PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

DEPARTAMENTO DE CIÊNCIAS ECONÔMICAS

MONOGRAFIA DE FINAL DE CURSO

INDUSTRIAL PRODUCTION FORECASTING WITH HIGH DIMENSIONAL MODELS COMPARING AGGREGATED VS DISAGGREGATED APPROACHES: EVIDENCE FROM BRAZIL

VICTOR ENES COTA

1710582

RIO DE JANEIRO – RJ

2022

DEPARTAMENTO DE CIÊNCIAS ECONÔMICAS

MONOGRAFIA DE FINAL DE CURSO

INDUSTRIAL PRODUCTION FORECASTING WITH HIGH DIMENSIONAL MODELS COMPARING AGGREGATED VS DISAGGREGATED APPROACHES: EVIDENCE FROM BRAZIL

Monografia de conclusão de curso apresentado ao Departamento de Economia da PUC-Rio como requisito para o recebimento do Bacharel em Ciências Econômicas.

Orientador: Francisco Luna

Coorientador: Gilberto Boaretto

VICTOR ENES COTA

RIO DE JANEIRO – RJ

2022

ACKNOWLEDGMENTS

First, I would like to thank my mother, Rosa Claudia, and my father, Manuel Júlio, for all the support that was given to me during the entirety of this academic course. The example they gave with all the hard work throughout their careers was a meaningful motivation and seeing them proud is an accomplishment in itself for me. I would also like to thank my brothers, Raphael and Eduardo, who made themselves present and helped me in difficult lessons, even though they're living in distant places now. And my sister, Larissa, who shared much love and support through all these years.

I would like to thank my friends that made this journey much more pleasing and joyful than it could've been without them. Their companionship through difficult times was essential to maintain good mental health and reduce stress.

Also my advisors, Francisco and Gilberto, who I have a great admiration for sharing with me their knowledge and guidance throughout the process of making this study. Their help significantly inspired me to keep pursuing information in this field.

And lastly, I would like to thank my teachers that guided me through this whole academic course and shared experiences that impacted my perceptions in this field of study.

Table of Contents

List of tables and figures

1. INTRODUCTION

It is well known that forecasts and expectations poses as relevant guidelines in the decision making process of policy makers. Central Banks and other financial institutions around the world dedicates a significant amount of resources to assure these forecasts are accurate. In this regard, the motivation for this study is to provide a better understanding of how high dimensional models, implemented in different approaches, can help improve these forecasts and thus the guidelines of policy makers and financial institutions.

There are plenty of macroeconomic variables that helps diagnose a country's economic situation. Some examples are employment, exchange rates, interest rates, inflation, fiscal situation and economic activity indicators such as the GDP. In this study, we're going to focus on an industrial production index that serves as a proxy for the industrial component of a country's GDP. In 2021, this component was responsible for almost 20% of Brazil's GDP, revealing the importance of having accurate forecasts in this spectrum.

Moreover, there are some reasons why studying an industrial production index might be as interesting as studying the GDP itself of a country. The main one is that the index is disclosed with a higher frequency than the GDP (monthly vs quarterly frequency), meaning that it's useful for having a closer grasp of a country's economic reality. Besides that, the index is released with a lag of two months, while the GDP has at least a three months lag.

In this research, we're addressing if recent Machine Learning models such as the LASSO, the Random Forest and factor models brings any improvements to these forecasts. Furthermore, we are examining whether using a disaggregated approach can increase these models performance in any way. By disaggregated, we mean that the main index can be decomposed into two categories. More specifically, the general industrial production index can be decomposed into the manufacturing industries index and the mining industries index. And this approach implies in computing forecasts for these sub-indexes and then aggregating them by their respective weights in the general industrial production.

1.1. Brief literature comparison

In the recent decades, we've seen a vast rise in the literature concerning high dimensional models used together with a large number of variables, also known as a scenario of big data. In this regard, we can name different studies that demonstrated how

these models are able to bring improvements in the accuracy of forecasts. Namely, for the case of U.S inflation, the LASSO and Random Forest showed a considerably better performance than the benchmark (Medeiros, Vasconcelos & Zilberman, 2021). We've also seen this for the case of Brazil's inflation (Garcia, Medeiros, Vasconcelos, 2017).

