Session Number: Poster Session 1
Time: Monday, August 23, PM

Paper Prepared for the 31st General Conference of
The International Association for Research in Income and Wealth

St. Gallen, Switzerland, August 22-28, 2010

Completing sub-national coverage of national accounts: an auxiliary information approach

Nicholas T. Longford, Grazia Pittau and Roberto Zelli

For additional information please contact:

Name: Roberto Zelli
Affiliation: Universit`a 'La Sapienza', Rome, Italy

Email Address: roberto.zelli@uniroma1.it

Completing sub-national coverage of national accounts: an auxiliary information approach

Nicholas T. Longford,
SNTL and Universitat Pompeu Fabra, Barcelona, Spain

Grazia Pittau and Roberto Zelli
Università 'La Sapienza', Rome, Italy

`Please do not quote without permission`

July 29, 2010

Abstract

This paper describes a method for estimating regional and provincial indicators when only their aggregate versions for the country are available. This and related problems have been dealt with by Chow and Lin (1971), who identified a unified regression-based framework for them. We develop an alternative approach which entails fewer assumptions and is easier to extend to more complex settings. This paper provides an example of predicting GDP per capita at regional and provincial-level in Italy. The method relies on auxiliary information in the form of the values of the indicators at the disaggregated level for the past years. The precision of the estimators is assessed by applying the method to data as they would be available in the past.

Research for this paper was supported by Grant No. 2-RISTO02_10 awarded by Istituto Tagliacarne, Rome, Italy, to the authors. Thanks to Alessandro Rinaldi for support and encouragement.

Address for correspondence: Roberto Zelli, Department of Statistics, Università 'La Sapienza', Piazzale Aldo Moro, 00185 Rome, Italy; email: `roberto.zelli@uniroma1.it` .

1 Introduction

Availability of coherent databases of national accounts indicators at different spatial disaggregation level is particularly relevant in the European Union. In fact, criteria for assignment and evaluation of European regional funds are essentially based on the availability of timely and reliable regional accounts aggregates. These figures are collected by Eurostat at different spatial levels according to a common classification of territorial units for statistics, the Nomenclature of Territorial Units for Statistics (NUTS). Throughout this hierarchical classification each Member State is subdivided into three levels: NUTS levels 1, 2 and 3 (Eurostat 2007). Currently, the Member States may go further in terms of hierarchical levels by subdividing NUTS3 units (local administrative units and municipalities). The Eurostat regional database provides several tables on economic accounts, but still displays missing data for some countries and some variables at the lowest levels of territorial disaggregation.

Moreover, one of the relevant dimension of quality of economic statistics is timeliness. Typically, national accounts aggregates are first released at national level, and, only with varying delay, at sub-national levels. EU Regulation No 1392/2007, amending Council Regulation (EC) No 2223/96 with respect to the transmission of national accounts data, specifies a time limit for the release of sub-national data. Specifically, the maximum delay for NUTS2 and NUT3 GDP figures is fixed to 24 months. For example, Istat, the Italian National Statistical Institute, releases national GDP annual figures of year t in March of year $t+1$. Regional, that is NUTS2 level of territorial disaggregation, GDP figures for year t are now released in October $t+1$, while provincial (NUTS3 sub-divisions) GDP data are available in December of year $t+2$. Regional household accounts of year t are also expected to be released by December of year $t+2$.

While this time schedule is compatible with the amount of information necessary to build up national account figures, policy makers and local authorities often demand more timely sub-national data as a support for their decisions. At this regard, it is relevant to explore a method able to produce good estimates, in terms of timeliness and reliability, of sub-national aggregates, maintaining the territorial constraints.

More formally, the focal problem investigated in this paper can be formulated as follows. The domain (a country or a region) is divided into several subdomains (provinces

or districts). For a particular variable, say a national account variable, only its total over the entire domain is known, but the within-subdomain totals are of interest. Some information is available about the subdomains, such as their (population) size and some socio-economic indicators. This auxiliary information may come from administrative and/or survey data. Information about the within-subdomain totals may be also available from the past. The problem is to draw on all this information by estimating the split of the domain-level total in a way that best conforms with the known split for the indicators.

The outlined problem was first addressed by Chow and Lin (1971) in the context of time series: e.g., the total of a variable is known for the whole year, and we wish to estimate the totals within the quarters. Bollino (1998) and Polasek and Sellner (2008) pointed out that the quarters in the time series and subdomains of a domain present similar problems; they differ only in the association structure (time series vs. spatial dependence) of the units of analysis. The solution proposed by Polasek and Sellner and by Llano et al.(2009), in a Bayesian setting, is limited to auxiliary information with a restricted format, as is the original solution by Chow and Lin. Instead of the regression framework they apply, we will estimate the division from conditional expectations of the subtotals given the total and the auxiliary information, and will represent the uncertainty about the estimates by a set of plausible solutions. The key technical device in the solution is evaluation of conditional expectations under the assumption of normality and a pattern of dependence of the subdomain totals. This pattern may take into account the neighbourhood (spatial) structure of the subdomains (Anselin, 1988).

