A set of state space models at an high disaggregation level to forecast Italian Industrial Production

Riccardo Corradini

7 February 2018

A set of state space models at an high disaggregation level to forecast Italian Industrial Production

Riccardo Corradini*

Abstract

Normally econometric models that forecast Italian Industrial Production Index do not exploit pieces of information already available at time t+1 for its own main industry groupings. A new strategy is sketched here using state space models and aggregating the estimates to obtain improved results. The variables available at time t+1 are Consumption of Electricity, Compressed Natural Gas distributed on its own net, Production of Compressed Natural Gas, Compressed Natural Gas for thermoelectric use, Registration of commercial vehicles for Italy, Germany, France and Spain. Unfortunately for the other main industry groupings there are not available variables not prone to high revisions. The issue coming out from holidays taken during Tuesday or Friday will be tackled. How to handle in-sample forecast with different aggregating weights will be considered for the period before the first of January of 2010 where is impossible to use the same structure for the base year 2010.

Keywords: Industrial Production Index, forecasting, Vector Autoregressive Models, disaggregation, Kalman filter, unobserved components models.
Journal of Economic Literature Classification: C43, C51, C53, Q47.

1 Introduction

Forecasting Industrial Production can be a very difficult task, but perhaps forecasting its own subcomponents at an high disaggregation level can be even more challenging for researchers. This happens because there are few pieces of information at disaggregated level, and unless such data have strong comovements with the subcomponents, the risk is to worsen forecast results simply exploiting the past. I analyzed gross data on the Italian Industrial Production Index at an higher disaggregated level, exploiting the high correlation with other times series when they are available. I found not an high forecasting performance on average when poor or no information is available, consistent with existing literature, but larger improvements for those disaggregated components who face a richer correlation with other variables on their past and on the one step ahead prediction.

To the best of my knowledge this is the only study that compares the short-term forecasting performance of a naive autoregressive model applied to Italian Industrial Production Index, with a set of state space models applied to its own main industry groupings. Such an effort is motivated to understand evolution of Italian economy in the short term. It is mandatory to continuously review the estimates to monitor evolution of quarterly estimates of Gross Domestic Product. There is a trade-off between the need to provide Industrial Production forecast closely related with the future evolution of Gross Domestic Product in a short time and doing such job exploiting as much is possible all the pieces of information available with the most suitable econometric tools.

Brunhes-Lesage and Darné (2012) point out analyzing the French Industrial Production Index also that:"The IIP, however, is characterized by a significant publication delay, around 40 days after the end of the reference month for the main European countries, and the first IIP estimations are often revised significantly. Thus, it is less useful for short-term forecasting exercises." Italian data are not so prone to high revisions, but they are also characterized by the same delay (40 days). The aim of this paper is, therefore, to propose many models designed to forecast the current-month Italian Industrial Production Index (nowcast), using data listed in table 1 and to forecast Italian Industrial Production Index for the next two months after the nowcasted month.

The remainder of the paper is organized as follows. Section 2 describes the data used and introduces the models. Section 3 describes the issues related to long weekends. Section 4 outlines some issues related to backcast some series and about constructing the weights where is impossible to use the structure of the base year 2010. Section 5 concerns about the model used to estimate simultaneously IPI for intermediate and capital goods. Section 6 describes the models for durable goods and non durable goods. Section 7 describes models for energy's disaggregation. Section 8 is about the forecasting performance. Section 9 concludes.
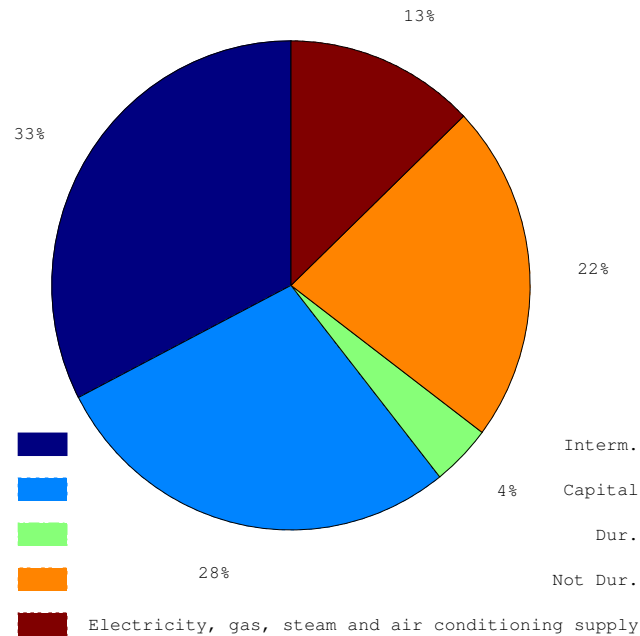
Figure 1: This pie graph shows the weights of Main Industry Groupings of Italian Industrial Production Index