Ultimately, in a recent study we saw that using a disaggregated approach for forecasting industrial production can significantly increase a model's performance (de Prince, Marçal & Pereira, 2022). This was the case for the ETS (exponential triple smoothing) model in forecasting industrial production in Brazil – using the same index we're going to use in this study.

2. DATA

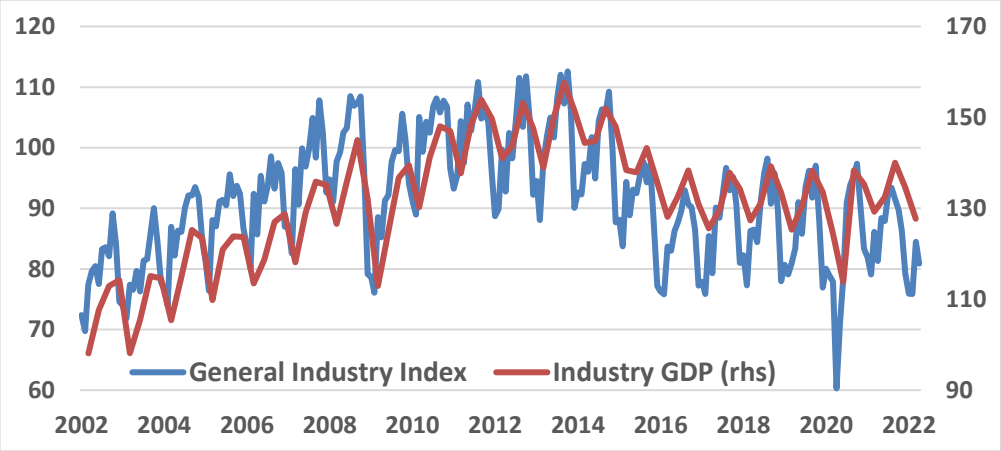
The data we are using comes from the Monthly Industrial Survey of Physical Production (PIM-PF) which is organized by the Brazilian Institute of Geography and Statistics (IBGE). It is a monthly data that ranges from January 2002 to April 2022 (241 observations) and we do not consider any seasonal adjustments or past revisions. It's worth mentioning that this indicator is usually released by the first week of the month, and since it is lagged by two months, the data for April would only be released in June. In order to maintain the dataset close to what the econometrician would have available by the time he would be making these forecasts, we adjusted all the series accordingly to the lags of its releases. For example in the case of PIM-PF, the data for April would appear as June in the dataset, since it was the month when the econometrician was actually able to observe it. As a result, the projection made in June would forecast the index for May (which wasn't yet available for the econometrician). The purpose of doing that is to maintain these forecasts as close as possible to what would be considered 'real-time'. However, it is not exactly 'real-time' forecasts because we didn't have access to the first release data (without any revisions). Moreover, we use a month over month percentage change to assure stationarity (this month compared to the immediately precedent month).

In this regard, we are considering that the econometrician would usually estimate their projection around the 20th day of the month, so all data available to him by that moment would be considered in the models. Moreover, every model is being estimated through a rolling window methodology, with each window having a range of 10 years, so the first one

would go from June 2002 to June 2012 (we start in June because it is the first month when the econometrician can observe two lags of the monthly change). The first forecast would be made for the data released in July 2012, that's referent for May 2012. This way, each model is being evaluated with a sample size of 121 observations and the last forecast is for April 2022 (released in June 2022), resulting in a total of 120 windows.

This period covers a range of scenarios for the Brazilian economy. The initial forecasts are being made with a somewhat positive macro outlook, with the GDP growing 1.9% in 2012 and 3.0% in 2013. However, looking to the industry GDP component, it showed negative a growth of -0.7% in 2012 and rebounded to 2.2% in 2013. Afterwards, from 2014 to 2016 the industry sector slowed down considerably with several years of recession. And from that point onwards, it has shown minor signs of recovery, but remained somewhat stable in lower than previous levels. In 2020, when the pandemic hit, the sector largely tumbled again, but recovered sharply by the end of the year and through the beginning of 2021. Finally, it has slowed a bit after hitting a peak in 2021, and in the most recent months of the sample, both the industry GDP and the PIM-PF index growth has been trending upwards.

Figure 1: PIM-PF vs Industry GDP Index NSA



Source: IBGE

We are using both the aggregated (general index) and the disaggregated indexes (mining and manufacturing industries indexes) for our forecasts, aiming to find out which method performs best at which given scenario. Regarding the aggregation method for the data divided by sectors, we are using the weights provided by the PIM-PF survey, which are fixed since 2002 – 11.2% to the mining sector and 88.8% to the manufacturing sector.

