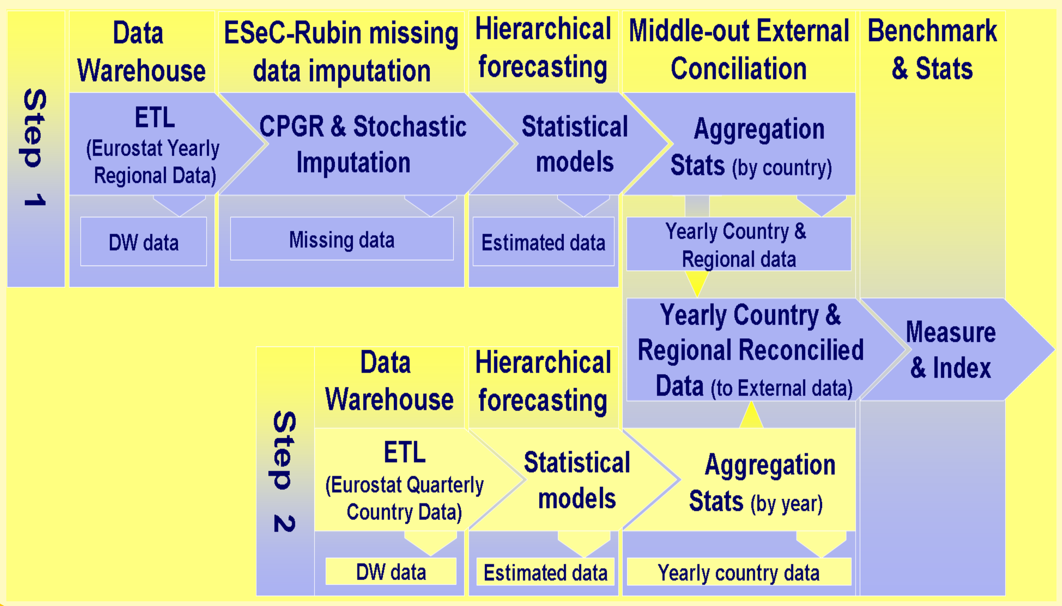
Infinite data and few information for regional forecast: an applied approach from this paradox

Paola M. Chiodini, Department of Statistics, University of Milano-Bicocca
Silvia Facchinetti, Department of Statistical Science, Catholic University of Milan
Giancarlo Manzi, Department of Economics, Business and Statistics, Università degli Studi di Milano
Flavio Verrecchia, ESeC no-profit Association

Introduction

In sectorial and regional forecasting it is customary to deal with situations where data are not suitable since: *i)* regional borders can possibly change; *ii)* the classification of economic activities is periodically revised; *iii)* the EU policy on updating methods and survey domains is frequently re-adapted. Statistical tools can be useful in strengthening the power of the models when multiple time series are handled at the same time. When data are analyzed at a regional level or with a limited history, the most used techniques are those of classical time series analysis and strategies are available for regional forecast aggregation but the results are successful when time series are longer and complete. However, if the data quality is not satisfactory, a different strategy can be adopted - namely the *External Middle-Out Hierarchical Forecasting* (EMOHF) - which is based on a joint use of multiple forecasts. It consists in performing one forecast at a regional level and another one at a national level (external data), and then obtaining from them a *hierarchical conciliation*, resulting in a national estimate.

Double-phase National Middle-Out Hierarchical Forecasting strategy



Conclusions

In order to get more reliable estimates, when there is few information for regional forecasting, it can be necessary the use of data other than that available for the specific analysis at hand. In this application, when multiple sources of data are managed, forecasts are better (the 2007 national APE from regional aggregation is 0.4%, the corresponding APE after reconciliation is 0.1%). Moreover, national data are often provided before the regional data, so that, in the case of the employment level, the proposed procedure allows for a reliable estimates before the official publication by national statistics agencies.

Acknowledgements

The authors are grateful to Professor Umberto Magagnoli for advice, comments and suggestions and to a SIS suggestions. This work is financially supported by ESeC and SAS institute.

Essential references

[1] R.J. Hyndman, A.B. Koehler, J.K. Ord, R.D. Snyder (2008), Forecasting with exponential smoothing, *Springer-Verlag, Berlin*

[2] L. Santamaria (2000), Analisi delle serie storiche economiche, *Vita e Pensiero, Milano*

[3] F. Verrecchia (2008), Previsione e selezione automatica dei modelli per serie storiche regionali: metodo bi-fase a conciliazione esterna, *SAS Business Analytics Gallery*

State Space models

Let ℓ, b, T_h and ϕ $(0 < \phi < 1)$ be respectively the level term, the growth term, the forecast term over the next h time periods, and the damping parameter. ℓ and b can be combined, giving five future trend patterns:

- None N: $T_h = \ell$
- Additive A: $T_h = \ell + bh$
- Additive damped A_d: $T_h = \ell + (\phi + \phi^2 + \ldots + \phi^h)b$
- Multiplicative M: $T_h = \ell b^h$
- Multiplicative damped M_d: $T_h = \ell b^{(\phi+\phi^2+\ldots+\phi^h)}$

The seasonal component is then matched with the trend component.