2 A light description for the models used for disaggregated components

In figure 3 is plotted the level of Industrial Production Index for Intermediate goods, for Capital goods and the sum of Registration of light commercial vehicles over Italy, Spain, France and Germany. The last variable is always available at time t+1 and it is used in vector autoregressive moving average model in first difference of first log seasonal differences shaped with the aforementioned endogenous variables. I am assuming that these goods are transported in these countries monthly and that persists a relationship among their seasonal growth rates. A slight different strategy will be considered for industrial production index of durable goods without using the log transformation. A bivariate seemingly unrelated time series equations (see example about car drivers accidents in Durbin and Koopman (2012)) will be composed by the total sum of light commercial vehicles and industrial production index of durable goods. In contrast, no pieces of information are available for industrial production index of non durable goods so here a popular benchmark namely the autoregressive model of order three over the seasonal differences with
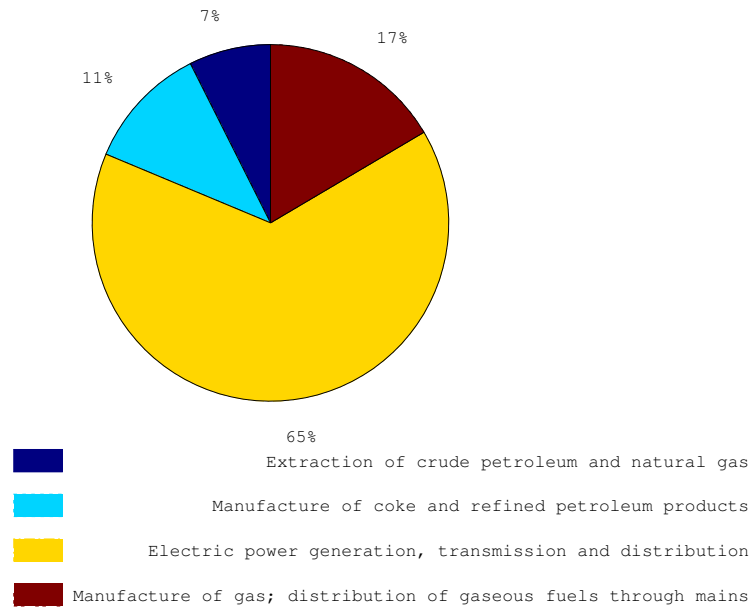
Figure 2: This pie graph shows the weights of every sub component concerning Electricity Production, CNG transportation, Extraction of CNG and Petroleum, Rafineries

no logarithmic transformation will be exploited. On the contrary many pieces of information are ready at the time of publication on ISTAT website concerning production electricity and other sub components. The daily data about consumption of electricity can be cumulated to obtain a preliminary estimate of the monthly data that will be released roughly fifteen - twenty days later. Data about compressed natural gas are less prone to revision and when they are cumulated to obtain their monthly value do not differ too much from their preliminary value. More in detail to obtain an estimate of main industry grouping Industrial Production Index of Electricity, gas, steam and air conditioning supply we have to consider five sub-components:

- Extraction of crude petroleum and natural gas;

- Manufacture of coke and refined petroleum products;

- Electric power generation, transmission and distribution;

- Manufacture of gas; distribution of gaseous fuels through mains;

