Forecasting composite indicators with anticipated information: an application to the industrial production index

Francesco Battaglia

University La Sapienza, Rome, Italy.

Livio Fenga

Istituto Nazionale di Statistica, Rome, Italy.

Summary. Many economic and social phenomena are measured by composite indicators computed as weighted averages of a set of elementary time series. Often data are collected by means of large sample surveys, and processing takes a long time, whereas the values of some elementary component series may be available some time before the others, and may be used for forecasting the composite index. This problem is addressed within the framework of prediction theory for stochastic processes. A method is proposed for exploiting anticipated information in order to minimise the mean square forecast error, and for selecting the most useful elementary series. An application to the Italian general industrial production index is illustrated, which demonstrates that knowledge of anticipated values of some, or even just one, component series may reduce the forecast error considerably.

Keywords: Forecasting; Industrial Production Index; Leading indicators; Multivariate Autoregressive Models

1. Introduction

Many phenomena in economic and social sciences are measured by composite indicators obtained as weighted averages of a set of univariate time series, for example prices or production indexes, or fertility rates. In most cases, the data come from large sample surveys and the recording, controlling and organising process takes a long time. Often, provisional values are published, and later revised. It is not unusual, however, specially when the set of component series is large, and each relates to different areas, that the values of a few of them may be available some time before the others. Thus one can attempt to forecast the composite index exploiting, in addition to the previous values of the entire set, the additional information given by the current value of some components.

This kind of problem is often addressed in the framework of leading indicators (Lahiri and Moore, 1991), in the disaggregation of econometric models (Barker and Pesaran, 1990), or in multivariate methods for time series — e.g. principal components (Brillinger (1981), Ch. 9), or canonical analysis (Box and Tiao, 1977).

The present paper deals with the proposed problem in the framework of the prediction theory of stochastic processes (e.g. Priestley (1981), Ch. 10), and tries to develop the best linear predictor (in a mean square forecast error sense) based on the entire set of available information, and to address essentially two questions:

- What is the best way of utilising the additional information to forecast the indicator?

- What components ensure the best improvement if known in advance?

We shall consider here the industrial production index data. They are collected by means of a monthly sample survey, involving more than 8,000 companies operating throughout Italy and producing goods which are organised into 592 categories according to the Commission of the European Communities classification of economic activities (NACE Rev. 1). Upper level classifications include classes (with four-digit codes, an example is DJ 2751, casting of iron); groups (with three-digit codes, for example DJ 275, casting of metals) and finally 16 main branches (subsections, one or two letter codes). A weighted average of such 16 branches, whose weights are determined according to their relative production values, yields the general industrial production index. Table 1 provides a description of each component and their weights.

The data production process is rather complex and requires efficient coordination among various local and central statistical offices, while the timeliness of the publication of the official data, mainly for the general index, is critical; also, the result is released on the same date each month. Though revisions are usually published (one and two months later), the general production index figures obviously have considerable impact on economic operators, therefore, timely precision is essential.

To this aim, the process may be organized in such a way that information on the most important branches is retrieved first, so that possible gaps or mistakes in the last observed components have a smaller influence on the general index. This also provides an efficient sequence of early index estimates.

We shall analyse, for the purpose of forecasting the general index, the use of a priori observed components, both at the top classification level (the 16 branches in Table 1) and at a lower hierarchy (three-digit and four-digit components).

Industrial production is a very important indicator of the business cycle, and its prediction is crucial, so it has attracted much attention in statistical literature. In addition to the more usual ARMA framework, many univariate methods have been proposed, including non linear models (Byers and Peel, 1995), and structural models (Thury, 1998). Multivariate techniques have also been adopted, mainly using relationships between the industrial production and different types of information: among others, survey data (Rahiala and Teräsvirta, 1993), energy consumption (Bodo and Signorini, 1987), and their combination (Marchetti and Parigi, 2000). Furthermore, the preliminary values of the index itself have been employed (Boucelham and Teräsvirta, 1990). A thorough discussion of the features of preliminary values and revisions has been recently proposed by Patterson (2002).

