Clustered Panel Data Models: An Efficient Approach for Nowcasting from Poor Data

Michel Mouchart [a] et Jeroen V.K. Rombouts[b]*

[a]*Institut de statistique and* [b]*CORE*
Université catholique de Louvain(B).

May 28, 2003

First draft in progress
Not to be quoted

Abstract

Nowcasting regards the inference on the present realization of random variables, on the basis of information available until a recent past. This paper proposes a modelling strategy aimed at a best use of the data for nowcasting based on panel data with severe deficiencies, namely short times series and many missing data. The basic idea consists of introducing a clustering approach into the usual panel data model specification. A case study in the field of R&D variables illustrates the proposed modelling strategy.

Keywords: Panel data, forecast, nowcast, missing data, clustering, R&D data

*The work underlying this paper stems from a research convention originated by the European Commission and Eurostat and commissioned through CAMIRE, Estadística y Análisis, SL (Luxembourg). The authors would like to thank Elisabeth Lamp and Christophe Zerr of CAMIRE, the participants of an R&D nowcasts meeting in Brussels and of the meeting of a Working Party on S&T and Innovation Statistics for constructive comments and useful questions from which this paper has substantially benefited. Moreover, the assistance of Christophe Zerr for preparing the data base has been deeply appreciated. The authors gratefully acknowledge the hospitality of CentER, at the Tilburg University, taking advantages of the privileged working conditions at the final writing of this paper.

Contents

1 Introduction

Macro-economic time series are typically available with some lags, different by variables and by countries. In particular, R&D variables are published with a delay often larger than for other macro-economic series such as GNP, unemployment or budget deficit. Moreover, designing economic policy should be based on the present state of the economy rather than on the state of the economy corresponding to the publication of macro-economic data. This raises the necessity of evaluating the present values of some series, as R&D data, on the basis of available, and delayed, data. Borrowing from the meteorological literature this problem is called "nowcasting" rather than "short-term prediction" in order to emphasize the fact that when nowcasting, the time of availability of the data is not the same for all variables, in particular for the possible predictors, and to emphasize that the horizon of prediction is "to-day" rather than "to-morrow". This paper puts forward a modelling strategy for nowcasting values from a panel of national macro-economic time series taking into account severe shortcomings in the data base, namely very short time series and many missing values. Marcelino, Stock and Watson (2001) considers a rather close problem of forecasting from panel data with severe deficiencies. The difference, with respect to our problems, are twofold: Marcelino, Stock and Watson (2001) are concerned with forecast rather than nowcast and work, among others, with monthly data rather than yearly data. Their objective is different as they want to compare the performances of several modelling strategies rather than to look for a "best" use of the available data in a specific case.

In classical panel data analysis one pools all the country data together and specifies a unique model holding for all the countries but allowing for some heterogeneity between the countries. This is typically done either by keeping a country specific intercept in the model (fixed effects) or by incorporating this heterogeneity in the innovation term (random effects). By so-doing one avoids estimating models country by country that, in the case of small samples, suffer from the problem of imprecise parameter estimates.

In this paper we propose an alternative method for pooling the countries together into one model by incorporating into a unique model clusters of countries specific to each coefficient to be estimated. This leads to a model that is both parsimonious and flexible in terms of adjusting to country specificities. This approach also allows that countries with deficient data may draw advantage from data of coefficient-similar countries. From a mean-squared error point of view, this approach may be interpreted as a trade-off between creating a bias, due to non exactly true restrictions, and reducing a variance by lowering the number of parameters and increasing the amount of data oriented toward a particular parameter. That is, we look for a flexible pooling approach that balances unwarranted restrictions of equality against unnecessary loss of degrees of freedom. Islam, Fiebig and Meade (2002) considers also a forecast model for deficient panel

data in the field of telecommunication demand. They propose a different pooling procedures that may accommodate an environment with still shorter and also unequally short time series, with no within series missing data and with an economic argument suggesting a between constancy of parameters.

A major emphasis is put on making the modelling strategy explicit in order to foster a clear interpretation of the empirical results, and on developping models, the results of which are comparable between countries and easily updatable as new data becomes available. This is done by distinguishing clearly those steps that are essentially computational and other steps that require careful examination of the intermediary results.

The paper is organised as follows. The next Section considers, within the framework of the mean-squared forecast error, the problem of specifying a conditional model for nowcasting purposes. The third section gives a detailed description of the modelling strategy. Emphasis is put on motivating each step of the strategy. The fourth section presents a case study to R&D expenditures. The work underlying this paper is motivated by a problem in the field of public R&D expenditures for the European Community countries. An application based on data from Eurostat exemplifies the suggested strategy. The last section provides some concluding remarks and some recommendations for the future use of the model.

2 Evaluating the Mean-Squared Forecast Error

A basic argument for building a forecast, or nowcast, from a conditional model rather than from a marginal model may be the conviction that the conditional model generating $y|Z$ enjoys better statistical properties (*e.g.* of structural stability) than the marginal y data generating process (DGP) and eventually allows to better accumulate the empirical knowledge on the parameters characterizing the y-DGP.

As an illustration, for the linear model, this boils down to compare two predictors for y_{n+1}, namely $\hat{\gamma}' y_{n-p}^{n}$ (where $y_{n-p}^{n} = (y_n, y_{n-1}, \ldots, y_{n-p+1})'$, a vector of available realizations) under a linear autoregressive specification of order p, or $\hat{\beta}' \hat{z}_{n+1}$ (where $\hat{z}_{n+1}$ stands for a predicted realization of z_{n+1}) under a conventional linear model specification. Note that the vector z_{n+1} may involve past or present realizations but we specifically focus the attention on conditional models that involve contemporaneous exogenous variables. In nowcasting problems, z_{n+1} may contain variables available with equal or lesser lags than the variable to be nowcasted. If $(y_i, z_i) \in R^2$ were *i.i.d.*, the predictors in the autoregressive case and the OLS-based prediction case are both equal to $\bar{y}$. Therefore, the actual issue with forecasting, or nowcasting, regards essentially the non *i.i.d.* case.

Let us now formalize these ideas and demonstrate that an adequate modelling strategy urges parsimonious models. Consider a standard linear model

$$y = Z\,\beta + \epsilon \tag{1}$$

where

1. y is an n-dimensional vector of n observations on a variable to be forecasted or nowcasted.
2. Z is a data matrix of dimension $n \times K$ of n observations on K predicting variables.
3. β is a K-dimensional parameter vector.
4. ϵ is an n-dimensional residual vector with $E(\epsilon) = 0$ and $V(\epsilon) = V(y|Z) = \sigma^2_{\epsilon|Z} I_n$.
   The notation $\sigma^2_{\epsilon|Z}$ refers to the fact that the conditional variance depends on the selection of the explanatory variables, and possibly on the values of the explanatory variables as would be the case with heteroskedasticity.

We are interested in building an inference on y_f, a value of y assumed to be generated by the same model as above (f for "future" relatively to the most recent available realisation of the variable y), namely

$$y_f = \beta' z_f + \epsilon_f \tag{2}$$

with $Cov(\epsilon_f, \epsilon) = 0$, $E(\epsilon_f) = 0$, $V(\epsilon_f) = \sigma^2_{\epsilon|Z}$. For this purpose, one may consider the bi-linear predictor

$$\hat{y}_f = \hat{\beta}' \hat{z}_f \tag{3}$$

where $\hat{\beta} = (Z'Z)^{-1} Z'y$ and $\hat{z}_f$ is some prediction for the value of z_f.

