Università di Pisa

Sant'Anna Scuola Universitaria Superiore Pisa

Consiglio Nazionale delle Ricerche

Università degli Studi di Cagliari

Book of Short Papers SIS 2021

SIS2021

PISA, 21-25 GIUGNO 2021

SIS Società Italiana di Statistica

Editors: Cira Perna, Nicola Salvati and Francesco Schirripa Spagnolo

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN 9788891927361

Contents

Hierarchical Forecast Reconciliation on Italian Covid-19 data

Riconciliazione gerarchica su time-series: applicazione ai dati Covid-19 italiani

Andrea Marcocchia, Serena Arima and Pierpaolo Brutti

Abstract Due to the spread of the Covid-19 pandemic, many different actions to limit personal freedom have been decided according to the trend of the epidemiological situation. Often these decisions were made at the national level or, in a second phase, at the regional one. In order to ensure the implementation of actions and behaviours more in line with the real epidemiological situation, it would be useful to decide and give specific information at a more granular territorial level (provinces) even if one of the main problem is the quality of the data. This paper presents how the use of forecast reconciliation techniques can help forecasts at a detailed territorial level, while exploiting the more robust information available at the aggregate level.
Abstract *Con la comparsa della pandemia da Covid-19 si sono rese necessarie azioni di limitazione delle libertà personali, decise in funzione dell'andamento della situazione epidemiologica. Tali decisioni sono state prese a livello nazionale o, in una seconda fase, regionale. Per garantire la messa in atto di azioni maggiormente in linea con la reale situazione epidemiologica di un territorio, sarebbe utile agire ad un livello di maggiore granularità territoriale (province). Uno dei principali problemi nel prendere decisioni a un livello più dettagliato risiede nella qualità dei dati. In questo lavoro si presenta come l'uso di tecniche di forecast reconciliation possono aiutare a migliorare le previsioni ad un livello granulare, sfruttando contemporaneamente le, più robuste, informazioni disponibili a livello aggregato.*

Key words: forecasting, time series, forecast reconciliation, hierarchical time series, Covid-19 data

---

Andrea Marcocchia
Sapienza University of Rome, e-mail: andrea.marcocchia@uniroma1.it

Serena Arima
University of Salento, e-mail: serena.arima@unisalento.it

Pierpaolo Brutti
Sapienza University of Rome, e-mail: pierpaolo.brutti@uniroma1.it

1 Introduction

The main idea behind forecast reconciliation in a hierarchy is that observed demands at each level will always add up to the observed demands at higher levels. It is usually desirable that the same holds true also for forecasts (that is the "forecasting coherence"). If forecasting at the different levels is done independently, we usually have forecast incoherence, meaning that the bottom level forecasts do not add up.
The various components of the hierarchy can interact in a variety of complex ways. A change in one series at one level, can have an impact on other series at the same level, as well as on series at higher and/or lower levels. By modeling the entire hierarchy of time series simultaneously, we obtain better forecasts of each of the component series. *Reconciliation* is the process that fix incoherent forecasts. Another important benefit of forecast reconciliation is that, although usually the research interest lays in the most disaggregated data, these are also noisier than the others; in contrast, the "total" series, although less interesting, is typically way more resilient to noise: forecast reconciliation is capable to extract the most relevant information from both series [1].

2 Models

This section presents the used models: the first part explains the forecast algorithms, while the second subsection adds the description of the reconciliation step.

2.1 Forecast models

Any type of prediction methods can be used in the first phase because the reconciliation models have no limits in this regard. The used forecast methods are:

- **ARIMA with and without external covariates**: considering the difficulty in identifying a single parameterization that guarantees a good performance on all the series of the hierarchy, we proceeded using an automatic selection method of the best set of parameters for each series. In an other case, an external variable has been added as additional information. The considered variable indicates for each date whether it is in a lockdown condition or not. The variable is the same at all levels of the hierarchy, therefore restrictions were not considered at the regional or the provincial level, but only at the national level. It should be borne in mind that in the considered period (up to the beginning of November 2020), the closure measures imposed by the Italian government were mainly national, with few measures at the regional level, as lately more frequently decided.
- **Exponential Smoothing (ETS)**: also with regard to this method, the best model was selected for each series.