This paper focuses on disaggregation of annual (per capita) GDP at regional (NUTS2) and provincial (NUTS3) levels in Italy and it is organized as follows. Section 2 gives the details of the method we propose. Section 3 presents the predicted results of the disaggregation, using as auxiliary variables the previous-year data of the same variable for one or a few years. We start estimating GDP per capita at NUTS2 level in order to have more timely estimates, coherently with the national value released by ISTAT, that represent in this step the domain-level total. In the second step, regional-level ad national information both contribute to estimating a province-level summary. The de-

layed estimates of provincial data coming from ISTAT allow us to evaluate our models. Section 4 presents a discussion of the preliminary results and a detailed research agenda for future research for a fuller exploitation of the auxiliary information is outlined.

2 The auxiliary approach

2.1 Method

Suppose a country comprises R regions. The regions are exclusive and exhaustive; that is, every unit (individual or household) belongs to exactly one region. Let θ_r, $r = 1, \ldots, R$, be the region-level summary of a variable of interest, such as GDP per capita, and $\theta = w_1\theta_1 + \cdots + w_R\theta_R$ the corresponding national summary. We assume that the weights (sizes of the regions) w_r as well as θ are available, but θ_r, $r = 1, \ldots, R$, are not. Further, some related summaries $\theta_r^{(\mathrm{x})}$ and $\theta^{(\mathrm{x})}$ are available; they may be summaries of the same variable for the previous years (e.g., GDP per capita in the past years), or of related variables, such as the rate of unemployment or the total value of exports, for the past and possibly even the current year. We refer to them as *auxiliary* variables (summaries). The vectors of all the summaries (focal and auxiliary) are denoted by $\boldsymbol{\theta}_r$ and $\boldsymbol{\theta}$.

We assume that the R vectors $\boldsymbol{\theta}_r$ are a random sample from a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The unknown components of $\boldsymbol{\theta}_r$ are estimated by their conditional expectations given the observed components. This leads to regression-like expressions, similar to Chow and Lin (1971). In our approach we have more flexibility because we can condition on variables that one would not contemplate as covariates in a (linear) regression.

To give details of the method, using a slightly different notation than above, we use the following partitioning:

$$\begin{pmatrix} \boldsymbol{\theta}^{(\mathrm{y})} \\ \boldsymbol{\theta}^{(\mathrm{x})} \\ \theta \end{pmatrix} \sim \mathcal{N}\left\{ \begin{pmatrix} \mathbf{1} \otimes \mu_{\mathrm{y}} \\ \mathbf{1} \otimes \boldsymbol{\mu}_{\mathrm{x}} \\ \mu_{\mathrm{y}}\, w_+ \end{pmatrix}, \begin{pmatrix} \sigma_{\mathrm{y}}^2 \mathbf{I} & \mathbf{I} \otimes \boldsymbol{\Sigma}_{\mathrm{xy}}^\top & \sigma_{\mathrm{y}}^2 \mathbf{w} \\ \mathbf{I} \otimes \boldsymbol{\Sigma}_{\mathrm{xy}} & \mathbf{I} \otimes \boldsymbol{\Sigma}_{\mathrm{x}} & \mathbf{w} \otimes \boldsymbol{\Sigma}_{\mathrm{xy}} \\ \sigma_{\mathrm{y}}^2 \mathbf{w}^\top & \mathbf{w}^\top \otimes \boldsymbol{\Sigma}_{\mathrm{xy}} & \sigma_{\mathrm{y}}^2 \mathbf{w}\mathbf{w}^\top + \sigma_\varepsilon^2 \end{pmatrix} \right\}$$

where $w_+ = w_1 + \cdots + w_R$ is the total of the weights, $\mathbf{w} = (w_1, \ldots w_R)^\top$ their (column) vector, $\mathbf{1}$ the $R \times 1$ vector of unities ($w_+ = \mathbf{w}^\top \mathbf{1}$) and $\mathbf{I}$ and the $R \times R$ identity matrix.

The unknown elements of $\boldsymbol{\theta}$ are collected in the vector $\boldsymbol{\theta}^{(\mathrm{y})}$ and the available elements in $\boldsymbol{\theta}^{(\mathrm{x})}$; both vectors are sorted by region, comprising elements $\theta_r = \theta_r^{(\mathrm{y})}$ and vectors $\boldsymbol{\theta}_r^{(\mathrm{x})}$, respectively. θ is the (known) national summary of the variable of interest. The symbol $\otimes$ denotes the Kronecker product (Magnus and Neudecker, 1998). Further, the matrix $\boldsymbol{\Sigma}$ has the partitioning

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathrm{x}} & \boldsymbol{\Sigma}_{\mathrm{xy}} \\ \boldsymbol{\Sigma}_{\mathrm{xy}}^{\top} & \sigma_{\mathrm{y}}^2 \end{pmatrix},$$

compatible with $(\boldsymbol{\theta}_r^{(\mathrm{x})\top}, \theta_r^{(\mathrm{y})})$. We allow for some discrepancy between the national summary θ and the regional summaries θ_r, due to rounding and other (administrative) errors. We represent the compendium of these (small) errors by the variance σ_ε^2.

