Methodological aspects of time series back-calculation*

Massimiliano Caporin†
University of Ca' Foscari Venezia

Domenico Sartore
University of Ca' Foscari Venezia

November 15, 2004

Abstract

This paper provides the theoretical and operational framework for estimating past values of relevant time series starting from a limited information set. We consider a general approach exploring the relevant problems and the possible solutions evidencing that linear models could be the preferred choice. An empirical example is presented: we analyse the back-calculation of Eu15 Industrial Production Index comparing our approach to Eurostat one.

Keywords: back-calculation, retropolation, historical reconstruction, back-forecasting

*This paper is part of the research project "Methodology for back-calculation" supported by EUROSTAT. We are grateful to Gian Luigi Mazzi, Giovanni Savio, Fabio Sartori, Tommaso di Fonzo and Monica Billio for hekpful comments.

†Corresponding author - Dipartimento di Scienze Economiche, Università di Venezia, San Giobbe 873, 30121 Venezia, ITALY - tel. +39-041-234-9142 - fax +39-041-234-9176 - email: mcaporin@unive.it

1 Introduction

Policy makers and practitioneers often requires long time series for model evaluation, policy analysis and macroeconomic studies. However, long time series are not necessarily available at the desired frequency, with the needed spatial or sectorial coverage or they are not available at all. This unavailability can depend on various motivations: the needed indicators have been subjected to an extraordinary revision process and past values have not been reconstructed; we are searching for indicators referring to new subjects (such as the Euro area or the enlarged European Union); production standards have been modified increasing the series frequency. Neverthless, time series users may need in any case an estimate of the needed variables for past years. The aim of this paper is twofold: on the one side, it defines the back-calculation as a process for estimating past values of a time series and it highlights the back-calculation relation with time series temporal disaggregation, time series aggregation and the construction of proxy variables; on the other side, it provides a general methodological scheme that includes the available approaches and the possible technical solutions for time series back-calculation, pointing out their dependence on the available information and on the characteristics of the neede series.

The need of having a clear definition and of rationalizing the approach derives from our practical experience with macroeconomic series; the discussion will be influenced by that, we will mainly refer to economic time series examples. However, the methodological treatment will be general enough to be used in any time series framework including biostatistics, climathology and social sciences.

The plan of the paper is the following: section 2 defines the back-calculation framework while section 3 introduces our methodological scheme for the back-calculation process; section 4 considers several statistical aspects of time series back-calculation; section 5 presents an empirical example and section 6 concludes.

2 Defining time series back-calculation

At first, let us define the problem whose methodological solution is the time series "back-calculation" as a "back-calculation problem". Then, we introduce some notation. Assume we are focusing on the time series x_t which

is available for $t = 0, 1, ...T$. We are interested in estimating the values of x_t for $t \in \{-M, -M+1, -M+2, ... -2, -1\}$ using the information set $Z = \left\{x_t, Y_t, K_{i(t)}, W_{j(t)}\right\}$ where: $t = -M, -M+1, ...T$ and all variables can have initial and/or final missing values (as an example a related series could be available from $-M < j < 0$ to $0 < i < T$ - internal missing values creates treatment problems which are not considered in this paper); Y_t is a set of variables with the same frequency of x_t (i.e. if x_t is a quarterly series Y_t contains quarterly variables); $K_{i(t)}$ contains series observed at a frequency lower than x_t (i.e. if x_t is a quarterly series $K_{i(t)}$ contains annual figures) and $i(t) = a + bt$ (in the quarterly-annual case with annual series observed at the end of the fourth quarter and $-M$ is the first quarter of an year $a = 0$ and $b = 4$); $W_{j(t)}$ contains series observed of a frequency higher than x_t (i.e. if x_t is a quarterly series $W_{j(t)}$ contains monthly figures) and $j(t) = c + dt$ (in the quarterly-monthly case with the quaterly series observed the third, sixth, nineth and twelth month of the year and $W_{j(t)}$ starting the first month of an year $c = 0$ and $d = 1/3$).

Definition 1 *A time series "back-calculation" is the estimation of x_t for $t \in \{-M, -M+1, -M+2, ... -2, -1\}$ when the following conditions are jointly satisfied:*

i) Y_t does not contain a complete and exaustive disaggregation of x_t for $t \in \{-M, -M+1, -M+2, ... -2, -1\}$ (i.e. if x_t is the industrial production index, Y_t does not contain the sectorial industrial production indices for the needed time span or if x_t is the Gross National Product, Y_t does not contain the regional gross product);

ii) $K_{i(t)}$ does not contain a complete and exaustive temporal aggregation of x_t (i.e. if x_t is the quarterly Gross National Product, Y_t does not contain the annual Gross National Product);

iii) $W_{j(t)}$ does not contain a complete and exaustive temporal disaggregation of x_t (i.e. if x_t is the quarterly industrial production index, Y_t does not contain the monthly industrial production index);

A violation of the previous conditions, which pose us outside the back-calculation framework, will be discussed in a following section when dealing with model choice.

Note that point i) may seem to be incomplete since it deas only with contemporaneous disaggregation and not with contemporaneous aggregation. In fact, the availability of an aggregated variable which contains x_t does not

necessarily provide optimal informations since its patterns may be biased by other components. As an example, consider the Industrial Production Index on the first level NACE Rev. 1 classification: it contains the main economic sectors, A agriculture and fishing, C mining, D manufacture and so on. If our purpose is the back-calculation of IPI for agriculture and we have the Total IPI, the we cannot safely use it since its behaviour is highly dependent on the Manufacture sector which weigths much more than agriculture in the Total IPI.