When the value of z_f is known, or corresponds to a controlled variable, one would specify $\hat{z}_f = z_f$ and the mean-squared forecast error (MSFE) is known to be equal to

$$\mathrm{E}\left[(\hat{y}_f - y_f)^2 | z_f, Z\right] = \sigma^2_{\epsilon|Z} \left[1 + z'_f (Z'Z)^{-1} z_f\right]. \tag{4}$$

In the case where there is a constant in the model:

$$Z = \begin{bmatrix} i & Z_2 \end{bmatrix} \qquad Z'Z = \begin{bmatrix} i'i & i'Z_2 \\ Z'_2 i & Z'_2 Z_2 \end{bmatrix} \qquad z_f = \begin{bmatrix} 1 \\ z_{2f} \end{bmatrix} \tag{5}$$

where $i'i = n$, $i'Z_2$ is a $(K-1)$-dimensional vector and $Z'_2 Z_2$ is a squared matrix of order $(K-1)$, it is also known that

$$\mathrm{E}\left[(\hat{y}_f - y_f)^2 | z_f, Z\right] = \sigma^2_{\epsilon|Z} \left[1 + \frac{1}{n} + (z_{2f} - \bar{z_2})' \left(Z'_2 N Z_2\right)^{-1} (z_{2f} - \bar{z_2})\right] \tag{6}$$

where $N = I_n - i(i'i)^{-1}i' = I_n - \frac{1}{n}ii'$ and $\bar{z_2} = n^{-1}Z_2'i$ is the column vector of the sample averages, which shows that the MSFE increases as z_{2f} goes further from $\bar{z_2}$ (details in the Appendix). For this case of known z_f, Danilov and Magnus (2002) performs a detailed analysis of the MSFE when the regressors are determined through pretesting procedures.

Suppose now that z_f is not anymore controlled but estimated through $\hat{z}_f$, liable to a prediction error $\hat{z}_f - z_f$. In such a case, we need more hypotheses. Let us consider the following ones:

H1: $\quad \epsilon_f \perp\!\!\!\perp \epsilon \perp\!\!\!\perp Z, z_f$

H2: $\quad \hat{z}_f = f(Z)$.

Under these two hypotheses the MSFE may be evaluated in two steps as follows:

$$\mathrm{E}\left[(\hat{\beta}'\hat{z}_f - \beta' z_f - \epsilon_f)^2 | z_f, Z\right] = \sigma^2_{\epsilon|Z}\left[1 + \hat{z}_f'(Z'Z)^{-1}\hat{z}_f + [\beta'(\hat{z}_f - z_f)]^2\right] \tag{7}$$

Integrating out the unknown value of z_f, we obtain

$$\mathrm{E}\left[(\hat{\beta}'\hat{z}_f - \beta' z_f - \epsilon_f)^2 | Z\right] = \sigma^2_{\epsilon|Z}\left\{\left(1 + \hat{z}_f'(Z'Z)^{-1}\hat{z}_f\right) + \beta'\left[\mathrm{E}(\hat{z}_f - z_f)(\hat{z}_f - z_f)'|Z\right]\beta\right\}. \tag{8}$$

or

$$\begin{aligned}\mathrm{E}\left[(\hat{y}_f - y_f)^2|Z\right] &= \sigma^2_{\epsilon|Z}\left\{\left(1 + \hat{z}_f'(Z'Z)^{-1}\hat{z}_f\right)\right. \\ &\quad \left. + \beta'\left[V(z_f|Z) + (\mathrm{E}(z_f|Z) - \hat{z}_f)\,(\mathrm{E}(z_f|Z) - \hat{z}_f)'\right]\beta\right\}. \end{aligned} \tag{9}$$

Consider now two sets of competing predictors. We want to know under which conditions one set of predictors will provide a smaller MSFE than the other one. For the sake of presentation, we concentrate the attention on the analysis of the situation leading to the MSFE in (9), *i.e.* bilinear predictor based on an OLS of the training sample, independent sampling (**H1**) and predicted values of the z variables based on the training sample (**H2**). From (9), it is seen that three aspects compete for a "best" forecast. Firstly the conditional variance $\sigma^2_{\epsilon|Z}$, which would tend to favour a larger set of predictors, secondly, the sampling error when estimating the coefficients from the training sample, *i.e.* $\hat{z}_f'(Z'Z)^{-1}\hat{z}_f$ and thirdly the sampling error when forecasting the predictors themselves, captured by the term $\beta'\left[V(z_f|Z) + (\mathrm{E}(z_f|Z) - \hat{z}_f)\,(\mathrm{E}(z_f|Z) - \hat{z}_f)'\right]\beta$.

The first aspect may be illustrated by considering whether to add, or not, a further predictor. Let us partition Z_2 in (25) into $[Z_{21} \;\; Z_{22}]$. When y, Z_2 are jointly normal, we have

$$V(y|Z_{21}) - V(y|Z_{21}, Z_{22}) = \mathrm{Cov}(y, Z_{22}|Z_{21})'V(Z_{22}|Z_{21})^{-1}\mathrm{Cov}(Z_{22}, y|Z_{21}) \tag{10}$$

This difference is always non-negative definite and suggests that a larger decrease in the residual variance is obtained when the covariance between y and Z_{22}, conditionally on Z_{21} is large, and the conditional variance of Z_{22} given Z_{21} is small. Evaluating a similar difference for the other two terms, and for the combined impact of them, as described in (9) is quite cumbersome.

Furthermore, the choice of the regressors should also take into account the quality of the data actually available for a potential regressor. To illustrate this aspect, suppose that a predictor Z is not available for a potential of N observations but only n of those data are available. A simple case is the following: denote the actually available part of Z by Z_0 and write $Z_0 = SZ$ where S is an $n \times N$ random selection matrix. Let us assume a rather natural missingness condition, namely

H3: $S \perp\!\!\!\perp z_f, Z$.

Under this condition the term $V(z_f|Z)$ becomes

$$V(z_f|SZ) = E\left[V(z_f|Z)|SZ\right] + V\left[E(z_f|Z)|SZ\right]. \tag{11}$$

(Indeed: $S \perp\!\!\!\perp z_f, Z \Rightarrow z_f \perp\!\!\!\perp S|Z \Rightarrow z_f \perp\!\!\!\perp S, Z|Z \Rightarrow z_f \perp\!\!\!\perp SZ|Z$, for details see *e.g.* Florens *et al.*, 1990, Section 2.2.) Furthermore, the predictor $\hat{z}_f$ is not anymore a function of Z but a function of Z_0 and the estimation will be based on Sy, SZ rather than on y and Z, eventually deteriorating the MSE of $\hat{\beta}$ unless some imputation device is used for some missing data, a possibility that creates further complexity in the evaluation of the MSFE. Therefore, $V(z_f|SZ)$ is unaffected by the missingness, *i.e.* $V(z_f|SZ) = V(z_f)$ in the independent case when $z_f \perp\!\!\!\perp Z$ holds along with **H3**, but in the dependent case $V(z_f|SZ) \geq V(z_f|Z)$ on SZ-average.

We conclude that the evaluation of MSFE ($\hat{\beta}'\hat{z}$) is substantially more complicated than MSFE ($\hat{\beta}'z$) and, that the inclusion of an additional regressor in the model has mixed effects on MSFE ($\hat{\beta}'\hat{z}$) if only because of the forecasting error of the additional $\hat{z}$. In case of missing data this additional predictor may behave poorly because of its missingness pattern and more dramatically, the missingness pattern of the additional predictor may spillover on the performance of the old predictors by impeding the use of data available for y and Z_1 but not for Z_2. It should be clear that the previous arguments call for extremely parsimonious models, an important message of practical interest for this paper.

3 Modelling Strategy

3.1 Basic Issues

The context of nowcasting R&D variables from poor panel data calls for specific desirable properties fo the modelling strategy, namely a proper account of the data quality, an easy updatability and an appropriate dynamic specification.

A proper account of the data quality

In order to face the particular data deficiencies, we suggest to combine two strategic options. *Firstly*, we keep the number of exogenous variables minimal. Indeed, nowcasts conditional on

uncontrolled explanatory variables require nowcasting those explanatory variables. Most of the time the missingness pattern for possible predicting (or, explanatory) variables is different among variables and from country to country. Therefore, increasing the number of predicting variables implies a drastic decrease in the number of complete observations and eventually leads nowcasting to become widely stepwise, in the sense of making nowcasts on the variable of interest dependent on nowcasting more predicting variables. *Secondly*, we use somewhat subtle panel data techniques. The basic idea is the following: Separate regressions for each country leads to extremely unreliable estimation because the number of degrees of freedom is never large and because some countries suffer at the same time from many missing data and large residual variances. The standard models for pooling panel data (see *e.g.* Baltagi, 1995) typically involve equality constraints among parameters corresponding to different countries: those constraints, once not carefully controlled, may deteriorate, rather than improve, the statistical quality of the final results.

