8}$$

where in the new vector state $\boldsymbol{\alpha}_t^*$ we have stacked $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$. The state evolves now as

$$\boldsymbol{\alpha}_{t+1}^* = \boldsymbol{T}_t\boldsymbol{\alpha}_t^* + \boldsymbol{\nu}_t \tag{9}$$

with $\boldsymbol{T}_t$ an upper block triangular matrix made of unit matrices $\boldsymbol{I}_p$ and $\boldsymbol{\nu}_t^T = \left[\boldsymbol{\beta}_t^T \boldsymbol{\delta}_t^T\right]^T$ (see details in Durbin and Koopman (2001) or Harvey (1989)). We shall assume $\boldsymbol{\nu}_t \sim N(\boldsymbol{0}, \boldsymbol{N}_t)$, with covariance matrix $\boldsymbol{N}_t$ that we can (and will) restrict conveniently, to prevent proliferation of parameters.

*Multiple Sources and Repeated Observations*

A common problem, conspicuously present in what follows, is the availability of price observations coming from different sources for the same item $i$ and time $t$.

If observations were entirely homogeneous, there would be no problem: we could take the average of all observed prices for each item and point of time. However, as discussed in Section 2.1, this is not the case here. Prices observed correspond to (possibly) different qualities and different contractual conditions: institutional prices akin to wholesale prices, retail prices, cooperative prices, somewhere in between the two previous categories, etc.

Thus, it is not practical or even possible, unless we have very detailed information of the different qualities, markups, etc, to try to manually synthesize all information from different sources to come up with a single price time series for each item. If, however, we can make the already mentioned assumption of proportionality of prices coming from different sources, a simple alteration in our model will let us use all available information, aligning redundant time series as part of the estimation process.

Consider, for instance, two different sources, $A$ and $B$, both giving prices for item $i$ at time $t$. Assume that source $A$ provides retail prices $y_{it}^A$ for the finest quality of item $i$, while $B$ provides wholesale prices $y_{it}^B$ for a lesser quality. We can imagine $y_{it}^A$ and $y_{it}^B$ being generated as follows:

$$y_{it}^A = 1 \cdot \alpha_{it} + \varepsilon_{it}^A$$

$$y_{it}^B = \delta_{iB} \cdot \alpha_{it} + \varepsilon_{it}^B$$

with $\delta_{iB}$ a coefficient to be estimated. (In a situation such as the one described, we would expect $\delta_{iB} < 1$.) In the model made up of equations (4)–(5) or (8)–(9) we only have to take an appropriate $\boldsymbol{Z}_t$. The coefficient $\delta_{iB}$ can be estimated with the rest of the parameters in the model.

The procedure sketched nicely generalizes some *ad hoc* methods used in price indices, when there are changes in the quality of the items surveyed (see OECD (1984, p. 16) for a description of common practices).

Various other models suggest themselves. One referee proposed a local linear trend model with prices in the log scale as a plausible *a priori* alternative, and this has also been considered in the following. When modeling the log of prices, however, the observation equation has to be modified somewhat: rather than $\boldsymbol{y}_t = \boldsymbol{Z}_t\boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t$ with adjusting parameters in $\boldsymbol{Z}_t$ multiplying elements of the state vector, *additive* adjusting parameters are required. Thus, the state space model becomes

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{T}_t\boldsymbol{\alpha}_t + \boldsymbol{\eta}_t \tag{10}$$

$$\boldsymbol{y}_t = \boldsymbol{Z}_t\boldsymbol{\alpha}_t + \boldsymbol{G}_t\boldsymbol{u}_t + \boldsymbol{\varepsilon}_t, \tag{11}$$

where entry $g_{ij}$ in matrix $\boldsymbol{G}_t$ is the additive parameter required to align the $i$-th component of $\boldsymbol{y}_t$. Matrix $\boldsymbol{Z}_t$ has a 1 in position $(i, j)$ if element $i$ of $\boldsymbol{y}_t$ is the price of one of the variants of item $j$, and $\boldsymbol{u}_t$ is a vector of dummy variables (Harvey (1989, § 8.6) offers examples of the multivariate state space model with explanatory variables).