- **Segmented regression**: this approach allows to divide the time series into different segments and to use a simple linear regression to estimate the trend of each segment. It was decided to use this approach after noting that there were purely linear trends within the time series. The number of segments that have been created is equal to 3. The reason is that, until November 2020, there was an initial period of severe growth in the number of daily infections, then a decrease during the summer and then a new growth phase after summer.

2.2 Reconciliation models

The reconciliation problem can be formalize as follows:

- Consider a multi-level hierarchy, where level 0 denotes the completely aggregated series and level k contains the most disaggregated time series and assume that the observations are recorded at times $t = 1,2,\dots,n$ and that we are interested in forecasting each series at each level at times $t = n+1, n+2, \dots, n+h$.
- The additive structure of the time-series involved can be exemplified in a three level hierarchy: $Y_t = \sum_i Y_{i,t}$ represents the most aggregated series, obtained as sum of all the series that belong to the first level ($Y_{i,t}$), that is $Y_{i,t} = \sum_j Y_{ij,t}$ so that the first level series is obtained starting from the second level. As final stage, $Y_{ij,t} = \sum_z Y_{ijz,t}$ represents the aggregation of the lowest level of the hierarchy (the third level, with $k = 3$). The notation refers to $Y_{ij,t}$ as the values of series ij at time t and to Y_t as the aggregation of all the series at time t.
- More generally, let $\mathbf{Y}_{i,t}$ be the vector that collects the values of all the series for $i = 1,\dots,k$, observed at a specific level i at time t. Now define $\mathbf{Y}_t = \left[\mathbf{Y}_{1,t},\dots,\mathbf{Y}_{k,t}\right]'$. According to this notation, we can define $\mathbf{Y}_t = \mathbf{S}\mathbf{Y}_{k,t}$, where $\mathbf{S}$ is a summing matrix used to aggregate the lowest level series. If we assume that the number of bottom level time series is m_k and that m is the total number of series (considering all the levels), then $\mathbf{S}$ is an $(m \times m_k)$ matrix. The $\mathbf{S}$ matrix can be partitioned by the levels of the hierarchy: the top row is a unit vector of length m_k and the bottom section is a $m_k \times m_k$ identity matrix. The symbols m and m_i, are the total number of series in the total hierarchy and at level i, so it is always true that $m_i > m_{i-1}$ and $m = m_0 + m_1 + \dots + m_k$.
- If we compute forecasts for each period $n+1, n+2, \dots, n+h$ for a generic series X, we can define this forecasted series $\widehat{Y}_{X,n}(h)$. Applying the same logic, $\widehat{Y}_n(h)$ denotes the h-step ahead prediction of the total. The final step is to define $\widehat{\mathbf{Y}}_n(h)$ as the vector consisting of these base forecasts, stacked in the same series order as for $\mathbf{Y}_t$;
- Bearing all this in mind, all existing hierarchical forecasting methods can be written as $\widetilde{\mathbf{Y}}_n(h) = \mathbf{S}\mathbf{P}\widehat{\mathbf{Y}}_n(h)$. The effect of the $\mathbf{P}$ matrix is to extract and combine the relevant elements of the base forecasts $\widehat{\mathbf{Y}}_n(h)$, which are then summed by $\mathbf{S}$ to give the final revised hierarchical forecasts, $\widetilde{\mathbf{Y}}_n(h)$.

