

Regional Disaggregation of National Data with Instrumental Variables

Disaggregazione Regionale di Dati Nazionali con Variabili Strumentali

Giuseppe Venanzoni
Dipartimento di Contabilità Nazionale
e Analisi dei Processi Sociali

Francesco Felici
CONSIP Spa - Roma
francesco.felici@tesoro.it

Università di Roma La Sapienza
giuseppe.venanzoni@uniroma1.it

Riassunto: La disaggregazione di serie nazionali con modelli ottimali migliora significativamente quando si dispone di un *panel* parziale di osservazioni disaggregate, che consente di accertare la presenza di eterogeneità dei parametri. Il modello proposto, utilizzabile per estrapolare serie disaggregate i cui totali sono noti, impiega variabili strumentali per correggere tale eterogeneità, e segnalare inoltre l'eventuale presenza di cambiamenti strutturali nelle relazioni. Nell'applicazione presentata si ottiene una riduzione media dell'errore di circa il 95% rispetto al modello standard.

Keywords: Modelli ottimali; Disaggregazione; Variabili strumentali

1. Optimal Disaggregation Models ¹

Optimal disaggregation models, where the estimation of the disaggregated values is consistent with the aggregated ones, are based on the specification of a functional relationship between the variable to be disaggregated and an available vector of indicators at area-level. The analysis of such relation is carried on as a forecasting problem, whose parameters are estimated under the equality constraint between the sum of the disaggregated values and the corresponding aggregated ones (Chow e Lin, 1971)²:

$$Y = X\beta + U, \quad E(U) = 0, \quad COV(U) = \Phi; \quad (1)$$

$$Y_A = X_A\beta + U_A, \quad E(U_A) = 0, \quad COV(U_A) = B\Phi B' = \Phi_A. \quad (2)$$

The optimal predictor of Y is the GLS estimator of (2) under the constraint $Y_A = B\hat{Y}$:

$$\hat{Y} = X\hat{\beta} + G\hat{U}_A, \quad (3)$$

$$\hat{\beta} = (X_A' \Phi_A X_A)^{-1} X_A' \Phi_A^{-1} Y_A, \quad (4)$$

$$G = \Phi B' (B\Phi B')^{-1} = \Phi B' \Phi_A^{-1}, \quad (5)$$

$$\hat{U}_A = (\hat{Y}_A - X_A \hat{\beta}) = (I - X_A (X_A' \Phi_A^{-1} X_A)^{-1} X_A' \Phi_A^{-1}) Y_A. \quad (6)$$

The aggregated errors covariance matrix ($\Phi_A = B\Phi B'$) depends on the disaggregated matrix, usually unknown. In order to estimate Φ , the standard assumption $\Phi = \sigma^2 I$ is too

¹ Research financed by MIUR-University of Rome "La Sapienza" and CNR.

² Y: (NTx1) vector of unobservable values; X: (NTxK) matrix of the observed K indicators; Y_A, X_A: aggregated vector and matrix; B: aggregation matrix; I: identity matrix; T: periods; N: areas.

strong in most empirical cases. The model (1-3), initially proposed for time series data, has been extended to spatial data by assuming only errors heteroschedasticity (Σ_n) and one-period autocorrelation (R_t). Under these hypotheses, the disaggregated covariance matrix can be written as $\Phi = R_t \otimes \Sigma_n$ (Bollino, 1999), and:

$$\Phi_A = \text{COV}(U_A) = B(R_t \otimes \Sigma_n)B' = \sigma_A^2 R_t \quad \text{with } \sigma_A^2 = \sum_i \sigma_i^2 = \text{tr}(\Sigma_n). \quad (7)$$

Under simplifying but realistic assumptions for heteroschedasticity (variance of area i proportional to its average population P_i) and autocorrelation (same autoregressive parameter ρ for every area) it is possible to estimate β , ρ , σ_A^2 , then obtaining feasible estimators for (4-7). A further extension of the model, which accounts for the spatial dependence of the disturbances, assumes a non-diagonal $\Phi = R_t \otimes \Sigma_n$, whose generic element $\sigma_{ij} = \sigma^2 P_i P_j$ can be estimated from $\hat{\sigma}_A^2$ (Venanzoni, 2001)³:

$$\Phi = \frac{\sigma^2}{1-\rho^2} \cdot \begin{bmatrix} 1 & \rho & \cdots & \rho^{T-1} \\ \vdots & \ddots & & \vdots \\ \rho^{T-1} & \cdots & \rho & 1 \end{bmatrix} \otimes \begin{bmatrix} P_1^2 & P_1 P_2 & \cdots & P_1 P_n \\ \vdots & P_2^2 & & \vdots \\ & & \ddots & \\ P_n P_1 & \cdots & & P_n^2 \end{bmatrix}. \quad (8)$$