In the model given by (10)–(11) or any of the variants previously discussed, it is assumed that time series giving prices for the same item follow trajectories that are approximately proportional to each other (or parallel, when working with log prices). The observation matrix $\boldsymbol{Z}_t$ (or $\boldsymbol{G}_t$) accounts for the differences in level, as discussed previously. It is also assumed that $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\eta}_t$ are zero mean uncorrelated Gaussian sequences (hence, $\boldsymbol{\alpha}_t$ are $\boldsymbol{y}_t$ are also Gaussian). In the absence of Gaussianity, the algorithms used (Kalman filter and smoother) still give estimates of the state vector which are optimal in the least squares sense among all linear estimators.

The model with equicorrelated matrix $\boldsymbol{Q}$ may seem unduly restrictive. While we feel that a common trend in prices is the overwhelming effect to account for, there is room for some more structure. A general $\boldsymbol{Q}$ is out of the question, as its specification would require far more parameters than the length of the series allows; but the hypothesis of groups of products displaying higher correlation among themselves than with the rest can be entertained.

Products can be grouped on a priori grounds (foods, energy, clothing . . . ) or on the basis of empirical correlations. We have followed the last route, fitting the simplest local level model, and computing estimates of the state disturbances $\hat{\boldsymbol{\eta}}_t = \hat{\boldsymbol{\alpha}}_{t+1} - \hat{\boldsymbol{\alpha}}_t$. The distance between items $i, j$ is then defined as $d_{ij} = 1 - \hat{\rho}_{ij}$, where $\hat{\rho}_{ij}$ is the estimated correlation between $\hat{\eta}_{it}$ and $\hat{\eta}_{jt}$. Then, an agglomerative cluster algorithm with complete linkage (see, for instance, Kaufman and Rousseeuw, 1990) is run to produce the dendrogram in Figure 4.

There seems to be some structure: most foodstuffs appear clustered together, with the exception of rice. Some non-food items like housing rents, cotton cloth and charcoal appear in a cluster by themselves. As a compromise, partly accounting for the correlation structure but keeping the number of new parameters to a minimum, we have taken three broad clusters, with foodstuffs and some non-food items in one cluster, rice and electricity in the second, and housing rents, cotton cloth and charcoal in the third.

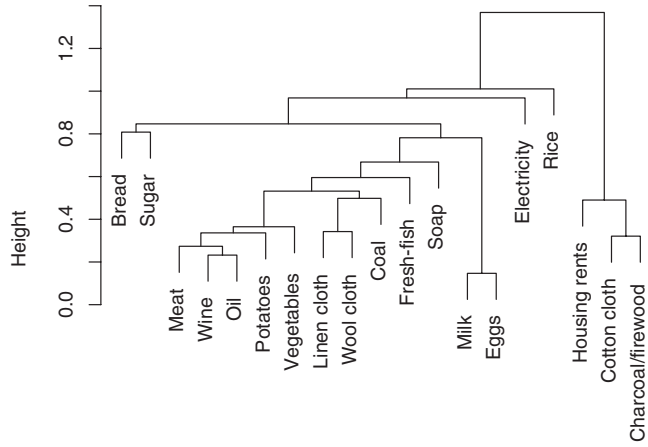


Figure 4. Cluster of Items. Distance Between Items *i, j* is Based on the Correlation of the Disturbances Driving The State Vector. The Clustering Method is Complete Linkage

## 4. Results

After discarding series too short to justify inclusion or clearly of very poor quality, we were left with 52 series extending over (part of) 79 years. If all time series were fully observed, there would be 4,108 observations; instead, we only have 1,849 non-missing observations, i.e. 45.01 percent of the total.

The number of items considered is 19 (see Table 2). This requires a state dimension of 19 for a local level model, or 38 for a local linear trend model, and that many variances in $\boldsymbol{Q}$ need to be estimated. Equicorrelation of group equicorrelation with three groups adds (respectively) one or three correlations to be estimated. In all cases, we have 52 variances to be estimated in $\boldsymbol{H}$. To this, we have to add the adjusting parameters in $\boldsymbol{Z}$ (or $\boldsymbol{G}$, if the model is specified in the log scale and the adjustment of varieties of the same item is through additive constants for each variety: these total $52 - 19 = 33$ parameters). Thus, for instance, we have for the local level model with simple equicorrelation a total of $19 + 1 + 52 + 33 = 105$ parameters. Group equicorrelated models use an extra two correlations. Local linear trend models use a state vector twice as large, hence 38 variances rather than 19 in $\boldsymbol{Q}$. Other than that, two correlations are used (one for the "level" components of $\boldsymbol{\alpha}_t$, another one for the slope components) instead of one, hence an extra 20 parameters over the local level counterparts.