Concerning the set of possible predictors, the variables used in the models covers a broad range of industrial production, activity indicators, labor market, inflation, energy consumption, mobility indicators, uncertainty index, exchange rates, credit statistics, trade balance and monetary data. Furthermore, we are considering two lags of each variable plus two lags of the PIM-PF index. We are also making any adjustments necessary to guarantee the stationarity of our data. Additionally, we do not consider any market consensus forecast, since these are generally not that accurate. This also poses as a motivation for this study in order to verify if these high dimensional models can contribute to improve the precision of forecasts in this sector. The full list of predictors counts with 51 variables and is available at the end of the paper in appendix A.

On top of that, it's worth mentioning that not every variable in the dataset is available at the exact same range as the PIM-PF index. This means that some of these series begins after January 2002, so they wouldn't be available as predictors for some of the initial windows. However, as soon as the series starts and we have enough observations to include them in the dataset for that specific window, they would be considered.

3. EMPIRICAL METHODS

As was already mentioned, we aim to compare 2 different approaches: a forecast of the general industry index vs. a forecast for each sector aggregated by its respective weights. Regarding the empirical methods utilized, consider the following model:

$$y_{t+h} = F(\boldsymbol{x_t}) + u_{t+h}; \quad t = 1, \dots, T$$

where y_{t+h} is the industrial production index in month t+h, F(.) is a linear or nonlinear mapping between covariates and industrial production, $\boldsymbol{x_t} = (x_{1t}, \dots, x_{nt})'$ is a n-vector containing the potential predictors, as well as lags of y_t and monthly dummies, and u_t is a zero-mean random error. Lastly, we are only forecasting for the horizon of one month ahead (h = 1).

3.1. LASSO and adaptive-LASSO

The Least Absolute Shrinkage and Selection Operator (LASSO) is a shrinkage method capable of dealing with a large quantity of parameters. Its idea is to shrink to zero any irrelevant variable, thus reducing the dimensionality of the problem. It is worth noting that

the LASSO requires that the "irrepresentable condition" is satisfied in order to select models that are consistent. Put simply, this means that there should be a limit to the correlation between selected and not selected variables. So, for example, if two variables from the dataset are relevant, but they are highly correlated, chances are LASSO would probably only include one of them (the more relevant one). The LASSO estimator is defined as:

$$\hat{\beta}_{LASSO}(\lambda) = \underset{\beta \in \mathbb{R}^n}{arg\,min} \left\{ \frac{1}{T} \sum_{i=1}^{T} (y_i - \beta x_i)^2 + \lambda \sum_{j=1}^{n} |\beta_j| \right\}$$

where $\beta = (\beta_1, ..., \beta_n)' \in \mathbb{R}^n$, $x_i = (x_{1t}, \ldots, x_{nt})$, and λ is the parameter that determines how strong the shrinkage will be and it is set by data-oriented techniques, namely cross-validation and information criteria (BIC or AIC). In this study, we're using the BIC.

The adaptive-LASSO is similar to the above except for the addition of one more component responsible to assign weights to each covariate. In order to compute these weights, it's necessary to follow a 2-step method which determines a first step coefficient (that can be achieved through LASSO or OLS) to define the importance of each regressor. It is defined as following:

$$\hat{\beta}_{adaLASSO}(\lambda) = \underset{\beta \in \mathbb{R}^n}{arg\,min} \left\{ \frac{1}{T} \sum_{i=1}^{T} (y_i - \beta x_i)^2 + \lambda \sum_{j=1}^{n} \omega_j |\beta_j| \right\}$$

where $\omega_j = \left(|\widehat{\beta_j^*}| + \frac{1}{\sqrt{T}} \right)^{-\tau}$ is the weight aforementioned, in which $\widehat{\beta_j^*}$ are the coefficients estimated in the first step through a LASSO model, and $0 < \tau < 1$. We will consider $\tau = 1$. The purpose of adding this component is to assign heavier weights to the variables that are less relevant and thus decreasing the chances of the second step LASSO selecting these variables, relieving the irrepresentable condition (which is called weighted irrepresentable condition in this case). Moreover, under some circumstances, the adaptive-LASSO has the oracle property, meaning that the estimation is asymptotically equivalent to the OLS estimation using only the relevant variables, if these were known.