Table 1: Information sources of data

Name	Number of series	Sources	Freq.	Time [a]
IPI	5	ISTAT	Monthly	-
IPI of Electricity, gas, steam and air conditioning supply	4	ISTAT	Monthly	-
GigaWatt Electricity net production	1	TERNA	Monthly	+50+55 days
GigaWatt Electricity Consumption	1	TERNA	Monthly	+50+55 days
GigaWatt Electricity Consumption	1	TERNA	Daily	Every Day
Total m3 of Compressed Natural Gas transported	1	SNAM	Daily	Every Day
Production of m3 of Compressed Natural Gas	1	SNAM	Daily	Every Day
M3 of Compressed Natural Gas for Thermoelectric use	1	SNAM	Daily	Every Day
Registration of Light Commercial Vehicles in Italy	1	ACI	Monthly	-5-7 days
Registration of Light Commercial Vehicles in Spain	1	ANFAC	Monthly	-5-7 days
Registration of Light Commercial Vehicles in France	1	CCFA	Monthly	-5-7 days
Registration of Light Commercial Vehicles in Germany	1	KBA	Monthly	-5-7 days

[a] Pubblication corresponds to amounts of days before the official release .

Most of the times extraction of crude petroleum happens together with natural gas. Unfortunately no pieces of information are available at time t+1 for crude petroleum but we do have daily data about extraction of natural gas from 2006 up to now. These data are cumulated and inserted into Seemly Unrelated Time Series Equations (see for instance the example about car drivers killed or serious injured in Great Britain from January 1969 to December 1984 in Durbin and Koopman (2012)). For t+2 roughly 10 daily observations are always available and will be appended to daily time series about production of compressed natural gas. Manufacture of coke and refined petroleum products is estimated by an autoregressive moving average identification procedure described in Gómez and Maravall (2000) applied to ISTAT's monthly data. Electric power generation, transmission and distribution is somewhat more complicated because two endogenous variables are available at time t+1 :
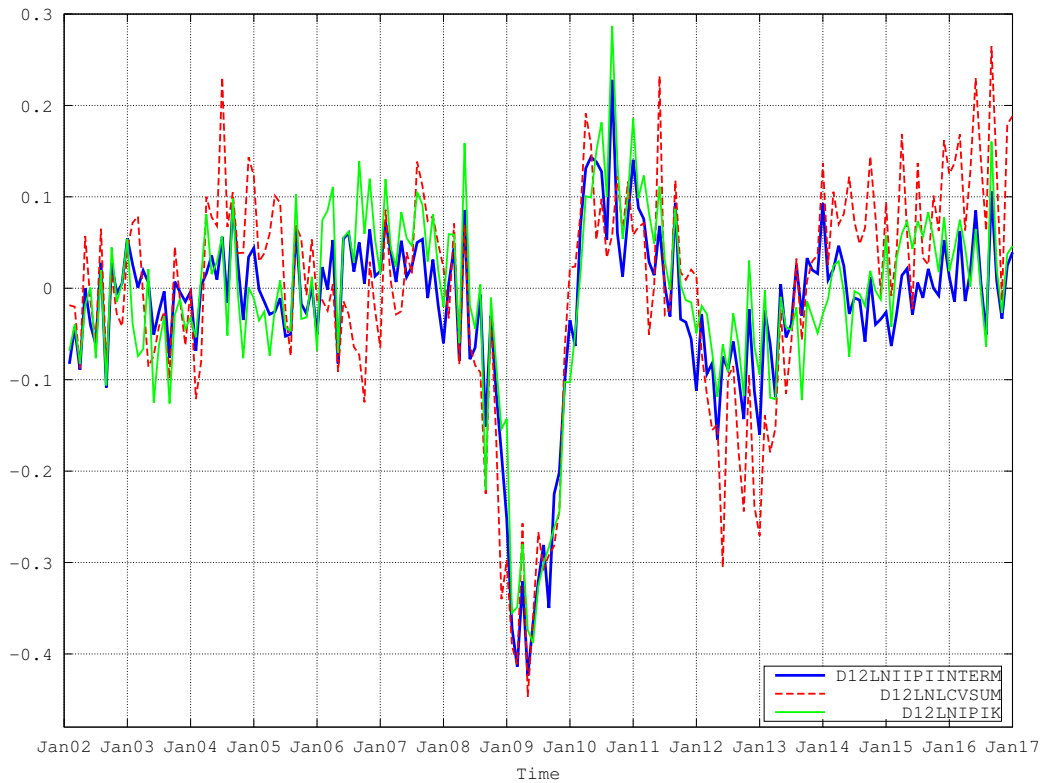
Figure 3: This graph plots first log seasonal differences of Industrial Production Index of Intermediate Goods, Capital Goods and Total sum of Light Commercial Vehicles registration over Italy, Germany, France and Spain