We fitted the models summarized in Table 3. All computations were programmed in R (described in R Development Core Team, 2007). Code is available from the second author.

In Table 3, the third column gives the value of the maximized likelihood. Although the numeric values are similar, they are only comparable among models specifying the response in the same scale (original or in logs). The fourth column $p$ gives the number of parameters as described above; we might call these hyperparameters. The fifth column, "edf," gives the "equivalent degrees of freedom" as described in the following.

TABLE 3
SUMMARY OF MODELS FITTED

| Model | Response | $\log f(x, \hat{\theta})$ | $p$ | edf | AICc |
|---|---|---|---|---|---|
| Local level<br>$\boldsymbol{Q}_t$ = equicorrelated,<br>$\boldsymbol{H}_t$ = diagonal | $y_t$ | 1,442.08 | 105 | 525 | –1,416.55 |
| Local level<br>$\boldsymbol{Q}_t$ = group equicorrelated,<br>$\boldsymbol{H}_t$ = diagonal | $y_t$ | 1,404.32 | 107 | 671.1 | –699.74 |
| Local level<br>$\boldsymbol{Q}_t$ = equicorrelated,<br>$\boldsymbol{H}_t$ = diagonal | $\log(y_t)$ | 1,543.04 | 105 | 738.3 | –625.49 |
| Local level<br>$\boldsymbol{Q}_t$ = group equicorrelated,<br>$\boldsymbol{H}_t$ = diagonal | $\log(y_t)$ | 1,651.47 | 107 | 691.8 | –1,090.33 |
| Local linear trend<br>$\boldsymbol{Q}_t$ = equicorrelated,<br>$\boldsymbol{H}_t$ = diagonal | $y_t$ | 1,576.09 | 125 | 627.9 | –1,249.22 |
| Local linear trend<br>$\boldsymbol{Q}_t$ = equicorrelated,<br>$\boldsymbol{H}_t$ = diagonal | $\log(y_t)$ | 1,763.22 | 125 | 811.1 | –633.75 |

Consider the simplest possible univariate state space model,

$$y_t = \alpha_t + \varepsilon_t \tag{12}$$

$$\alpha_{t+1} = \alpha_t + \eta_t. \tag{13}$$

If $\sigma_\eta = 0$ and $\sigma_\varepsilon > 0$, clearly $\alpha_1 = \alpha_2 = \ldots = \alpha_T = \alpha$ say, and the best fit of $y_t$ for $t = 1, \ldots, T$ is $\hat{\alpha}$: we are using a model which effectively reduces to fitting a single parameter to all $y_t$ and thus consumes a single degree of freedom. On the other hand, if $\sigma_\eta$ is large compared to $\sigma_\varepsilon$, $\alpha_t$ can closely track $y_t$. In the limit, when $\sigma_\eta/\sigma_\varepsilon \to \infty$, we would be effectively fitting one $\hat{\alpha}_t$ to each $y_t$ for a total of $T$ effective degrees of freedom used (and none left).

Similarly, for a general state-space model, stacking $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$ on a vector $\boldsymbol{y}$, doing likewise with $\tilde{\boldsymbol{\alpha}}_1, \ldots, \tilde{\boldsymbol{\alpha}}_T$, and making $\boldsymbol{Z}$ a block-diagonal matrix with blocks $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T$, it can be shown (cf. Durbin and Koopman, 2001, § 4.6) that,

$$\tilde{\boldsymbol{y}} = \boldsymbol{Z}\tilde{\boldsymbol{\alpha}} = \boldsymbol{S}\boldsymbol{y} \tag{14}$$

for a certain smoother matrix $\boldsymbol{S}$ which is dependent on the covariance matrices $\boldsymbol{Q}$ and $\boldsymbol{H}$. If $\boldsymbol{S}$ were a unit matrix, we would have a perfect fit—and no degrees of freedom left. We can take trace ($\boldsymbol{S}$) as a reasonable measure of the "equivalent degrees of freedom" used by the state-space model (for the rationale of that, see Hodges and Sargent (2001); see also Hastie and Tibshirani (1991 § 3.5).