3.2. Factor models

Considering that our set of predictors consists of $\boldsymbol{x_t} = (x_{1t}, \ldots, x_{nt})$ and $n > t$ (or at least close to it), we wouldn't be able to estimate our regression through the ordinary least squares method, or the coefficients would show a very large variance. The reason for that is because

the number of covariates is larger, or close, to the sample size, and in these high dimension scenarios the OLS doesn't work.

In order to surpass this, a possible method for estimating the regression is through a factor model. These models consists in extracting principal components from the set of predictors and use them as covariates in our regression instead of $\boldsymbol{x_t}$ (these principal components can also be called 'factors' in this case). Since the principal components analysis is a way to summarize the variability of the dataset, we're able to use a number k < n of factors and thus solve the high dimension problem. It's worth mentioning that we're able to create as many principal components as the minimum of min{n,t}, however, as each factor is calculated, the amount of the original dataset's variance explained by it is being reduced. In other words, the first principal component is able to summarize a bigger portion of the data's variability than the second, and so on. In this regard, consider the following model:

$$y_{t+h} = \beta_0 + \sum_{i=1}^{k} \gamma_i f_{t,i} + \beta_1 y_{t-1} + \beta_2 y_{t-2} + u_{t+h}; \quad t = 1, \dots, T; \; h = 1$$

where β_0 is the intercept, $f_{i,t}$ is a vector of k principal components extracted from $\boldsymbol{x_t}$ (considering only non-lagged variables), in which k is significantly smaller than n, and $\gamma = (\gamma_1, \dots, \gamma_k)$ is the set of coefficients that approximates these factors to the variable of interest. We do not consider any lags of the principal components, but we do consider two lags of the industrial production index (dependent variable). Furthermore, the value for k is given by a formal procedure (Onatski, 2010) that maximizes the following:

$$k = \arg\max_{1 \le j < n} \frac{\lambda_j}{\lambda_{j+1}}$$

where λ_j is the variance of the j principal component. Since each factor is responsible to summarize only a portion of the data's variability, by this way we can select the number of principal components that maximizes the change of variance explained by one component to the one immediately after.

3.2.1. Factor models with target

The target factor model is similar to the factor model, however it adds a precedent step before computing the principal components. Instead of calculating these components through the whole dataset, we first isolate only the variables that were selected by the

LASSO and them computes the principal components. All the other variables that weren't selected would be discarded since their presence in the dataset could create noise when computing the factors. In the case of the first step LASSO selecting a lagged variable, we would consider the contemporary version of this variable (without lag). This way, we consider the correlation between $\boldsymbol{x_t}$ and y_t in the first step, and then only summarize the variance of the data that's relevant for explaining y_t. Other studies shows that this targeting is able to considerably reduce the forecasting errors of these models.

Moreover, after computing the principal components, we're using a LASSO in a dataset with every component, as well as two dependent variable lags. We do this in order for letting the LASSO estimator select which components would be considered in the model and if it's worth to add any lags into it.

3.3. Random forests

The Random Forest model was brought up by Breiman (2001) as a method to stabilize regression trees by applying the concept of bootstrap aggregation (also known as bagging) in randomly constructed regression trees, leading to a reduction of their variance.

Concerning regression trees, they are a flexible nonparametric model which aims to approximate an unknown non-linear function to a set of observations. This is done by using recursive binary partitioning, in the form of nodes, which defines rules in the space of the covariates. In each partition, the idea is to calculate a simple local model, usually the mean of all observations that belongs to the subspace, and then compute its residuals (difference between actual observations and the mean) to quantify the quality of these predictions. This way, it can determine the sum of squared residuals that each threshold would give and then, by minimizing the sum of squared residuals, it would find the best possible threshold that would give the best predictions inside each partition it creates. The first threshold would be the root of the tree, and by doing this recursively (same procedure inside each subspace), we can create an array of nodes dividing the dataset into a variety of partitions. This could be done repeatedly until there's one observation in each subspace, however this would lead to an overfitting situation, so the model wouldn't perform well with new data (bias-variance trade-off). In order to prevent that, there's a minimum number of observations required in each partition of the space. All in all, the model's main advantage is its high interpretability and usually low bias, but, on the other hand, it is generally very unstable and presents a high variance.