Let $\mathbf{z} = (\boldsymbol{\theta}^{(\mathrm{x})\top}\theta)^{\top}$; $\mathbf{z}$ is observed, whereas $\boldsymbol{\theta}^{(\mathrm{y})}$ is not. The expressions for the conditional expectation and variance matrix of $\boldsymbol{\theta}^{(\mathrm{y})}$ given $\mathbf{z}$ involve the inverse of $\mathrm{var}(\mathbf{z})$. This is equal to

$$\{\mathrm{var}(\mathbf{z})\}^{-1} = \frac{1}{d}\begin{pmatrix} d\,\mathbf{I}\otimes\boldsymbol{\Sigma}_{\mathrm{x}}^{-1} + \left(\mathbf{w}\mathbf{w}^{\top}\right)\otimes\left(\boldsymbol{\Sigma}_{\mathrm{x}}^{-1}\boldsymbol{\Sigma}_{\mathrm{xy}}^{\top}\boldsymbol{\Sigma}_{\mathrm{x}}^{-1}\boldsymbol{\Sigma}_{\mathrm{xy}}\right) & -\mathbf{w}\otimes\boldsymbol{\Sigma}_{\mathrm{x}}^{-1}\boldsymbol{\Sigma}_{\mathrm{xy}} \\ -\mathbf{w}^{\top}\otimes\boldsymbol{\Sigma}_{\mathrm{xy}}^{\top}\boldsymbol{\Sigma}_{\mathrm{x}}^{-1} & 1 \end{pmatrix}, \tag{1}$$

where $d = \mathbf{w}^{\top}\mathbf{w}\left(\sigma_{\mathrm{y}}^2 - \boldsymbol{\Sigma}_{\mathrm{xy}}^{\top}\boldsymbol{\Sigma}_{\mathrm{x}}^{-1}\boldsymbol{\Sigma}_{\mathrm{xy}}\right) + \sigma_\varepsilon^2$. Note that this is equal to $\mathbf{w}^{\top}\mathbf{w}\sigma_{\mathrm{res}}^2 + \sigma_\varepsilon^2$, where σ_{res}^2 is the residual variance in the regression of $\theta^{(\mathrm{y})}$ on $\boldsymbol{\theta}^{(\mathrm{x})}$. The identity in (1) can be proved directly by multiplication, although it can also be derived using the expression for the inverse of a partitioned matrix.

The conditional distribution of $\boldsymbol{\theta}^{(\mathrm{y})}$ given $\mathbf{z}$ is multivariate normal, with the expectation

$$\mathrm{E}\left(\boldsymbol{\theta}^{(\mathrm{y})}\,|\,\mathbf{z}\right) = \mathbf{1}\otimes\mu_{\mathrm{y}} + \mathrm{cov}\left(\boldsymbol{\theta}^{(y)},\mathbf{z}\right)\{\mathrm{var}\,(\mathbf{z})\}^{-1}\begin{pmatrix} \boldsymbol{\theta}^{(\mathrm{x})} - \mathbf{1}\otimes\boldsymbol{\mu}_{\mathrm{x}} \\ \theta - \mu_{\mathrm{y}}\,w_{+} \end{pmatrix}$$

and variance

$$\mathrm{var}\left(\boldsymbol{\theta}^{(\mathrm{y})}\,|\,\mathbf{z}\right) = \sigma_{\mathrm{y}}^2 - \mathrm{cov}\left(\boldsymbol{\theta}^{(y)},\mathbf{z}\right)\{\mathrm{var}\,(\mathbf{z})\}^{-1}\mathrm{cov}\left(\mathbf{z},\boldsymbol{\theta}^{(y)}\right).$$

For both of these objects, we require the identity

$$\mathrm{cov}\left(\boldsymbol{\theta}^{(\mathrm{y})},\mathbf{z}\right)\{\mathrm{var}\,(\mathbf{z})\}^{-1} = \left\{\mathbf{1}\otimes\boldsymbol{\beta}^{\top} - \frac{\sigma_{\mathrm{res}}^2}{d}\left(\mathbf{w}\mathbf{w}^{\top}\right)\otimes\boldsymbol{\beta}^{\top}, \frac{\sigma_{\mathrm{res}}^2}{d}\,\mathbf{w}\right\}$$

where $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\mathrm{x}}^{-1}\boldsymbol{\Sigma}_{\mathrm{xy}}$. Hence

$$\begin{aligned}
\mathrm{E}\left(\boldsymbol{\theta}^{(\mathrm{y})} \,|\, \mathbf{z}\right) &= \mu_{\mathrm{y}}\mathbf{1} + \frac{\sigma_{\mathrm{y}}^2}{d}\left(\theta - \mu_{\mathrm{y}}\mathbf{w}^\top\mathbf{1}\right)\mathbf{w} + \left\{\left(\mathbf{I} - \frac{\sigma_{\mathrm{y}}^2}{d}\,\mathbf{w}\mathbf{w}^\top\right) \otimes \boldsymbol{\beta}^\top\right\}(\mathbf{x} - \mathbf{1}\otimes\boldsymbol{\mu}_{\mathrm{x}}) \\
\mathrm{var}\left(\boldsymbol{\theta}^{(\mathrm{y})} \,|\, \mathbf{z}\right) &= \sigma_{\mathrm{y}}^2\left(\mathbf{I} - \frac{\sigma_{\mathrm{y}}^2}{d}\mathbf{w}\mathbf{w}^\top\right). \qquad (2)
\end{aligned}$$

The estimates of these moments serve as estimate of the value (realization) of $\boldsymbol{\theta}^{(\mathrm{y})}$ and of its sampling variation.

