Selection of lags and univariate models to forecast industrial production in an Emerging country: is disagreggation useful?

Diogo de Prince*
Emerson Fernandes Marçal†

Abstract

A point to address is if the use of disaggregated data helps in forecasting variables. Our point is whether the prediction of the disaggregated components of Brazilian industrial production improves the accuracy of the forecast of aggregate Brazilian industrial production. Our contribution is that we do not know articles that address the contribution of the disaggregated data of the weighted lag adaptative least absolute shrinkage and selection operator (WLadaLASSO) methodology or the exponential smoothing (selecting the most appropriate model). We estimate a rolling window of 100 fixed observations and we evaluate the forecast from 1 to 12 months ahead for Brazilian industrial production, in which we estimate 61 rolling windows. Our results point to a better performance of the exponential smoothing model with disaggregated data for the forecast of 1 to 7 months ahead for Brazilian industrial production by mean square error (MSE). But WLadaLASSO disaggregated forecasts with more accuracy for 8 to 12 months ahead by MSE.

Abstract

Um ponto do trabalho é abordar se o uso de variáveis desagregadas auxilia na previsão de variáveis. Nosso ponto é se a previsão de componentes desagregados para a produção industrial brasileira melhora a acurácia da previsão para a produção industrial brasileira agregada. Nossa contribuição é que não conhecemos artigos que tratam do uso de dados desagregados com a metodologia weighted lag adaptative least absolute shrinkage and selection operator (WLadaLASSO) ou suavização exponencial (no qual seleciona-se o modelo mais apropriado). Nós estimamos uma janela móvel de 100 observações fixas e nós avaliamos a previsão de 1 a 12 meses a frente, no qual nós estimamos 61 janelas móveis. Nosso resultado aponta para uma melhor performance do modelo de suavização exponencial com dados desagregados para prever de 1 a 7 meses a frente pelo erro quadrático médio (EQM). Mas o WLadaLASSO desagregado prevê com maior acurácia de 8 a 12 meses a frente pelo EQM.

JEL Codes: C53, E27, C52
Key Words: industrial production, forecasting, model selection.
Palavras-chave: produção industrial, previsão, seleção de modelos.

Área 4 - Macroeconomia, Economia Monetária e Finanças

*Assistante Professor at Federal University of Sao Paulo
†Head of Center for Applied Macroeconomic Research at Sao Paulo School of Economics

Selection of lags and univariate models to forecast industrial production in an Emerging country: is disagreggation useful?

Abstract

A point to address is if the use of disaggregated data helps in forecasting variables. Our point is whether the prediction of the disaggregated components of Brazilian industrial production improves the accuracy of the forecast of aggregate Brazilian industrial production. Our contribution is that we do not know articles that address the contribution of the disaggregated data of the weighted lag adaptative least absolute shrinkage and selection operator (WLadaLASSO) methodology or the exponential smoothing (selecting the most appropriate model). We estimate a rolling window of 100 fixed observations and we evaluate the forecast from 1 to 12 months ahead for Brazilian industrial production, in which we estimate 61 rolling windows. Our results point to a better performance of the exponential smoothing model with disaggregated data for the forecast of 1 to 7 months ahead for Brazilian industrial production by MSE. But WLadaLASSO disaggregated forecasts with more accuracy for 8 to 12 months ahead by MSE.

JEL Codes: C53, E27, C52

Key Words: industrial production, forecasting, model selection.

1 Introduction

Economic agents decide based on a global view of how the economy behaves at that moment. The general level of output, employment, interest rates, exchange rates and inflation are examples of important economic indicators that help in the country's diagnosis. One of the most comprehensive and important macroeconomic indicators of the economy is the Gross Domestic Product (GDP), as it is a proxy for a country's economic performance. In the present work, the proxy used for GDP is industrial production, since the indicator of industrial production is monthly (higher frequency than GDP) and is released with a lag of about one month, which is therefore smaller than GDP (with a delay of more than two months). Therefore, the proposition and evaluation of econometric models of forecasting are relevant and bring benefits to build better forecasts to enable more accurate expectations of economic agents.

A point to address is if the use of disaggregated data helps in forecasting variables. Data disaggregation is an alternative to lead to more accurate forecasts. This alternative refers to the decomposition of the main variable into several subcomponents, which have different weights for the aggregate series. These subcomponents are estimated individually and then grouped to obtain a forecast of the aggregate series. This technique can increase the quality of the forecast because we model the subcomponents taking into account their individual characteristics. We use this alternative in the present work to understand if there are gains to predict the aggregate series if we estimate each subcomponent and we use the weight of this subcomponent.

The accuracy of this alternative for forecasting has been discussed in some studies. The theoretical studies indicate that when the data generating process (DGP) is known, it is preferable to first use the data disaggregated in multiple series to later aggregate them, than to directly predict the already aggregated series. However, the author acknowledges that in most cases DGP is unknown and therefore the use of aggregate series may be preferable due to variability of specification and estimation of the model. Some examples of contributions to the theoretical literature on aggregate or disaggregate forecasting are Lutkepohl (1984, 1987), Granger (1987), Pesaran, Pierse, and Kumar (1989), Garderen, Lee, and Pesaran (2000), and Giacomini and Granger (2004). As DGP is not known, then the question (if aggregating the disaggregated forecasts improves the accuracy of the aggregate forecast) becomes an empirical question.

Our point is whether the prediction of the disaggregated components of Brazilian industrial production improves the accuracy of the forecast of aggregate Brazilian industrial production. Our contribution is that we do not know articles that address the contribution of the disaggregated data of the weighted lag adaptative least absolute shrinkage and selection operator (WLadaLASSO) methodology or the exponential smoothing (selecting the most appropriate model). In addition, there are few papers that analyze the contribution of disaggregated data to industrial production and we intend to fill this gap. From monthly data from 2002 to 2017, we select the best unviariate model, estimate a rolling window of 100 fixed observations and we evaluate the forecast from 1 to 12 months ahead for Brazilian industrial production, in which we estimate 61 rolling windows. We consider as naive models the first order autoregressive model (AR(1)), AR(1) with time-varying parameters (TVP AR(1)), the Stock and Watson (2007) unobserved components with stochastic volatility (UC-SV), estimated based on Kroese and Chan (2014). We consider the following methods for selecting the best model: exponential smoothing based on

Hyndman et al (2002), Hyndman and Khandakar (2008) and Hyndman et al (2008), the least absolute shrinkage and selection operator (LASSO), adaptive LASSO (adaLASSO), and the WLadaLASSO. We use the LASSO and its variants to select the lags from an AR (15). We compare the prediction performance between the models based on the mean square error (MSE) and the Diebold and Mariano (1995) test. Our results point to a better performance of the exponential smoothing model with disaggregated data for the forecast of 1 to 7 months ahead for Brazilian industrial production by MSE. But WLadaLASSO disaggregated forecasts with more accuracy for 8 to 12 months ahead by MSE. To analyze whether there is a better statistical performance, we use the exponential model with disaggregated data a benchmark with the Diebold and Mariano (1995) test. We can consider that the disaggregated exponential smoothing model presents better statistical performance for prediction in relation to the naive models (AR(1), AR(1) with time-varying parameters, UC-SV) considered with aggregate or disaggregated data and the aggregate exponential smoothing model.