A forecasting approach, similar to that employed in the present paper, has recently been introduced for deriving optimal aggregate linear and non linear models (see van Garderen et al. (2000) and references therein). Coccia and Iafolla (2000) address the problem of anticipated estimates using a combined strategy based on (static) principal components and Braun operators (Braun, 1973) to obtain a synthetic indicator whose values are used as additional information.

In the following Section we present the proposed method and discuss the choice of the most important univariate components. The resulting forecasts may be computed by fitting a multivariate linear model. In Section 3 we discuss the frequently occurring case of an existing predictor with a fixed functional form which is improved by adding a linear combination of the additional observations. Some possible choices are discussed, including forecasting the univariate composite index series, or fitting univariate ARMA models to each component series. In Section 4 we illustrate the results of our application to the industrial

Table 1. Components of the industrial production index and their weights.

Component	Description	weight β_i
C	Mining and Quarrying	0.019
DA	Food, Beverages and Tobacco	0.086
DB	Textiles and Textile Products	0.098
DC	Leather and Leather Products	0.028
DD	Wood and Wood Products	0.021
DE	Pulp, Paper and Paper Products	0.056
DF	Coke, Refined Petroleum and Nuclear Fuel	0.024
DG	Chemicals, Chemical Products and Man-Made Fibres	0.070
DH	Rubber and Plastic Products	0.039
DI	Other Non-Metallic Mineral Products	0.052
DJ	Basic Metals and Fabricated Metals Products	0.131
DK	Machinery and Equipment	0.103
DL	Electrical and Optical Equipment	0.083
DM	Transport Equipment	0.054
DN	Manufacturing, not elsewhere classified	0.038
E	Electricity, Gas and Water Supply	0.098

production index. Some conclusions are drawn in the last Section.

2. The best linear forecast based on additional information

We formalise the problem as follows. Let $X(t) = [X_1(t), X_2(t), \ldots, X_m(t)]'$, for integer values of t, be a multivariate second-order stationary process with zero means, autocovariance matrices $\Gamma(h)$, and consider the univariate process $Y(t)$ defined by:

$$Y(t) = \sum_{j=1}^{m} \beta_j X_j(t) \tag{1}$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_m)'$ is a vector of known positive constants.

Denote by $I_t = \{X_j(s), s \leq t, j = 1, 2, \ldots, m\}$ the whole information at time t, and the best predictor in mean square error of $Y(t+1)$ by

$$Y_t(1) = E\{Y(t+1)|I_t\}. \tag{2}$$

We suppose that some components of $X(t)$ may be observed at time $t+1$, and use them to forecast $Y(t+1)$. We denote by O the set of indices j such that $X_j(t+1)$ is observed, and by U the set of indices relating to the unobserved components, so that $O \cap U = \emptyset, O \cup U = \{1, 2, \ldots, m\}$. Accordingly, all vectors and matrices will be partitioned with a similar notation:

$$X(t) = \begin{bmatrix} X_O(t) \\ X_U(t) \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_O(t) \\ \beta_U(t) \end{bmatrix} \qquad \Gamma(h) = \begin{bmatrix} \Gamma_{OO}(h) & \Gamma_{OU}(h) \\ \Gamma_{UO}(h) & \Gamma_{UU}(h) \end{bmatrix}.$$

Given the additional information $X_O(t+1)$, the best predictor of $Y(t+1)$ may be written:

$$Y_t^*(1) = E\{Y_{t+1}|I_t, X_O(t+1)\}. \tag{3}$$

Assuming that $X(t)$ is multivariate gaussian, conditional expectations are linear. Denoting by $X_{j,t}(1) = E\{X_j(t+1)|I_t\}$ the best predictor of $X_j(t+1)$, the forecast (2) of $Y(t+1)$ becomes

$$Y_t(1) = \sum_{j=1}^{m} \beta_j X_{j,t}(1). \tag{4}$$

We use the following lemma (for a proof see e.g. Reinsel (1993), p. 14- 15).