If the observations are of dimension $k$, $\boldsymbol{S}$ is a $kT \times kT$ matrix whose direct computation involves an inverse of possibly very large size. Fortunately, only the trace is needed, and it can be shown that the diagonal blocks of $\boldsymbol{S}$ can be obtained

as a by-product of running the Kalman filter and smoother (Fahrmeir and Wagenpfeil, 1997). In our case, the diagonal blocks of $\boldsymbol{S}$ are given by $\boldsymbol{Z}_t\boldsymbol{V}_t\boldsymbol{Z}_t^T\boldsymbol{H}_t$, where $\boldsymbol{V}_t$ is the covariance matrix of the smoothed state given by the Kalman smoother.

The fifth column of Table 3 gives the equivalent degrees of freedom used by the different models computed as

$$\text{trace}(\boldsymbol{S}) = \sum_{t=1}^{T} \text{trace}\left(\boldsymbol{Z}_t\boldsymbol{V}_t\boldsymbol{Z}_t^T\boldsymbol{H}_t\right).$$

As can be seen, models are very heavily parameterized, using between 525.0 and 811.1 equivalent degrees of freedom.

From the value of maximized likelihood and the equivalent degrees of freedom used (or "equivalent parameters"), we could compute the AIC criterion (Akaike, 1972), to help choose a model (at least, among those with the response in the same scale). AIC is known to perform poorly with heavily parameterized models (which is the case here). Hence, we have turned to a similar criterion, AICc (Hurvich and Tsai, 1989); see also Bengtsson and Cavanaugh (2006). This is the value in the last column of Table 3.

We settled for the local level equicorrelated model with $\boldsymbol{y}_t$ as response. It looks clearly best in terms of AICc among models having $\boldsymbol{y}_t$ as response. Regarding the models fitted to $\log(\boldsymbol{y}_t)$, they all use a larger number of equivalent parameters, which is uncomfortable since the effective size of the sample is relatively small. Aside from that, while having some advantages, use of the log scale also has inconveniences. One is that the smoothed state vectors $\tilde{\boldsymbol{\alpha}}_t$ give estimates of the underlying *log prices* and their variances. Given a set of weights $w_i$, there is no problem in computing the index as:

$$I_t = \frac{\sum_{i=1}^{p} w_i \exp(\tilde{\alpha}_{it})}{\sum_{i=1}^{p} w_i \exp(\tilde{\alpha}_{i0})}.$$

(The expression above would introduce a small bias both in numerator and denominator, due to Jensen's inequality, of no consequence whatsoever.) However, we have the covariance matrices for $\tilde{\boldsymbol{\alpha}}_t$, not for $\exp(\tilde{\boldsymbol{\alpha}}_t)$. Thus we are forced to some manipulation to come up with approximate variances and confidence intervals for the index itself.

After the parameters in the selected local level model are estimated, a smoothing algorithm (Durbin and Koopman, 2001, § 4.3) gives an estimate of the trajectory of the state vector and the covariance matrices at each point in time.

The components of the state vector model the "underlying prices" of the items in the indices. For each item $i$, the estimated "underlying price" at time $t$—the $i$-th component of $\tilde{\boldsymbol{\alpha}}_t$—takes into account the observed price(s) at time $t$ of item $i$, if any, the observed prices at time $t$ of all other items and the full set of prices observed at times other than $t$: $\tilde{\boldsymbol{\alpha}}_t$ is the best estimate (in the least squares sense) of $\boldsymbol{\alpha}_t$ given all information (and best linear if the hypothesis of Gaussianity is dropped).