The article structure contains six sections in addition to this introduction. The following is a brief review of the literature. Next, we address the methodology of the models considered in the paper. Section 4 presents the data used, the empirical forecasting strategy and the Diebold and Mariano (1995) test to compare the performance of the models. In section 5, we discuss the results of the study. Finally, we show the final considerations of the research and the next steps.

2 Literature Review

The present section address some empirical articles that address the difference in forecast accuracy between aggregating the disaggregated forecasts or modeling only the aggregate variable.

Stock and Watson (1998) compare 49 univariate projection models to forecast industrial activity and inflation in the United States from 215 monthly series from the years 1959 to 1996. One conclusion of the paper is the finding that the aggregation of forecasts exceeded the performance of separate forecasts. The authors also find that gains from the use of forecast combinations are significant enough to justify their use by a risk-averse analyst. The authors also point to the importance of performing the unit root test to reduce errors substantially in the estimates. Tobias and Zellner (1998) seek to determine the effects of aggregation and disaggregation to predict the average annual growth rate of 18 countries. In general, the disaggregation leads to more observations to estimate the parameters, besides the authors obtain better predictions for the aggregate variable (growth rate).

Marcellino, Stock and Watson (2003) find evidence that the individual estimation of inflation in each euro area country and the subsequent aggregation of projections increases the accuracy of the final result of the estimation, in relation to the option to predict this variable only in aggregate level. Hubrich (2005) obtains that aggregating forecasts by component of inflation does not necessarily better predict year-on-year inflation in the euro area one year ahead.. Espasa, Senra and Albacete (2002) have similar results indicating, however, that the disaggregation leads to better projections for periods longer than one month. Carlos and Marçal (2016) compare the prediction from models for aggregate inflation and aggregating the forecasts for the groups and items of the Brazilian inflation index. The authors obtain that there are gains in the accuracy of the forecast with disaggregated data.

Barhoumi et al (2010) analyze the forecasting performance of the France GDP between alternative factor models. The point of the work is whether it is more appropriate to extract data factors with aggregates or disaggregated to predict. Rather than using 140 disaggregated series, Barhoumi et al (2010) show that the static approach of Stock and Watson (2002) with 20 aggregate series leads to better prediction results. The next section presents the methodology we use in this paper.

3 Methodology

We consider three naive estimators to compare our model forecasts: UC with stochastic volatility, autoregressive model and time-varying parameters autoregressive. Our point is select the lag variables that are relevant based on different univariate methodologies. We use LASSO and two variants (LASSO, ADALASSO and WLADALASSO) and exponential smoothing method with ETS.

3.1 Time-varying parameters autoregressive of first order

In this section, we present the methodology we use to the autoregressive first order model with time-varying parameters. This model serves as a naive predictive alternative and therefore we use only one lag. The methodology

with time-varying parameters seeks to contemplate the changes that can occur in the economy over time (Kapetanio et al, 2017). Consider the time-varying paramenter AR(1) model, where we can write the measurement equation as

$$y_t = \beta_{0t} + \beta_{1t} y_{t-1} + \varepsilon_t \tag{1}$$

where $\varepsilon_t \sim N\left(0, \sigma^2\right)$ for $t = 1, ..., T$, and y_0 is an initial observation. We can write the autoregressive coefficients $\beta_t = (\beta_{0t}, \beta_{1t})^\top$ with the following transition equation

$$\beta_t = \beta_{t-1} + u_t \tag{2}$$

where $u_t \sim N(0, \Omega)$ and the transition equation has the initial value $\beta_1 \sim N(\beta_0, \Omega_0)$.

We will use Bayesian estimation following Kroese and Chan (2014) and for this we consider the equation (1) with matrix notation

$$y_t = x_t^\top \beta_t + \varepsilon_t \tag{3}$$

where $x_t^\top = (1, y_{t-1})$. If we stack the observations over all times t, we have

$$y = X\beta + \varepsilon \tag{4}$$

where $y = (y_1, ..., y_T)^\top$, $\beta = \left(\beta_1^\top, ..., \beta_T^\top\right)^\top$, $\varepsilon = (\varepsilon_1, ..., \varepsilon_T)^\top \sim N\left(0, \sigma^2 I\right)$, and $X = \begin{pmatrix} x_t^\top & 0 & \dots & 0 \\ 0 & x_t^\top & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & x_t^\top \end{pmatrix}$. The logarithm of the joint density function of y (omitting the initial observation y_0) is

$$\ln f\left(y|\beta, \sigma^2\right) = -\frac{T}{2} ln\sigma^2 - \frac{1}{2\sigma^2}\left(y - X\beta\right)^\top \left(y - X\beta\right) + const \tag{5}$$

where *const* is the constant term. Now we can stack the transition equation (2) over t. We consider $\beta_0 = 0$ for simplification. We can write the transition equations in matrix form as

$$H\beta = u \tag{6}$$

where $u \sim N(0, S)$, $u = (u_1, ..., u_T)^\top$, $H = \begin{pmatrix} I & 0 & \dots & 0 & 0 \\ -I & I & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & -I & I \end{pmatrix}$, and $S = \begin{pmatrix} \Omega_0 & 0 & \dots & 0 \\ 0 & \Omega & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Omega \end{pmatrix}$.

Given that $|H| = 1$ and $|S| = |\Omega_0||\Omega|^{T-1}$, the logarithm of the joint density function of β is given by

$$\ln f(\beta|\Omega) = -\frac{T-1}{2} \ln|\Omega| - \frac{1}{2}\beta^\top H^\top S^{-1} H\beta + const \tag{7}$$

We can reduce the number of parameters assuming that Ω is diagonal. So consider $\omega^2 = \left(\omega_0^2, \omega_1^2, ..., \omega_p^2\right)^\top$ as the vector of diagonal elements of Ω.

We can obtain the posterior density specifying the prior for σ^2 and ω^2. Assume an independent prior $f\left(\sigma^2, \omega^2\right) = f\left(\sigma^2\right) f\left(\omega^2\right)$, where $\sigma^2 \sim IG\left(\alpha_{\sigma^2}, \lambda_{\sigma^2}\right)$ and $\omega_i^2 \sim IG\left(\alpha_{\omega_i^2}, \lambda_{\omega_i^2}\right)$. We specify the constants α_{σ^2}, λ_{σ^2}, $\alpha_{\omega_i^2}$, and $\lambda_{\omega_i^2}$.