In order to reduce this elevated variance, the Random Forest model applies the concept of bagging in regression trees. Initially, it creates a bootstrap dataset by randomly selecting samples from the original dataset, with the possibility of picking the same sample more than once. Next, it must also select only a few random variables from each bootstrap dataset in order to create very distinct trees. So overall, there's two random selections happening: one in the observations and another one in the variables selected. As a result, we create a wide range of different regression trees which makes the random forest more effective than one individual tree. Afterwards, we can run the data down in every tree created and keep track of the predictions each tree gives, so that later we can aggregate these projections and find the prediction the random forest made. This is aggregate decision using bootstrapped datasets is what is called 'bagging'.

The total number of decision trees in a random forest can be monitored by the out-of-the-bag error, which uses the samples that were not selected to the bootstrap dataset in order to evaluate the quality of the random forest predictions. The idea is to select the number of trees that's able to minimize the out-of-the-bag error. In this regard, the model in itself is already working on a way to minimize its forecasts out-of-sample errors, so it has good overall predictions properties. However, it's worth noting that due to the bootstrapped samples, random regression trees, and bagging the model has a more difficult interpretability.

4. MAIN RESULTS

All models described above were estimated considering a short term forecast with the horizon of only one month ahead (h = 1). The initial window ranges from June 2002 to June 2012, and the first forecast is for July 2012, referent for the industrial production at May 2012.

4.1. Aggregated approach: general industry index

In table 1, we can see the comparison of the results regarding the models for the general industry index. We use out of sample parameters in order to evaluate the model's quality. Namely, we're computing the forecasting MSE (mean squared errors), RMSE (root mean squared errors) and MAE (mean absolute errors). The values in bold represent the lowest errors in the column. Furthermore, the last two columns of the table puts every model in comparison against the benchmark used, which is the AR(1). The lower the ratio, the better the model quality will be. In these columns, we can verify that every model using machine learning estimates was able to beat the benchmark providing a significant reduction of these measures of out-of-sample forecasting errors.

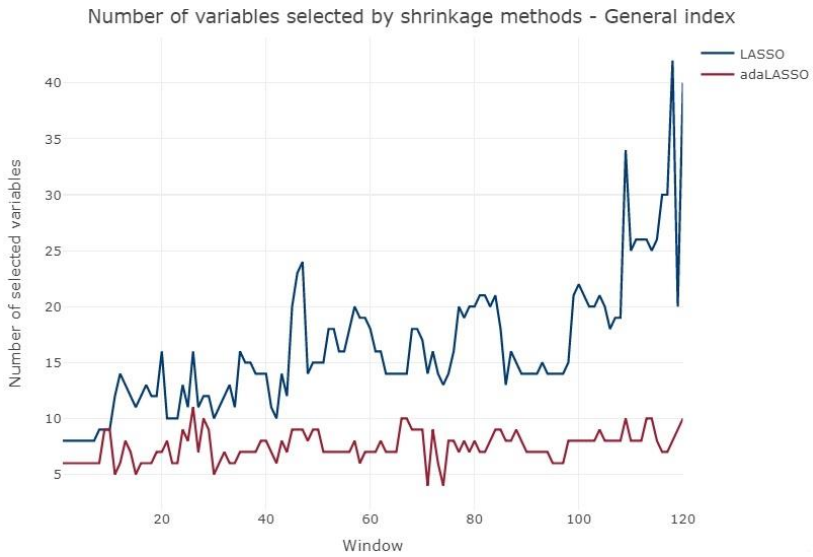
Table 1: Results for the general industry index

Results for General Industy Index					
	MSE	RMSE	MAE	RMSE/ RMSE(AR1)	MAE / MAE(AR1)
LASSO	**0.00039**	**0.01974**	**0.01444**	**0.28259**	**0.26426**
ada-LASSO	0.00053	0.02308	0.01674	0.33034	0.30633
Factor Model	0.00343	0.05854	0.03839	0.83782	0.70240
Target Factor Model	0.00121	0.03485	0.02501	0.49883	0.45773
Random Forest	0.00073	0.02704	0.01757	0.38701	0.32152
AR(1)	0.00488	0.06987	0.05465	1.00000	1.00000

*Values in bold represent the smallest forecasting errors in each column.