In the univariate case, the conditional expectation of $\boldsymbol{\theta}^{(\mathrm{y})}$ given $\mathbf{z}$ assumes the form:

$$\mathrm{E}\left(\boldsymbol{\theta}^{(\mathrm{y})} \,|\, \boldsymbol{\theta}^{(\mathrm{x})}, \theta\right) = \mu_{\mathrm{y}}\mathbf{1} + \beta(\boldsymbol{\theta}^{(\mathrm{x})} - \mu_{\mathrm{x}}\mathbf{1}) + \frac{\sigma_{\mathrm{res}}^2}{d}\{\theta - \mu_{\mathrm{y}}\mathbf{1}^\top\mathbf{w} - \beta(\boldsymbol{\theta}^{(\mathrm{x})} - \mu_{\mathrm{x}}\mathbf{1})^\top\mathbf{w}\}\mathbf{w}\,, \qquad (3)$$

where $\beta = \frac{\sigma_{xy}}{\sigma_x^2}$ is the regression coefficient and $\sigma_{\mathrm{res}}^2 = \{\sigma_y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2}\}$ the residual variance in the regression of $\boldsymbol{\theta}^{(\mathrm{y})}$ on $\boldsymbol{\theta}^{(\mathrm{x})}$.

When $\sigma_\epsilon^2 = 0$, that is no discrepancy is allowed to arise in the aggregation, object (3) simplifies as:

$$\mathrm{E}\left(\boldsymbol{\theta}^{(\mathrm{y})} \,|\, \boldsymbol{\theta}^{(\mathrm{x})}, \theta\right) = \mu_{\mathrm{y}}\mathbf{1} + \beta(\boldsymbol{\theta}^{(\mathrm{x})} - \mu_{\mathrm{x}}\mathbf{1}) + \frac{1}{\mathbf{w}^\top\mathbf{w}}\{\theta - \mu_{\mathrm{y}}\mathbf{1}^\top\mathbf{w} - \beta(\boldsymbol{\theta}^{(\mathrm{x})} - \mu_{\mathrm{x}}\mathbf{1})^\top\mathbf{w}\}\mathbf{w}\,.$$

When also $\mathbf{w} = \mathbf{1}$, that is the country aggregate is the sum of the regional values, it further simplifies as:

$$\mathrm{E}\left(\boldsymbol{\theta}^{(\mathrm{y})} \,|\, \boldsymbol{\theta}^{(\mathrm{x})}, \theta\right) = \mu_{\mathrm{y}}\mathbf{1} + \beta(\boldsymbol{\theta}^{(\mathrm{x})} - \mu_{\mathrm{x}}\mathbf{1}) + \frac{1}{R}\{\theta - R(\mu_{\mathrm{y}} - \beta\mu_{\mathrm{x}}) - \beta\theta^{(\mathrm{x})}\}\,, \qquad (4)$$

being $\theta^{(\mathrm{x})} = \mathbf{1}^\top\boldsymbol{\theta}^{(\mathrm{x})}$ the national total of the auxiliary variable. From this simplification, it clearly appears that the estimation of $\theta_r^{(\mathrm{y})}, r = 1, 2, \ldots R$, amounts to an allocation of the total θ, in accordance with the dependence of regional-level $\boldsymbol{\theta}^{(\mathrm{y})}$ on $\boldsymbol{\theta}^{(\mathrm{x})}$. In a regression setting, the first two terms of RHS of equation (4) gives the predicted regional values based on observed regional indicators and the regression coefficient β. What the third term does is to allocate the national residual to the regions, such that the national total of the interpolated values equal the observed national value.

2.2 Structural parameters

In a typical example, μ_{y}, σ_{y}^2 and d are not known and cannot be estimated directly because θ_r are not observed. The vector of population sizes (weights) $\mathbf{w}$ is known and $\boldsymbol{\mu}_{\mathrm{x}}$ and $\boldsymbol{\Sigma}_{\mathrm{x}}$ are estimated from $\boldsymbol{\theta}_r^{(\mathrm{x})}$.

In this preliminary paper, we used as auxiliary variables the previous-year version of the same variable for one or a few years. The performance of lagged variables represents a useful benchmark with which any other auxiliary indicator should be compared.

In our approach it is explicitly stated that the assumption of normality is entirely technical, because all the quantities considered, the summaries (e.g., of GDP per capita in one or several years) are fixed; that is, we do not have a (hypothetical) replication scheme in which different values of some of these summaries would be realised. There is no sampling variation in this setting, because all the recorded quantities are precise (not estimated).