Finally, a comment on the word "back-calculation": at our best knowledge, no other published papers used that name; some referred to similar problems using related words such as "retropolation" or "reconstruction" (Di Fonzo, 2003b) while the ARIMA literature uses "back-casting" (Box and Jenkins, 1970) for estimation of past values of a time series. In the following, we consider "back-calculation", "back-recalculation", "retropolation", "back-forecasting" and "backward estimation" as synomims.

3 Back-calculation methodology: a step-by-step guide

Up to this point we have just defined the environment, but in the real life specific problems occur and some questions may naturally arise: how is x_t chosen? What does $K_{i(t)}$ contains? and $W_{j(t)}$? and Y_t? what about the technical aspects of x_t estimation? The following section will discuss the technical and statistical aspects while this section focuses on the various steps that must be considered in order to solve a back-calculation problem.

Step 1: Planning

Before considering the technical aspects of the back-calculation problem, some points must be clarified. First of all, we must define x_t and not in terms of the variable which we want to back-calculate (e.i. say the GDP, which is assumed to be known) but in its definition: are we interested in, say, the unemploymnet compiled with the International Labour Office standards or on a different basis? Alternatively, are we interested in the Harmonised Index of Consumer Prices on EUROSTAT standards or on tha Consumer Price Index based on the national definition? Last example, are we interested in the raw series or in the seasonally adjusted series? Therefore, we should

specify which is the series methodological definition since different sources may measure the same quantity or index with different methodology and different adjustments. The point may seem obvious but it has to considered in order to prevent the back-calculated series from possible errors.

Furthermore, we should define M, the back-calculation "optimal" horizon. Two cases maybe considered: M is fixed a-priori given some specific needs of policy makers or users of the back-calculated x_t series; alternatively we could consider a data-driven specification of M. The second option is the preferred one, since it allows the researcher to extract all the information from the available data. Moreover, whenever M is fixed a priori and is large, there is no reason to assume that the available information will allow a complete back-calculation.

Step 2: Data availability

Once we defined x_t, the second step focuses on the information set $Z = \{x_t, Y_t, K_{i(t)}, W_{j(t)}\}$. The object of this task is to collect all the available information related to x_t. Several aspects must be considered:

i) the sources: we have to search for data at National Statistical Institutes, National Banks, international organisations (European Central Bank, OECD, UN...), data providers (Datastream, Reuters...) and in general all instututions which may provide some information. The search must not be limited to the time series figures but should also collect the production methodologies (including informations on the series definition, the possible adjustments for seasonality, working days, outliers...);

ii) x_t series dimension: we have to search for temporal, spatial and sectorial aggregated or disaggregated figures of x_t as well as for series measuring the same quantities but on different definitions; in this last case, we include both different standards (such as for the unemployment case) as well as different data adjustments (such as for seasonality, outliers, or working days);

iii) series related to x_t: we have to search for proxies of x_t which could be used when the analysis of x_t series dimension produced poor results. These series may belong to Y_t, $K_{i(t)}$ or $W_{j(t)}$ and may measure a larger or smaller geographical area, one of the components of x_t or they can be referred to series containing x_t; furthermore, these related series may be suggested by an (economic) theory that postulate a relation between the searched indicator and x_t.

The collected data must be subjected to a first screening, at least graphical, in order to identify possible errors: typing, superimposion of series on

different scales or reference years etc. These errors should not be present in official data but they are not so rare. This point takes into consideration the data reliability which is generally attached to the source. National statistical institutes and interantional organisations are generally reliable sources, while companies that distribute data on the internet without specifying their source are generally more unreliable. In this last case, data errors could be more frequent. Obviously, unrealiable data and sources should not be considered.

Step 3: Strategy

Once we have collected all the available informations and series we can proceed to a comparative analysis whose final purpose is the strategy definition. That is, we must decide which model to use. Furthermore, only at this stage we have all the information necessary to specify if we are dealing with a "pure" back-calculation problem. This step is the core of the methodology and will be analysed in the following section. The model choice strictly depends on the available data, in particular on their coverage and on their quality. Finally, the strategy can include a preliminary data analysis and also a sensitivity analysis of the chosen model.

Step 4: Production

We collected the data and we specified the model, next step is straighforward: apply the model to the information set and get the back-calculated series.

4 Data driven model choice

The model definition is a fundamental step in any statistical problem. However, in the back-calculation case, model choice is mostly based on the available data and not fixed a-priori. Furthermore, the data themselves define if we are considering a pure back-calculation problem or something different. In the following paragraphs we will consider several aspects related to the back-calculation strategy which jointly considered allow a correct model definition.

4.1 Aggregation and disaggregation

Only once the data have been collected we can define if we are in a back-calculation problem. Using the conditions of Definition 1 we can state the following:

a) whenever Z contains a complete and exaustive disaggregation of x_t for $t \in \{-M, -M+1, -M+2, ... -2, -1\}$ (violating condition i) we are facing an aggregation problem (spatial or sectorial);

b) whenever Z contains a complete and exaustive temporal aggregation of x_t for $t \in \{-M, -M+1, -M+2, ... -2, -1\}$ (violating definition ii) we are facing a temporal disaggregation problem;

c) whenever Z contains a complete and exaustive temporal disaggregation of x_t for $t \in \{-M, -M+1, -M+2, ... -2, -1\}$ (violating definition iii) we are facing a temporal aggregation problem.