An easy updatability

The updating of the nowcasts should be made fast and easy because, as frequently happens in nowcasting problems, data are flowing continously with variable lags. When new data become available it should accordingly be straightforward to incorporate this new information into the model and to update the nowcasts.

As a matter of fact, the actual difficulty in model updating concerns the judgemental aspects of the model specification: the purely computational steps are indeed much alleviated by the recent developments of statistical and econometric packages. Therefore, easy updatability crucially depends on making the process of model building explicit, on clearly separating the steps of judgment from those of computations and on keeping the need for judgemental ingredients minimal. These requirements ground the model building process to be exposited.

An appropriate dynamic specification

From an empirical point of view, most macro-economic series under analysis are not stationary; series in first differences are typically closer to stationarity than absolute levels, a well known issue in the litterature on co-integration, see *e.g.* Hamilton(1994). From a structural-economic point of view, the R&D data are not likely to be mutually independent (even conditionally on explanatory variables). In particular, because the feasibility of actual decisions crucially depends on (the level of) previous decisions. For instance, governments often tend to justify budgetary decisions in terms of rate of growth, even if the final balance of the budget is operated in absolute levels. It is accordingly unpalatable to stick to a purely *i.i.d.* framework: a minimum of dynamics should be striven for, taking nevertheless into account the very short time span of the series.

Furthermore, the countries under analysis are characterized by widely different sizes (*e.g.* USA and Belgium). Thus the variables of interest are to be expected of different orders of magnitude. Taking into account these different aspects, we develop a model where the variables are taken in first log differences ($\approx$ percentage differences), so as to ensure comparability of different sizes of the countries, to provide a workable amount of dynamics and to avoid artificially low residual variances (as compared with models in level). Notice also that differencing renders the models' forecasts robust to shifts in the coefficients (see *e.g.* Hendry and Clements, 2001, p. 9) and allows to use ordinary least squares estimation procedure.

To conclude, the main issue is a "best use" of the available data. It has to be accepted that poor data are unlikely to produce rich models. Too sophisticated a model typically "overfits" the sample and usually deteriorates the quality of the forecasts/nowcasts performance because it is not robust with respect to structural changes whereas, at the same time, it gives an unduly optimistic feeling through artificially low residual variances.

3.2 Dealing with missing data and ruptures

Let us consider two important data deficiencies. Firstly, the data can be missing for some periods and secondly there maybe abrupt changes in the time series due to structural economic changes or to redefinitions of macro-economic variables, just to give two examples.

Imputation of within-sample missing values.
In general, time series in levels on macro-economic or on policy time series are rather smooth. This feature suggests a simple method for the imputation of the missing values, based on a nearest neighbour method: draw a line between the two nearest points covering the missing value(s) and interpolate.

This method uses the time dimension of the data to provide a simple and obvious solution to the missing value problem. Indeed, the treatment of missing values by means of this mechanical and atheoretical rule is motivated by the nature of the data. For policy variables and macroeconomic time series, the occurence of missing values can be reasonably regarded as independent of the realized value of the aggregates. This contrasts with the unit non response of microeconomic data (individuals, households, firms, ...), where the probability of scoring a missing value may be sensitive to different realized values.

It should be stressed that this method does not apply to extrapolation, *i.e.* imputation of missing values outside of the time span of the available data. Extrapolation with this method, meaning in this context backcasting, could be inadequate because it unwarrantedly supposes that

the time trend goes back into the past. For data missing at the extremes of the sample, we make no imputation. We perform no backcasting of the data missing at the beginning of the sample. For data missing at the end of the sample, we use our model of nowcasting for the variables of interest, in this case the R&D variable, and another method, explained below, for the predictors. Thus, after performing this imputation procedure, we are left, for each variable and country, with data missing at the ends of the sample only.

Control of the ruptures.

In general, the forecasting performance of a predictive model crucially depend on an adequate account of the eventual ruptures in the series under analysis. Given the short time span of the series, and the substantial presence of missing data, we look for sturdy and simple procedures for detecting (significant) ruptures in the series.

For each individual time series we proceed as follows:

1. Adjust, by OLS, a simple quadratic trend model:

$$y_t = \alpha_1 + \alpha_2\, t + \alpha_3\, t^2 + \epsilon_t \tag{12}$$

and define the corresponding residuals:

$$e_t = y_t - \hat{\alpha}_1 - \hat{\alpha}_2\, t - \hat{\alpha}_3\, t^2. \tag{13}$$

Note that the quadratic trend is added to gain some protection against possibly explosive or damping series.

2. Control whether the first differences in the sequence of the residual terms suggest some ruptures. More specifically, define

$$e_t - e_{t-1} = d_t' e \tag{14}$$

where d_t is the vector transforming the vector of LS residuals, e, into the first difference of its t-th component.

A simple diagnosis of rupture is then obtained by considering whether $|e_t - e_{t-1}|$ is significantly different from its expected value, namely 0. For this purpose we evaluate

$$\begin{aligned} \hat{V}(e) &= s^2 M \\ \hat{v}_t &= \hat{V}(e_t - e_{t-1}) = s^2\, d_t' M d_t \end{aligned} \tag{15}$$

where M is the symmetric idempotent matrix that projects onto the complement of the space generated by the explanatory variables.

3. Construct a rupture indicator in the form of a binary diagnosis as follows

$$r_t = \mathbb{1}_{\{|e_t - e_{t-1}| > 2\sqrt{v_t}\}}. \tag{16}$$

When dealing with the series transformed into first log-differences, we shall consider that the first difference corresponding to a significant rupture ($r_t = 1$) of the series in arithmetic terms is likely to be outlying.

3.3 Selection of the variables and country-wise estimation

For each variable to be nowcasted, we examine country by country the correlation structure with all reasonable candidates of prediction variables and try to detect common patterns shared by all or most countries. More specifically we first look, within each country, at the absolute value of the correlation coefficient to decide whether it is likely to be a "good" regressor. Next we compare between all countries the signs and the absolute values of the correlation coefficients to check for common parameters. Note that each correlation is computed on the basis of the available pairs of observations, eventually accepting a different number of observations for the different correlations. Then, we select a set of explanatory variables for each variable to be nowcasted.

In a next step we continue to treat each variable to be nowcasted separately and estimate country by country regressions based on different choices of the explanatory variables selected in the previous step. We then look for specifications the coefficients of which are as stable and/or significant among countries as possible.

3.4 Pooling and clustering

The regressions in the previous step are typically based on a small number of observations. As we will see in the application for some countries the number of observations may be as small as 5. Nowcasting on so small samples would eventually be unreliable. It is accordingly important to try to spill over the information available for one country to another country. Crude pooling, *i.e.* a unique model with identical parameters for all countries, would be bound to dramatic specification errors. Thus an intermediate solution consists of a group-wise pooling for each regression coefficient seperately.

We start with the following model:

$$y_{it} = \beta_{1,i} + \beta_{2,i}\, x_{2,it} + \beta_{3,i}\, x_{3,it} + \ldots + \beta_{K,i}\, x_{K,it} + \epsilon_{it} \tag{17}$$

where $i = 1, \ldots, n$ and $t = 1, \ldots, T_i$. This is a linear model that has common exogenous variables with country specific parameters and can be estimated by Ordinary Least Squares (OLS). For each

regression coefficients of the selected specification, $\beta_{k,i} \quad k = 0, \ldots, K$, we cluster the countries into G_k groups. The clusters may be defined either by means of the quantiles of the empirical distribution of the estimates or by dividing the range of the estimates. More formally the model can be written as follows.

$$y_{it} \quad = \quad \beta_{1,g_1(i)} + \beta_{2,g_2(i)}\, x_{2,it} + \beta_{3,g_3(i)}\, x_{3,it} + \ldots + \beta_{K,g_K(i)}\, x_{K,it} + \epsilon_{it} \tag{18}$$

$$\begin{aligned} g_1(i) \in \{1, \ldots, G_1\} \quad &: \quad \text{group of country } i \text{ for } \beta_1 \\ g_2(i) \in \{1, \ldots, G_2\} \quad &: \quad \text{group of country } i \text{ for } \beta_2 \\ &\vdots \\ g_K(i) \in \{1, \ldots, G_K\} \quad &: \quad \text{group of country } i \text{ for } \beta_K \end{aligned}$$

This results in a model that is still flexible ($\prod_{k=1}^{K} G_k$ potentially different countries) but has at most ($\sum_{k=1}^{K} G_k$) parameters whereas a crude pooling (all countries identical) would involve (K) coefficients, standard panel (country specific intercepts and identical slopes) $n+K-1$ coefficients, and models different for each country n K coefficients. In the case study we will specify $G_k \leq 4$. By so-doing we specify 4 different values for each coefficient. Thus, with K coefficients and n countries we obtain an overall model with 4 K coefficients. Note however that those 4 K coefficients provide the possibility of recognising 4^K different types of countries, a very flexible specification indeed. In some cases, the homogoneity of the estimates may lead to less than 4 groups.