- Consumption of electricity

- Compressed natural gas for thermoelectric use

A univariate model over the seasonal differences with no logarithmic transformation and cumulated daily data with consumption of electricity and compressed natural gas for thermoelectric use as exogenous variable will be set up and finally for period t+1 the two variables mentioned above will be used to forecast electric power generation, transmission and distribution. For t+2 and t+3 a naive autoregressive model of order three using first seasonal weekly differences applied to daily data about consumption and thermoelectric, national natural gas production and transportation will allow us to have more reliable estimates for production of electricity, extraction of petroleum and natural gas and finally distribution of gaseous fuels through mains at time t+2 and t+3.

3   Long weekends

During some months there are fixed holidays that come on Tuesdays or on Fridays. A typical example is the Immaculate Conception the 8th of December of yearly. Roughly speaking, the 7th of December of 2015 many workers requested to go on holidays. In this way they used four days (Saturday the 5th of December, Sunday the 6th of December, Monday the 7th of December and finally the day of the Immaculate Conception). The raw data about Industrial Production Index are affected by such slowdown on Monday the 7th. Nevertheless, ISTAT does not consider such group of people that do not work during that day. The calendar and seasonal adjustment is strict. Alternatively most of the times a calendar adjusted time series is more predictable than a gross one.

A time series not affected by strikes or unexpected holidays could be better forecasted than a gross one. To tackle this problem I propose to trace on the calendar from the first of January 2001 the following days:

- Epiphany the 6th of January;

- Italian Republic day the 2nd of June;

- Immaculate Conception the 8th of December;

- International Workers' Day the 1st of May;

To trace the long weekends I propose the following strategy:

- for Epiphany, Italian Republic day, Immaculate Conception find the years where they happen on Tuesdays and Thursday;

- create a binary dummy variable to inform the model there will be a discrepancy between what it expects and what happens in the real world and place it on Mondays or Fridays for daily data;

- for the first of May let the computer decide; if the month of April is affected when Labour Day happens on Tuesday or Thursday

- Finally, compute dummy variables for the months over the years detected from the procedure mentioned above.

4   Some reflections over the period from January 2006 up to December 2009

The period from January 2001 up to December 2009 deserves an ad hoc treatment because I want to compare the performance of the disaggregated composed model

and a naive benchmark model .i.e. autoregressive model of order three over the log seasonal differences over the whole index, I have to tackle two issues:

- backcast consumption of electricity daily data from 2006 up to 2013 exploiting compressed natural gas for thermoelectricity daily data;

- since the weights for every subcomponent are not fixed for this period like from January 2010 up to nowadays (.i.e. they were rebuilt in terms of growth rate at the expenditure of fixed weight base year structure), it is mandatory to compute the discrepancy between the data published and the data obtained holding constant the structure of base year weight (i.e. 2010).

$$
\begin{aligned}
IPI &= \alpha IPI_x + \beta IPI_y + discrepancy : \\
discrepancy &= discrepancy_\alpha + discrepancy_\beta; \\
discrepancy_\alpha &= discrepancy \cdot \alpha; \\
discrepancy_\beta &= discrepancy \cdot \beta
\end{aligned}
\tag{1}
$$

5   Conditional Vector Autoregressive Moving Average Model for Intermediate Goods, Capital Goods, Total Sum of Light Commercial Vehicles Registration over Italy, Spain, France and Germany