Case a) is considered by (??) in the construction of aggregated serie for the Euro area dealing in particular with the exchange rate problem. The point can be generalised including the aggregation of sectorial series in the estimation of total figures. Case c) is similar to case a) the only difference is in the dimension which is here in the time domain. Finally, case b) is the most interesting. Temporal Disaggregation problems have been considered since the seminal work of Chow and Lin (1971, 1976) recently extended by Fernandez (1981), Litterman (1983), Santos Silva and Cardoso (2001) and Di Fonzo (2003a and 2003c). In general we can distinguish two subcases:

b1) $Z = \left\{x_t, K_{i(t)}\right\}$ - i.e. there is no information available at the x_t frequency but only at a lower frequency. The following strategies can be considered: 1) using purely statistical temporal disaggregation approaches such as the Denton moving preservation principle; 2) using proportional distributions estimating weights with the available x_t observations; 3) extract with a structural or linear model the components of x_t, project them into the past and then use a Chow-Lin approach;

b2) $Z = \left\{x_t, Y_t, K_{i(t)}\right\}$ - i.e. there is something more than in case b1), there exist some information on series related to x_t. This setup is generally named constrained retropolation and is partially discussed in Di Fonzo (2003b). The related series included in Y_t could be used to back-calcuate x_t but we have also some information on a lower frequency that can be used for benchmarking the back-calculated series. The preliminary estimate of x_t past values can be used as the best related indicator in a constrained retropolation framework. We refer to the combined used of back-calculation and other

methods (aggregation and disaggregation) as "mixed approaches". The availability of an aggregated estimated or in general of additional informations at different time frequencies can be considered as a plus with respect to the standard back-calculation problem. The derivation of a preliminary back-calculated series is strictly related to the construction of a proxy variable: in fact, the use for the estimation of a desired series of related series with different coverage and frequency or suggested by some theoretical linkage is the standard problem of constructing a proxy variable.

4.2 Back-calculation

Assume that Definition 1 is satisfied; we are then considering a pure back-calculation problem. However, several cases may realise:

a) $Z = \{x_t\}$ - there is nothing more than the series we are considering;

b) $Z = \{x_t, Y_t\}$ - there are related series withouth missing values;

c) $Z = \{x_t, Y_t\}$ - there are related series with missing values.

In case a) the only possibility is the use of an ARIMA approach. Two solutions are available: estimate an ARIMA on the current series, reverse the model and use it to produce some forecasts; reverse the series, fit an ARIMA model and use it to produce forecasts. Both approaches presents some problem: forecasts can be made only for a limited number of steps into the past otherwise either we will estimate only a tendency or the forecasts will converge to the long run level or they will explode. Focus at first on time series temporal reversion: this approach must be carefully considered. A time reversibility test exists, Ramsey and Rothman (1996), but it requires simmetry of the series (i.e. no trends, no asymmetric seasonal components and no asymmetric cycles). This last hypothesis is know to be rejected by economic time series which are influenced by the business cycle that is known to be asymmetric, see (??). In addition, reversing the time path of a series does not necessarily have a clear economic interpretation, despite from a statistical point of view it could be strongly supported.

Differently, the reversion of an ARIMA model can be used to produce very few forecasts. A simple example may clarify the point. Consider the simple AR(1) model

$$y_t = \phi y_{t-1} + \varepsilon_t$$

with $|\phi| < 1$. If we estimate the model and then we reverse it the back-

calculating equation is

$$\hat{y}_{t-j} = \left(\frac{1}{\hat{\phi}}\right)^{j} y_t \qquad j > 0$$

resulting in a explosive pattern. Differently, the reversion of MA terms is useless since past values of the innovation term are not available. In that case a stochastic simulation approach could be considered: we could use a bootstrap sampling from the estimated residuals of the fitted MA model. However, the reliability of the resulting series could be questionable.

Cases b) and c) can be treated within a common regression framework. Assume that the general model to be used for a back-calculatios is the linear regression model possibly extended with ARMA terms in the residuals. The general model is the following

$$\Delta^m x_t = \beta \Delta^m Y_t + \delta D_t + \gamma T_t + \Phi^{-1}(L)\,\Theta(L)\,\varepsilon_t \tag{1}$$

where Y_t contains the related indicators possibly including the constant, D_t a set of seasonal dummies, T_t the time trend and ε_t is an innovation process. Furthermore, the objective series and the contermporaneous indicators may be considered in their m-th difference, possibly including even seasonal differences. Estimation can be made by maximum likelihood and standard tests can be used to evaluate the estimated coefficients and the residuals. The estimation is possible if the relevant series and the related indicators are available on an (even limitted) overlapping sample (in case this sample is very small, the back-calculation could provide unreliable past observations). Once the coefficients have been estimated, the back-calculated series is obatined as follows

$$\Delta^m x_t = \hat{\beta} \Delta^m Y_t + \hat{\delta} D_t + \hat{\gamma} T_t \qquad t < 0$$

Case b) directly fits with equation (1) while case c) requires a preliminar step. The information set contains related series with initial or final missing values, i.e. the various related series have a different time and/or spatial and/or sectorial coverage (the case of internal missing values is not considered here). Some example may claify the point: if we are interested in estimation past values for the Euro 15 GDP from 1980 and the information set contains the past GDP values for 10 out of the 15 countries with 6 series starting in 1980 while the remaining start in 1985; alternativaly, again in the GDP case, we have 15 series for private consumption back to 1980 and only 10 for

investments back to 1982. We face then a problem, we should decide among these two aternative approaches: c1) to extract a proxy from the available information and then to use in equation (1); c2) to proceed with several back-calculation steps progressively extending the back-calculation horizon.