Lemma 1. Let $x = (x_1, x_2, \ldots, x_p)'$, $y = (y_1, y_2, \ldots, y_q)'$, $z = (z_1, z_2, \ldots, z_r)'$ be multivariate gaussian random variables and assume that (x, y, z) is also multivariate gaussian. Then

$$\begin{aligned} E(y|x,z) &= E(y|x) + Cov(y,z|x)Var(z|x)^{-1}\{z - E(z|x)\} \\ Var(y|x,z) &= Var(y|x) - Cov(y,z|x)Var(z|x)^{-1}Cov(z,y|x). \end{aligned}$$

On identifying y with $Y(t+1)$, x with I_t and z with $X_O(t+1)$ the lemma gives

$$Y_t^*(1) = Y_t(1) + G\{X_O(t+1) - X_{O,t}(1)\}$$

where

$$G = Cov\{Y(t+1), X_O(t+1)|I_t\}Var\{X_O(t+1)|I_t\}^{-1}$$

which shows how the additional information is linearly incorporated into the updated forecast. To simplify notation, we denote by $u_j = X_j(t+1) - X_{j,t}(1)$ the forecast errors, and write $u = (u_O', u_U')'$ for the corresponding vector. The variance covariance matrix of the u_j's will be written Σ and partitioned accordingly: the quantities $Cov\{Y(t+1), X_O(t+1)|I_t\}$ and $Var\{X_O(t+1)|I_t\}$ may be easily written in terms of Σ. In summary, we obtain the following result.

Theorem 1. The best predictor of $Y(t+1)$ based on the information at time t, I_t, and the values at time t+1 of the subvector X_O, may be written:

$$Y_t^*(1) = Y_t(1) + \beta_O' u_O + \beta_U' \Sigma_{UO}\Sigma_{OO}^{-1} u_O \tag{5}$$

and the mean square forecast error is

$$\begin{aligned} E\{Y_t^*(1) - Y(t+1)\}^2 &= \beta_U'\Sigma_{UU}\beta_U - \beta_U'\Sigma_{UO}\Sigma_{OO}^{-1}\Sigma_{OU}\beta_U \\ &= E\{Y_t(1) - Y(t+1)\}^2 \\ &\quad -(\Sigma_{OO}\beta_O + \Sigma_{OU}\beta_U)'\Sigma_{OO}^{-1}(\Sigma_{OO}\beta_O + \Sigma_{OU}\beta_U). \end{aligned} \tag{6}$$

Note that the second term on the right hand side of both equations of (6) is positive due to the positive definiteness of Σ_{OO}. The second term on the right hand side of (5) corresponds to plugging the observed values $X_O(t+1)$ into the forecast of $Y(t+1)$ in place of the predicted values $X_{O,t}(1)$, while the third term explains the influence of the additional information on the unobserved components at time t+1. Accordingly, the first row of (6) decomposes the mean square forecast error into the amount we would obtain by simply substituting observed values of X_O to their forecasts, and the additional advantage obtained exploiting the relationship between observed and unobserved components. The second row of (6) displays the improvement in terms of mean square forecast error that may be attained using the additional information $X_O(t+1)$.

If only one component may be observed at time $t+1$, say $X_k(t+1)$, then

$$\begin{aligned} Y_t^*(1) &= Y_t(1) + u_k \sum_{j=1}^{m} \beta_j \sigma_{jk}/\sigma_{kk} \\ E\{Y(t+1) - Y_t^*(1)\}^2 &= \beta'\Sigma\beta - \big(\sum_{j=1}^{m} \beta_j \sigma_{jk}\big)^2/\sigma_{kk}. \end{aligned} \tag{7}$$

The reduction in mean square forecast error is proportional to the correlation between u_k and $Y(t+1)$: this suggests which component should be observed in advance, if possible. If $|O| > 1$ components are observable at time t+1, and one may select them, the choice of the subset X_O is neither trivial nor simple by means of analytical methods. In addition, this task cannot be accomplished in an iterative (stepwise) fashion, since the best components to be observed when $|O| = \nu$ do not necessarily remain in the optimal choice for $|O| = \nu+1$. For example, if only one series may be obtained in advance, the mean square reduction, from (7) is $G_k^2 = (\sum \beta_j \sigma_{jk})^2/\sigma_{kk}$, and let k^* be the optimal choice. If two components, i and j, say, may be observed at time $t+1$, the reduction in mean square error may be written:

$$(1-\rho_{ij}^2)^{-1}(G_i^2 + G_j^2 - 2\rho_{ij}G_iG_j) \tag{8}$$

where ρ_{ij} denotes the correlation between u_i and u_j. Maximization of (8) does not necessarily imply that i, or j, equals k^*. However, in the particular case that the data are collected in a fixed sequence with updating of estimates after each new item, conditionally optimal subsets may be considered appropriate. On the other hand, when the components to be observed in advance may be chosen, the problem is similar to that of variable selection in regression analysis and may be addressed by one of the related criteria, using (6) instead of the residual sum of squares as a measure of performance. When the number of components is large, genetic algorithms have been proposed for problems of this nature (Chatterjee et al., 1996).

The results of the present Section rely on the linear predictors $X_{j,t}(1) = E\{X_j(t+1)|I_t\}$ which may be estimated by fitting a multivariate autoregressive moving average to the data. This also provides estimates of the parameters of the linear representation (see e. g. Hannan (1970), pp. 157-158)

$$X(t) = \sum_{j=0}^{\infty} \Psi(j) U_{t-j}$$

which allows us to use additional information for lead-ℓ forecasts as well. Let

$$X_t(\ell) = E\{X(t+\ell)|I_t\}, \quad u(\ell) = X(t+\ell) - X_t(\ell) = \sum_{j=0}^{\ell-1} \Psi(j) U_{t+\ell-j}, \quad Y_t(\ell) = \beta' X_t(\ell).$$

Suppose that the values of a subvector X_O at time $t+h$ $(1 \le h \le \ell)$ are available, then we may form the predictor

$$Y_t^*(\ell) = Y_t(\ell) + M u_O(h)$$

and minimise the mean square forecast error with respect to the elements of the matrix M by means of similar arguments as those used before. The solution is

$$Y_t^*(\ell) = Y_t(\ell) + \{\beta_O'\Sigma_{OO}(h,\ell) + \beta_U'\Sigma_{UO}(h,\ell)\}\Sigma_{OO}^{-1} u_O(h)$$

where the matrices $\Sigma_{OO}(h,\ell)$ and $\Sigma_{UO}(h,\ell)$ are formed with the covariances between $u_i(h)$ and $u_j(\ell)$ and are obtained from:

$$\left[\begin{array}{cc} \Sigma_{OO}(h,\ell) & \Sigma_{OU}(h,\ell) \\ \Sigma_{UO}(h,\ell) & \Sigma_{UU}(h,\ell) \end{array} \right] = \Sigma(h,\ell) = \sum_{j=1}^{h} \Psi(\ell - j)\Sigma\Psi(h-j)'.$$

Often, however, building a vector ARMA model may be impossible or impractical because of the large number of series or computational difficulties. For example, building a vector model for 16 branches is relatively easy, while adopting a similar strategy for the 120 groups (three-digit series) would be impractical. Furthermore, in some cases, prediction of $Y(t+1)$ is obtained from a linear combination of the component series with pre-determined fixed weights suggested by previous experience, tradition or independent guidelines, so that the forecaster is not prepared to modify them considerably. Under such circumstances, we propose that the components observed in advance be exploited anyway by combining them linearly with the given predictor in order to minimise the mean square forecast error, as indicated in the next Section.

3. Exploiting additional information with a given predictor form

Let

$$\hat{Y}_t(1) = \sum_{s \le t} c(s)'X(s) \tag{9}$$

where $c(s) = [c_1(s), c_2(s), \ldots, c_m(s)]'$, $s \le t$, are vectors of fixed known constants, be the linear combination of I_t which is used for predicting $Y(t+1)$. Suppose that the q $(< m)$ values of the subvector $X_O(t+1)$ are observed, and consider the form

$$Y_t^*(1) = \hat{Y}_t(1) + \alpha' X_O(t+1)$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_q)'$. We derive the values of α that minimise the mean square error $E\{Y(t+1) - Y_t^*(1)\}^2$.

Let $\Gamma(h) = E\{X(t)X(t+h)'\}$ denote the autocovariance matrices of $\{X(t)\}$, and partition them as usual. An application of standard regression theory provides the following result.