Bruno and Lupi (2004) use the following unrestricted Vector Autoregressive Model:

$$
\Delta\Delta_{12}y_t = \beta\Delta_{12}y_{t-1} + \Sigma_{j-1}^{13}\gamma_j\Delta\Delta_{12}y_{t-j} + \phi d_t + \varepsilon_t
\tag{2}
$$

where $\Delta = (1\text{-}L)$, $\Delta_{12} = (1 - L^{12})$, L is the usual lag operator such that $L^p z_t = z_{t-p}$, $y_t = (log(IIPI_t); log(TONFSt); PPt)$ (PPt represents production prospects from ISTAT surveys and TONFSt stands for tons of raw material transported by Italian railways for intermediate goods. Bruno and Lupi (2004) apply to PPt a logistic transformation, d_t contains some deterministic components (constant, specific impulse dummies). The endogenous variables available at the time t+1 of forecast are $TONFS_t$ and PP_t. For the next periods the forecast is unconditional since there is no further information. Since together intermediate goods plus capital goods account for 61 percent of the whole index and they are transported on light commercial vehicles I take the sum of their Registration over Italy, France, Spain and Germany. So the new model is

$$
\Delta\Delta_{12}y_t = \phi d_t + \Sigma_{j=1}^{13}\gamma_j\Delta\Delta_{12}y_{t-j} + \Sigma_{k-1}^{13}\theta_k\varepsilon_{t-k}
\tag{3}
$$

where $\Delta = (1\text{-}L)$, $\Delta_{12} = (1 - L^{12})$, L is the usual lag operator such that $L^p z_t = z_{t-p}$, $y_t = (log(IPIINTERMEDIATE_t); log(IPICAPITAL_t); log(SUMLCV_t))$ ($SUMLCV_t$ represents the total level of Registrations of light commercial vehicles over the four mentioned European countries, while d_t contains the long weekend dummies, the growth rates of number of days worked over the four countries and the growth rates of number of days worked in Italy computed as first differences as first log year to year differences).

6 Seemly Unrelated Time Series Equations for Durable Goods, seasonal differences autoregressive model for non durable goods

For Durable Goods I exploit a similar system of equations mentioned above (see Durbin and Koopman (2012)) without using logarithmic transformation. More in detail here $y_t = ((IPIDURABLEGOODS)_t; SUMLCV_t)$. On the contrary at the time of Italian Industrial Index publication on ISTAT website there won't be any pieces of information available for non-durable consumer goods index. I will use a popular autoregressive model without the logarithmic transformation.

$$\Delta_{12} y_t = d_t + \Sigma_{j=1}^3 \Delta_{12} \gamma_j y_{t-j} + \varepsilon_t \tag{4}$$

where $y_t = (IPINONDURABLEGOODS)_t$ and d_t contains the same dummies and same working days number variables mentioned in equation 3.

7 Which models for energy's disaggregation?

Here the picture is somewhat more complicated by information available on monthly and daily basis. More in detail the daily observations about consumption of electricity downloadable from Terna website (www.terna.it) start from the first January of 2013 up to nowadays. For the past only chunks of observations are available. It is mandatory to backcast this time series using daily information of m^3 of compressed natural gas for thermoelectric use available from Snam website (www.snam.it). In the past most of production of electricity was generated by compressed natural gas so it sounded fair to exploit these series with no missing values as an exogenous regressor and put it in the benchmark model with weekly frequencies in place of monthly frequencies, and hoc set of daily dummies (see the appendix for further details) and no logarithmic transformation using data from 2013 up to now. Once the model is identified and casted in Kalman Filter equations the chunks available from Terna website will be inserted into the equations of

Kalman filter smoother to backcast the whole time series of daily consumption from January 2006 up to now. Such data finally can be cumulated on monthly basis and compared with published monthly data from 2006 up to now. For the sake of simplicity such discrepancies are not reported here since they are negligible. At this point three endogenous variables are available:

- index of production of electricity until time t

- consumption of electricity until time $t + 1$

- meter cubes of compressed natural gas for thermoelectric use until time $t + 1$

Since at the time of publication on ISTAT website roughly ten daily observations for t+2 are already available and they correspond to roughly one third of a month, the benchmark model on daily data mentioned above is used to compute t+2 and t+3. Again with a similar equation 3 it will be estimated index of production of electricity at time $t + 1$, $t + 1$, $t + 3$. The endogenous variables available for the forecast window will be consumption of electricity cumulated on daily basis, volumes of natural gas for thermoelectricity. The same strategy will be exploited with only two variables for the index correlated with transportation of meter cubes of compressed natural gas. Extraction of crude petroleum and natural gas requires a model more flexible, so a seemingly unrelated equations is used. See Gómez (2015) for all the details about the implementations. Manufacture of crude petroleum suffers from insufficient information somewhat compensated by its own relative low weight (only 11 per cent on the Main Index Grouping Index, see fig.2). Only for this last component an automated model identification procedure will be used (see Gómez and Maravall (2000)).