The posterior density function is given by

$$f\left(\beta, \sigma^2, \omega^2|y\right) \propto f\left(y|\beta, \sigma^2\right) f\left(\beta|\omega^2\right) f\left(\sigma^2\right) f\left(\omega^2\right) \tag{8}$$

where $f\left(y|\beta, \sigma^2\right)$ and $f\left(\beta|\omega^2\right)$ is given by (5) and (7) respectively. We can obtain posterior draws using Gibbs sampler. We draw from $f\left(\beta|y, \sigma^2, \omega^2\right)$ followed by a draw from $f\left(\sigma^2, \omega^2|y, \beta\right)$. As $f\left(\beta|y, \sigma^2, \omega^2\right)$ is a normal density, if we determine the mean vector and the precision matrix, we can apply the algorithm described below to obtain a draw from it efficiently. Using (5) and (7), we write

$$\ln f\left(\beta|y, \sigma^2, \omega^2\right) = \ln f\left(y|\beta, \sigma^2\right) + \ln f\left(\beta|\omega^2\right) + const \tag{9}$$

as

$$\ln f\left(\beta|y, \sigma^2, \omega^2\right) = -\frac{1}{2}\left(\beta - \hat{\beta}\right)^\top K_\beta \left(\beta - \hat{\beta}\right) + const \tag{10}$$

where $K_\beta = \frac{1}{\sigma^2} X^\top X + H^\top S^{-1} H$ and $\hat{\beta} = K_\beta^{-1} \left(\frac{1}{\sigma^2} X^\top\right)$. This means that $\beta|y, \sigma^2, \omega^2 \sim N\left(\hat{\beta}, K_\beta^{-1}\right)$. Then with the algorithm described soon we can draw $f\left(\beta|y, \sigma^2, \omega^2\right)$.

The algorithm obtains the multivariate normal vector generation using the precision matrix. The algorithm generates N independent draws from $N\left(\mu, \Lambda^{-1}\right)$ of dimensions n with the following steps:

1. We obtain the lower Cholesky factorization $\Lambda = DD^\top$.
2. We draw $Z_1, ..., Z_n \sim N(0, 1)$.
3. We determine Y from $Z = D^\top Y$.
4. We obtain $W = \mu + Y$.
5. We repeat steps 2-4 independently N times.

The next point is to be able to draw $f\left(\sigma^2, \omega^2|y, \beta\right)$. Given y and β, σ^2 and ω^2 are conditionally independent. From (8), we have $f\left(\sigma^2|y, \beta\right) \propto f\left(y|\beta, \sigma^2\right) f\left(\sigma^2\right)$ and $f\left(\omega^2|y, \beta\right) \propto f\left(\beta|\omega^2\right) f\left(\omega^2\right)$. Both conditional densities are inverse-gamma densities, which shows that

$$\sigma^2|y, \beta \sim IG\left(\alpha_{\sigma^2} + \frac{T}{2}, \lambda_{\sigma^2} + \frac{1}{2}(y - X\beta)(y - X\beta)^\top\right) \tag{11}$$

and

$$\omega_i^2|y, \beta \sim IG\left(\alpha_{\omega_i^2} + \frac{T-1}{2}, \lambda_{\omega_i^2} + \frac{1}{2}\sum_{t=2}^{T}(\beta_{it} - \beta_{it-1})^2\right) \tag{12}$$

Following Kroese and Chan (2014), we set small values for the shape parameter of the inverse gamma distribution so that the prior is more non-informative. That is, $\alpha_{\sigma^2} = \alpha_{\omega_i^2} = 5$, $i = 0, 1$. Also we set $\lambda_{\sigma^2} = (\alpha_{\sigma^2} - 1)$, $\lambda_{\omega_0^2} = 0.5^2\left(\alpha_{\omega_0^2} - 1\right)$, and $\lambda_{\omega_1^2} = 0.1^2\left(\alpha_{\omega_1^2} - 1\right)$ as the prior. Finally, we set the covariance matrix Ω_0 to be diagonal with diagonal elements equal to five, in line with Kroese and Chan (2014).

3.2 UC with stochastic volatilty

Stock and Watson (2007) include stochastic volatility in an unobserved component model. The authors show that UC-SV presents a great performance to forecast US inflation. The UC-SV is defined as

$$y_t = \beta_t + \sigma_t^{\frac{1}{2}} v_t \tag{13}$$

$$\beta_t = \beta_{t-1} + \omega_t^{\frac{1}{2}} e_t \tag{14}$$

where $ln\sigma_t$ is the log stochastic volatility, β_t is the trend, $ln\sigma_t = ln\sigma_{t-1} + e_{1t}$, and $ln\omega_t = ln\omega_{t-1} + e_{2t}$, in which the variances of e_{1t} and e_{2t} are respectively g_1 and g_2. We estimate the model using Markov chain Monte Carlo (MCMC) algorithm with Gibbs sampling method following Barnett et al (2014).

Our first step is to establish the priors and starting values. We define the prior for the initial value of the $ln\sigma_t$ as $ln\sigma_0 \sim N(\mu_0, 10)$ where μ_0 is the variance of $y_{t0} - \beta_{t0}$ and $t0$ refers to the training sample of 40 observations and β_{t0} is an initial estimate for the trend from the Hodrick–Prescott filter. In a similar way, $ln\omega_0 \sim N(\omega_0, 10)$ where $\omega_0 = \Delta\beta_{t0}$. We use the priors for g_1 and g_2 from an inverse gamma, we set the prior scale parameter equal to 0.01 and 0.0001 respectively with one degree of freedom as Barnett et al (2014).

So the next step is to simulate the posterior distributions. We draw σ_t and ω_t conditional on the value for g_1 and g_2 with the Metropolis algorithm based on Jacquier et al (2004). We draw β_t using the Carter and Kohn (2004) algorithm. We generate the sample for g_1 and g_2 from the inverse gamma distribution. We consider 10,000 replications of the MCMC algorithm and we keep with the last 1,000 replications for inference.

Below we detail how we calculate the marginal likelihood. We use a particle filter to calculate the log likelihood function for the UC-SV. We define Ξ as all parameters of the model. Based on Chib (1995), we consider the log marginal likelihood as:

$$lnP(y_t) = lnF\left(y_t|\hat{\Xi},\right) + lnP\left(\hat{\Xi}\right) - lnG\left(\hat{\Xi}|y_t\right) \tag{15}$$

where $lnP(y_t)$ is the log marginal likelihood that we want to calculate, $lnF\left(y_t|\hat{\Xi}\right)$ is the log likelihood function, $lnP\left(\hat{\Xi}\right)$ is the log prior density, and $lnG\left(\hat{\Xi}|y_t\right)$ is the log posterior density of the model parameters. The three elements on the right hand side of (15) are evaluated at the posterior mean for the model parameters $\hat{\Xi}$.

We calculate the log likelihood function for this model using a particle filter following Barnett et al (2014). We evaluate the prior density easily. But to get the term $lnG\left(\hat{\Xi}|y_t\right)$, we need an additional step. $lnG\left(\hat{\Xi}|y_t\right)$ can be factorized into conditional and marginal densities of various parameter blocks and we use Gibbs and Metropolis algorithm to approximate these densities according to Chib (1995) and Chib and Jeliazkov (2001). The posterior density is defined as $G\left(\hat{\Xi}|y_t\right) = G\left(\hat{g}_1, \hat{g}_2\right)$ and we drop the dependence on y_t to simplify the notation. The factorization of this density can be described by

$$G\left(\hat{g}_1, \hat{g}_2\right) = H\left(\hat{g}_1|\hat{g}_2\right) H\left(\hat{g}_2\right) \tag{16}$$

where $H\left(\hat{g}_1|\hat{g}_2\right) = \int H\left(\hat{g}_1|\hat{g}_2, \Theta\right) H\left(\Theta|\hat{g}_2\right) d\Theta$ and $H\left(\hat{g}_2\right) = \int H\left(\hat{g}_2|\Theta\right) H\left(\Theta\right) d\Theta$, in which $\Theta = \{\beta_t, \sigma_t, \omega_t\}$ is the state variables in the model. These two densities $H\left(\hat{g}_1|\hat{g}_2\right)$ and $H\left(\hat{g}_2\right)$ can be obtained as a 'weighted average' across state variables. We can approximate $H\left(\hat{g}_1|\hat{g}_2\right)$ and $H\left(\hat{g}_2\right)$ with additional Gibbs runs and we can integrate over the states. We consider 10,000 iterations in these additional Gibbs samplers and we remain with the last 3,000 like Barnett et al (2014).