Therefore, we cannot rely on the adopted model of normality of $\boldsymbol{\theta}_r$, even though estimation of the unobserved quantities is based on it. The standard errors obtained from the squared roots of the conditional variances are problematic; we apply an entirely empirical approach in which we estimate the analogous summaries at various dates in the past, such as every year, using the data available at that time.

For example, the estimator of the region-level GDP per capita for 2009 that uses the national value of GDP per capita for 2009 and the regional-level values for 2008 is assessed by the estimates of the region-level GDP per capita for $2009 - i$ that use the national value of GDP per capita for $2009 - i$ and the region-level GDP per capita for $2008 - i$ for $i = 1, \ldots, 13$; the relevant annual time series are available since 1995 till 2008, so the estimator can be evaluated with the time 'rolled back' by i years, and the performance of this estimator can be assessed empirically.

We assume that the nature of the summaries has not changed substantially over the last few years, so the performance of the estimator in the last few years is a good indicator of the performance for the current year. Figure 1 presents the time series of GDP per capita for the Italian regions and provinces from 1995 till 2008 and 2007, respectively. The principal feature of the diagram is the high correlation of values across the years.

In such cases, we estimate μ_{y} by assuming that the ratios of the weighted and unweighted means over two recent years are approximately constant; that is

$$\frac{\mu_{\mathrm{y}}}{\mu_{\mathrm{y}}^{(\mathrm{pr})}} \doteq \frac{\theta}{\theta^{(\mathrm{pr})}} \quad (= \rho)\,, \tag{5}$$

Figure 1: The time series of the values of GDP per capita for the Italian regions and provinces.

where 'pr' refers to one or two years behind the current (prediction) year. The variance σ_{y}^2 is estimated as the square of the ratio in (5):

$$\hat{\sigma}_{\mathrm{y}}^2 \,=\, \rho^2 \sigma_{\mathrm{y,\,pr}}^2 \,.$$

More generally, the variance matrix of the values of GDP per capita for a set of years $\mathbf{m}$, which contains at least for one year for which GDP per capita is not available, is estimated by borrowing the correlation structure from the past. That is, let Δm be the smallest integer such that the values of GDP per capita is available for all years $\mathbf{m} - \Delta m$. Then we estimate the variance matrix by scaling the correlation matrix for years $\mathbf{m} - \Delta m$ so that its variances would agree with its true values when they are available. Otherwise we rely on the proportionality of the covariances for values a given number of years apart. For example, prediction of GDP per capita for regions in 2009 is based on the GDP per capita in 2008. The national value of GDP per capita is available for both 2008 and 2009, so we use the ratio $\rho = \theta^{(2009)}/\theta^{(2008)}$ to estimate the mean of the region-level values of GDP per capita as $\hat{\mu}_{\mathrm{y}}^{(2009)} = \rho \mu_{\mathrm{y}}^{(2008)}$. Further, we estimate the region-level variance in 2009 by $\hat{\sigma}_{\mathrm{y}}^{(2009)} = \rho^2 \sigma_{\mathrm{y}}^{(2008)}$ and the region-level covariance for 2008 and 2009 by $\hat{\sigma}_{2008,2009} \,=\, \rho \sigma_{2007,2008}$. Figure 2 plots the annual national values of GDP per capita against the standard deviations of GDP per capita for the regions and provinces. The strong linear association is transparent.

The constant d is a function of the variance matrix $\mathbf{\Sigma}$; we set the discrepancy variance σ_ε^2 to 1000 (Euro^2), to reflect some minor disagreements between the data for the regions (provinces) and their national weighted total.

2.3 Assessment of the estimators

We use an empirical method for assessing the quality of the predictors, based on their application in the past using the information available at that time. For example, the predictor for one year ahead (2009), using the previous years (2007 and 2008), applied in 2006, would predict the values in 2007 based on data for 2005 and 2006. Each set of estimates for $R = 20$ regions (or $P = 107$ provinces) is summarised by the root mean

Figure 2: The annual national values of GDP per capita and the standard deviations of the region- and province-level values. The capitals indicate the years from 1995 (A) to 2008 (N).

squared error of estimation,

$$\text{rMSE} = \sqrt{\sum_{r=1}^{R} \left(\hat{\theta}_r - \theta_r\right)^2},$$

$\hat{\theta}_r$ is the estimate of θ_r. (For the analysis at the province level, replace R (=20) in these expressions with P (=107), and the index r with p.) This approach relies on having a reasonably long history of data the outcome and the variables used for prediction; we have annual time series for GDP per capita for regions and provinces since 1995. We have found no useful variables for prediction other than the past-year versions of the outcome variable.

A predictor is more efficient than (superior to) its competitor if its rMSE is smaller for all the past years. Of course, the rMSE of a predictor may be smaller for some years but not for others. In such a case, we may declare an impasse, or weigh our choice by the comparisons for the more recent years.

We obtain also a model-based estimator of rMSE, the square root of the conditional variance of the prediction. However, this estimator (or its average) is very poor, and

differs from the empirical assessment substantially, both for regions and provinces.