The choice of $\mathbf{P}$ is a crucial step: the effect of its choice is to extract and combine the relevant elements of the base forecasts according to: $\mathbf{SP}\widehat{\mathbf{Y}}_n(h)$.
According with this formulation, the following reconciliation approaches have been used in order to reconcile the base forecasts introduced in the previous paragraph. Net of some exceptions, all the reconciliation methods were applied for each prediction method:

- **Top-Down**: this method entails the forecasting of the completely aggregated series (higher level of the hierarchy), and then the disaggregation of the forecasts based on historical proportions [5]. In this approach $\mathbf{P} = \left[\mathbf{p} \mid \mathbf{0}_{m_k \times (m-1)}\right]$, where $\mathbf{p}$ is a vector of proportions that sum to one. Different methods of top-down forecasting lead to different proportionality vectors $\mathbf{p}$. Some possible choices are the *Average historical proportions* that reflects the average of the historical proportions of the bottom-level series, the *Proportions of the historical averages* that captures the average historical value of the bottom-level series $Y_{j,t}$ relative to the average value of the total aggregate Y_t or the *Forecast proportions* that uses proportions based on forecasts rather than historical data.
- **Bottom-Up**: the idea is to forecast each of the disaggregated series at the lowest level of the hierarchy, and then to obtain forecasts at higher levels of the hierarchy by using simple aggregations. In this case $\mathbf{P} = \left[\mathbf{0}_{m_k \times (m-m_k)} \mid \mathbf{I}_{m_k}\right]$, so the $\mathbf{P}$ matrix extracts only bottom level forecasts from $\widehat{\mathbf{Y}}_n(h)$, which are then summed by $\mathbf{S}$ to give the bottom-up forecasts.
- **Mint**: the forecast reconciliation through trace minimization is an approach that incorporates the information coming from the full covariance matrix of forecast errors in order to obtain a set of coherent forecasts. It minimizes the mean squared error of the coherent forecasts across the entire collection of time series under the assumption that they are unbiased [3].
- **Bayesian approach**: as typical of this inferential paradigm, the central point is to explore and exploit the posterior distribution of the reconciled forecast. The distribution of the reconciled forecasts can be obtained by sampling from the posterior predictive distribution using a Gibbs sampler. Using such techniques it is possible to keep into account some prior belief about the component time series, so that during the reconciliation step, more importance is given to the more robust time series by taking into account the in-sample variance of each base forecast [2].
- **Ordinary or Weighted Least Square**: the revised forecast at each node will be a weighted average of the forecasts from all nodes. The weights are obtained using linear regression, where all the independent forecasts (from all nodes) are regressed against a set of dummy variables indicating which of the bottom-level series contribute to each node[4]. The mean squared reconciliation error, computed using the differences between the reconciled and independent forecasts, is as small as possible. In the OLS case the $\mathbf{P}$ matrix is equal to $(\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'$.

3 Data

The data used are the Italian Covid-19 incidence (new daily confirmed cases) obtained from the Github repository of the *Protezione Civile* agency. The data are daily updated and the information are released on a province basis. The first considered observation is recorded on 24^{th} February 2020 and the last one on 3^{rd} Novembrer 2020 for a total of 254 observations. The data are divided into two groups: the training set (from the first day until 14^{th} October 2020 for 234 observations) and the test set (from 15^{th} October 2020 until the last day for 20 observations). The dataset is made by 107 bottom-level time series, related to the Italian provinces.
Starting from these bottom level time series, a hierarchical structure has been created looking at the italian administrative organization, so the provinces have been aggregated into regions, the regions into the five macrozones partion and the macrozes summed up to obtain the total. The original data are transformed applying a logarithmic function. Due to the fact that the data are forecasted using models not well performing on count data. In detail, for a single value x_i the following transformation is applied: $x_i = log(x_i + \varepsilon)$.

4 Results

Each forecast and reconciliation method leads to different performances depending on the level of the hierarchy on which we focus. The most interesting level, for the purposes of the research, is the bottom-level series, that is the provinces. In over 70% of the 107 Italian provinces, the prediction adjusted by the reconciliation step is more accurate, in terms of RMSE, than the basic prediction.