Thus, we have estimations for the full set of prices at each moment, which make use of all available information. Even at times when no direct observations

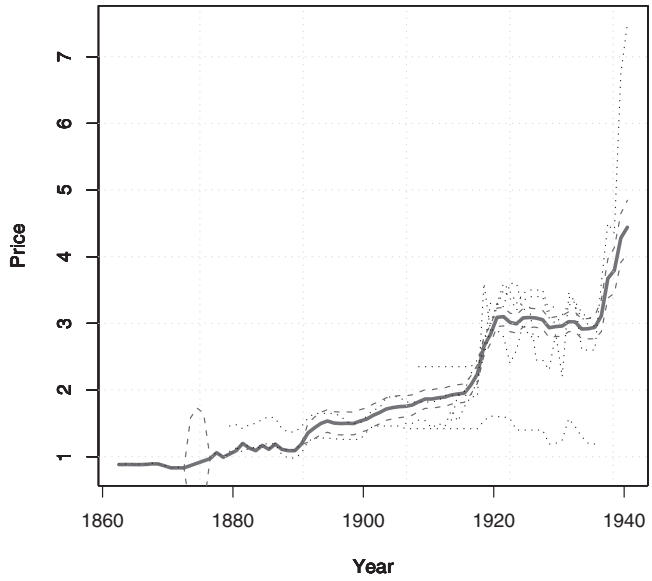


Figure 5. Beef Prices from Different Sources (dotted lines), State Estimation or "Underlying Price" (thick line) and 95% Confidence Interval for the State (dashed lines)

of the price of an item are available, we have an estimation of the underlying price on which to base the computation of indices.

As an illustration, Figure 5 shows the different time series of prices available for beef, the state or "underlying price" and the 95 percent confidence interval for the same. It can be seen that the width of the confidence interval is quite variable, reflecting the lack of data for some periods (like 1873–75) and the varying volatility over time of the price series available.

The prices of items thus estimated can be scaled to value 100 at the base year (which we have chosen to be 1913) and multiplied by the coefficients in the chosen weighting scheme to compute a price index. The confidence interval of the index is derived in the obvious manner.

As mentioned previously, we have used two different sets of weights, to reflect the changing expenditure patterns over time. The results using the two sets of weights are difficult to tell apart, and can be seen in the two panels of Figure 6. Their general appearance is similar to results obtained for the whole of Spain by other researchers: see Maluquer (2005, figure 16.5, p. 1267) or Prados de la Escosura (2003), for instance. What the method used adds is an explicit indication of the uncertainty associated with the index at each point in time, supplying a confidence interval.[9]

Intervals in Figure 6 show that there is little that can be confidently said prior to 1900: the fluctuations observed are non-significant in view of the width

[9]This confidence interval is optimistic in that it ignores the fact that the parameters in the model generating estimates of $\alpha_t$ have themselves been the object of estimation, which adds (unaccounted) uncertainty. Work in progress seeks to account also for this uncertainty by using multiple imputation (see Little and Rubin, 2002).