Equation (18) provides a unique model across countries that may be written in the usual form (1) (details given in Appendix I). Thus, under an *i.i.d.* specification for the residuals, all the coefficients $\beta_{k,g_2(i)}$ may be estimated by OLS at once.

3.5 How to nowcast?

Given that the models are built in log-differences:

$$\Delta \ln y_{it} \quad = \quad \beta_{1,g_1(i)} + \beta_{2,g_2(i)}\, \Delta \ln x_{2,it} + \beta_{3,g_3(i)}\, \Delta \ln x_{3,it} + \ldots + \beta_{K,g_k(i)}\, \Delta \ln x_{K,it} + \epsilon_{it}$$

some care has to be taken for transferring nowcasts to the arithmetic scale of the variables. Because of the scarcity of the data, no account for the nonlinearity of the exp-log transformation will be taken. Let us denote by $\hat{Y}_{i,T_i+h}$, the nowcasted value of endogenous variable for the i-th country, h periods ahead from the sample available up to time T_i. After some simple algeabraic manipulations, we obtain the following nowcasters (for the sake of exposition we only consider one explanatory variable):

Nowcast at time $T_i + 1$

$$\begin{aligned} \hat{y}_{i,T_i+1} &= y_{i,T_i} \exp\left(\hat{\beta}_{1,g_1(i)} + \hat{\beta}_{2,g_2(i)} \Delta \ln x_{i,T_i+1})\right) \\ &= y_{i,T_i} e^{\hat{\beta}_{1,g_1(i)}} \left[\frac{\hat{x}_{i,T_i+1}}{x_{i,T_i}}\right]^{\hat{\beta}_{2,g_2(i)}} \end{aligned} \tag{19}$$

Nowcast at time $T_i + h$

$$\begin{aligned} \hat{y}_{i,T_i+h} &= \exp\left(\hat{\gamma}_{g_1(i)} + \hat{\beta}_{g_2(i)} \hat{E}[\Delta \ln x_{i,T_i+h}|\mathcal{I}_{T_i}]\right) \times \hat{E}[y_{i,T_i+h-1}|\mathcal{I}_{T_i}] \\ &= e^{\hat{\gamma}_{g_1(i)}} \left\{\hat{E}\left[\frac{x_{i,T_i+h}}{x_{i,T+h-1}}|\mathcal{I}_{T_i}\right]\right\}^{\hat{\beta}_{g_2(i)}} \times \hat{E}[y_{i,T_i+h-1}|\mathcal{I}_{T_i}] \end{aligned} \tag{20}$$

where $\mathcal{I}_T$ is the information set up to time T. As already pointed out for the nowcasts we need to estimate $E[\Delta \ln X_{i,T+h}|\mathcal{I}_T]$.

The nowcasts of the explanatory variables are done by using the best of the following four models

$$\begin{aligned} x_t &= \gamma_1 + \epsilon_t \\ x_t &= \gamma_1 + \gamma_2\, t + \epsilon_t \\ x_t &= \gamma_1 + \gamma_2\, t + \gamma_3\, t^2 + \epsilon_t \\ x_t &= \gamma_1 + \gamma_2 x_{t-1} + \epsilon_t \end{aligned}$$

where best means the model that minimizes

$$\frac{|e_{\,T}| + |e_{\,T-1}| + |e_{\,T-2}|}{R^2}, \tag{21}$$

i.e. a model is deemed to be "good" if the sum of the last three residuals (in absolute values) is small and if the fit is good.

3.6 Model validation

The poor quality of the available data also raises substantial problems for validating a proposed methodology. Indeed most hypotheses involved in the model building are not amenable to formal testing. The validity of the proposed models rests largely on a careful examination of two aspects.

First, the models should have an acceptable economic interpretation. More specifically, the model should be partly, but not completely, structural; *i.e.* partly because a pure reduced form approach is likely to be unstable in time, due to structural changes, and among countries; but

not completely structural because a purely structural approach would require much a richer data base. Thus a compromise is required. Furthermore, economic meaningfulness reinforces a better comparability among countries. In the case study, a particular attention is paid to exogeneity considerations.

Second, the nowcasting performance should be good, or at least acceptable, for every country. A particular attention should be paid to the following issues:
i) Evaluate the adopted clustering of parameters by controlling the structures of the residuals country-by-country and comparing these structures between an unreduced model and the final clustered model.
ii) Control in the final model whether the flexible constraints approach produces reasonably precise estimations (in particular the standard errors and the t-statistics).
iii) Control the fit of the final model with a particular attention to the right end of the sample.
iv) Control the nowcasting quality of the exogneous variables, in particular, the quality of the fit at the right end of the sample and the overall fit.

From a pragmatic point of view, the poor data quality suggests that only visual checkings rather than formal tests are workable but that those checks should be made carefully.

4 A Case Study

4.1 Introduction

This section presents a case study that motivated the development of the modelling strategy just sketched. The general framework of interest is the nowcast of macro-economic variables related to the field of R&D (Research and Development) for 18 countries (14 of the European Community, along with Iceland, Japan, Norway and USA) and is the object of Mouchart and Rombouts (2003). The presentation in this section concerns only the case of the so-called "Total Government Budget Appropriations or Outlays for R&D " listed as GBAORD in the Eurostat data system. This Section makes the proposed modelling strategy more explicit for a particular case and illustrates numerically the main steps.

4.2 Correlations structures

We start by surveying the economic literature bearing on R&D (Mouchart and Rombouts, 2003). From this examination, we first consider the macro-economic implications of the micro-economic analysis of R&D and innovation variables. Several issues are of particular concern. *Firstly*, at the aggregate level, we determine which variables are plausibly exogenous: in a simultaneous equation approach, we look for a reduced form equation rather than a purely (multi-equation) structural

form model. Indeed, an economically meaningful explanatory variable is more likely to be associated with a structurally stable effect than a variable selected on the basis of a purely random sampling effect of statistical association, as would be obtained, for example, from a stepwise regression algorithm in a data-mining environment. *Secondly*, we try to evaluate which of the exogenous variables are likely to characterize the economies of most countries under analysis. In view of the conclusions of Section 2 above, we also pay a particular attention to data availability: the effect of a poorly available variable can only be measured with poor statistical precision and eventually deteriorates the nowcasting performance, as compared with a model where such variables would be deleted.

From the economic literature, we draw a list of five variables that may be reasonably considered as exogenous, namely growth in real GDP (denoted GDP for simplicity), general government net balance as a % of GDP ($eb060$), growth in General government consolidated gross debt as a % of GDP ($dleb070$), growth in total employment-Fulltime ($dlemp$) and growth in the employment indicator ($dlemp64$). Real means that the data are first divided by the country specific deflator. We select a very small number, namely one or two, for nowcasting purposes. This selection is based on Table 1 of correlations between the variable to be nowcasted, namely GBAORD, and each of the potentially reasonable regressors. Table 1 also gives (in parentheses) the sample sizes corresponding to each correlation: a quick glance at those numbers gives a first idea of the magnitude of the missing data issue.

Table 2, below, summarises the results of Table 1 in terms of the between country stability of absolute values and of signs.

4.3 Country by country regressions

From Table 2, we propose in Table 3 different regressions to further explore the data at hand, on a country by country basis. This selection takes also into account the data availability.

We now have to examine the estimations corresponding to Table 3, namely seven regressions for each of the eighteen countries. These results are given extensively in Mouchart and Rombouts (2003) and, because of space limitations, are summarized as follows.