3 Predicting GDP per capita for Italian regions and provinces

3.1 Regions

As the benchmark, we use the estimator which inflates the outcome variable from the previous year by the (multiplicative) increase of the national summary. For example, if the GPD per capita has increased from one year to the next by 2.5%, then the prediction is formed by increasing each regional (or provincial) GDP per capita by 2.5%.

Any estimator that is less efficient than the benchmark is unsatisfactory. In prediction for regions one year ahead, we have surpassed the benchmark estimator by only a small margin, but have done so uniformly for all years 1996 – 2008. The results are displayed in Table 1.

Table 1: Comparison of the benchmark and univariate one-year-ahead predictors for the Italian regions; root mean-squared errors (rMSE).

	Year					
	2008	2007	2006	2005	2004	2003
Benchmark	130.74	194.74	273.06	192.81	289.59	218.48
Method	128.31	194.60	271.19	189.08	288.89	218.37
	2002	2001	2000	1999	1998	1997
Benchmark	239.42	216.53	193.48	241.52	154.67	325.69
Method	238.46	215.05	193.42	240.38	154.23	322.66

The reason for the close agreement of our method with the benchmark is that the values of GPD per capita in the regions are very highly correlated from one year to the next (correlations in excess of 0.998), and the regression with zero intercept and a slope slightly greater than 1.0 fits the data very well.

3.2 Provinces

For provinces, the relevant prediction is for two years ahead. The prediction with GDP for the past two years is superior to the univariate prediction (based only on the most recent year), although there are two exceptions, for 2001 and 2002.

The results are displayed in Table 2. Method 21 uses the province-level values of GDP per capita only for the current year and method 22 uses the current and the preceding years. Method 21 is more efficient than the benchmark for every year, but by only a narrow margin (less than 10 Euro) in years 2007, 2004, 2001 and 2000. In contrast, method 22 is much more efficient for 2007, 2006, 2003 and 2000, but much less efficient in 2002 and 2001. An explanation for the poor performance in these two years remains a challenge, together with finding a way of anticipating it in the future.

For completeness, we list in the row marked Method 2R1 the results for estimating province-level GDP per capita separately within each region, using the province-level data from the previous year. The region-level information is effective, because the regions differ substantially. The estimate for province Vall d'Aosta is omitted from the summary because it forms a region on its own. The estimates with province-level data from the last two years are less efficient than the estimates with method 2R1.

Table 2: Comparison of the benchmark, univariate and bivariate predictors for the Italian provinces for two years ahead; root mean-squared errors (rMSE).

	Year							
	2007	2006	2005	2004	2003	2002	2001	2000
Benchmark	833.76	719.88	470.49	664.01	755.98	617.66	607.86	584.36
Method 21	831.59	708.27	453.94	658.72	744.00	602.64	605.27	582.52
Method 22	628.51	409.10	511.62	643.51	492.48	999.06	937.35	557.48
Method 2R1	515.63	468.72	356.73	356.15	303.12	486.50	347.35	349.36

3.3 Estimates for 2009

Tables 3 and 4 list the estimates of GDP per capita for regions and provinces, respectively. The estimates for the regions are based on the region-level values of GDP per

Table 3: Estimates of the region-level values of GDP per capita in 2009.

	Region	Estimate (2009)	2008		Region	Estimate (2009)	2008
PIE	Piemonte	24 828.19	25 828.54	MAR	Marche	23 044.22	23 975.42
VAO	Vall d'Aosta	26 271.38	27 335.89	LAZ	Lazio	26 819.65	27 899.67
LOM	Lombardia	29 421.14	30 602.07	ABR	Abruzzo	18 761.41	19 518.37
TAA	Trentino-A. Adige	28 108.28	29 246.54	MOL	Molise	17 298.47	17 997.07
VEN	Veneto	26 567.73	27 638.41	CMP	Campania	14 313.55	14 883.69
FBG	Friuli-Ven.-Giu.	25 466.73	26 497.07	PUG	Puglia	14 892.25	15 488.12
LIG	Liguria	23 435.18	24 382.26	BAS	Basilicata	16 557.09	17 225.14
ERO	Emilia-Romagna	28 021.06	29 151.64	CAL	Calabria	14 420.83	15 000.06
TOS	Toscana	24 751.97	25 750.11	SIC	Sicilia	14 590.21	15 172.59
UMB	Umbria	21 075.56	21 927.37	SAR	Sardegna	17 220.33	17 914.07

capita in 2008 (univariate estimation with one-year lag), and the province-level values of GDP per capita for the provinces in 2007 and 2006 (bivariate estimation with two-year lag). Both sets of estimates require the value of the (national) GDP per capita in 2009 and the population sizes of the regions (provinces) in 2009. For orientation, the tables contain also the values for the last year available. The estimated standard errors are not listed because they are better summarized by their (narrow) ranges. The standard errors for the regions are in the range 110 – 121; the smallest value is for Lombardia and the largest for Valle d'Aosta. The standard errors for the provinces are in the range 664 – 743; they are lowest for the most populous provinces (Roma, Milano, Napoli, Torino and Bari), and largest for the least populous provinces (Ogliastra, Isernia, Medio-Campidano, Valle d'Aosta and Carbonia-Iglesias). For the sake of uniformity, we have not exploited the fact that Vall d'Aosta is a region with a single province, so the estimation for it is greatly improved when based on the region-level data for 2008.