For the case of the general industry index, the LASSO was the best performing model showing a reduction of roughly 74% of the RMSE and 72% of the MAE when compared to the benchmark. In second place, was the adaptive-LASSO model, which corroborates with the idea that both of these estimators have good short-term prediction properties. This is aligned with the results in (Medeiros, Vasconcelos & Zilberman, 2021), in which the LASSO and adaptive-LASSO performed well in short-term horizons. Nonetheless, their quality were close to each other with the LASSO presenting a RMSE of roughly 15% less than the adaptive-LASSO. Additionally, as theory suggests, the adaptive-LASSO was more restrictive than the LASSO in its variable selection.

Figure 2: Number of variables selected by shrinkage methods in each window: General industry index



One might find strange that the number of variables selected by the LASSO is increasing with the window number. For clarification purposes, a possible reason for that is because, as was mentioned in the data section, some of these series doesn't start at the same time as the PIM-PF index. So, as the windows move forward, we're including more variables in the dataset, thus the LASSO has the possibility to select a higher number of predictors.

Furthermore, in the table at appendix B, we can see the proportion that each variable was selected in both shrinkage models. The 1st lag of the ABCR heavy vehicles (trucks and buses) mobility indicator was selected in every window by the LASSO, which makes sense considering the high correlation of this variable with industrial production ($\rho = 0.83$) – more truck's traffic in the highways in the previous month suggests that more inputs are circulating and reaching factory gates. Other relevant predictors were oil and gas production (industrial inputs), total of hours worked in the manufacturing industries sector (labor supply), usage of installed industrial capacity, and others.

Continuing looking at the models ranking, the random forest placed third in terms of quality, but it's still presenting a considerable decrease of 62% and 68% from the benchmark RMSE and MAE, respectively. It seems that in this case, the non-linearity of the random forest didn't improve much the accuracy of these forecasts when compared to other linear models such as the LASSO and adaptive-LASSO. Moreover, it's probable that the good predictor selection from these shrinkage methods also explains the difference in

these results. Lastly, it's worth mentioning that other studies suggests that the random forest usually is not the best model in any specific horizon, but its overall performance can be consistently good (Garcia, Medeiros and Vasconcelos, 2017).

Looking at the factor model and the target factor model we verify that using a targeting method before computing the principal components can significantly increase the model performance, showing a reduction of almost 40% of the RMSE. Nonetheless, both of these performed worse than the other high dimensional models tested.

4.2. Disaggregated approach

Looking at table 2, which has similar format as table 1, we can see the results regarding the manufacturing industries index.

Table 2: Results for the manufacturing industries index

Results for Manufacturing Industries					
	MSE	RMSE	MAE	RMSE/RMSE(AR1)	MAE/MAE(AR1)
LASSO	0.00064	0.02529	**0.01593**	0.33304	**0.27346**
ada-LASSO	**0.00049**	**0.02216**	0.01617	**0.29185**	0.27754
Factor Model	0.00393	0.06268	0.04002	0.82552	0.68694
Target Factor Model	0.00186	0.04317	0.03027	0.56857	0.51952
Random Forest	0.00084	0.02895	0.01835	0.38122	0.31491
AR(1)	0.00577	0.07593	0.05827	1.00000	1.00000

*Values in bold represent the smallest forecasting errors for each column.

Once again, every model was able to beat the benchmark with a considerable reduction of the forecasting errors. Overall, these results are very similar to the results for the general industry index, because the manufacturing industries sector is responsible for the largest participation in the general index.

In this case, the adaptive-LASSO showed the smallest RMSE, presenting a 71% reduction when compared to the AR(1). However, the LASSO presented the smaller MAE, with a 72% decrease against the benchmark. The reason for the best models changing between parameters is because the MSE is more sensitive to outliers than the MAE, thus bigger error in some specific windows might have inflated the MSE of the LASSO model. Anyhow, their results are pretty similar, likewise as in table 1.

Maintaining the same format, in table 3 we can verify the outcomes for the mining & quarrying industries index. For this sector, although every model was able to beat the benchmark, we have relatively worse results. The first reason for that is because the AR(1)

model was able to perform better than in the other segments. And secondly, every other model, except for the factor model, showed a considerably worse performance, thus resulting in higher forecasting errors ratios in the last two columns of the table.