8 Performance of the index compared to a naive benchmark over three steps

In table 2 are summarized the performances of the models applied to disaggregated components mentioned in figure 1. An expanding window starting from the January of the year 2001 and moving from the January of 2008 up to the November of 2017 [1] is considered.
The mean absolute forecast error (https://en.wikipedia.org/wiki/Mean_absolute_ error) is computed over the three step ahead horizon using the first logarithmic seasonal differences. The column MAE-BEN shows the results for the autoregressive of order three over the seasonal logarithmic differences.

---

[1] The dataset is updated on monthly basis and is not prone to high revisions (see dati.istat.it/).

Table 2: Comparison of mean absolute error with expanding window.

Index	Step 1		Step 2		Step 3	
	MAE-BEN	MAE-DIS	MAE-BEN	MAE-DIS	MAE-BEN	MAE-DIS
IPI-INT	0.0822	0.0414	0.0831	0.0541	0.0848	0.0576
IPI-CAP	0.0503	0.0459	0.0529	0.0568	0.0574	0.0614
IPI-DUR	0.0572	0.0559	0.0589	0.0589	0.0606	0.0592
IPI-NDUR	0.0358	0.0251	0.0357	0.0260	0.0355	0.0269
IPI-NRG	0.0452	0.0170	0.0496	0.0322	0.0516	0.0776
IPI-WHOLE	0.0397	0.0269	0.0382	0.0577	0.0397	0.1044
IPI-BENAGG	0.0362	NaN	0.0373	NaN	0.0518	NaN

Table 3: Comparison of mean absolute error with expanding window for electricity.

Index	Step 1		Step 2		Step 3	
	MAE-BEN	MAE-DIS	MAE-BEN	MAE-DIS	MAE-BEN	MAE-DIS
IPI-EXTRA	0.0619	0.0488	0.0928	0.1918	0.1283	0.2209
IPI-PETROL	0.0509	0.0391	0.0599	0.0481	0.0632	0.0518
IPI-PRODELE	0.0408	0.0251	0.0466	0.0305	0.0483	0.0462
IPI-CNG	0.0813	0.0061	0.0852	0.0650	0.0877	0.1101
IPI-WHOLE	0.0390	NaN	0.0444	NaN	0.0470	NaN
IPI-BENAGG	0.0452	0.0170	0.0496	0.0322	0.0516	0.0776

Table 4: Comparison of mean absolute error of benchmark model with autoregressive model of order three with no logarithmic transformation plus dummies for long weekends

Index	Step 1	Step 2	Step 3
	MAE	MAE	MAE
IPI-BEN	0.0418	0.0417	0.0455
IPI-NOLOG-AR3LONGWEEKEND	0.0352	0.0376	0.0446
IPI-NOLOG-AR3LONGWEEKEND-CONDDIS	0.0269	0.0321	0.0404

Alternatively the column MAE-DIS shows the performances of the aforementioned models. The disaggregated model outperforms the benchmark at one step. On the contrary the benchmark beats the disaggregated model on the second and third step (see the line IPI-WHOLE). Disaggregating simply using the past does not help. The line IPI-BENAGG shows that aggregating the benchmarks applied to the subcomponents worsen the situation. This concept is confirmed looking at the results shown in table 3. Simply using ten days of data we have better results not only at step 1, but also at step 2 (see the column MAE-DIS for the row IPI-BENAGG). These observations point out that we may use the autoregressive model with no logarithmic transformation and the long weekend dummies for the step 2 and the step 3 (see table 4 at the last line).