3.3 Lasso-type penalties

We present three lasso-type penalties in this subsection: LASSO, adaLASSO, and WLadaLASSO.

3.3.1 Lasso

Tibshirani (1996) proposes the LASSO method that is based on the following minimization problem

$$\hat{\beta}^{LASSO} = argmin_{\beta_0,\beta_1,\ldots,\beta_k} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ji} \right)^2 + \lambda \sum_{j=1}^{k} |\beta_j| \tag{17}$$

where $\lambda \geq 0$ is a tunning parameter and lasso requires a method to obtain a value for λ, that we explain soon. The first term is the sum of square of residuals and the second term is a shrinkage penalty. $\sum_{j=1}^{k} |\beta_j|$ is the ℓ^1 norm of a coefficient vector β. The ℓ^1 penalty forces some of the coefficients estimates to be equal to zero when λ is sufficiently large. When $\lambda = 0$, LASSO estimates are equal to ordinary least squares estimates. So, lasso technique performs variable selection.

Cross-validation is usually the method to obtain the λ value. With time series data, we use the Bayesian information criterion (BIC) to choose λ, following Konzen and Ziegelmman (2016). We consider a grid of λ values.

3.3.2 Adalasso

Zou (2006) states that with LASSO we can obtain an inconsistent selection of variables that keep noisy variables for example for a given λ that leads to optimal estimation rate. Also the author shows that LASSO can lead to the right selection of variables with biased estimates for large coefficients and this take to suboptimal prediction rates.

So, Zou (2006) introduces the adaptive LASSO, which considers weights ω_j that adjust the penalty to be different for each coefficient. The adaptive LASSO seeks to minimize

$$\hat{\beta}^{adaLASSO} = argmin_{\beta_0,\beta_1,\ldots,\beta_k} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ji} \right)^2 + \lambda \sum_{j=1}^{k} \omega_j |\beta_j| \tag{18}$$

where $\omega_j = \mid \hat{\beta}_j^{ridge} \mid^{-\tau}$, $\tau > 0$. The adaptive LASSO considers that large (small) coefficients have small (large) weights - and small (large) penalties. The coefficients estimated by ridge regression lead to get the weight ω_j. Ridge regression shrinks the vector of coefficients by penalizing the sum of the squares of the residuals:

$$\hat{\beta}^{ridge} = argmin_{\beta_0,\beta_1,\ldots,\beta_k} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ji} \right)^2 + \lambda \sum_{j=1}^{k} \beta_j^2 \tag{19}$$

where the penalty is the ℓ^2 norm of the β vector. Ridge regression is not a method of variable selection because this regression obtains non-zero estimates for all coefficients.

3.3.3 WLadalasso

In case we consider adalasso with time series data, each coefficient associated with a lagged variable is penalized according to the size of the Ridge's estimate. As less distant lag variables should usually raise the time series forecast, the coefficients of more lagged variables should be penalized according to the lag of the variable.

Park and Sakaori (2013) propose some types of penalties for different lags. Konzen and Ziegelmann (2016) present the adalasso with weighted lags based on Park and Sakaori (2013). The wladalasso is given by

$$\hat{\beta}^{wladaLASSO} = argmin_{\beta_0,\beta_1,...,\beta_k} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ji} \right)^2 + \lambda \sum_{j=1}^{k} \omega_j^w |\beta_j| \tag{20}$$

where $\omega_j^w = \left(\mid \hat{\beta}_j^{ridge} \mid e^{-\alpha l} \right)^{-\tau}$ is the weight, $\tau > 0$, $\alpha \geq$ and l is the order of the variable's lag. We set the parameter τ equal to one as Konzen and Ziegelmann (2016) for the adalasso and wladalasso. We consider a grid for α, where the set of possible values for α is $[0, 0.5, 1, ..., 10]$. We calculate the optimal λ among those possible for that model with the lowest BIC value for each value of α. We choose the α value from that model that produces the smallest BIC value among all possible α values, following Konzen and Ziegelmann (2016).

3.4 Exponential smoothing

The name exponential smoothing comes from the weights decreasing exponentially when the observation becomes older. We can represent the exponential smoothing methods as state space models (Ord et al, 1997, Hyndman et al, 2002, Hyndman et al, 2005). The exponential smoothing method is an algorithm which only produces point forecasts. The stochastic state space model associated to this method provides a framework which leads to the same point forecast but also estimates the prediction intervals for example (Hyndman et al, 2008).

Economical series exhibit some features that can be worked on. We can break down the economic series into certain components, such as trend (T), cycle (C), seasonality (S), and irregular or error (E). The exponential smoothing method that we use consider the decomposition into these components. Only the cycle component is not decomposed separately so that we model along with the trend component, following Hyndman et al (2005), Hyndman and Khandakar (2008), and Hyndman et al (2008). Thus, we can combine the components of trend, seasonality and error additionally or multiplicatively for example.

Hyndman et al (2002), Hyndman and Khandakar (2008) and Hyndman et al (2008) propose 15 different combinations between trend and seasonality components. The trend component can present five different possibilities: none, additive, additive damped, multiplicative, and multiplicative damped. The trend component is a combination between the level (ℓ) and growth (b) parameter. Consider T_h the forecast trend over the next h periods, and ϕ is a damping parameter ($0 < \phi < 1$). So if there is no trend component, $T_h = \ell$. If the trend component is additive, $T_h = \ell + bh$. If the trend component is additive damped, $T_h = \ell + \left(\phi + \phi^2 + \cdots + \phi^h\right) b$. If the trend component is multiplicative, $T_h = \ell b^h$. If the trend component is multiplicative damped, $T_h = \ell b^{\left(\phi+\phi^2+\cdots+\phi^h\right)}$. If growth rate at the end of the series is unlikely to continue, the damped trend seems to be a reasonable option.

After we present the types of trend component, the next step is to detail the types of seasonal component. The seasonal component can be none, additive or multiplicative. Also the error component can be additive or multiplicative, but this distinction is not relevant to make point forecast (Hyndman et al, 2008).

So, we consider the combination of five types of trend component and three types of seasonal component which leads to a total of 15 types of exponential smoothing methods that we consider in this paper, following Hyndman and Khandakar (2008). We present these 15 possibilities in the Table 1, in which the first entry refers to the trend component and the second to the seasonal component. Some of these exponential smoothing methods are known by other names. For example, cell N, N represents the simple exponential smoothing method, cell A, N refers to the Holt's linear method, and cell A_d, N is associated with the damped trend method. The cell A, A describes the additive Holt-Winter's method, the cell A, M refers to the multiplicative Holt-Winter's method.