1. In general there is not a severe missing data problem with the GBAORD variable. The problem lies with the availability of explanatory variables except GDP growth.
2. Often, GDP is a good predictor but its quality is unstable over the different countries.
3. The other explanatory variables are either bad or unstable.
4. Some countries, Greece for example, seem to have no potential candidates as regressors.

Table 1: GBAORD Correlations

Country	GDP	$eb060$	$dleb070$	$dlemp$	$dlemp64$
Belgium	-0.038 (20)	0.478 (11)	-0.211 (10)	-0.305 (19)	0.038 (10)
Germany	0.248 (9)	0.189 (9)	-0.029 (9)	0.451 (9)	0.644 (8)
Denmark	0.338 (17)	0.173 (8)	-0.811 (6)	0.184 (17)	0.818 (7)
Spain	0.402 (19)	0.658 (4)	-0.757 (9)	0.333 (19)	0.799 (9)
Greece	0.259 (20)	0.324 (11)	-0.131 (10)	0.040 (16)	0.025 (10)
France	0.202 (20)	0.720 (11)	-0.759 (10)	0.513 (17)	0.746 (10)
Finland	-0.014 (20)	-0.001 (11)	0.105 (10)	0.135 (20)	-0.069 (10)
Italy	0.246 (18)	0.375 (9)	-0.525 (8)	0.361 (18)	0.592 (5)
Ireland	0.328 (10)	0.207 (10)	-0.263 (10)	0.204 (10)	-0.098 (10)
Netherlands	0.569 (19)	0.361 (10)	-0.695 (9)	0.789 (12)	0.591 (9)
Portugal	0.284 (12)	0.084 (11)	-0.695 (10)	0.302 (12)	0.333 (10)
Sweden	-0.048 (17)	0.012 (5)	0.589 (3)	0.100 (10)	0.028 (7)
UK	-0.398 (20)	0.100 (11)	0.036 (10)	0.121 (6)	0.274 (10)
Iceland	0.451 (9)	0.094 (9)	-0.168 (9)	NA	NA
Norway	-0.144 (18)	-0.376 (8)	-0.874 (4)	NA	NA
Austria	0.236 (15)	-0.021 (11)	0.151 (10)	0.379 (12)	0.495 (6)
US	0.408 (18)	-0.228 (18)	0.354 (18)	NA	NA
Japan	0.185 (12)	-0.420 (12)	0.310 (12)	NA	NA

NA means Not Available.

Table 2: **Interpretation of GBAORD correlations**

variables	stability of absolute values	stability of signs
GDP	ok except BE, FI, SE, NO	ok except UK, NO
$eb060$	ok except FI, PT, IS, AT	ok except NO, US, JP
$dleb070$	ok except DE, UK, FI	ok except SE, US, JP
$dlemp$	ok except GR	ok except BE
$dlemp64$	ok except B, GR, FI, IE, SE	ok

ok in the stability of absolute values means not bad. ok for the stability of signs means either that all correlations have the same sign or that values for those countries different from the most frequent signs are small.

Table 3: **GBAORD: Potential regressions**

Model	GDP	$eb060$	$dleb070$	$dlemp$	$dlemp64$
1	x				
2	x	x			
3	x		x		
4	x	x		x	
5	x		x	x	
6	x	x			x
7	x		x		x

5. Some countries, for example Norway and Iceland, have barely any results available due to the missingness of the data.

6. US and Japan have a lot of missing values for the explanatory variables. For comparable models they have similar coefficients.

Given these conclusions, we decide to push further model 1 (GBAORD nowcasted by GDP growth). It should be mentioned that when nowcasting other R&D variables this step of the modelling strategy was rather intricate in view of the missing data pattern of other regressors. Even for the GBAORD model, we also have explored other models, namely models 2 and 4 but these explorations systematically convinced us to focus the attention on model 1.

4.4 Pooled regressions

We now explore three "pooled" models with GDP growth as the only explanatory variable in order to investigate whether somehow we can find clusters of coefficients. The three models are:

$$\text{Model 1:} \quad GBAORD_{it} = \beta_{1,i} + \beta_{2,i}\, GDP_{it} + \epsilon_{it} \tag{22}$$

$$\text{Model 2:} \quad GBAORD_{it} = \beta_1 + \beta_{2,i}\, GDP_{it} + \epsilon_{it} \tag{23}$$

$$\text{Model 3:} \quad GBAORD_{it} = \beta_{1,i} + \beta\, GDP_{it} + \epsilon_{it} \quad i = 1, \ldots, 18\,;\, t = 1, \ldots, T_i \tag{24}$$

The estimation results for Equation (22), (23) and (24) can be found in Appendix 2 in Tables 8 to 10.

We have to examine carefully the estimations of the 74 (=36+19+19) parameters produced by these three models based on a sample size of $n = 282$. A overall look suggests the following remarks: (i) Both the constants ($\beta_{1,i}$) and the slopes ($\beta_{2,i}$) are not likely to be close together

among countries. On this ground, model 2 and model 3 embody unjustified restrictions ($\beta_{1,i} = \beta_1$ in model 2 and $\beta_{2,i} = \beta_2$ in model 3). (ii) The sign and t-statistics of the estimation of $\beta_{1,i}$ (in models 1 and 2) and the estimation of $\beta_{2,i}$ (in models 1 and 3) suggest that these three models are not far from each other, after making due allowance for unjustified restrictions. (iii) In spite of a substantial degree of coherence between the estimations, none of these models should be considered as a satisfactory final model. In particular, the estimated standard errors are mostly large, except in four to six cases in each model. The overall fit, measured by the R^2's do not display drastic differences among the models but the adjusted $\bar{R}^2$'s stress the role of the degrees of freedom.

As a consequence, we now look for a pooling model more flexible that Model 2 or 3 but consuming less degrees of freedom than Model 1. As shown in (18), the basic idea is to look for a country clustering separate for each parameter. From the results of Model 1, Table 4, displays these two clusterings obtained as follows. We distribute the coefficients of the slopes and constants into four groups according to their value. For the sake of illustration, we first evaluted the three quartiles and therafter made some marginal adjustments in order to obtain more homogenous groups. Mouchart and Rombouts (2003) consider other clusterings.

Table 4: **GBAORD: Regroupment of countries**

Group	I	II	III	IV
Constant	IT, FR, DE NL	UK, US, IS DK, SE, AT	ES, BE, IE GR, JP	FI, NO, PT
Slope	NO, SE, BE FI	UK, JP, DE IE, US	AT, NL, GR PT, FR	DK, ES, IT IS

4.5 Validation of the Final Model

The estimation results corresponding to the regroupments of Table 4 can be found in Appendix 2, Table 11. The coefficients for D_2 and X_1 are not significant and D_2 is close to zero. In the final model we will leave out this dummy. We nevertheless do not eliminate X_1 in view of the nowcasting purpose of the model. In general, it is not advisable to follow the rather frequent use of deleting variables the coefficients of which are not formally significant at a given level. Such a practice would tend to produce undesirable clustering of countries.

We obtain the final model, the results of which are summarized in Table 5. Notice that the overall fit of Table 5 is close that of the unconstrained model 1 (Table 9 in the Appendix) but the adjusted $\bar{R}^2$ has been improved by the clustering procedure. Working in first log-differences

eventually produces regular overall fits. Graphs 1 to 3 show that the implied fit per country in level is pretty good. In other words, the clustering modelling strategy succeeds in mimicking rather well the past behaviour of the country GBAORD series. Furthermore, Mouchart and Rombouts (2003) also compares, country-wise, the results of the unconstrained model 1 with the residuals of the clustered model and show that the pattern of residuals is barely affected by the constraints incorporated in the clustered models.