Exponential Smoothing methods

Trend component	Seasonal component N	A	M
N	N, N	N, A	N, M
A	A, N	A, A	A, M
A_d	A_d, N	A_d, A	A_d, M
M	M, N	M, A	M, M
M_d	M_d, N	M_d, A	M_d, M

Then the automatically selected models are:

- Simple, Double Exponential Smoothing (Brown) - (N, N);
- Linear Exponential Smoothing (Holt) - (A, N);
- Damped Additive Trend - (A_d, N);
- Additive Seasonal Smoothing (Winters) - (A, A).

Let l_t, b_t, s_t and m denote respectively the series level at time t, the slope at time t, the seasonal component at time t and the number of seasons. Then is possible to express the Exponential Smoothing equations (where α, β^*, γ, ϕ are constants, $\phi_h = \phi + \phi^2 + \ldots + \phi^h$ and $h_m^+ = [(h-1) \bmod m] + 1$).

Exponential Smoothing formulae

Methods	Equations
N, N	$l_t = \alpha y_t + (1-\alpha) l_{t-1}$
	$\hat{y}_{t+h\|t} = l_t$
A, N	$l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1})$
	$b_t = \beta^*(l_t - 1_{t-1}) + (1-\beta^*) b_{t-1}$
	$\hat{y}_{t+h\|t} = l_t + h b_t$
A_d, N	$l_t = \alpha y_t + (1-\alpha)(l_{t-1} + \phi b_{t-1})$
	$b_t = \beta^*(l_t - 1_{t-1}) + (1-\beta^*) \phi b_{t-1}$
	$\hat{y}_{t+h\|t} = l_t + \phi_h b_t$
A, A	$l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1})$
	$b_t = \beta^*(l_t - 1_{t-1}) + (1-\beta^*) b_{t-1}$
	$s_t = \gamma(y_t + l_{t-1} - b_{t-1}) + (1-\gamma) s_{t-m}$
	$\hat{y}_{t+h\|t} = l_t + h b_t + s_{t-m+h_m^+}$

The State Space general equations are:

$y_t = w(\mathbf{x}_{t-1}) + r(\mathbf{x}_{t-1})\varepsilon_t$

$\mathbf{x}_t = f(\mathbf{x}_{t-1}) + g(\mathbf{x}_{t-1})\varepsilon_t$

where $\mathbf{x}_t = (l_t, b_t, s_t, \ldots, s_{t-m+1})'$, $\mu_t = w(\mathbf{x}_{t-1})$ and with additive error $r(\mathbf{x}_{t-1}) = 1$. Assuming additive *i.i.d.* errors $\varepsilon_t \sim N(0, \sigma^2)$, let $\mu_t = \hat{y}_t$ denote the one-step forecast of y_t and $\varepsilon_t = y_t - \mu_t$ the one-step forecast error at time t.

Considering the triplet E, T, S (Error, Trend, Seasonality), we can find the State Space models for each Exponential Smoothing method (to simplify the notation, we use $\beta = \alpha\beta^*$).

State Space equations with additive error

Models	Equations
$ETS(A, N, N)$	$l_t = l_{t-1} + \alpha\varepsilon_t$
	$\mu_t = l_{t-1}$
$ETS(A, A, N)$	$l_t = l_{t-1} + b_{t-1} + \alpha\varepsilon_t$
	$b_t = b_{t-1} + \beta\varepsilon_t$
	$\mu_t = l_{t-1} + b_{t-1}$
$ETS(A, A_d, N)$	$l_t = l_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$
	$b_t = \phi b_{t-1} + \beta\varepsilon_t$
	$\mu_t = l_{t-1} + \phi b_{t-1}$
$ETS(A, A, A)$	$l_t = l_{t-1} + b_{t-1} + \alpha\varepsilon_t$
	$b_t = b_{t-1} + \beta\varepsilon_t$
	$s_t = s_{t-m} + \gamma\varepsilon_t$
	$\mu_t = l_{t-1} + b_{t-1} + s_{t-m}$

Applications

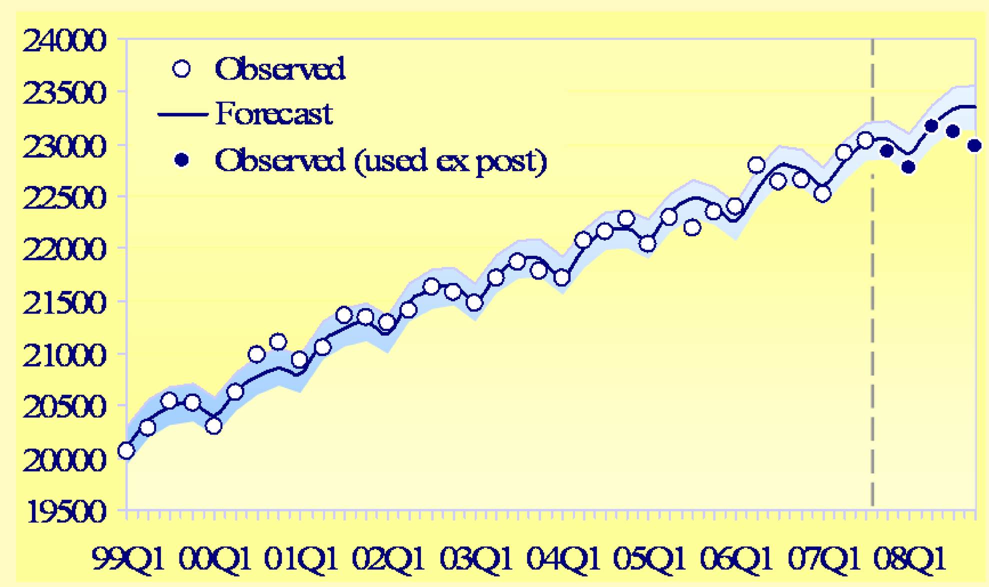
The Italian employment level derived from the aggregation of regional estimates is overestimated (+0.3%) if compared with the forecast obtained as a quarterly aggregation of external national estimates.