In the first case the literature on the construction of proxy variables can be recalled, among other we refer to (??).

Case c2) maybe easily understood referring again to the GDP example. Assume now that the available Euro 15 GDP series starts in 1990 and that we have the following available data: Germany, France, Spain, Italy and UK GDP from 1980; Belgium, The Netherlands and Finland GDP from 1985 (these coverages do not correspond to the reality - the problems referred to the GDP prices, its base year, the seasonal adjustment and the system of account are not considered for the sake of exposition). Our purpose is the back-calculation of Euro 15 GDP back to 1980. We have two groups of variables available with different coverages. We can forget about Belgium, Finland and The Netherland and use only the six countries available for the whole back-calculation sample. By this way the back-calculation procedure will turn out to be simpler but not all the available information will be used. Alternatively we can split the back-calculation into two parts: at first back-calculate Euro 15 GDP back to 1985 using the available information and then back-calculate the Euro 15 from 1980 to 1984 using a different (and smaller) information set. In this second case, we use of all the available data and the back-calculation turns out to be more efficient. As a general rule, we prefer the second approach.

In the following we focus on the two cases only in order to highlight some specific aspects and problem of the back-calculation. Consider the first case, the back-calculation of Euro 15 using the five series available from 1980. The Euro 15 GDP is computed as the sum of the national GDP, therefore two solutions are available: we can compute an Euro 5 GDP (summing the series available up to 1980) or use directly the national series. In the first case a proxy for Euro 15 GDP is computed, while in the second case we make directly use of the available series. The choice between the two approaches can be statistically tested within a regression approach if we focus on the levels (it is just a Wald test on a multiple coefficient restriction). However, this is not true if we consider the returns (a following paragraph will address the issue of a back-calculation based on levels, returns or differenced series) or if we consider an aggregated series which is computed by a weighted average (in this case the test is not so immediate and will include the weights to be

used in the aggregation), or, finally, if the back-calculation does not focus on components of the relevant series (i.e. if we are considering a series and some related indicator which does not consider a sector, an area or a component of the relevant series). In all cases, the problem is included in the formalisation of equation (1). In this case, we just have to apply the model computed the coefficients on the overlapping sample period (i.e. for the range 1990 to last available points) and the use the estimated coefficients applying them backward to back-calculate Euro 15 GDP.

Assume now we follow the second approach building two proxies: Euro 5 available from 1980 and Euro 7 available from 1985 (alternatively we could make use of two variable sets). Even in this case, we can follow two approaches: we can back-calculate Euro 15 on Euro 7 back to 1985 obtaining a new Euro 15 series and the in a second step back-calculate the newly available Euro 15 series back to 1980 using the Euro 5 series; alternatively, we can back-calculate the Euro 7 series back to 1980 using the Euro 5 series and then in the second step back-calculate the Euro 15 series back to 1980 using the newly available Euro 7 series.