Table 5: **GBAORD final regression**

Variable	Coefficient	stand. dev.	t-stat	P-value
D1	-.0243	.0105	-2.299	[.022]
D3	.0269	.9561E-02	2.814	[.005]
D4	.0556	.0133	4.176	[.000]
X1	-.2395	.3219	-.744	[.457]
X2	.4738	.2446	1.936	[.054]
X3	1.4615	.3405	4.291	[.000]
X4	2.9906	.3470	8.616	[.000]

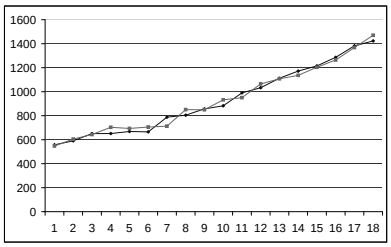
$X_i = D_i \times GDP$ where D_i is a dummy defined in Table 4. The Durbin-Watson test statistic is 1.98, $R^2 = .255$ and $\bar{R}^2 = .239$. $n = 282$

Thus the final model, with only 7 parameters instead of 19 or 38 as in models 1,2 or 3, compares favorably, in terms of the quality of fit and of the use of degrees of freedom, thanks to its flexibility. Indeed, Table 6 regroups the countries with identical coefficients and shows that the 7 coefficients of the final model distribute the 18 countries into 13 different types.

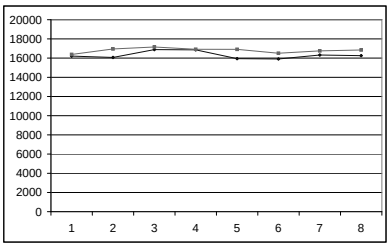
Table 6: **GBAORD: Regroupment of similar countries**

	D1	D2	D3	D4
X1		SE	BE	FI, NO
X2	DE	UK, US	IE, JP	
X3	NL, FR	AT	GR	PT
X4	IT	IS, DK	ES	

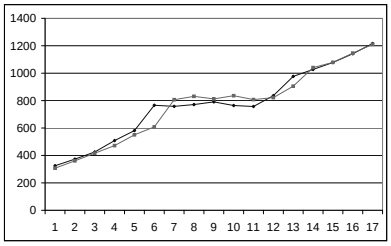
Equations (22) and (23) are models containing 19 coefficients leaving the possibility for 18 different types of countries. The model in (24) contains 36 coefficients also leaving the possibility for 18 different types of countries. The clustered model we propose contains only 7 parameters but still 16 different types of countries are possible, 13 of which are effectively used.



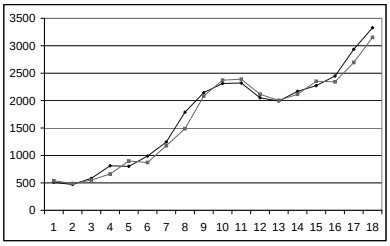
(a) BE



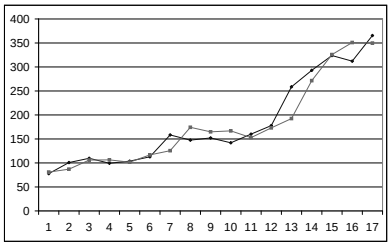
(b) DE



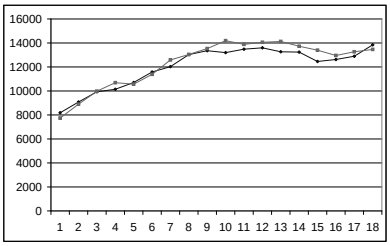
(c) DK



(d) ES

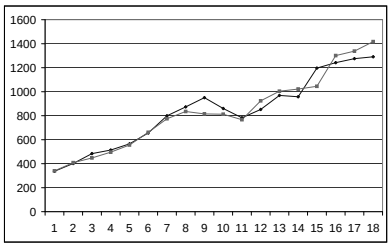


(e) GR

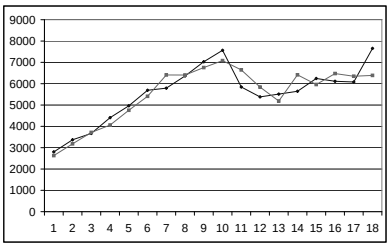


(f) FR

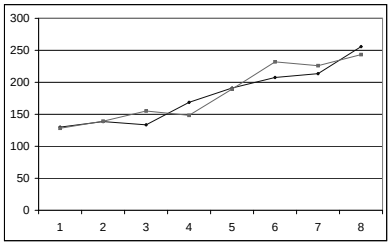
Figure 1: GBAORD fit1



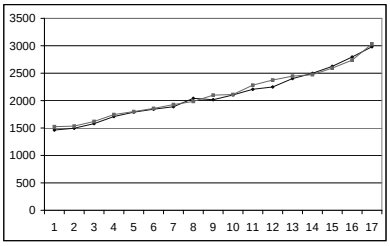
(a) FI



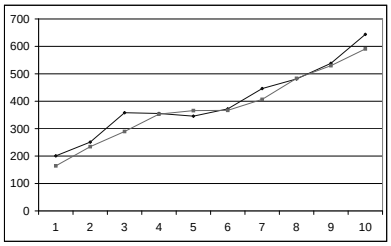
(b) IT



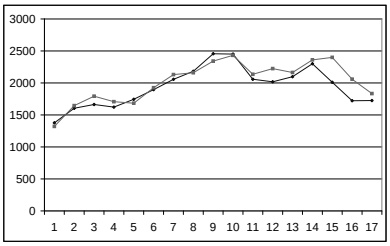
(c) IE



(d) NL

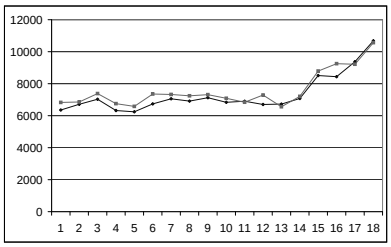


(e) PT

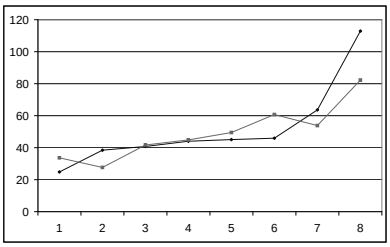


(f) SE

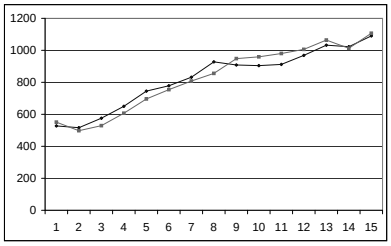
Figure 2: GBAORD fit2



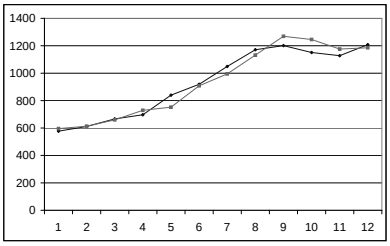
(a) UK



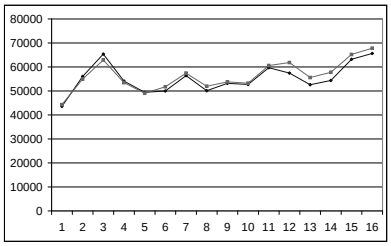
(b) IS



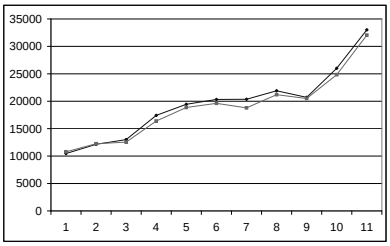
(c) NO



(d) AT



(e) US



(f) JP

Figure 3: GBAORD fit3

4.6 Nowcasts

Table 7 presents the nowcasts in nominal terms. It calls for several comments: (i) These nowcasts are first built from data in real terms, using Equations (19) and (20), and thereafter transformed into nominal terms, using the country specific price deflator always available until the current period (*i.e.* nowcasting is never necessary for the price deflator). Thus actual nowcasts depends not only on the rate of growth of the GDP but also on the price deflator. (ii) Italicized entries refer to data available at the time of nowcasting whereas unitalicized items refer to actually nowcast values. (iii) At the time of nowcasting, the predictor GDP is available until 2002 for all countries. Therefore, nowcasting the predictor is unnecessary. For other variables than GBAORD, Mouchart and Rombouts (2003) report cases where nowcasting the predictor is necessary, at least, for some countries. (iv) Even though the data on the predictor are available until 2002, the availability of GBAORD varies substantially among countries. This feature impedes to supervise the modelling strategy. Indeed, the fact that for several countries GBAORD should be nowcasted as early as in 1999 makes problematic to construct the model with data available until 1997, nowcast for 1998 and compare with the realization on 1998 because such a trial would have faced a dramatic decrease of the usable data.