Table 3: Results for the mining & quarrying industries index

Results for Mining & Quarrying Industries					
	MSE	RMSE	MAE	RMSE/RMSE(AR1)	MAE/MAE(AR1)
LASSO	0.00183	0.04276	0.02913	0.67456	0.61229
ada-LASSO	**0.00161**	**0.04009**	**0.02818**	**0.63240**	**0.59224**
Factor Model	0.00389	0.06234	0.04692	0.98330	0.98613
Target Factor Model	0.00248	0.04979	0.03333	0.78536	0.70056
Random Forest	0.00166	0.04078	0.02941	0.64333	0.61810
AR(1)	0.00402	0.06340	0.04758	1.00000	1.00000

*Values in bold represent the smallest forecasting errors for each column.

A possible explanation for these overall worse results in this sub-index is because the sector is majorly composed by some large specific companies, like Vale and Petrobrás. So perhaps, a more suitable approach for this case is using these companies' production estimations and reports as predictors, instead of the general macroeconomics dataset used. This is aligned with the outcomes regarding the average number of variables selected by shrinkage methods in each different sector. Since these set of covariates seems to be less relevant for explaining the mining & quarrying sector, it's also plausible that the shrinkage models in this segment would select a lower number of variables. And that's what we can verify by looking at figures 3 and 4, in which the average number of selected variables by the LASSO for the manufacturing sector is 15, while for the mining & quarrying sector is 9 (the averages for the adaptive-LASSO are 7 and 6, respectively).

All in all, for the mining & quarrying index, the adaptive-LASSO was the model that presented the smallest forecasts errors with a 37% RMSE reduction and a 41% MAE decrease compared to the AR(1). Since this model is more restrictive in its variable selection, this result also corroborates with the idea that a more company specific approach, driven by production reports and guidance, could be more suitable for modeling this sector than the general approach with macroeconomics data.

Furthermore, this was the only segment in which the random forest was able to beat the performance of a shrinkage method model, namely the LASSO in this case. A possible justification for that might be because the LASSO selected a number of variables that aren't

that much relevant to explaining the sector performance, thus creating noise in the model and producing worse forecasts.

Figure 3: Number of variables selected by shrinkage methods in each window in the manufacturing industries sector

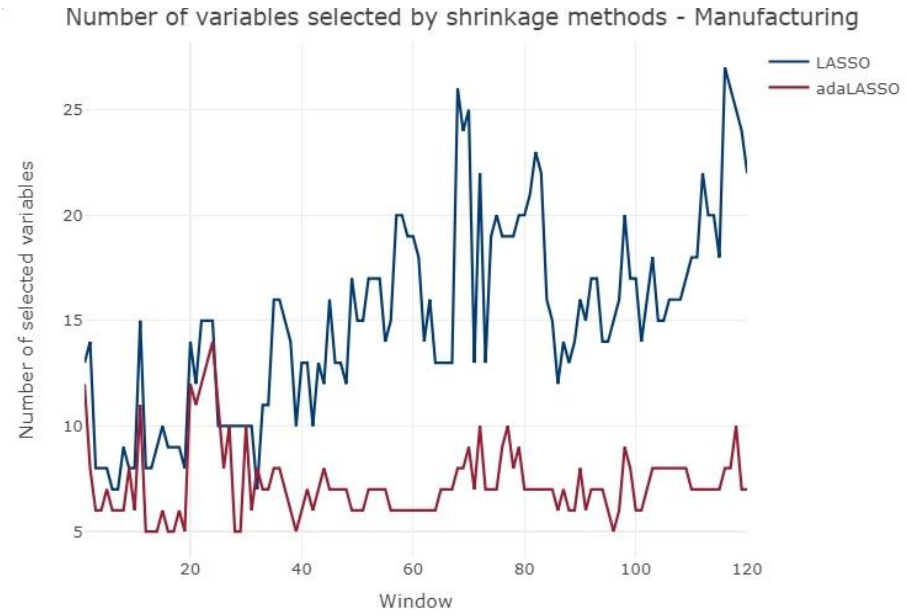
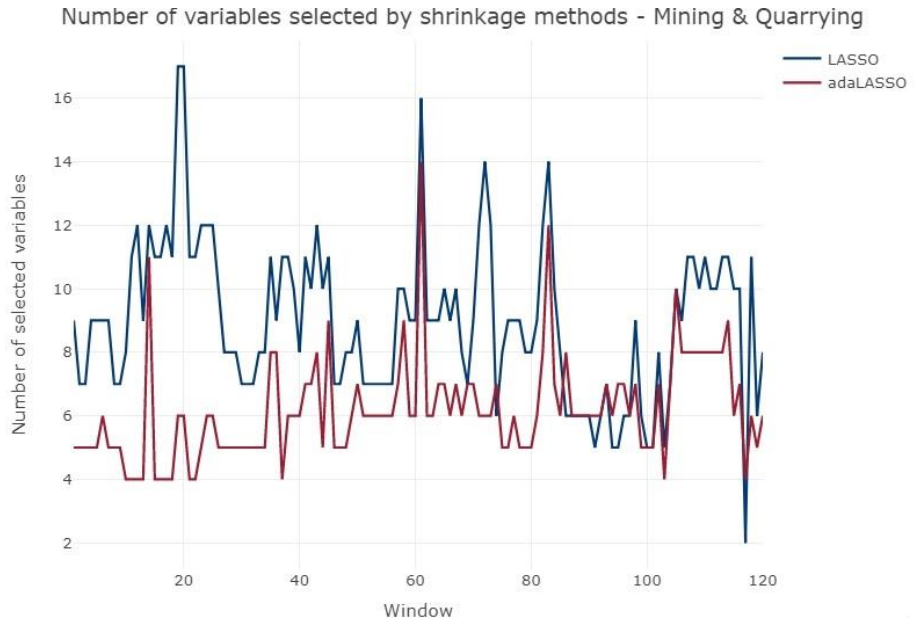