Table 1 Count of bottom-level series where a method over-performs the others

Method	Count	Method	Count	Method	Count
Segmented	27	Ets + TD-GSF	4	Arima with cov. + TD-FP	2
Ets + OLS	13	Segmented + WLS	4	Segmented + TD-FP	1
Arima + TD-FP	10	Ets + WLS	3	Arima with cov. + OLS	1
Arima + Mint	9	Arima with cov.	3	Arima + TD-GSA	1
Ets + Mint	8	Arima	3	Arima + OLS	1
Ets + TD-FP	8	Ets + TD-GSA	3	-	-
Arima with cov. + Mint	4	Ets	2	-	-

Analyzing Table 1, it is possible to observe that the prediction method that most often has the best performance is the segmented regression. However, it must be considered that there are some Italian provinces in which the trend of Covid-19 daily cases in the considered period has remained stable and close to 0 even in the test period. Focusing on some of the larger Italian provinces, such as Rome (see Figure 1), the best score obtained without the reconciliation was of an RMSE equal to 0.59

(obtained applying segmented regression), while by inserting the reconciliation step, a value of 0.15 is reached (the ETS method was used for the base forecast). This behaviour is true for most of the bigger Italian provinces.

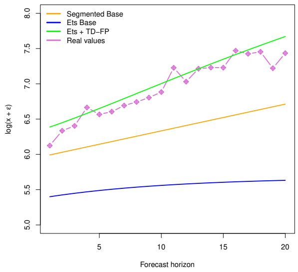

Fig. 1 Forecasted and real values for Covid-19 time series in Rome

It is important to note that among the reconciliation approaches there is no technique that systematically performs better than others. It must be borne in mind that the performance of the various series is profoundly different. Even at a higher level of aggregation (regions, macrozones, ...) there are important differences in the behaviour, which make it difficult to identify a single method valid all over the series.

5 Future works

Given the high variability of methods that over-perform others on the bottom-level time series, the construction of a method that, using a validation frame, is able to identify the best model would be very useful. Reconciliation techniques that make machine learning have recently been presented in the literature, and this could help in solving the problem by providing more accurate predictions.

References

1. G. Athanasopoulos and R. Ahmed and R. J. Hyndman: Hierarchical forecasts for Australian domestic tourism. International Journal of Forecasting **25**, 146-166 (2009)
2. F. Eckert and R. Hyndman and A. Panagiotelis: Forecasting Swiss Exports using Bayesian Forecast Reconciliation
3. S. Wickramasuriya and G. Athanasopoulos and R. Hyndman: Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization. Journal of the American Statistical Association - **114**, 1-45 (2018)
4. G. Athanasopoulos and R. Hyndman and R. Ahmed and H.L. Shang: Optimal combination forecasts for hierarchical. Computational Statistics & Data Analysis - **55**, 2579-2589 (2011)
5. Charles W. Gross and Jeffrey E. Sohl: Disaggregation methods to expedite product line forecasting (1990)

Link between Threshold ARMA and tdARMA models

Relazione tra modelli a soglia ARMA e modelli ARMA con parametri dipendenti dal tempo

Guy Mélard and Marcella Niglio

Abstract In the present contribution, we propose a link between Threshold Autoregressive Moving Average (TARMA) and Time-Dependent ARMA (tdARMA) models. We show that a proper parametrization allows to include the TARMA model in the large class of tdARMA structures. The main advantage that can be obtained from this result is the derivation of the asymptotic properties of the estimators of TARMA parameters that can be obtained under weaker conditions with respect to those in the available literature.
Abstract *Nel presente contributo proponiamo una relazione tra modelli a soglia autoregressivi media mobile (TARMA) e modelli ARMA con parametri dipendenti dal tempo (tdARMA). Diamo evidenza che una opportuna parametrizzazione consente di includere il modello TARMA nell'ampia classe dei modelli tdARMA. Il principale vantaggio che può essere ottenuto da questo risultato, è la derivazione delle proprietà asintotiche degli stimatori dei parametri TARMA che possono essere ottenuti sotto condizioni più deboli di quelle disponibili in letteratura.*

Key words: Threshold model, time-dependent ARMA model

1 Introduction

In time series analysis, the dynamic of data has often been modelled through the introduction of time-dependent coefficient structures. Starting from [8] different proposals have been given in this domain.

Guy Mélard
Université Libre de Bruxelles, Solvay Brussels School of Economics and Management, ECARES, Bruxelles, Belgique e-mail: gmelard@ulb.ac.be

Marcella Niglio
Department of Economics and Statistics, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano (SA), Italy e-mail: mniglio@unisa.it