Theorem 2. Let $X(t)$ be a zero mean second order stationary multivariate process, $Y(t) = \beta' X(t)$ and

$$Y_t^*(1) = \sum_{s \le t} c(s)'X(s) + \alpha' X_O(t+1).$$

The minimum of $E\{Y(t+1) - Y_t^*(1)\}^2$ with respect to α is attained by

$$\alpha^* = \Gamma_{OO}(0)^{-1}\left\{\Gamma_{OO}(0)\beta_O + \Gamma_{OU}(0)\beta_U - \sum_{s \le t}[\Gamma_{OO}(s-t-1)\,,\, \Gamma_{OU}(s-t-1)]c(s)\right\} \tag{10}$$

and the minimum is

$$E\{Y(t+1) - Y_t^*(1)\}^2 = E\{Y(t+1) - \hat{Y}_t(1)\}^2 - (\alpha^*)'\Gamma_{OO}(0)\alpha^*.$$

If only one component may be observed in advance, $X_k(t+1)$ say, then α is scalar and

$$\alpha^* = \frac{1}{\gamma_{kk}(0)} \sum_{j=1}^{m} \{\gamma_{kj}(0)\beta_j - \sum_{s \le t} \gamma_{kj}(s-t-1)c_j(s)\}$$

which represents the regression coefficient of $X_k(t+1)$ on $Y(t+1)-\hat{Y}_t(1)$. If the component series are completely uncorrelated with each other, then

$$\alpha^* = \beta_k - \sum_{s\leq t} r_{kk}(s-t-1)c_k(s)$$

where $r_{kk}(.)$ is the autocorrelation of X_k. If, furthermore, they are serially uncorrelated, then obviously $\alpha^* = \beta_k$.

Since the expression (9) used for $\hat{Y}_t(1)$ is completely general, the results may be applied in several cases. For example, the choice of a univariate autoregressive model

$$Y(t) = \sum_{i=1}^{p} \phi_i Y(t-i)$$

for forecasting $Y(t+1)$ corresponds to weights $c_j(s) = \beta_j \phi_{t+1-s}$ for $j = 1, 2, \ldots, m$ and $s = t-p, \ldots, t$.

A more precise result may be obtained if forecasts of the single components are first computed by means of the predictors

$$\hat{X}_{j,t}(1) = \sum_{s\leq t}\sum_{i=1}^{m} k_{ji}(s)X_i(s) = \sum_{s\leq t} k_j(s)'X(s)$$

and then the predictor of $Y(t+1)$ is obtained by linearly combining them:

$$\hat{Y}_t(1) = \sum_{j=1}^{m} \beta_j \hat{X}_{j,t}(1) = \sum_{j=1}^{m} \beta_j \sum_{s\leq t} k_j(s)'X(s). \tag{11}$$

In that case, if $X_O(t+1)$ is known, we first substitute the known values to their forecasts, obtaining

$$\bar{Y}_t(1) = \beta_O' X_O(t+1) + \sum_{j\in U} \beta_j \sum_{s\leq t} k_j(s)'X(s)$$

and then regress $X_O(t+1)$ on the $\hat{X}_{j,t}(1)$, giving

$$Y_t^*(1) = \bar{Y}_t(1) + \gamma' X_O(t+1).$$

The optimal choice in terms of γ may be directly obtained from Theorem 2, letting $c(s) = \sum_{j\in U} \beta_j k_j(s)$ and $\gamma^* = \alpha^* - \beta_O$:

$$\gamma^* = \Gamma_{OO}(0)^{-1}\big\{\Gamma_{OU}(0)\beta_U - \sum_{s\leq t}[\Gamma_{OO}(s-t-1)\,,\,\Gamma_{OU}(s-t-1)]\sum_{j\in U}\beta_j k_j(s)\big\}.$$

In particular, fitting different univariate autoregressive models to each component:

$$X_j(t) = \sum_{i=1}^{p} \phi_j(i) X_j(t-i)$$

corresponds to the previous results setting $k_{ji}(s) = 0, j \neq i, \forall s$, and $k_{jj}(s) = \phi_j(t+1-s), t-p+1 \leq s \leq t, k_{jj}(s) = 0$ otherwise.