9 Conclusions

The aim of this paper is to understand if it does worth to exploit pieces of information already available at highest diasaggregation level at time t+1 and at time t+2. Of course when the forecasting errors propagate over time, even a naive autoregressive model performs better than a more disaggregated one. Yet, using an hybrid approach, i.e. conditioning a model exploiting the long weekends persistence over time with the aggregated one step, the results improve over three steps horizon.

10 About this document

This document is released under the Creative Commons Attribution 4.0 International (CC BY 4.0)
See https://creativecommons.org/licenses/by/4.0/.

11 Appendix

11.1 A model to backcast daily consumption of electricity from 2006:01:01 up to 2012:12:01

From www.terna.it and from www.snam.it is possible to download daily data from January of 2013 up to nowadays. If on one hand we do have daily data from 2006 up to 2012 of production of natural gas for thermoelectric use, however only chunks are available for consumptions of electricity during this period in the past. I assume that in the remote past renewable sources of energy (like wind, solar

power, hydroelectricity, geothermic source) had been less relevant.

$$\Delta_7 log(C_t) = const_t + \phi_1 \Delta_7 log(C_{t-1}) + \phi_2 \Delta_7 log(C_{t-2}) + \\ \phi_3 \Delta_7 log(C_{t-3}) + \beta_1 \Delta_7 log(CH4_t) + \gamma D_t + \varepsilon_t \tag{5}$$

Equation 5 will be casted in state space form to backcast daily consumption data in the remote past. The equation is a univariate autoregressive model over weekly frequencies with a set of exogenous regressors composed be weekly seasonal log differences daily data of natural gas for thermoelectric use and by a year fixed seasonal cycle over a year (see Gómez (2015)).

More in detail D_t is obtained by the following equation according to Gregorian calendar with a period of 365.25:

$$s_t = a \cdot cos(wt) + b \cdot sin(wt), \\ w = 2\pi \cdot \frac{k}{n} \tag{6}$$

For a more complex example considering two fixed seasonal patterns (according to the Gregorian and the Hiri calendar) see Livera et al. (2011). For all the explanations concerning filtering and smoothing (used here to backcast data) is possible to have a look at chapter four of Gómez (2016).

References

Brunhes-Lesage, V., and Darné, O. (2012). Nowcasting the French index of industrial production: A comparison from bridge and factor models. *Economic Modelling*, 29: 2174–2182. ISSN 0264-9993. DOI 10.1016/j.econmod.2012.04. 011. URL http://dx.doi.org/10.1016/j.econmod.2012.04.011.

Bruno, G., and Lupi, C. (2004). Forecasting industrial production and the early detection of turning points. *Empirical Economics*, 29(3): 647–671. ISSN 1435-8921. DOI 10.1007/s00181-004-0203-y. URL https://doi.org/10.1007/ s00181-004-0203-y.

Durbin, J., and Koopman, S. (2012). *Time Series Analysis by State Space Methods*. Oxford University Press. URL https://books.google.it/books?id=lGyshsfkLrIC& dq=durbin+koopman+Time+series+analysis+by+state+space+models&lr= &hl=it&source=gbs_navlinks_s.

Gómez, V. (2015). SSMMATLAB: A Set of MATLAB Programs for the Statistical Analysis of State Space Models. *Journal of Statistical Software, Articles*, 66(9): 1–37. ISSN 1548-7660. DOI 10.18637/jss.v066.i09. URL https://www.jstatsoft. org/v066/i09.

Gómez, V. (2016). *Multivariate Time Series With Linear State Space Structure*. Springer. DOI 10.1007/978-3-319-28599-3. URL http://www.springer.com/gp/book/9783319285986.

Gómez, V., and Maravall, A. (2000). *Automatic Modeling Methods for Univariate Series, in A Course in Time Series Analysis*. John Wiley & Sons, Inc., Hoboken, NJ, USA. URL http://onlinelibrary.wiley.com/doi/10.1002/9781118032978.ch7/summary.

Livera, A. M. D., Hyndman, R. J., and Snyder, R. D. (2011). Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing. *Journal of the American Statistical Association*, 106(496): 1513–1527. DOI 10.1198/jasa.2011.tm09771. URL https://doi.org/10.1198/jasa.2011.tm09771.