Table 7: **GBAORD nowcasts**

Country	1999	2000	2001	2002
BE	*1382.1*	*1423.2*	1492.3	1560.3
DE	*16322.3*	*16253.0*	16107.9	16010.6
DK	*1216.4*	1275.8	1322.7	1366.2
ES	*3328.1*	3685.7	4192.2	5210.2
GR	*365.5*	423.3	499.9	591.3
FR	*12891.8*	*13842.1*	14073.4	14257.0
FI	*1275.2*	*1290.6*	1392.3	1486.3
IT	*6079.4*	*7656,73*	8082.5	8408.6
IE	*255.7*	288.4	320.8	350.1
NL	*2982.2*	3049.6	3074.6	3120.5
PT	*643.8*	681.7	715.0	735.9
SE	*1724.8*	1802.0	1673.0	1097.7
UK	*9373.7*	*10680.9*	10834.2	11339.8
IS	*112.9*	101.8	148.2	163.9
NO	*1090,03*	1365.1	1474.7	1561.7
AT	1286.4	1366.3	1467.0	1572.2
US	71335.6	85829.7	90948.5	98818.8
JP	*26020.5*	*33016.6*	32799.8	32816.6

Italicized data correspond to years with actually available data.

5 Concluding Remarks

Let us conclude this paper by evaluating the achieved contributions. We comment on the design of the modelling strategy, on the general meaning of the final results and on the interpretation of the empirical results for the case under study.

1. *An explicit modelling strategy*

This report presents a modelling strategy totally made explicit, with two objectives: Firstly, to help reproducing and updating the model construction and secondly to clearly display which steps are essentially computational whereas other steps require careful examination of the intermediary results. If we label computational steps by (C) and steps of thinking and reflexion by (R), the modelling strategy may be summarized as follows:

(R1) List all potential explanatory or exogenous variables taking into account the structure of the missing values when necessary, impute in-sample missing values for these variables and check also for ruptures.

(C2) Compute the simple correlation between the variable to be nowcasted and each variable of the first step, repeat this computation for each country.

(R3) Examine the results of (C2), look for structural stabilities and deduce models of possible interest.

(C4) For each of the models retained in (R3), estimate regressions with country specific parameters.

(R5) Examine the results of (C4) and select a model of final interest.

(C6) For each parameter of the model selected in (R5) cluster the country specific estimates in 4 groups. The regroupment of countries is therefore coefficient specific. Estimate the final model where all the countries are treated within a unique final equation where the coefficients are now group specific.

(R7) Validate the final equation by (i) comparing the structure of the residuals between the purely heterogenous model and the final clustered model, (ii) examine the final equation characteristics: the estimates, the standard errors, the t-statistics and summary statistics such as R^2 and $\bar{R}^2$. If unsatisfactory, return to (R1).

(C8) Nowcast the exogenous and then the endogenous variables.

2. *Understanding the final results*

Two features help to understand why the final results are satisfactory in spite of the many shortcomings in the available data base. *Firstly*, the underlying model is built in terms of rate of growth (more precisely: first differences of the log values) whereas the nowcast is expressed in nominal values of the variables of interest. Thus the final nowcast is implicitly based on an efficient dynamic approach. *Secondly*, the country specific data are modelled through a flexible, and efficient, panel approach, by clustering the countries into 4 groups, differently for each parameter, instead of introducing country specific parameters, as in standard fixed effects models. By so-doing we obtain a final model that is both very economical in terms of degrees of freedom and very flexible in terms of adjusting to country specificities. This approach also allows that countries with deficient data may draw advantage from data of coefficient-similar countries. Hendry and Clements (2002) consider the pooling of forecasts obtained from several models (see also Clemen, 1989) and analyze why even simple averages often work as well as more elaborate rules and why a combined forecast provide surprising improvements over a forecast based on a unique model. They notice that "pooling can also be viewed as an application of Stein-James shrinkage estimation" which may in turn be interpreted as randomly restricted estimation. A clustered panel data model is also a way of randomly restricting the coefficients in terms of within cluster equality. It should be of interest to further analyze, and better understand the connections between forecast combination and clustered panel data models.

3. *Empirical results*

For GBAORD, the fit of the final model is satisfactory and eventually suggest that a "best use" has been made of the available data. Mouchart and Rombouts (2003) have shown that, in general, modelling and nowcasting of variables related to expenditures are more satisfactory than for variables related to personnel; private and public variables provide results of similar quality. Finally, variables of the total level give the best results. A first explication of these findings lies in the structure of the available data. Thus for total variables, times series are longer and there are no missing values for the corresponding explanatory variable.

However, comparing the structure of the residuals between the purely heterogenous pooling and the clustered regression may be unsatisfactory for some countries. Interpreting this occurence should take into account two aspects. *Firstly*, in most cases, those countries with unsatisfactory comparisons are also countries with a substantial amount of missing data and it may be reasonable to consider that countries with more missing data are likely to be also countries with data of lesser quality. *Secondly*, for every country, and every variable to be nowcasted, the number of observations is low; therefore the sampling fluctuations of the country specific regressions are substantial. More specifically, the clustered regression may also be viewed as a set of country

specific regressions, *i.e.* a restriction of the purely heterogenous model; under this approach we actually compare the results of a regression with a small sample size (between 4 or 5 and 20 or 21 observations) and a regression with larger number of observations (between approximately 150 and 300 observations). It is accordingly reasonable to consider that a substantial difference in the structure of the residuals reflect the higher sample fluctuations of the country specific models.

Comparing the results between countries suggest to further analyze how far the quality of the empirical results is, or not, associated with the quality of the national data and/or with the degree of economic development and of the integration within the european economic system.

References

Baltagi, H.B. (1995), *Econometric Analysis of Panel Data*, New York: John Wiley nand Sons.

Clemen, R.T. (1989), Combining forecasts: A review and annotated bibliography, *International Journal of Forecasting*, **5**, 559-583.

Danilov, D, and J.R. Magnus (2002), Forecast accuracy after pretesting with an application to the stock market, CentER, Tilburg University.

Hamilton, J.D.(1994), *Time Series Analysis*, Princeton University Press.

Hendry, D.F. and M.P. Clements (2001), Economic Forecasting: Some Lessons from Recent Research, European Central Bank Working Paper, Nr. 82 (October).

Hendry, D.F. and M.P. Clements (2002), Pooling of Forecasts, *Econometrics Journal*, **5**, 1-26.

Islam, T, Fiebig, D.G. and N. Meade (2002), Modelling multinational telecommunications demand with limited data, *International Journal of Forecasting*, **18**, 605-624.

Little, R.J.A and D.B. Rubin (1987), *Statistical Analysis with missing data*, New York: John Wiley nand Sons.

Marcelino, M., Stock, J.H. and M.W. Watson (2002), Macroeconomic forecasting in the EURO area: Country specific versus area-wide information, *European Economic Review* (forthcoming).

Mouchart, M. and J.V.K. Rombouts (2003), Econometric Models for Nowcasts on R&D variables, Consulting Report, Louvain la Neuve (B):Institut de Statistique, UCL and Luxembourg: CAMIRE, Estadística y Análisis, SL (Confidential Report available upon request at CAMIRE).