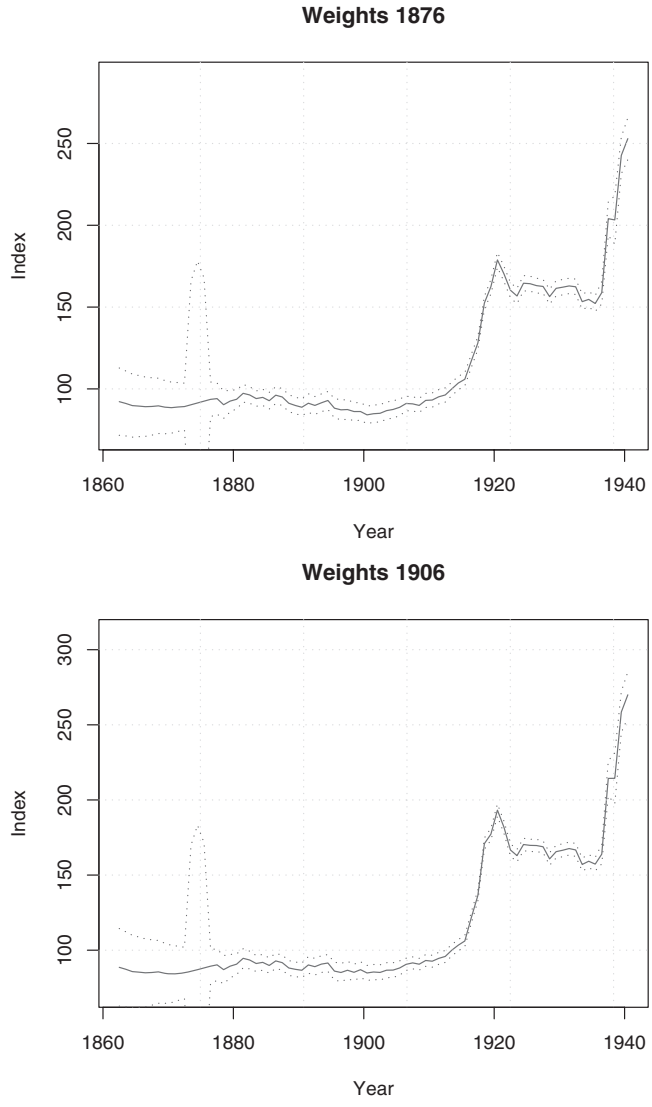


Figure 6. Cost of Living Indices for Biscaye with Two Different Weightings, Reflecting Consumption Patters in 1876 and 1906. Base 1913 = 100 for both indices

of the intervals. A constant level of prices is essentially compatible with the estimated indices. From 1914 onwards, both indices display unmistakably the phenomenal inflation brought about by World War I, nearly doubling prices, then the smooth decline of prices which extended for the Depression period following 1929; a decline, however, which failed to return prices to their pre-war level. The Spanish Civil War 1936–39 ignited inflation again, roughly doubling prices once more.

## 5. Discussion

We have described a technique for the construction of indices from time series allowing both overlapping observations and missing data. The technique resorts to state space modelling and is demonstrated by constructing price indices for Biscaye, with two different weightings.

Although motivated by the problem of constructing price indices, the technique is much more general, and apt to find use in many other settings: whenever we find any combination of overlapping sources that we need to reconcile and/or missing data.

Antecedents to our work can be found in the literature (see, for instance, Harvey and Pierse, 1984; Dagum *et al.*, 1998; Harvey and Chung, 2000; Feder, 2001, and references therein). We have presented a novel use of existing tools, exploiting the flexibility of state space models to provide an approach to index construction. A benefit of our approach is that it provides an explicit indication of the uncertainty in the index values. In particular, for the case analyzed, it shows that the fluctuations in our price indices until the end of the 19th century are rather small relative to their variance, and hence do not support any elaborate interpretation.

There is an issue that needs further discussion: the nature of missing observations. We made the assumption that they were missing completely at random (MCAR). This implies that the fact that an observation is missing is entirely unrelated to the value we might have observed, or any others: it is not the case, for instance, that high values are more likely to be missing than low values.

This is a hypothesis that may be suspect in some cases. For instance, there is a total absence of data for the period 1873–75, coincident with the last Carlist war. Our model deals with this lack of data, doing what is in essence an interpolation. The implicit assumption is that prices in that period were generated by the same mechanism than prices before or after. However, the analyst must be aware that the lack of data for those years is not merely incidental: it is no doubt related to the war which, we might hypothesize, brought about scarcity, hardship and higher prices. Therefore, data absence here probably correlates with the values we would have observed, invalidating the MCAR assumption (and even the weaker MAR assumption).

Nevertheless, we are using a statistical model which makes up for missing data exploiting observed regularities: contemporaneous correlations between prices of different items and correlations between prices of the same and different items over time. The model, though, cannot detect, nor account for, exceptional periods where these regularities break.

There is not much we can do about this. But the analyst should be fully aware of both the power and limitations of the tool used. In the case of our indices, we feel we should at least warn the reader that the values for the years 1873–75 should be read with circumspection.

It has been pointed out to the authors that a model containing fewer unobservable components might be easier to estimate. A plausible hypothesis would be that prices are cointegrated (Engle and Granger, 1987; Harvey, 1989). At the very

least one should be able to achieve a considerable reduction in the dimension of the problem, fitting only a few unobservable components.

This is no doubt true. However, we want to compute an index *for a specific set of weights*. This requires the estimation of a full set of prices to multiply by those weights. If we fit a reduced rank model we could end up with estimated trajectories for, say, three unobservable components which are not in clear connection with items or groups of items. There is no obvious way to weight those unobservable components. Thus, we are trading simplicity in exchange for interpretability and usability of our index for purposes like deflating nominal wages.