2. Distortion Bias and Instrumental Variables

In the standard optimal model only (2) is estimated on aggregated data; the estimated β (4) are used in (3) under a spatial homogeneity hypothesis: $\beta_i = \beta$ (no bias). This way, if a partial panel of disaggregated values of the variable is available, this information is lost; moreover, the assumption $\beta_i = \beta$ is often too strong in most empirical cases. This feature, already underlined by Bollino (1999), has been treated by Berrettoni et al. (1999) by superimposing in the same matrix for (4-7) the two different sources of information used for the estimation (the disaggregated panel and the aggregated data). However, this procedure does not eliminate the bias (the homogeneity assumption is not tested), and moreover leads to two new problems: an ambiguous meaning of β , estimated under a form intermediate between an aggregated specification and a disaggregated one; an increased risk of structural breaks from merging two different sources of information. This approach is therefore recommended only if such problems can be ruled out, as for instance when the available indicators are just a transformation of the variable under analysis⁴.

When this is not the case, as for the extrapolation of disaggregated series using a structural or behavioural relationship between the variable and the indicators⁵, we propose to use in model (1-3) an instrumental variable, which controls for the variability of the parameters across areas. The estimation of area-specific relationships between the variable and the available disaggregated indicators allows us to construct an instrument (NxT) by suppressing the error term. By extrapolating this instrumental variable we get a new series of disaggregated data, which can be used in (2) instead of the indicators.

³ If the spatial dependence is investigated only among neighbouring areas, a connection matrix is used.

⁴ Bollino (1999) estimates Private Regional Consumption from Household Survey Budget data. Berrettoni et al. (1999) use the optimal model only for Regional Foreign Trade using monetary data.

⁵ This is for instance the case for the revision of territorial estimates no longer consistent with the updated aggregated data, as for the regional economic accounts usually lagged behind the national ones.

The hypothesis of no distortion, a reasonable one for the instrument, which summarises the area-level information about the variable of interest, can be verified by testing $\beta_i = \beta$. It is also possible to test the presence of structural breaks affecting the estimators of ρ ⁶, an often neglected procedure in this kind of analysis.

3. Regional Disaggregation of Public Administrations Consumption.

The model with instrumental variable (MIV) has been used to disaggregate the Public Administrations (PA) Consumption across Italy's regions, and its performance has been checked against the standard model with indicators (MWI). We used the following informations: 1) regional values of the variable under analysis for the period 1983-92 (Istat, 1996); 2) national values of the same variable, and regional and national GDP data for the period 1980-96 (Istat, anni vari)⁷.

The construction of the instrumental variables is based on just one regression (1983-92) with regional dummies (two, slope and intercept, for each region, as differentials from Lombardy, the omitted one), under the assumption of homoschedastic and uncorrelated errors⁸. The relationship between PA Consumption and GDP turns out to be variable across regions, both for the intercept and the slope. The extrapolation of the instrumental variable until 1996 gives preliminary estimates of the regional PA Consumptions which are not consistent with the available national values, and do not consider the national dynamics outside the original sample period. The use of these values instead of the GDP in (2) leads to identical results. At the aggregated level the outcome of the instrument is the same as the original indicators, whose values it is but a linear combination.

But the instrument is crucial at step (3), where the regional estimates are made under the national constraint. The comparison between the different results shows the gain from MWI (model 1 of Table 1) to MIV (model 2). The error reduction is patent within the sample period (from 15% to 1% on average); moreover, the crucial link between the true values in 1992 and the estimated ones in 1993 (year of first disaggregated forecasts outside the sample period) is improved.

For instance, the average national increment is +3.5% for 1993 on 1992. For Lombardia MWI gives an estimate of 44.560 Lit. billions, against a true value of 37.972 in 1992 (+ 17,3%); MIV gets a more credible +3,3%. For Campania too MIV produces a more realistic +3,4% versus -7,5% with MWI.

⁶ Its symptoms are a large ρ with a non-unit coefficient in the aggregated regression without intercept between the variable and the instrument. The autoregressive residuals may suggest a structural break which is not properly modelled, for instance by using dummy variables.

⁷ Data are available for a longer time period, especially at a national level. The choice of this interval avoids the problems arising from change of data definitions and classifications (from SEC79 to SEC95).

⁸ The number of dummies should require an ex-post reduction: different clusters (according to different slopes, intercepts or both) may however arise from the elimination of non-significant dummies. Given the regression purpose (i.e. the instrument construction), the lack of synthesis may therefore be justified. Pooling NxT observations leads to more efficient estimators than separate regional regressions. The methodology is similar to a panel data fixed effects model: further assumptions on the residuals lead to a SUR model. The introduction of the instrument is a preliminary step to the optimal model: more general error structures are assumed later (the aggregate estimation and the error redistribution steps). In the optimal model, the consistent aggregation condition requires a static linear no-constant form: MIV can, but the standard MWI cannot, use different specifications, such as non-linear or dynamic models (e.g. VAR etc).