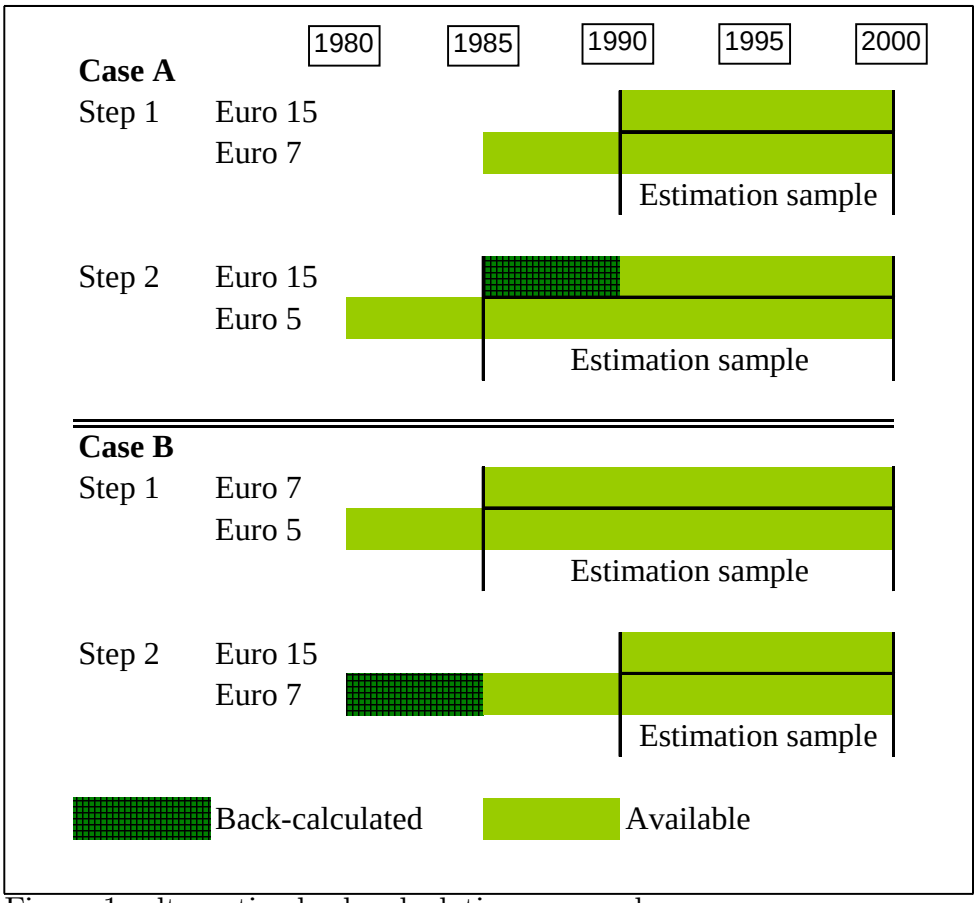

Figure 1: alternative back-calculation approaches

The differences among the two approaches are highlighted by Figure 1. In case A, step 2 involves a regression based on estimated values while case B uses only readily available data. Differently, the back-calculation of Euro 15 is based on readily available data in case A while in case B it involves also back-calculated data (for the range 1980-1985). Which should be the preferred one? We do not have a final answer: assuming that the specification of the regression model is optimal in both cases, the two will provide very close results. In our experience we mainly used the second solution back-calculating several proxies for different coverages; the motivation is that Euro 5 is preferred as indicator in the retropolation of Euro 7, while Euro 7 is a preferred indicator for the retropolation of Euro 15. This is not the case if we use Euro 5 in estimating Euro 15; Euro 7 contains more information than Euro 5.

Both approaches may include estimation errors: in case A due to the bias in the regression of step 1 which may affect coefficient estimates of step 2; in case B due to multiplicative effects in the back-calculation of the range 1980-1985. Our experience tells us that none of the two is prevalent, and we leave the proof to an extensive analysis in future researches.

Our primary focus was on linear models, however, alternative models could be considered. We refer to the classical time series approach and the structural models. In fact, the decomposition of a time series into its cycle-trend, seasonal and irregular components could be of some relevance for the back-calculation. In that case we can transform the structural approach in a linear model extracting from the information set the structural components and then reconstructing them. In a second step, the model used for the structural decomposition of the relevant series can then be used to back-calculate it starting from the retropolated structural comoonents. Within this approach we can use also the common factors models. In fact, the extraction of structural components from the information set could consider a common factor approach.

4.3 Statistical properties - Levels, logs or differences?

In order to perform a back-calculation, statistical properties of the series should be analysed. A first point should be stressed: if we have some related series, a proxy for the series x_t, in order to safely perform a back-calculation the indicator must be cointegrated with x_t; furthermore, as far as the cointegrating vector of (x_t, y_t) is closer to $(1, 1)$ the back-calculation will produce better results. This presume to have common trends in the series. Similarly and evidently, if we perform a back-calculation using several indicators the cointegration vector would not be composed of ones. Before proceeding to the back-calculation we must decide if we will work on the levels, the logs, or on any difference of the series. The choice must be based on the purposes of the back-calculation and on the series characteristics. If series are not integrated of any order, the levels provide a good back-calculation of the tendency of the series, while the difference or the log-returns provide a good back-calculation of the growth rates. If series have a seasonal pattern, seasonal (log-)differences can be considered in order to back-calculate yearly growth rates if series are integrated at the seasonal level.

4.4 Combining back-calculations

In the previous paraghap we evidenced that back-calculations could be made on the levels, on the 1-period growth rates, on the seasonal growth rates and on the 1-period and seasonal difference. In principle, we could choose one of these cases to perform the analysis, alternatively, we can compute all the back.calculations and then consider a combined back-calculation approach. In this case we are mirroring the literature of combined forecast, see (??). As with combined forecasts, the alternative back-calculated series will likely be collinear, making the estimation of combining coefficient problematic. In order to reduce this effect, we suggest to include a constant in the regression and not imposing any bound on the coefficients. Assuming that the series x_1, x_2 and x_3 are three alternative back-calculations of the series y we suggest estimating the following equation

$$y_t = \alpha_0 + \alpha_1 x_{1,t} + \alpha_2 x_{2,t} + \alpha_3 x_{3,t} + \varepsilon_t$$

over an a-priori fixed in-sample period (that is, over a fixed range $0, 1...j < T$).

Finally, back-calculation models would not be necessary stable over time, in the sense that the available indicators may change for past ranges. As an example, focus again on the Euro 15 case: in that problem, the model changes over time, and we are practically combining over time different models. Several problems may surge related to multiplicative or additive errors as already mentioned. The important thing to note is that the selection of related indicators is not constrained to be fixed for the back-calculation, instead we can change the relevant indicator as they become available. Our final goal is the back-calculation over the fixed range (-M,-1) with a reliable model, but not necvessarily with a fixed model.

4.5 Sensitivity analysis

A further point emerge having different alternative back-calculated series: how could we choose one of them or a combination of some of them? In sample sensitivity analysis can be used to choose model specification and possible forecast combination. In this case, RMSE, AMSE and Theil U index could be used as measures of back-calculation accuracy. These measures should be calculated again over an a-priori fixed in-sample period.

5 A case study: retropolation of EU15 Industrial Production Index

In order to provide an empirical example of our back-calculation approach we focus on the estimation of the EU15 Industrial Production Index for Total Industry excluding construction (NACE Rev. 1.1 sectors C, D and E - mining and quarrying, manufacture and energy), seasonally unadjusted, working day adjusted, in 2000 base year. The EU15 and national series are available on the NewCronos database (EUROSTAT database) with the following coverage:

Table 1: IPI Total Industry excluding construction (WDA - 2000 base year)

Country	First Obs.	Weigth	Country	First Obs.	Weigth
Austria	January 1996	2.5	Italy	January 1990	13.9
Belgium	January 1970	3.2	Luxembourg	January 1970	0.2
Danemark	January 1985	1.9	Portugal	January 1990	1.3
Finland	January 1990	2.0	Spain	January 1980	6.9
France	January 1990	14.2	Sweden	January 1990	3.2
Germany	January 1978	26.4	The Netherlands	January 1970	4.1
Greece	January 1995	0.8	United Kingdom	January 1986	17.5
Ireland	January 1980	2.1	EU15	January 1986	100.0