Table 4: Estimates of the province-level values of GDP per capita in 2009, based on province-level values for 2006 and 2007.

Region PROVINCE	Estimate (2009)	2007	PROVINCE	Estimate (2009)	2007
Piemonte					
TORINO	24 953.39	25 724.40	NOVARA	24 930.60	25 614.10
VERCELLI	25 627.08	26 349.60	CUNEO	26 676.29	27 353.70
BIELLA	24 616.54	25 245.30	ASTI	22 282.09	22 777.10
VERBANO-CUSIO-OSS.	20 698.01	21 251.40	ALESSANDRIA	24 108.40	24 760.60
Vall d'Aosta					
VALL D'AOSTA	25 703.62	26 642.50			
Lombardia					
VARESE	26 732.04	27 367.90	BRESCIA	28 061.02	28 801.10
COMO	25 570.39	25 861.70	PAVIA	24 129.68	24 461.40
LECCO	26 849.36	27 443.70	LODI	23 687.70	24 549.30
SONDRIO	26 400.02	26 846.40	CREMONA	25 919.25	26 399.90
MILANO	33 073.73	34 228.00	MANTOVA	28 711.65	29 447.50
BERGAMO	28 676.98	29 475.20			
Trentino-Alto Adige					
BOLZANO	29 267.95	30 233.50	TRENTO	26 735.31	27 406.10
Veneto					
VERONA	26 655.65	27 537.80	VENEZIA	26 613.97	27 469.40
VICENZA	27 429.42	28 066.90	PADOVA	26 569.19	27 441.30
BELLUNO	26 713.17	27 435.60	ROVIGO	23 846.11	24 429.00
TREVISO	26 255.39	26 968.50			
Friuli-Venezia-Giulia					
PORDENONE	25 606.59	26 518.20	GORIZIA	22 934.17	23 671.20
UDINE	25 601.57	26 091.40	TRIESTE	26 801.08	27 593.50
Liguria					
IMPERIA	23 018.43	22 991.50	GENOVA	23 963.21	24 456.00
SAVONA	24 160.43	24 613.30	LA SPEZIA	21 696.96	22 360.60
Emilia-Romagna					
PIACENZA	26 469.51	27 064.30	FERRARA	24 837.17	25 194.00
PARMA	27 780.05	28 631.30	RAVENNA	25 460.82	26 372.70
REGGIO NELL' EMILIA	27 328.59	28 233.30	FORLI-CESENA	27 991.06	28 563.30
MODENA	29 905.02	30 613.30	RIMINI	26 880.16	27 516.60
BOLOGNA	29 834.47	30 976.60			

Table 4: continued

Region PROVINCE	Estimate (2009)	2007	PROVINCE	Estimate (2009)	2007
Toscana					
MASSA-CARRARA	19 881.88	20 382.20	LIVORNO	22 579.18	23 539.70
LUCCA	25 542.32	25 742.20	PISA	25 245.35	25 934.20
PISTOIA	23 528.20	23 795.10	AREZZO	23 471.48	24 239.20
FIRENZE	26 783.13	27 842.70	SIENA	25 032.48	25 867.00
PRATO	24 418.18	25 052.60	GROSSETO	22 737.17	23 329.40
Umbria					
PERUGIA	21 481.85	22 063.40	TERNI	20 496.52	20 976.10
Marche					
PESARO E URBINO	22 602.19	23 306.00	MACERATA	21 585.99	22 406.10
ANCONA	25 515.15	26 098.90	ASCOLI PICENO	21 865.19	22 255.70
Lazio					
VITERBO	20 819.35	20 879.00	LATINA	22 152.79	22 518.70
RIETI	20 314.11	20 402.80	FROSINONE	21 171.03	21 459.30
ROMA	28 748.03	29 649.80			
Abruzzo					
L'AQUILA	19 266.38	19 499.90	PESCARA	18 415.07	18 803.20
TERAMO	18 593.18	19 000.50	CHIETI	18 681.22	19 301.80
Molise					
ISERNIA	16 239.19	16 765.20	CAMPOBASSO	17 531.17	17 787.00
Campania					
ASERTA	13 970.16	14 178.30	AVELLINO	15 644.80	15 889.10
BENEVENTO	15 014.69	15 020.30	SALERNO	15 738.57	15 954.20
NAPOLI	14 192.14	14 394.70			
Puglia					
FOGGIA	13 259.41	13 470.70	BRINDISI	13 912.48	14 281.20
BARI	15 830.31	16 135.90	LECCE	14 128.89	14 362.10
TARANTO	14 840.44	15 196.30			
Basilicata					
POTENZA	16 838.59	17 146.30	MATERA	15 981.52	16 326.60
Calabria					
COSENZA	14 826.50	15 007.40	VIBO VALENTIA	13 637.41	13 911.40
CROTONE	12 831.84	13 181.10	REGGIO DI CAL.	14 046.88	14 414.30
CATANZARO	15 642.94	16 095.50			