Employment-persons, regional and national forecasts, aggregation and conciliation ratio, Italy, 2007-08

Year	Aggregation (000) Region	Quarter	Conciliation ratio
2007	22,941	22,875	0.997
2008	23,275	23,193	0.996

At European regional level the forecast EMOHF is proportionally adjusted according to the national conciliation ratios.

Employment-persons aged 15-64, national forecasts, Italy, 1999-08



Employment-persons aged 15-64, regional models, external conciliated forecasts, ex-post APE, by regions, 2007-08

	2007		2008				Level		Trend		Weight/Damping/Seasonal	
NUTS	$\hat{y}_C$ (000)	APE expost	$\hat{y}_C$ (000)	APE expost	Model ETS(...)	MAPE	Par. estim.	P-value	Par. estim.	P-value	Par. estim.	P-value
ITC1	1,829	0.1%	1,842		A, A, N	0.82%	0.132	0.323	0.001	0.997		
ITC2	55	1.6%	56		A, N, N	0.93%					0.971	0.000
ITC3	620	2.3%	624		A, A, N	1.06%	0.001	0.998	0.001	1.000		
ITC4	4,250	0.4%	4,308		A, A, N	0.26%	0.276	0.049	0.001	0.975		
ITD1	223	0.3%	225		A, A, N	1.11%	0.999	0.013	0.001	0.996		
ITD2	215	2.4%	215		A, N, N	1.56%	0.999	0.006				
ITD3	2,084	0.1%	2,110		A, A, N	0.46%	0.205	0.158	0.001	0.992		
ITD4	510	0.7%	515		A, A, N	0.85%	0.129	0.342	0.001	0.998		
ITD5	1,885	1.4%	1,907		A, A, N	0.45%	0.193	0.177	0.001	0.993		
ITE1	1,519	0.3%	1,537		A, A, N	0.51%	0.188	0.204	0.001	0.994		
ITE2	348	3.5%	353		A, A, N	0.86%	0.047	0.730	0.001	1.000		
ITE3	640	0.1%	648		A, A, N	0.38%	0.267	0.166	0.001	0.986		
ITE4	2,120	2.7%	2,152		A, A, N	0.55%	0.999	0.020	0.001	0.996		
ITF1	499	0.9%	507		A, A, N	0.91%	0.216	0.221	0.001	0.993		
ITF2	108	2.7%	108		A, A, N	1.04%	0.001	0.993	0.001	1.000		
ITF3	1,766	3.8%	1,769		A, A_d, N	1.24%	0.209	0.685	0.001	0.999	0.999	0.000
ITF4	1,276	0.5%	1,311		A, A, N	1.05%					0.999	0.000
ITF5	195	1.4%	197		A, A, N	1.35%	0.048	0.723	0.001	1.000		
ITF6	622	4.3%	634		A, A_d, N	1.20%	0.192	0.665	0.001	0.999	0.999	0.000
ITG1	1,496	1.6%	1,519		A, A, N	0.58%	0.179	0.182	0.001	0.994		
ITG2	613	1.3%	627		A, A, N	1.48%	0.179	0.162	0.001	0.993		
IT	22,857	0.1%	23,193	0.8%	A, A, A	0.37%	0.314	0.001	0.001	0.977	0.001	0.991

Even if these models present non-significant parameters, an *ex post* investigation shows that when the regional estimates are conciliated with the national estimates, the Absolute Percentage Error (APE) is less than 4.3%. In a forecasting framework the use of exponential weights is a form of prudential behaviour. When breakpoints are detected, significant parameters are obtained by truncation of the time series (a model with intervention variable can also be used).

Employment-persons aged 15-64, national model (data 04Q2-07Q3), forecasts, ex post APE, by regions, 2007-08

	2007		2008				Level		Trend		Seasonal	
NUTS	$\hat{y}_C$ (000)	APE expost	$\hat{y}_C$ (000)	APE expost	Model ETS(...)	MAPE	Par. estim.	P-value	Par. estim.	P-value	Par. estim.	P-value
IT	22,863	0.1%	23,144	0.6%	A, A, A	0.38%	0.078	0.379	0.001	0.995	0.001	0.997