Actual national and EU15 coverage as reported in NewCronos (first available observation - availability at the end of August 2004) - country weigth on total EU15

EUROSTAT computes the EU15 series by a weighted average of national indices using weigths reported in Table 1. Whenever, one of more countries are missing, lower geographical coverage level indices are determined (that is, if one country is missing for dates before January 1996, a weighted EU14 series is computed). Then the drift and level shift of these proxies are matched to the available total coverage index (see http://europa.eu.int/newcronos/suite/info/notmeth/en/theme4/ebt/ebt.htm?action=notmeth#updss for a description of the process). The process requires that the weight of the considered countries should be at minimum 60% of total EU15. For this reason, the EU15 series is available from January 1986 when United Kingdom series starts, whereas some national series are available on a longer range. Furthermore, the EU15 series is exactly EU15 series only from January 1996, when Austria

series starts. Practically, EUROSTATA is not properly considering the data availability step and is not making any estimationm, but simply some adjustments. This inconsistency produces incongruencies in the official EU15 series before 1996 where trend and seasonal patterns are referred to a different geographical area. As a result, the EU15 series is not comparable through time since after 1996 is an EU15 computed on total coverage while before 1990 is an EU15 series computed (and not estimated) on 7 countries. The inconsistency will be evident in a few steps.

The following excercise is simply an example of our approach and should not be considered as the best solution for back-calculating EU15 IPI. In fact, by using data available from the OECD and National Statistical Institutes, and data available in NewCronos on the NACE classification with 4 digit precision, a different and more efficient strategy can be designed. In this excercise we assume that the only information is given by the NewCronos national series. Given the available data, we plan to back-calculate the EU15 series from 1980. A further extension could be considered but it would be based on a very limited information set that could excessively bias the results (the coverage is of of the 31.9% from 1978 to 1980 and 6.3% from 1977 back to 1970). In order to evidence the properties and the reliability of our approach we will assume that the actually available EU15 series starts in January 1996, when all national series are available.

By using the weights of Table 1 we compute the following related indicators (they correspond to a partial measure to the needed index, or, to a measure on a different geographical area): EU6 starting in 1980, EU7 starting in 1985, EU8 available from January 1986, EU13 from January 1990 and EU14 from January 1995. Figure 2 reports the coverage over total Euro15 for the back-calculation range.

Table 2 reports the set of linear regressions fitted on the logarithmic differences with the inclusion of seasonal dummies and ARMA terms in the residuals used for the back-calculation process. In turn, EU6 is used to retropolated EU7 series, the estimated EU7 series is used as a related indicator for the EU8 series, EU13 series back-calculation uses the estimated EU8 series, EU14 series retropolation uses the estimated EU13 series and finally EU15 series reconstruction is based on the estimated EU14 series.

Figure 3 reports a the back-calculated EU15 series. Furthermore, Figures 4 and 5 reports on the range from January 1986 to December 1995, the official EU15 logarithmic differences, the retropolated EU15 series logarithmic differences and the discrepancies between the two series, respectively. Fig-

ures 6 and 7 reports the discrepancies between the levels of the official and back-calculated series and the comparison between the seasonal logarithmic differences of the official and retropolated EU15 series.