For example, consider the Holt's linear method (cell A, N) that can be described as

$$\ell_t = \alpha y_t + (1 - \alpha)\left(\ell_{t-1} + b_{t-1}\right) \tag{21}$$

$$b_t = \beta^* \left(\ell_t - \ell_{t-1}\right) + \left(1 - \beta^*\right) b_{t-1} \tag{22}$$

Trend component	Seasonal component		
	N (None)	A (Additive)	M (Multiplicative)
N (None)	N,N	N,A	N,M
A (Additive)	A,N	A,A	A,M
A_d (Additive damped)	A_d,N	A_d,A	A_d,M
M (Multiplicative)	M,N	M,A	M,M
M_d (Multiplicative damped)	M_d,N	M_d,A	M_d,M

Table 1: The different combinations of exponential smoothing methods

$$\hat{y}_{t+h|t} = \ell_t + b_t h \tag{23}$$

where the first equation (21) shows the model to the level of the series at time t ℓ_t, the second equation (22) describes the growth rate (estimate the slope) of the series at time t b_t. b_t is given by a weighted average of the estimate of growth obtained by the difference between successive levels and the previous growth b_{t-1}. Finally, the equation (23) presents the prediction for the variable y h periods ahead using information available at time t. This equation describes that the forecast for the variable h periods ahead is given by the level in the current time ℓ_t adding the growth b_t for the h periods. α is the smoothing parameter for the level with $0 < \alpha < 1$, and β^* is the smoothing parameter for the trend with $0 < \beta^* < 1$.

Consider a model with the seasonality component, when seasonal variations are constant throughout the series, the additive method for the seasonal component is preferred. When seasonal variations change proportionally to the level of the series, the multiplicative method for the seasonal component is preferred. Consider for example the Holt-Winters method of additive trend with additive seasonal component (cell A, A), the equations of this method are given by

$$\ell_t = \alpha\left(y_t - s_{t-m}\right) + (1-\alpha)\left(\ell_{t-1} + b_{t-1}\right) \tag{24}$$

$$b_t = \beta^*\left(\ell_t - \ell_{t-1}\right) + (1-\beta^*)\, b_{t-1} \tag{25}$$

$$s_t = \gamma\left(y_t - \ell_{t-1} - b_{t-1}\right) + (1-\gamma)\, s_{t-m} \tag{26}$$

$$\hat{y}_{t+h|t} = \ell_t + b_t h + s_{t-m+h_m^+} \tag{27}$$

The equations (24), (25), (26), and (27) describe, respectively, the level ℓ_t, the growth rate b_t, the seasonality s_t, and the forecast h periods ahead of the series. m is the length of seasonality (e.g., number of months or quarters in a year), is the seasonal component, and $h_m^+ = [(h-1)\ mod\, m] + 1$. The parameters of the Holt-Winters method $(\alpha, \beta^*, \gamma)$ are restricted to lie between 0 and 1.

Table 2 presents the equations for the level, growth, seasonality, and forecast of the series for h periods ahead for the 15 cases considered. Some values for exponential smoothing parameters lead to interesting specific cases. Some examples are: the level remains constant over time if $\alpha = 0$, the slope is constant over time if $\beta^* = 0$, and the seasonal behavior is the same over time if $\gamma = 0$. Finally, the methods A and M for the trend component are particular cases of A_d and M_d with $\phi = 1$.

The next point is to discuss the state space models that underlie exponential smoothing methods. Each of the 15 models considered consist of a measurement equation that describes the data, and state equations that represent how the unobserved components (level, trend, seasonal) modify over time. The measurement equation together with the state equations are known by state space models.

Hyndman et al (2002) propose to differentiate the behavior of the model with additive errors in relation to the multiplicative errors. If the estimated parameters are the same, the point forecast is not affected if the error is multiplicative or additive, only the prediction interval. When considering the behavior of the error, Hyndman et al (2002) have the triplet (E, T, S) that refers to the three components: error, trend, and seasonality. The model $ETS(A, A, N)$ means that the errors and the trend are additive and that there is no seasonality.

Trend	Seasonal		
	N	A	M
N	$\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}$	$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)\ell_{t-1}$	$\ell_t = \alpha(y_t/s_{t-m}) + (1-\alpha)\ell_{t-1}$
		$s_t = \gamma(y_t - \ell_{t-1}) + (1-\gamma)s_{t-m}$	$s_t = \gamma(y_t/\ell_{t-1}) + (1-\gamma)s_{t-m}$
	$\hat{y}_{t+h\|t} = \ell_t$	$\hat{y}_{t+h\|t} = \ell_t + s_{t-m+h_m^+}$	$\hat{y}_{t+h\|t} = \ell_t s_{t-m+h_m^+}$
A	$\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1})$	$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$	$\ell_t = \alpha(y_t/s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$
	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$
		$s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}$	$s_t = \gamma(y_t/(\ell_{t-1} - b_{t-1})) + (1-\gamma)s_{t-m}$
	$\hat{y}_{t+h\|t} = \ell_t + hb_t$	$\hat{y}_{t+h\|t} = \ell_t + hb_t + s_{t-m+h_m^+}$	$\hat{y}_{t+h\|t} = (\ell_t + hb_t)s_{t-m+h_m^+}$
A_d	$\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$	$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$	$\ell_t = \alpha(y_t/s_{t-m}) + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$
	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$	$b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$
		$s_t = \gamma(y_t - \ell_{t-1} - \phi b_{t-1}) + (1-\gamma)s_{t-m}$	$s_t = \gamma(y_t/(\ell_{t-1} - \phi b_{t-1})) + (1-\gamma)s_{t-m}$
	$\hat{y}_{t+h\|t} = \ell_t + \phi_h b_t$	$\hat{y}_{t+h\|t} = \ell_t + \phi_h b_t + s_{t-m+h_m^+}$	$\hat{y}_{t+h\|t} = (\ell_t + \phi_h b_t)s_{t-m+h_m^+}$
M	$\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}b_{t-1}$	$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)\ell_{t-1}b_{t-1}$	$\ell_t = \alpha(y_t/s_{t-m}) + (1-\alpha)\ell_{t-1}b_{t-1}$
	$b_t = \beta^*(\ell_t/\ell_{t-1}) + (1-\beta^*)b_{t-1}$	$b_t = \beta^*(\ell_t/\ell_{t-1}) + (1-\beta^*)b_{t-1}$	$b_t = \beta^*(\ell_t/\ell_{t-1}) + (1-\beta^*)b_{t-1}$
		$s_t = \gamma(y_t - \ell_{t-1}b_{t-1}) + (1-\gamma)s_{t-m}$	$s_t = \gamma(y_t/(\ell_{t-1}b_{t-1})) + (1-\gamma)s_{t-m}$
	$\hat{y}_{t+h\|t} = \ell_t b_t^h$	$\hat{y}_{t+h\|t} = \ell_t b_t^h + s_{t-m+h_m^+}$	$\hat{y}_{t+h\|t} = \ell_t b_t^h s_{t-m+h_m^+}$
M_d	$\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}b_{t-1}^{\phi}$	$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)\ell_{t-1}b_{t-1}^{\phi}$	$\ell_t = \alpha(y_t/s_{t-m}) + (1-\alpha)\ell_{t-1}b_{t-1}^{\phi}$
	$b_t = \beta^*(\ell_t/\ell_{t-1}) + (1-\beta^*)b_{t-1}^{\phi}$	$b_t = \beta^*(\ell_t/\ell_{t-1}) + (1-\beta^*)b_{t-1}^{\phi}$	$b_t = \beta^*(\ell_t/\ell_{t-1}) + (1-\beta^*)b_{t-1}^{\phi}$
		$s_t = \gamma(y_t - \ell_{t-1}b_{t-1}^{\phi}) + (1-\gamma)s_{t-m}$	$s_t = \gamma(y_t/(\ell_{t-1}b_{t-1}^{\phi})) + (1-\gamma)s_{t-m}$
	$\hat{y}_{t+h\|t} = \ell_t b_t^{\phi_h}$	$\hat{y}_{t+h\|t} = \ell_t b_t^{\phi_h} + s_{t-m+h_m^+}$	$\hat{y}_{t+h\|t} = \ell_t b_t^{\phi_h} s_{t-m+h_m^+}$