Table 4: continued

Region PROVINCE	Estimate (2009)	2007	PROVINCE	Estimate (2009)	2007
Sicilia					
TRAPANI	13 563.32	13 850.10	ENNA	13 695.63	13 721.20
PALERMO	15 040.43	15 374.80	CATANIA	14 183.33	14 472.60
MESSINA	15 429.45	15 688.20	RAGUSA	15 205.79	15 770.30
AGRIGENTO	12 597.59	12 725.40	SIRACUSA	15 275.19	15 747.20
CALTANISSETTA	15 036.14	15 241.10			
Sardegna					
SASSARI	16 464.13	16 816.90	OLBIA-TEMPIO	20 214.31	21 073.90
NUORO	16 875.76	17 207.30	OGLIASTRA	15 297.41	15 383.40
ORISTANO	15 289.90	15 700.00	MEDIO-CAMPIDANO	12 696.66	12 660.30
CAGLIARI	19 366.51	19 989.00	CARBONIA-IGLESIAS	13 596.80	13 631.70

4 Discussion and future research

Alternatives to our method are based on regression and time series analysis (Polasek and Sellner, 2008). Our method is similar to regression but does not require a model specification. With the standard criteria used for model selection, we would end up with much more complex models but much lower efficiency, because the selected model would be too complex. In any case, the data and the predicted values entail no randomness, and so the interpretation of the residual variance in the regression is problematic. The model-based estimates of the standard errors are very optimistic for complex models. We have found that the model-based standard errors bear little relation to the empirical root mean squared errors for both regions and provinces. The model-based rMSEs keep decreasing with complexity, whereas the empirical rMSEs indicate that substantially simpler models yield more efficient estimators. Time series models use the data for the entire history (up to 14 years), but the series for the regions (and provinces) are so highly correlated (co-integrated) that the information about the nature of the time series is far less important than the changes of the other regions in the last year.

There are many points in our future agenda to be explored. Concerning the GDP per capita, the method described in Section 2 will be adapted for the following settings:

- estimation of GDP per capita within industrial sectors (agriculture, manufacturing, construction and services), for regions and provinces. For estimating the summaries for industrial sectors we will explore how the other three sectors can contribute to the estimation of any given sector;

- using region-level summaries for year $t-1$ (e.g., 2008) for estimating the province-level summaries for year t (2009) when province-level data are available only up to year $t-2$ (2007), since we have found that provinces within regions are more similar, and have more similar progressions, than provinces in general. This requires a hierarchical specification of provinces within regions;

- further exploration of the value of the summaries of other indicators (e.g., export and import) as auxiliaries;

- estimation of annual change (in the value of GDP per capita) instead of levels.

Our method can be also adapted to other aggregates of national accounts. Specifically, it is in our agenda to focus on the disaggregation of households accounts data, at regional (NUTS2) and provincial (NUTS3) levels in Italy. ISTAT releases regional (NUTS2) households accounts with a delay of 24 months after the end of the reference period (e.g. regional households accounts for the period 2001–2007 were released by ISTAT on February 2010), in line with the EU transmission programme of national accounts data. Currently, ISTAT does not provide households aggregates at NUTS3 level.

An improvement of our method in the evaluation of conditional expectations is to explicitly take into account the neighbourhood (spatial) structure of the subdomains (Anselin, 1988) in the pattern of dependence of the subdomain totals. For spacial dependence, or similarity, we have to specify first the elements of the variance matrix $var(\boldsymbol{\theta}^{(y)})$ as functions of the distance between the subdomains involved. Apart from specifying $var(\boldsymbol{\theta}^{(y)})$ in this way, we also require a parametrisation for the covariance matrices. Analytical expressions for the conditional expectations and variance matrices are unlikely to exist, but approximations by $var(\mathbf{z})$ with compound symmetry may be useful, especially when most pair of districts are neighbours.

References

Anselin, L. (1988). *Spatial econometrics: methods and models.* Kluwer, London, UK.

Bollino, A. (1988). Econometric interpolation of regional time series. *Statistica Applicata (Journal of Applied Statistics)*, n.10.

Chow, G. C., and Lin, A. (1971). Best linear unbiased interpolation, distribution and extrapolation of time series by related series. *The Review of Economics and Statistics* **53**, 372–375.

Llano, C., Polasek W., and Sellner R. (2009). Bayesian methods for completing data in space-time panel models. Economic Series n.241, Institute for Advanced Studies, Vienna.

Magnus, J. R., and Neudecker H. (1998). *Matrix Differential Calculus with Applications in Statistics and Econometrics.* John Wiley and Sons, Chichester, UK.

Polasek, W., and Sellner R. (2008). Spatial Chow-Lin methods: Bayesian and ML forecast comparisons. Working Paper 38–08, the Rimini Centre for Economic Analysis, Rimini, Italy.