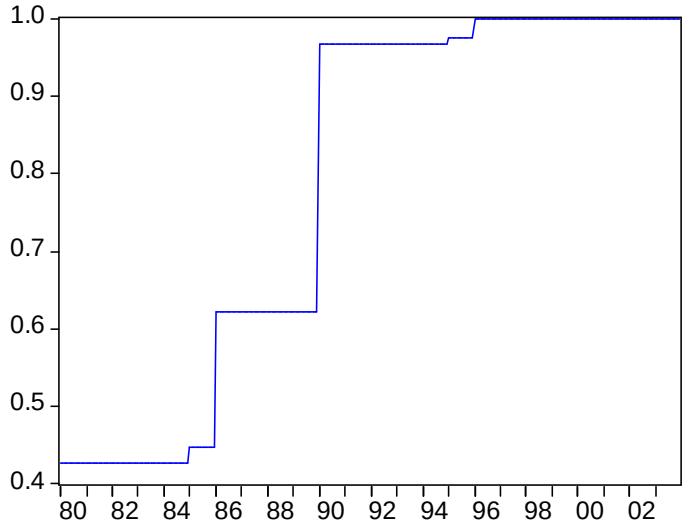

Figure 2: IPI coverage over Euro15

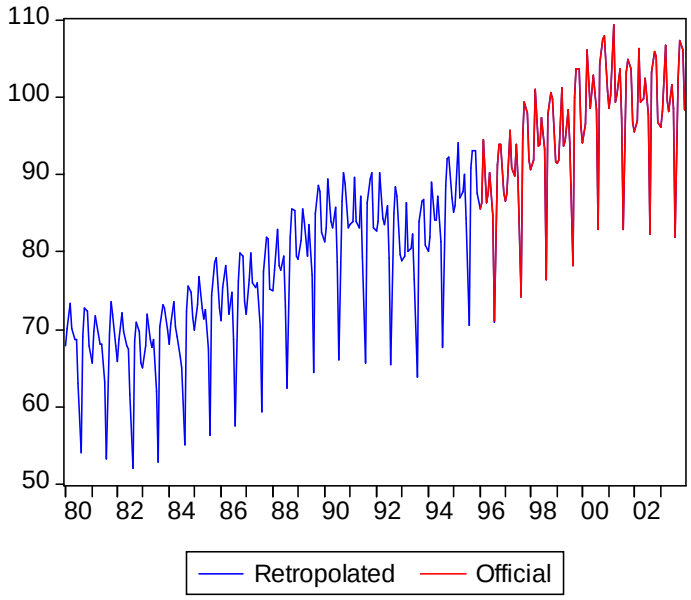

Figure 3: retropolated Euro15 series

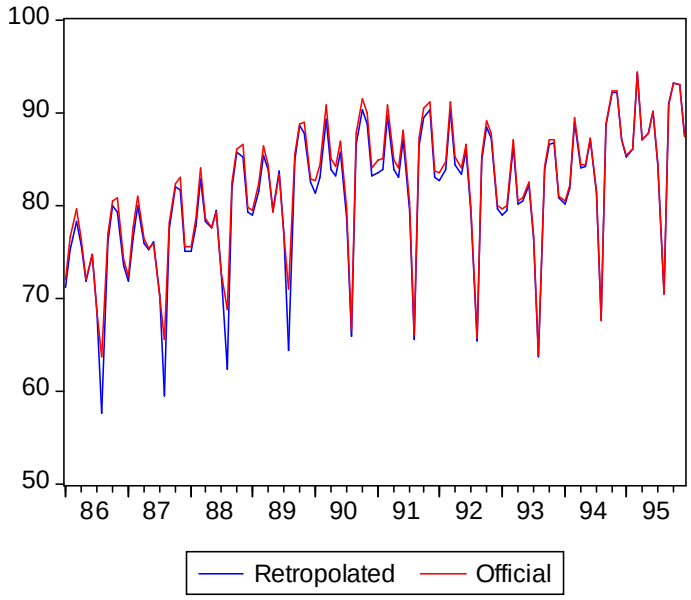

Figure 4: retropolated Euro15 and official series over the range 1986-1995

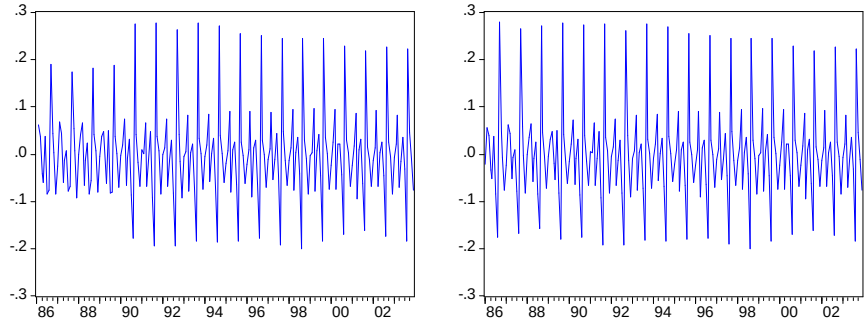

Figure 5: log-differences of official (left) and retropolated (right) Euro15 series over the range 1986-2003

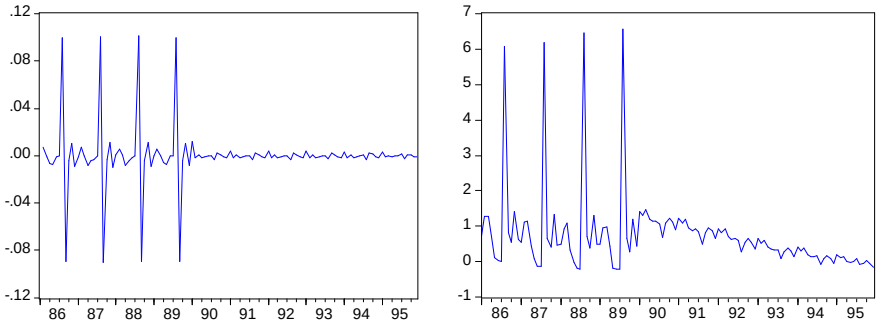

Figure 6: discrepancies between official and retropolated series over the range 1986-1995 - log-differences (left) and levels (rigth)