Table 2: The equations for the level, growth, seasonality, and forecast of the series for h periods ahead for the 15 cases considered

If we return to the Holt's linear method presented by the equations (21), (22), and (23) with additive errors for example, we have $ETS(A, A, N)$. Assuming that error $\varepsilon_t = y_t - \ell_{t-1} - b_{t-1} \sim N\left(0, \sigma^2\right)$ and independently distributed, we can write the error correction equations as

$$y_t = \ell_{t-1} + b_{t-1} + \varepsilon_t \tag{28}$$

$$\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t \tag{29}$$

$$b_t = b_{t-1} + \beta\varepsilon_t \tag{30}$$

where $\beta = \alpha\beta^*$. Equation (28) refers to the measurement equation and equations (29) and (30) describe the state equations for the level and growth respectively. Similarly, all 15 exponential smoothing methods presented in the table 2 can be rewritten in the form of state space model with additive or multiplicative errors, see Hyndman et al (2008) for example.

Basically, the estimation procedure is based on estimating the smoothing parameters α, β, γ, ϕ and the initial state variables ℓ_0, b_0, s_0, s_{-1}, ..., s_{-m+1} maximizing the likelihood function. The algorithm proposed by Hyndman et al. (2002) also determines which of the 15 ETS models is most appropriate by selecting the model based on the information criterion. The information criterion used to select the most appropriate model are Akaike information criterion (AIC), AIC corrected for small sample bias (AICc), and Bayesian information criterion (BIC), according to Hyndman et al (2008).

4 Data and Empirical Strategy

Our data is the Brazilian industrial production index at general level and your desagreggation by sectors. The data source is the Monthly Industrial Survey of Physical Production (PIM-PF) of the Brazilian Institute of Geography and Statistics (IBGE). We use monthly data from January 2002 to August 2017 without seasonal adjustment. We consider the first difference of data to have stationary series, with exception of the UC-SV and ETS models.

The forecast is made for the general industry to analyze the use of aggregate data and from the industry sectors we aggregate to get the general industry forecast. Thus we use the disaggregated data for the extractive industry and the 25 sectors of the manufacturing industry. However, two sectors printing and reproduction of recordings; and maintenance, repair and installation of machines and equipment only present data from January 2012 and therefore we do not include these sectors in the estimates.[1] Thus, our disaggregated sample includes the category

[1] The two sectors have combined share of 2.3% of the index of industrial production.

of extractive industries and 23 sectors of the manufacturing industry. Next we detail the empirical strategy and how we compare the predictions obtained.

4.1 Empirical Strategy and Forecast Comparison

We consider the model with up to 15 lags of the dependent variable and the selection of variables is based on this set of covariates. Thus, we use a rolling window of 100 fixed observations for each estimation for a forecast horizon of 1 to 12 months ahead. With this window size, we estimate with 61 samples for the general industry series as for each of the sectors. We estimate the forecast for each sector and we use the of each sector to obtain the forecast for the general industry from its components, thus obtaining the forecasts of the general industry from the disaggregated data. For such recomposition, we consider the forecast of each sector by its weight and then we sum to obtain the forecast from the disaggregated data. The forecast window for analysis ranges from September 2011 to August 2017.

We compare our forecasts with the estimates of three naive models. The first is the autoregressive model of the first order, the second is the time-varying parameters autoregressive model of the first order, and the third is the UC with SV. We estimate all models with the data for the general industry (aggregate forecast) and the forecast for the general industry from the disaggregated data (disaggregated forecast). Thus, we have two naive predictions from the AR (1) model (aggregate and disaggregated), two naive predictions from the time varying AR(1) (aggregate and disaggregated), and two naive predictions from UC with SV (aggregate and disaggregated). The naive model serves as reference or benchmark for the other forecasts. The performance analysis for the prediction of the models will be through the mean square error (MSE) and from the test of Diebold and Mariano (1995) to determine if there is a model with more accurate prediction for the general industrial production of Brazil for the period considered. Next we present the test of Diebold and Mariano (1995).

4.1.1 Diebold and Mariano test

The MSE is the sum of the difference squared between the actual value of the data and the estimated value, weighted by the number of terms. The test of Diebold and Mariano (1995) shows if there is any model with statistically more accurate prediction for the general industrial production of Brazil for the considered period. Consider a loss function g based on forecasting errors, in which we consider a quadratic loss function. Assuming two models (1 and 2), in which the forecasting errors $\hat{\varepsilon}_{t+h|t}$ for h period ahead are given by $\hat{\varepsilon}^1_{t+h|t} = y_{t+h} - \hat{y}^1_{t+h|t}$ and $\hat{\varepsilon}^2_{t+h|t} = y_{t+h} - \hat{y}^2_{t+h|t}$, where $\hat{y}^2_{t+h|t}$ is the forecast from model 2 for h periods ahead for example. The test of Diebold and Mariano (1995) is based on the difference between the forecasting error of the models.

The null hypothesis of the test is that there is equality of forecasting performance between the two models, that is, the models have statistically equal deviations. On the other hand, the alternative hypothesis (one-sided test) defines that the model used as a reference leads to more accurate forecasts than the other. The Diebold and Mariano (1995) test is given by:

$$S = \frac{\bar{d}}{\sqrt{\left(\widehat{Avar}\left(\bar{d}\right)\right)}} \tag{31}$$

where $\bar{d} = \frac{1}{T_0}\sum_{t=t_0}^{T} d_t$, $d_t = g\left(\hat{\varepsilon}^1_{t+h|t}\right) - g\left(\hat{\varepsilon}^2_{t+h|t}\right)$, and $\widehat{Avar}\left(\bar{d}\right)$ is an estimate of the asymptotic variance (large samples) of $\bar{d}$ for the sample selected. Thus, the S statistic follows a Student t-distribution (Diebold, Mariano, 1995). We consider the quadratic loss function for the test performed in the present work.