Extension to lead-ℓ forecasting is immediate. Denoting by

$$\hat{Y}_t(\ell) = \sum_{s \le t} k(s)' X(s)$$

the predictor of $Y(t+\ell)$ at time t, and supposing that we may observe the components of X_O at time $t+h$, $(1 \le h \le \ell)$, we form the improved predictor

$$Y_t^*(\ell, h) = \hat{Y}_t(\ell) + \alpha' X_O(t+h).$$

In a similar way to Theorem 2, it may be shown that the minimum mean square forecast error is obtained by choosing

$$\alpha_\ell^* = \Gamma_{OO}(0)^{-1}\big\{\Gamma_{OO}(\ell-h)\beta_O + \Gamma_{OU}(\ell-h)\beta_U - \sum_{s \le t}[\Gamma_{OO}(s-t-1)\,,\,\Gamma_{OU}(s-t-1)]k(s)\big\}$$

and the minimum is

$$E\{Y(t+\ell) - Y_t^*(\ell)\}^2 = E\{Y(t+\ell) - \hat{Y}_t(\ell)\}^2 - (\alpha_\ell^*)'\Gamma_{OO}(0)\alpha_\ell^*.$$

4. Results

The Italian general industrial production index is a linear combination of the indexes relating to 16 industrial branches according to the NACE–Revision 1 classification (see Table 1). Monthly data from January 1990 to December 1999 were analysed, and a multivariate second order autoregressive model identified for the twelfth differences. The parameters were estimated using the IMESTIM procedure of the SCA Statistical System (Liu and Hudak, 1992), which employs an iterative constrained least squares method. Each parameter whose estimated standardised value does not exceed 1.96 in modulus is set to zero, and the estimation stage is iterated until all parameters are significant.

The resulting model has 58 parameters. We obtained R^2 univariate values ranging from 0.82 to 0.95 and a multivariate R^2 larger than 0.99. The usual univariate and multivariate portmanteau tests do not reject the null hypothesis of white residuals at the 0.05 level.

The observed residuals were used for estimating Σ, and the minimum of the mean square forecast error (6) using 1 to 6 anticipated components was computed. Since the number of component series was relatively small, we have found the optimal solution by enumerating all possible choices of $|O| = \nu$ series out of 16 (for $\nu = 1, \ldots, 6$). Results are shown in Table 2.

The variance of the (differenced) general index series $Y(t)$ is 21.8, and a univariate AR(2) model fitted to $Y(t)$ has a residual variance of 17.35, while the forecasting variance using the multivariate model, i.e. the variance of $Y_t(1)$ in (4) is 16.33. Thus, it may be seen from Table 2 that the use of just one anticipated component drastically reduces the mean square error (the most favorable component is DL, which reduces it to 3.54), and knowledge of the anticipated values of just a few components may reduce the error to very small figures. However, the choice of components is important because they may have very different effects: for example using component C provides a mean square reduction of only about 0.3, and even using as many as six anticipated components, if badly chosen, may result in only a small reduction (about 8.5 in the worst case).

As an alternative, univariate AR(2) models were fitted to differenced data of the single components and the resulting predictor of $Y(t+1)$ as in (11) was computed. Its mean

Table 2. Mean square forecast error of the industrial production index fitting a multivariate autoregressive model and using the best ν anticipated components ($\nu = 0, \ldots, 6$).

ν	Selected components	Mean Square Forecast Error
0	$\emptyset$	16.33
1	DL	3.54
2	DJ, DL	1.39
3	DI, DJ, DL	1.03
4	DB, DI, DJ, DL	0.72
5	DB, DE, DI, DJ, DK	0.50
6	DA, DB, DI, DJ, DK, E	0.34

Table 3. Mean square forecast error of the industrial production index fitting univariate autoregressive models to the components, and using the best ν anticipated components ($\nu = 0, \ldots, 6$).

ν	Selected components	Mean Square Forecast Error
0	$\emptyset$	16.56
1	DL	6.48
2	DJ, DL	4.24
3	DJ, DL, DM	2.77
4	DJ, DK, DL, DM	1.84
5	DE, DJ, DK, DL, DM	1.27
6	DB, DE, DJ, DK, DL, DM	0.87

square error on the entire observed period was 16.56. We have computed the improved predictor using anticipated values of the best ν components, for $\nu = 1, \ldots, 6$; results are exposed in Table 3. It may be observed that the mean square errors are considerably larger than when using a multivariate model, as expected, but also here the knowledge of a few components may provide a large reduction of the forecasting error. Furthermore, the selected components are slightly different.