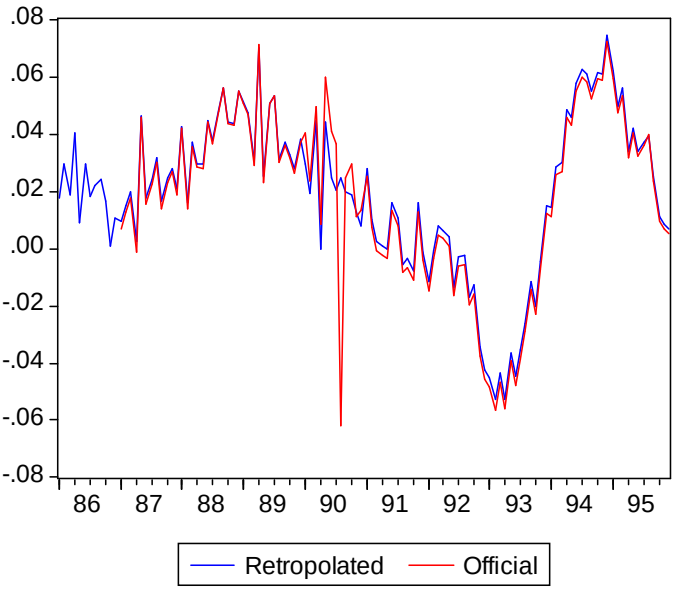

Figure 7: comparing annual log-differences over the range 1986-1995

Estimation	1985m2	1986m2	1990m4	1995m2	1996m2
range	2003m12	2003m12	2003m12	2003m12	2003m12
Dependent	Eu07	Eu08	Eu13	Eu14	Eu15
Explanatory	Eu06	Eu07	Eu8	Eu13	Eu14
β	0.97483	0.85192	0.94078	0.99565	0.99620
	0.00228	0.01033	0.01086	0.00060	0.00122
δ_0	-0.00165	-0.00398	0.00293	-0.00014	0.00104
	0.00018	0.00067	0.00044	0.00003	0.00010
δ_1					-0.00450
					0.00025
δ_2	0.00145	0.00508	-0.00707	0.00061	
	0.00062	0.00160	0.00171	0.00010	
δ_3	0.00495	0.02429			-0.00096
	0.00061	0.00165			0.00029
δ_4	-0.00197	-0.02282		0.00030	
	0.00059	0.00135		0.00008	
δ_5	0.00377	0.00666		0.00031	-0.00071
	0.00059	0.00146		0.00009	0.00021
δ_6	0.00505			0.00046	-0.00118
	0.00060			0.00009	0.00024
δ_7	-0.01128	0.00501	-0.00696	0.00052	-0.00197
	0.00060	0.00141	0.00158	0.00009	0.00021
δ_8	0.01642	0.00669	-0.11019	-0.00051	
	0.00055	0.00152	0.00189	0.00013	
δ_9		0.00582	0.09938	0.00040	-0.00072
		0.00260	0.00257	0.00018	0.00035
δ_{10}	0.00205	0.01385	0.00441	-0.00021	-0.00171
	0.00068	0.00144	0.00181	0.00009	0.00026
δ_{11}		0.00371	-0.01430		
		0.00150	0.00158		
θ_1	-0.75679	-0.61446		-0.71824	-0.75178
	0.04427	0.05406		0.07028	0.07796
θ_3		0.14879			
		0.05747			
θ_{12}		0.26810	0.39801		
		0.05619	0.07630		
φ_1			-0.64773		
			0.07987		
φ_2			-0.16358		
			0.08060		
$\overline{R}^2$	0.999	0.998	0.999	0.999	0.999
DW	1.577	2.109	2.066	1.608	1.984

Table 2: back-calculation regressions

It is evident that the discrepancies have a seasonal component in the

range 1986-1989; less evidently a trend is present in the whole range, as well as a seasonal effect in the second part of the back-calculated range. The discrepancies are due to the production process adopted by EUROSTAT: a trend and shift adjustment made on the EU8 series to chain it to the available EU13 series is not equivalent to the estimation of EU13 (and then EU15) using EU8 and a set of deterministic and stochastic components. In fact, the correction used by EUROSTAT does not recover the seasonal component of EU15 which is evidently different from the one of EU8 given the missing data of 5 countries (France and Italy included) whose weigth on EU15 is aroung the 34%. Similarly, the trend of the series is affected. As a result, the EU15 IPI series actually available is internally inconsistent and present discrepancies with respect to the true unavailable series both on the trend and on the seasonal components. Comparing the seasonal logarithmic differences one can note that the cyclical seems not to be affected, apart the inclusion of an outlier.

6 Concluding remarks

This paper presents a methodological approach for back-calculation problems, that is for the estimation of past values of relevant series by using a limited information set. We consider a general framework that includes a set of possible cases ranging from the temporal and/or spatial aggretation, the temporal and/or spatial disaggregation, the retropolation and constrained retropolation. We provide a scheme to be used for back-calculation problems and an empirical example showing the advantages of our approach compared to the one actually used by EUROSTAT in the reconstruction of the Euro 15 Industrial Production Index

References

[1] Box, G.E.P. and G.M. Jenkins (1970) Time series analysis: Forecasting and control, San Francisco: Holden-Day.

[2] Chow G. and A.L. Lin (1971), Best linear unbiased interpolation, distribution and extrapolation of time series by related series, The Review of Economics and Statistics, 53: 372-375.

[3] Chow G. and A.L. Lin (1976), Best linear unbiased estimation of missing observations in an economic time series, Journal of the American Statistical Association, 71: 719-721.

[4] Denton F.T., (1971), Adjustment of monthly or quarterly series to annual totals: an approach based on quadratic minimization, Journal of the American Statistical Association, 1971, 66, 99-102

[5] Di Fonzo, T., 2003a, Temporal disaggregation using related series: log-transformation and dynamic extension, Rivista Internazionale di Scienze Economiche e Commerciali, 50, 3, pp. 371-400.

[6] Di Fonzo, T., 2003b, Constrained retropolation of high-frequency data using related series. A simple dynamic model approach, Statistical Methods & Applications, 12, pp. 109-119.

[7] Di Fonzo, T., 2003c, Temporal disaggregation of economic time series: towards a dynamic extension, European Commission (Eurostat) Working Papers and Studies, Theme 1, General Statistics (pp. 41).

[8] Fernández R.B. (1981), A methodological note on the estimation of time series, The Review of Economics and Statistics, 63: 471-478.

[9] Litterman R.B. (1983), A random walk, Markov model for the distribution of time series, Journal of Business and Economic Statistics, 1: 169-173.

[10] Ramsey, J.B., Rothman, P., 1996. Time irreversibility and business cycle asymmetry. Journal of Money, Credit and Banking 28, 3–20.

[11] Santos Silva J.M.C. and F.N. Cardoso (2001), The Chow-Lin method using dynamic models, Economic Modelling, 18: 269-280.

[12] Stram D.O. and W.W.S. Wei (1990), Disaggregation of time series models, Journal of the Royal Statistical Society, 1990, 52, 453-467