5 Results

This section presents the prediction of 1 to 12 months ahead for the naive models (AR(1), AR(1) with time-varying parameters, UC with SV), besides the lag selection from the LASSO and its two variants (adaLASSO and WLadaLASSO), and exponential smoothing ETS to predict the general industry from aggregated and disaggregated data. We compare the results in two parts, from the MSE and the Diebold and Mariano (1995) test to obtain the model with the most accurate forecast.

Table 3 presents the MSE of 1 to 12 months ahead for the different models. In general, we can analyze that the MSE is smaller for the disaggregated models in relation to the aggregates independent of the forecast horizon,

with the exception of AR (1) with time-varying parameters and UC with SV. The model with the lowest MSE for the prediction of one to seven months ahead is disaggregated ETS. Increasing the forecast time horizon mainly from one to three periods ahead does not greatly affect the MSE of the disaggregated ETS. The WLadaLASSO model for disaggregated data presented lower MSE for the time horizon of 8 to 12 months ahead. The difference in MSE between the LASSO models and variants is small mainly using disaggregated data, in which the disaggregated WLadaLASSO performs better than the other LASSO methods disaggregated for the different forecast horizons. In line with Konzen and Ziegelmann (2016), the WLadaLASSO disaggregated model performed better than the LASSO and adaLASSO for disaggregated data. However, in the case of estimating the aggregate model, the LASSO model presents better prediction performance in relation to its variants. An interesting point would be to try to find out why WLadaLASSO does not perform better than its variants for aggregate data or because there is this performance difference between aggregated and disaggregated data. Since Konzen and Ziegelmann (2016) point out that WLadaLASSO would perform better when the sample is small and the greater the number of lags used but we are not varying these two points.

	AR(1)		TVP AR(1)		UC-SV		ETS		LASSO		adaLASSO		WLadaLASSO	
	Ag	Disag	Ag	Disag	Ag	Disag	Ag	Disag	Ag	Disag	Ag	Disag	Ag	Disag
1	7,96	7,68	9,87	10,30	9,18	9,38	1,61	**1,42**	2,72	2,25	2,93	2,21	2,80	2,20
2	7,93	7,77	9,98	10,46	8,94	9,10	1,71	**1,48**	2,74	2,36	2,95	2,36	2,94	2,34
3	7,98	7,86	10,75	11,11	9,18	9,31	1,77	**1,51**	2,79	2,41	2,96	2,42	3,01	2,41
4	8,10	8,02	11,29	11,53	9,46	9,66	2,07	**1,73**	2,82	2,43	3,02	2,44	3,09	2,42
5	8,12	8,05	11,34	11,56	9,49	9,64	2,23	**1,89**	2,84	2,47	3,06	2,47	3,17	2,44
6	8,10	8,05	11,43	11,64	9,54	9,70	2,38	**2,06**	2,85	2,48	3,09	2,49	3,22	2,46
7	8,15	8,11	11,41	11,68	9,51	9,72	2,67	**2,31**	2,87	2,50	3,10	2,50	3,26	2,48
8	8,16	8,13	11,37	11,64	9,45	9,67	2,93	2,54	2,90	2,53	3,13	2,53	3,32	**2,51**
9	8,19	8,17	11,14	11,41	9,30	9,56	3,20	2,78	2,95	2,58	3,19	2,58	3,39	**2,55**
10	8,19	8,17	10,75	11,01	9,06	9,28	3,51	3,08	2,99	2,62	3,21	2,62	3,43	**2,59**
11	8,19	8,17	10,34	10,56	8,88	9,12	3,83	3,36	3,02	2,66	3,24	2,66	3,46	**2,63**
12	8,19	8,17	9,94	10,13	8,72	8,86	4,15	3,68	3,05	2,69	3,26	2,69	3,48	**2,67**

Table 3: MSE results for forecasting from 1 to 12 months ahead for different models

However, the MSE difference does not allow to state if statistically the disaggregated ETS is higher in the accuracy of the forecast in relation to the others. Therefore, we analyze the results of the Diebold and Mariano (1995) test. In order to perform the test, we establish the disaggregated ETS as the benchmark model because it has lower MSE of 1 to 7 months ahead compared to the rest of the models, indicating predictive power gains. We compare the predictive errors of all models with those of the benchmark, having as null hypothesis the equality of predictive power. Under the null hypothesis, the disaggregated ETS predicts as well as the analyzed model. The alternative hypothesis indicates that the prediction of the disaggregated ETS is statistically better.

Table 4 presents the Diebold and Mariano (1995) statistic and the associated p-value for each of the models in relation to the benchmark. From the results of the test of Diebold and Mariano (1995), we can consider that the disaggregated ETS presents better statistical performance for prediction in relation to the naive models (AR(1), AR(1) with time-varying parameters, UC with SV) considered with aggregate or disaggregated data and the aggregate ETS. However, the disaggregated ETS model does not present better accuracy in relation to the LASSO models and their variants, regardless of being aggregated or disaggregated.

Conclusions

The present work seeks to analyze two points about the prediction of industrial production for Brazil. The first is to compare different univariate models for selection of lags like LASSO and two variants, in addition to the most suitable model for exponential smoothing. Among these models, which type of model best forecasts

	AR(1)		TVP AR(1)		UC-SV		ETS	LASSO		adaLASSO		WLadaLASSO	
	Ag	Disag	Ag	Disag	Ag	Disag	Ag	Ag	Disag	Ag	Disag	Ag	Disag
1	-10,38	-10,37	-11,84	-12,00	-10,82	-10,82	-5,99	2,40	3,96	1,56	4,00	0,72	4,11
	0,00	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	0,99	1,00	0,94	1,00	0,76	1,00
2	-9,51	-9,48	-11,05	-11,18	-10,00	-10,00	-4,23	1,66	2,68	1,08	2,72	0,49	2,78
	0,00	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	0,95	1,00	0,86	1,00	0,69	1,00
3	-9,12	-9,06	-10,17	-10,30	-9,29	-9,29	-3,53	1,40	2,25	0,91	2,29	0,41	2,34
	0,00	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	0,92	0,99	0,82	0,99	0,66	0,99
4	-8,27	-8,21	-9,67	-9,77	-8,68	-8,68	-3,06	1,27	2,05	0,82	2,09	0,37	2,13
	0,00	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	0,90	0,98	0,79	0,98	0,65	0,98
5	-8,03	-7,99	-9,58	-9,70	-8,52	-8,52	-2,87	1,21	1,94	0,78	1,98	0,35	2,02
	0,00	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	0,89	0,97	0,78	0,98	0,64	0,98
6	-8,20	-8,13	-9,79	-9,82	-8,67	-8,67	-2,71	1,17	1,88	0,76	1,92	0,34	1,95
	0,00	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	0,88	0,97	0,78	0,97	0,63	0,97
7	-8,53	-8,42	-10,50	-10,46	-9,19	-9,19	-2,58	1,15	1,83	0,74	1,87	0,33	1,90
	0,00	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,01**	0,87	0,97	0,77	0,97	0,63	0,97
8	-8,34	-8,25	-10,50	-10,43	-9,01	-9,01	-2,49	1,11	1,76	0,71	1,81	0,32	1,84
	0,00	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,01**	0,87	0,96	0,76	0,96	0,63	0,97
9	-7,87	-7,79	-10,57	-10,47	-8,81	-8,81	-2,38	1,05	1,69	0,68	1,73	0,31	1,76
	0,00	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,01**	0,85	0,95	0,75	0,96	0,62	0,96
10	-7,74	-7,65	-10,48	-10,43	-8,69	-8,69	-2,29	1,00	1,59	0,64	1,63	0,29	1,66
	0,00	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,01**	0,84	0,94	0,74	0,95	0,61	0,95
11	-7,57	-7,47	-10,45	-10,33	-8,44	-8,44	-2,21	0,93	1,49	0,60	1,52	0,27	1,55
	0,00	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,01**	0,82	0,93	0,73	0,94	0,61	0,94
12	-5,45	-5,41	-6,96	-7,02	-6,00	-6,00	-2,09	0,83	1,34	0,54	1,37	0,24	1,40
	0,00	**0,00**	**0,00**	**0,00**	**0,00**	**0,00**	**0,02**	0,80	0,91	0,70	0,91	0,60	0,92