Finally, in order to verify such results, we have actually computed one–step–ahead forecasts of the general production index for January to December 2000, using each of the possible proposed forms with one to four anticipated components. The observed average square forecast errors on the twelve months are reported in Table 4. If no anticipated components are employed, the average square forecast error is about 25 using both multivariate and univariate autoregressive models, owing to an unexpectedly large figure for May, which accounts for more than 40 per cent of the total error. Results using one to four anticipated components are progressively more precise. The actual differenced data and the forecasts with no anticipated components and the best three anticipated components are shown in Figure 1 for the multivariate model. Figure 2 refers to the case of univariate models fitting.

The proposed method may also be applied to exploit anticipated information concerning specific sectors, whose data may be elaborated more rapidly. We have taken into account the group (three–digit) and class (four–digit) information, and employed each individual series to linearly modify the prediction of the general index based on the multivariate autoregressive model that we have built on the 16 sub–sections of Table 1.

The series were evaluated according to their observed correlation with the prediction errors $Y(t+1) - \hat{Y}_t(1)$ in the period 1996–1999; the results for one–step–ahead forecasting for the twelve months of the year 2000 are shown in Table 5, where the average square forecast error is exhibited for a few of the most useful three–digit and four–digit series. It

Table 4. Average square forecast error for one–step–ahead forecasts, January to December 2000, using different predictors for the general production index.

no. of anticipated components ν	use of multivariate model	use of univariate models
0	25.86	25.08
1	7.65	11.74
2	2.90	7.20
3	1.73	5.40
4	1.42	3.64

Table 5. Average square forecast error for one–step–ahead forecasts, January to December 2000, using anticipated values of some three and four–digit classification series and multivariate predictor.

code	description	avg. sq. error
DE 212	articles of paper and paperboard	9.42
DJ 275	casting of metals	10.91
DH 252	plastic products	12.09
DJ 2751	casting of iron	12.26
DE 2121	container of paper and paperboard	14.10

may be seen that the advantage in terms of forecasting precision is considerably smaller than using the entire data of one or more section series as done before. However, knowledge of a single four–digit series (casting of iron or containers of paper) allows us to halve the observed square forecast error, and a three–digit series, such as the index of articles of paper and paperboard, reduces the error to almost one third.

5. Conclusions

The proposed method provides a way to exploit information as soon as it is available to estimate the values of the general production index in an optimal and iterative fashion, before the final correct figure is published. To this aim, we also note that the method may be equally applied without any difficulty in reverse order, allowing for a decision on what component (or what subset of two, three, or four components) may be ignored with the minimum square error.

Thus, and most importantly, our results may suggest how to organise the collecting and processing activities in order to privilege the timeliness of the most useful components, and therefore improve the accuracy of early published data and reduce the amount of revision.

In principle the method may be employed with any composite index. A particularly interesting case seems to be that of spatial averages, where the global index is obtained as an average of the corresponding indexes for different areas, regions or countries. In some cases local figures are published in sequence prior to the global ones, and depending on the order of their appearance, may induce misleading expectations.

Acknowledgements

Research supported by Ministero della Istruzione, Universitá e Ricerca Scientifica and by Consiglio Nazionale delle Ricerche.

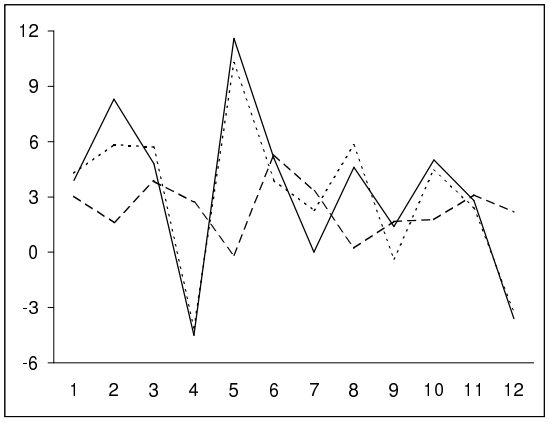

Fig. 1. Year 2000 forecasts with multivariate model. Solid line: actual data (differenced); dashed line: pure forecast; dotted line: forecast with the best 3 anticipated components.