## References


Abadir, K. M. and J. R. Magnus, *Matrix Algebra*, Cambridge University Press, 2005.

Akaike, H., "Use of An Information Theoretic Quantity for Statistical Model Identification," in *Proceeding of the 5th Hawaii International Conference on System Sciences*, 249–50, 1972.

Ballesteros, E., "Una estimación del coste de la vida en españa, 1861–1936," *Revista de Historia Económica*, XV(2), 363–95, 1997.

Bengtsson, T. and J. E. Cavanaugh, "An Improved Akaike Information Criterion for State-Space Model Selection," *Computational Statistics and Data Analysis*, 50, 2635–54, 2006.

Dagum, E. B., P. A. Cholette, and Zhao-Guo Chen, "A Unified View of Signal Extraction, Benchmarking, Interpolation and Extrapolation of Time Series," *International Statistical Review*, 66(3), 245–69, 1998.

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

Engle, R. F. and C. W. J. Granger, "Co-integration and Error Correction: Representation, Estimation and Testing," *Econometrica*, 55, 251–76, 1987.

Fahrmeir, L. and S. Wagenpfeil, "Penalized Likelihood Estimation and Iterative Kalman Smoothing for Non–Gaussian Dynamic Regression Models," *Computational Statistics and Data Analysis*, 24, 295–320, 1997.

Feder, M., "Time Series Analysis of Repeated Surveys: The State-Space Approach," *Stat. Neerlandica*, 55(2), 182–99, 2001.

Feinstein, C. H., "Pessimism Perpetuated: Real Wages and the Standard of Living in Britain during and after the Industrial Revolution," *Journal of Economic History*, 58, 625–58, 1998.

Harvey, A. C., *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, 1989.

Harvey, A. and Chia-Hiu Chung, "Estimating the Underlying Change in Unemployment in the U. K.," *Journal of the Royal Statistical Society, Series A*, 163(3), 303–39, 2000.

Harvey, A. C. and R. G. Pierse, "Estimating Missing Observations in Economic Time Series," *Journal of the American Statistical Association*, 79(385), 125–31, 1984.

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

Hastie, T. J. and R. J. Tibshirani, *Generalized Additive Models*, 2nd edition, Chapman & Hall, London, 1991.

Hodges, J. S. and D. J. Sargent, "Counting Degrees of Freedom in Hierarchical and Other Richly-Parameterised Models," *Biometrika*, 88(2), 367–79, 2001.

Hurvich, C. and C. Tsai, "Regression and Time Series Model Selection in Small Samples," *Biometrika*, 76, 297–307, 1989.

Kaufman, L. and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, 1990.

Little, R. J. A. and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd edition, Wiley, 2002.

Maluquer, J., "Consumo y precios," in A. Carreras and X. Tafunell (eds), *Estadísticas Históricas de España. Siglos XIX y XX*, Vol. III, Fundación BBVA, Madrid, 1249–73, 2005.

OECD, *Consumer Price Indices: Sources and Methods and Historical Statistics 1960–1983*, OECD, Department of Economics and Statistics, 1984.

Prados de la Escosura, L., *El progreso económico de España (1850–2000)*, Fundación BBVA, Madrid, 2003.

Pérez Castroviejo, P. M., "Poder adquisitivo y calidad de vida de los trabajadores vizcainos, 1876–1936," *Revista de Historia Industrial*, XV(30), 103–41, 2006.

Pérez-Castroviejo, P. M. and I. Martínez-Mardones, *La alimentación de los pobres. Estrategias de gasto alimentario y la dieta en la Santa Casa de Misericordia de Bilbao, 1840–1940*, Ayuntamiento de Bilbao, Bilbao, 1996.
R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2007, http://www.R-project.org.
Reher, D. and E. Ballesteros, "Precios y salarios en Castilla la Nueva: La construcción de un índice de salarios reales, 1501–1991," *Revista de Historia Económica*, XI(1), 101–51, 1993.
Vogt, A. and J. Barta, *The Making of Tests for Index Numbers*, Physica-Verlag, 1996.