Table 4: Diebold and Mariano (1995) test resultsfrom 1 to 12 months ahead for different models

the industrial production in Brazil. The second point is to consider whether the disaggregated data contribute to predict the aggregate level of industrial production. Basically our result points to a better performance of models that use disaggregated data. The exponential smoothing model with disaggregated data in which we obtain the best specification performs better in the forecast from 1 to 7 months ahead. The WLadaLASSO model with disaggregated data offers better forecasting performance from 8 to 12 months ahead. However, the difference in prediction performance between the LASSO and its variants (adaLASSO, WLadaLASSO) is small when we consider the disaggregated data.

This is an ongoing research. The next steps are to contemplate two additional models for prediction, which is the selection between different models of neural networks of Crone and Kourentzes (2010) and Kourentzes et al (2014), and the dynamic model averaging/selection framework of Koop and Korobilis (2012) and Raftery et al (2010). Also we will compare the forecasts with model confidence set of Hansen et al (2011).

References

Barhoumi, K., Darné, O., and Ferrara, L. (2010), "Are disaggregate data useful for factor analysis in forecasting French GDP", Journal of Forecasting, 29, 132-144.

Barnett, A., Mumtaz, H., and Theodoridis, K. (2014), "Forecasting UK GDP growth and inflation under structural change. A comparison of models with time-varying parameters", *International Journal of Forecasting*, 30, 129-143.

Carlos, T., and Marçal, E. (2016), "Forecasting Brazilian inflation by its aggregate and disaggregated data: a test of predictive power by forecast horizon", *Applied Economics*, 48, 4846-4860.

Carter, C., and Kohn, P. (2004), "On Gibbs sampling for state space models", *Biometrika*, 81, 541–53.

Chib, S. (1995), "Marginal likelihood from the Gibbs output", *Journal of the American Statistical Association*, 90, 1313–21.

Chib, S., and Jeliazkov, I. (2001), "Marginal likelihood from the Metropolis-Hastings output", *Journal of the American Statistical Association*, 96, 270–81.

Crone, S., and Kourentzes, N. (2010), "Feature selection for time series prediction – a combined filter and wrapper approach for neural networks", *Neurocmputing*, 73, 1923-1936.

Duarte, C., and Rua, A. (2007), "Forecasting through a Bottom-up approach: how bottom is bottom?", *Economic Modelling*, 24, 941-953.

Espasa, A., Senra, E., and Albacete, R. (2002), "Forecasting inflation in the European Monetary Union: a disaggregated approach by countries and by sectors", *European Journal of Finance*, 8, 402-421.

Giacomini, R., and Granger, C. W. J. (2004), "Aggregation of space-time processes", *Journal of Econometrics*, 118, 7-26.

Granger, C. W. J. (1987), "Implications of aggregation with common factors", *Econometric Theory*, 3, 208-222.

Hansen, P., Lunde, A., and Nason, J. (2011), "The Model Confidence Set", Econometrica, 79(2), 453-497.

Hendry, D. F., and Hubrich, K. (2011), "Combining disaggregate forecasts or combining disaggregate information to forecast an aggregate", *Journal of Business and Economic Statistics*, 29, 216-227.

Hyndman, R., Koehler, A., Snyder, R., and Grose, S. (2002), "A State Space Framework for Automatic Forecasting Using Exponential Smoothing Methods", *International Journal of Forecasting*, 18, 439-454.

Hyndman, R., Koehler, A., Ord, J., and Snyder, R. (2005), "Prediction Intervals for Exponential Smoothing Using Two New Classes of State Space Models", *Journal of Forecasting*, 24, 17-37.

Hyndman, R., and Khandakar, Y. (2008), "Automatic Time Series Forecasting: the forecast package for R", *Journal of Statistical Software*, 27, 1-22.

Hyndman, R., Koehler, A., Ord, J., and Snyder, R. (2008), *Forecasting with exponential smoothing*, Springer, Berlin.

Hubrich, K. (2005), "Forecasting euro area inflation: Does aggregating forecasts by HICP component improve forecast accuracy?", *International Journal of Forecasting*, 21, 119-136.

Jacquier, E, Polson, N, and Rossi, P (2004), "Bayesian analysis of stochastic volatility models", *Journal of Business and Economic Statistics*, 12, 371–418.

Kapetanios, G., Marcellino, M., and Venditti, F. (2017), "Large time-varying parameter VARs: a non-parametric approach", Bank of Italy working paper no. 1122.

Koop, G., and Korobilis, D. (2012), "Forecasting Inflation Using Dynamic Model Averaging", *International Economic Review*, 53(3), 867–886.

Kourentzes, N., Barrow, D., and Crone, S. (2014), "Neural network ensembles operators for time series forecasting", *Expert Systems with Applications*, 41, 4235-4244.

Kroese, D., and Chan, J. (2014), *Statistical Modeling and Computation*, Springer, Berlin.

Lütkepohl, H. (1984), "Linear transformation of vector ARMA processes", *Journal of Econometrics*, 26, 283-293.

Lütkepohl, H. (1987), *Forecasting Aggregated Vector ARMA Processes*, Springer-Verlag, Berlin.

Marcellino, M., Stock, J. H., and Watson, M. W. (2003), "Macroeconomic forecasting in the Euro area: Country specific versus area-wide information", *European Economic Review*, 47, 1-18.

Ord, J., Koehler, A., and Snyder, R. (1997), "Estimation and Prediction for a Class of Dynamic Nonlinear Statistical Models", *Journal of the American Statistical Association*, 92, 1621-1629.

Raftery, A., Karny, M., and Ettler, P. (2010), "Online Prediction Under Model Uncertainty via Dynamic Model Averaging: Application to a Cold Rolling Mill", *Technometrics*, 52(1), 52–66.

Stock, J., and Watson, M. (2007), "Why has U.S. inflation become harder to forecast?", *Journal of Money, Credit and Banking*, 39, 3-33.

Tibshirani R. (1996), "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Weber, E. and Zika, G. (2016), "Labour market forecasting in Germany: is disaggregation useful?", *Applied Economics*, 48, 2183-2198.