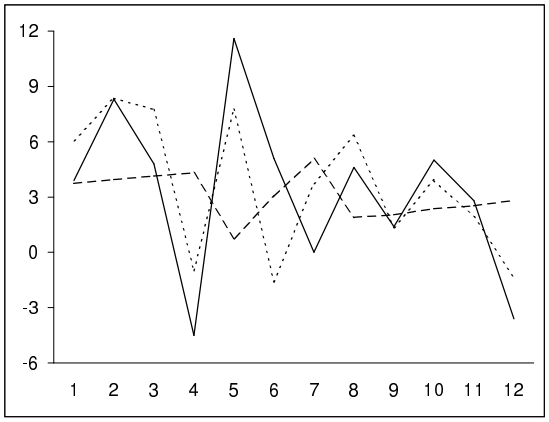

Fig. 2. Year 2000 forecasts with univariate autoregressive models. Solid line: actual data (differenced); dashed line: pure forecast; dotted line: forecast with the best 3 anticipated components.

We are indebted to the Joint Editor, an Associate Editor and an anonymous Referee for valuable comments and helpful hints.

References

Barker, T. and M. Pesaran (Eds.) (1990). *Disaggregation in Econometric Modelling*. London: Routledge.

Bodo, G. and L. Signorini (1987). Short-term forecasting of the industrial production index. *International Journal of Forecasting 10*, 285–299.

Boucelham, J. and T. Teräsvirta (1990). Use of preliminary values in forecasting industrial production. *International Journal of Forecasting 6*, 463–468.

Box, G. E. P. and G. C. Tiao (1977). A canonical analysis of multiple time series. *Biometrika 64*, 335–365.

Braun, J. M. (1973). Series chronologiques multiples: recherche d'indicateurs. *Revue de Statistique Appliquee 21*(1), 81–106.

Brillinger, D. R. (1981). *Time Series Data Analysis and Theory* (expanded ed.). S. Francisco: Holden-Day.

Byers, J. and D. Peel (1995). Forecasting industrial production using non linear models. *Journal of Forecasting 14*, 325–336.

Chatterjee, S., M. Laudato, and L. A. Lynch (1996). Genetic algorithms and their statistical applications: an introduction. *Computational Statistics and Data Analysis 22*, 633–651.

Coccia, M. and P. Iafolla (2000). Un metodo per ridurre le informazioni di base a scopo previsivo attraverso l'analisi fattoriale e l'analisi di serie storiche. Available at http://ser.sta.uniroma1.it/cofin98/Rt5-2000.ps.

Hannan, E. J. (1970). *Multiple Time Series*. New York: Wiley.

Lahiri, K. and G. Moore (Eds.) (1991). *Leading Economic Indicators. New Approaches and Forecasting Records*. Cambridge: Cambridge University Press.

Liu, L. and G. B. Hudak (1992). *Forecasting and Time Series Analysis using the SCA Statistical System*. Oak Brook, Illinois: Scientific Computing Associates.

Marchetti, D. J. and G. Parigi (2000). Energy consumption, survey data and the prediction of industrial production. *Journal of Forecasting 19*, 419–440.

Patterson, K. D. (2002). Modelling the data measurement process for the index of production. *J. R. Statist. Soc A 165*, 279–296.

Priestley, M. B. (1981). *Spectral Analysis and Time Series*. London: Academic Press.

Rahiala, M. and T. Teräsvirta (1993). Business survey data in forecasting the output of swedish and finnish metal and engineering industries: a kalman filter approach. *Journal of Forecasting 12*, 255–271.

Reinsel, G. C. (1993). *Elements of Multivariate Time Series*. New York: Wiley.

Thury, G. (1998). Forecasting industrial production using structural time series models. *Omega 26*, 751–767.

van Garderen, K., K. Lee, and M. Pesaran (2000). Cross-sectorial aggregation of non linear models. *J. Econometrics 95*, 285–331.