Graph 1 Campania: Public Administrations Consumption (Lit. bn., current prices, 1984-96)

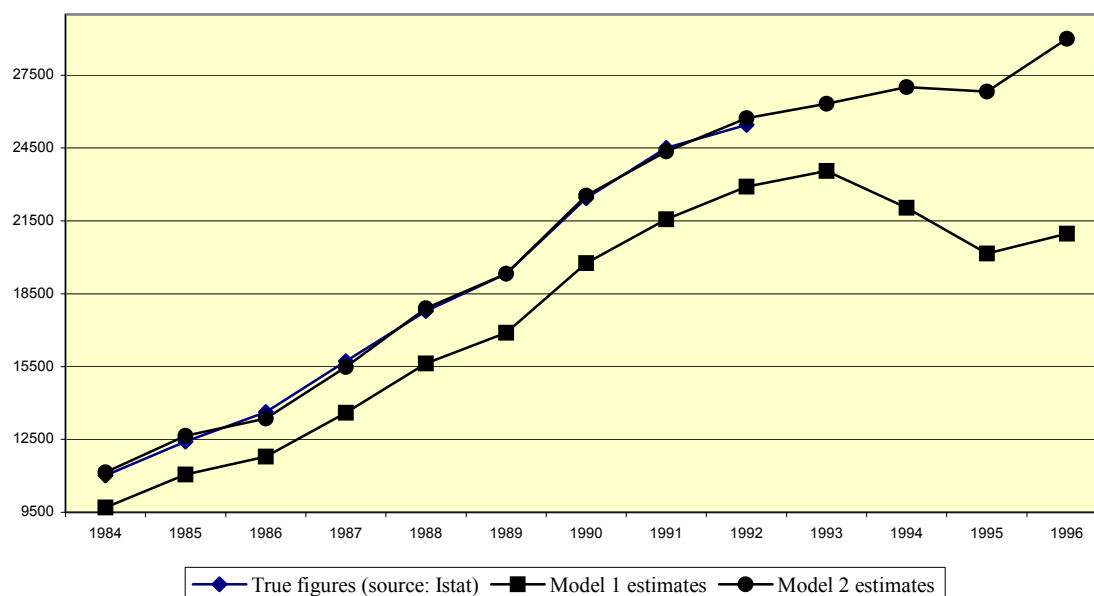


Table 1: PA Consumption (Lit. billions, current prices) - Performance of the two models

Year	Lombardia					Campania				
	Istat	Model 1 (Mwi)	Model 2 (Miv)	Error 1 %	Error 2 %	Istat	Model 1 (Mwi)	Model 2 (Miv)	Error 1 %	Error 2 %
1984	17315	19778	17025	14.2	-1.7	11019	9703	11141	-11.9	-1.1
1985	19442	22794	19746	17.2	1.6	12412	11060	12646	-10.9	-1.9
1986	21296	24719	21346	16.1	0.2	13629	11787	13356	-13.5	2.0
1987	23885	27609	23974	15.6	0.4	15719	13601	15475	-13.5	1.6
1988	26757	31427	27220	17.5	1.7	17791	15623	17908	-12.2	-0.7
1989	28623	33481	28875	17.0	0.9	19331	16891	19323	-12.6	0.0
1990	32880	38604	33479	17.4	1.8	22442	19758	22542	-12.0	-0.4
1991	35550	41183	35550	15.8	0.0	24502	21562	24366	-12.0	0.6
1992	37972	42885	37497	12.9	-1.3	25451	22907	25728	-10.0	-1.1
1993		44560	39215				23553	26320		
1994		49554	40091				22037	27010		
1995		52605	39706				20164	26820		
1996		58359	43088				20975	28994		

BIBLIOGRAFIA

- BERRETTONI P., DELOGU R., PAPPALARDO C., PISELLI P., (1999), Una ricostruzione omogenea di dati regionali: conti economici e reddito disponibile delle famiglie 1970-1995, *Tem di discussione del Servizio Studi*, n.346, Banca d'Italia.
- BOLLINO C.A., (1999), L'utilizzo delle tecniche di disaggregazione con indicatori per le stime di serie economiche territoriali, in "Modelli e strumenti per l'analisi economica a breve termine", *Annali di Statistica*, S. X, vol. 17., Roma.
- CHOW G., LIN A.L., (1971), Best Linear Unbiased Interpolation, Distribution and Extrapolation of Time Series by Related Series, *The Review of Economics and Statistics*, vol.53, n.4, 372-375.
- ISTAT, (Anni vari), *Conti economici nazionali*, Roma.
- ISTAT, (1996), *Conti economici regionali delle Famiglie e della Pubblica amministrazione*, Roma.
- VENANZONI G., (2001), *Modello di disaggregazione multiregionale dei Conti economici delle Amministrazioni pubbliche*, dattiloscritto, Roma, Gennaio.