Appendix 1: Technical details

Details on (6)

Let

$$Z = \left[\begin{array}{cc} i & Z_2 \end{array} \right] \qquad Z'Z = \left[\begin{array}{cc} i'i & i'Z_2 \\ Z_2'i & Z_2'Z_2 \end{array} \right] \qquad z_f = \left[\begin{array}{c} 1 \\ z_{2f} \end{array} \right] \tag{25}$$

where $i'i = n$, $i'Z_2$ is a $(k-1)$-dimensional vector and $Z_2'Z_2$ is a squared matrix of order $(k-1)$ and evaluate explicitly

$$(Z'Z)^{-1} = \left[\begin{array}{cc} A_{11} & A_{12} \\ A_{21} & A_{22} \end{array} \right] \tag{26}$$

with

$$\begin{aligned} A_{11} &= n^{-1} + n^{-1}\, i'Z_2 \left(Z_2'Z_2 - Z_2'i(i'i)^{-1}i'Z_2 \right)^{-1} Z_2'i\, n^{-1} \\ &= n^{-1} + \bar{z_2}' \left(Z_2'NZ_2 \right)^{-1} \bar{z_2} \end{aligned} \tag{27}$$

$$A_{22} = \left(Z_2'NZ_2 \right)^{-1} \tag{28}$$

$$A_{12} = -n^{-1}i'Z_2 \left(Z_2'NZ_2 \right)^{-1} = -\bar{z_2}' \left(Z_2'NZ_2 \right)^{-1} = A_{21}' \tag{29}$$

where $N = I_n - i(i'i)^{-1}i' = I_n - \frac{1}{n}ii'$ and $\bar{z_2} = n^{-1}Z_2'i$ is the column vector of the sample averages. Equation (6) rests on the following manipulation

$$\begin{aligned} & \left(\begin{array}{cc} 1 & z_{2f}' \end{array} \right) (Z'Z)^{-1} \left(\begin{array}{c} 1 \\ z_{2f} \end{array} \right) \\ = \;& n^{-1} + \bar{z_2}' \left(Z_2'NZ_2 \right)^{-1} \bar{z_2} + z_{2f}' \left(Z_2'NZ_2 \right)^{-1} z_{2f} - 2\bar{z_2}' \left(Z_2'NZ_2 \right)^{-1} z_{2f} \\ = \;& n^{-1} + (z_{2f} - \bar{z_2})' \left(Z_2'NZ_2 \right)^{-1} (z_{2f} - \bar{z_2}) \,. \end{aligned} \tag{30}$$

Details on (18)

In order to write (18) in the form of (1), we first stack the data by countries, and develop the K exogenous variables into K blocks of G_K variables:

$$y = \left[\begin{array}{c} y_{(1)} \\ y_{(2)} \\ \vdots \\ y_{(n)} \end{array} \right] \qquad y_{(i)} : T_i \times 1 \qquad y : \sum_{i=1}^{n} T_i \times 1 \tag{31}$$

$$Z = [Z_1 \dots Z_K] \qquad Z_k : \sum_{i=1}^{n} T_i \times G_k \qquad Z : \sum_{i=1}^{n} T_i \times \sum_{k=1}^{K} G_k \tag{32}$$

$$Z_k = \begin{bmatrix} Z_{k,(1)} \\ Z_{k,(2)} \\ \vdots \\ Z_{k,(n)} \end{bmatrix} \qquad Z_{k,(i)} : T_i \times G_k \tag{33}$$

$$Z_{k,(i)} = [z_{k,tu}] \qquad t = 1, \dots, T_i \qquad u = 1, \dots, G_k \qquad z_{k,tu} = x_{k,it} \mathbb{1}_{\{g_k(i)=u\}} \tag{34}$$

$$\beta = [\beta_{1,1}, \dots, \beta_{1,G_1}, \beta_{2,1}, \dots, \beta_{2,G_2}, \dots, \beta_{K,1}, \dots, \beta_{K,G_K}]' : \quad \sum_{k=1}^{K} G_k \times 1 \tag{35}$$

Appendix 2: Further numerical results

Table 8: GBAORD pooled regression 2

Variable	Coefficient	stand. dev.	t-stat	P-value
GDP	.959322	.251572	3.81331	[.000]
D1	.2478E-02	.018632	.133005	[.894]
D2	-.024085	.026137	-.921485	[.358]
D3	.026543	.019011	1.39623	[.164]
D4	.054682	.019026	2.87416	[.004]
D5	.046709	.018668	2.50205	[.013]
D6	-.012524	.018622	-.672547	[.502]
D7	.030748	.018965	1.62127	[.106]
D8	.8814E-02	.018454	.477666	[.633]
D9	.5320E-02	.030696	.173336	[.863]
D10	-.8742E-02	.019462	-.449219	[.654]
D11	.071069	.024616	2.88709	[.004]
D12	-.023370	.018940	-1.23392	[.218]
D13	-.032776	.019003	-1.72478	[.086]
D14	.097755	.026511	3.68731	[.000]
D15	.010338	.020915	.494286	[.622]
D16	.013610	.022433	.606693	[.545]
D17	-.013533	.020358	-.664757	[.507]
D18	.030186	.022956	1.31498	[.190]

The Durbin-Watson test statistic is 2.0062, $R^2 = .195$ and $\bar{R}^2 = .140$

Table 9: GBAORD pooled regression 1

Variable	Coefficient	stand. dev.	t-stat	P-value
D1	.0315	.0332	.946	[.345]
D2	-.0156	.0429	-.364	[.716]
D3	.3286E-02	.0308	.106	[.915]
D4	.0154	.0322	.478	[.633]
D5	.0436	.0221	1.976	[.049]
D6	-.0228	.0350	-.652	[.514]
D7	.0565	.0222	2.541	[.012]
D8	-.0656	.0354	-1.853	[.065]
D9	.0396	.0575	.688	[.492]
D10	-.0140	.0384	-.364	[.716]
D11	.0629	.0390	1.613	[.108]
D12	.6040E-02	.0254	.237	[.812]
D13	-.7970E-02	.0345	-.230	[.818]
D14	-.4360E-03	.0328	-.013	[.989]
D15	.0628	.0396	1.582	[.115]
D16	.9505E-02	.0506	.187	[.851]
D17	-.7385E-02	.0340	-.216	[.829]
D18	.0439	.0321	1.366	[.173]
X1	-.3552	1.2943	-.274	[.784]
X2	.4200	2.2528	.186	[.852]
X3	2.0744	1.2187	1.702	[.090]
X4	2.4194	1.0209	2.369	[.019]
X5	1.1582	.8866	1.306	[.193]
X6	1.4308	1.3944	1.026	[.306]
X7	-.0259	.5458	-.047	[.962]
X8	4.7632	1.5859	3.003	[.003]
X9	.4385	.7889	.555	[.579]
X10	1.1572	1.2862	.899	[.369]
X11	1.2232	1.0405	1.175	[.241]
X12	-.5250	.9301	-.564	[.573]
X13	.0270	1.1313	.023	[.981]
X14	5.1291	.9174	5.590	[.000]
X15	-.7191	1.1221	-.640	[.522]
X16	1.1207	1.8176	.616	[.538]
X17	.7612	.9322	.816	[.415]
X18	.2826	1.1786	.239	[.811]

The Durbin-Watson test statistic is 2.0062, $R^2 = .308$ and $\bar{R}^2 = .209$

Table 10: GBAORD pooled regression 3

Variable	Coefficient	stand. dev.	t-stat	P-value
C	.0175	.7701E-02	2.276	[.024]
X1	.1115	.7124	.156	[.876]
X2	-1.0024	1.3423	-.746	[.456]
X3	1.6113	.7368	2.186	[.030]
X4	2.3638	.5783	4.087	[.000]
X5	1.7980	.7288	2.466	[.014]
X6	.0258	.7308	.035	[.972]
X7	.5890	.4364	1.349	[.178]
X8	1.4987	.8225	1.822	[.070]
X9	.7118	.3533	2.014	[.045]
X10	.2183	.6298	.346	[.729]
X11	2.2153	.6220	3.561	[.000]
X12	-.8297	.6737	-1.231	[.219]
X13	-.6986	.6001	-1.164	[.245]
X14	4.7994	.7070	6.787	[.000]
X15	.4118	.5603	.734	[.463]
X16	.8574	.7835	1.094	[.275]
X17	.1827	.5253	.347	[.728]
X18	1.0037	.8153	1.230	[.219]

The Durbin-Watson test statistic is 1.89, $R^2 = .257$ and $\bar{R}^2 = .206$

Table 11: GBAORD pooled regression 4

Variable	Coefficient	stand. dev.	t-stat	P-value
D1	-.0251	.0109	-2.295	[.022]
D2	-.2735E-02	.9295E-02	-.294	[.769]
D3	.0259	.0101	2.549	[.011]
D4	.0547	.0136	4.001	[.000]
X1	-.2101	.3376	-.622	[.534]
X2	.5113	.2762	1.850	[.065]
X3	1.4956	.3603	4.151	[.000]
X4	3.0376	.3825	7.940	[.000]

The Durbin-Watson test statistic is 1.98, $R^2 = .255$ and $\bar{R}^2 = .236$