Figure 4: Number of variables selected by shrinkage methods in each window in the mining & quarrying industries sector



Lastly, in table 4, we can verify the RMSE and MAE after aggregating back these sub-indexes forecasts into the general index forecast, by their respective weights. In the last two columns of this table we can see these parameters as a ratio against the benchmark model for the general industry index (e.g. RMSE of a model using the disaggregated approach versus the RMSE of the general industry index AR(1)).

Table 4: Results after aggregating the sub-indexes forecasts

Results After Aggregating Sub-indexes					
	MSE	RMSE	MAE	RMSE/RMSE(AR1)	MAE/MAE(AR1)
LASSO	0.00058	0.02414	0.01549	0.34550	0.28348
ada-LASSO	**0.00043**	**0.02086**	**0.01543**	**0.29849**	**0.28230**
Factor Model	0.00344	0.05865	0.03864	0.83942	0.70702
Target Factor Model	0.00158	0.03974	0.02789	0.56882	0.51042
Random Forest	0.00070	0.02640	0.01696	0.37780	0.31039
AR(1)	0.00488	0.06985	0.05457	0.99975	0.99856

*Values in bold represent the smallest forecasting errors for each column.

** Last two columns used the RMSE and MAE from AR(1) model in table 1 to compute the ratios.

Through this approach, the adaptive-LASSO was the best performing model, achieving a 70% reduction of the RMSE and 72% of the MAE, when compared to the general industry AR(1). Comparing the results above with table 1, we can confirm an improvement in the performance of the adaptive-LASSO and the random forest models. However, all the other models (LASSO, and both factor models) showed relatively worse forecasting errors when using the disaggregated approach.

Moreover the lowest ratio, comparing the results of these high dimensional models with the aggregated approach benchmark, is coming from the general industry index LASSO model at table 1. Followed by the adaptive-LASSO through the disaggregated method. Nonetheless, it's worth mentioning that, since these 2 models showed very similar levels of RMSE and MAE, it's hard to define exactly which one is the best, and probably there's no significant difference in their performance. All things considered, every model tested was able to beat the benchmark AR(1). And, in this study, the disaggregated procedure was able to bring improvements in some models, but the overall best performing model still came from the aggregated method LASSO.

5. CONCLUSION

In this study, we show that, for the case of forecasting industrial production in Brazil through the PIM-PF index, high dimensional models, namely the LASSO, adaptive-LASSO, random forest, factor model and target factor model, are all able to produce more accurate forecasts than the benchmark model, AR(1), in terms of out-of-sample MSE, RMSE and MAE. This is done considering a wide dataset with over 51 potential predictors and 2 lags of each.

For the purpose of calculating forecasting errors, we used a rolling window scheme and then computed out-of-sample forecasts for the horizon of one month ahead, in order to compare the model's predictions with the actual observation.

Furthermore, we also compared the models through two different approaches: an aggregated one, using the general industry index, and a disaggregated, using the manufacturing industries sub-index and the mining & quarrying sector sub-index. In this regard, the most accurate forecast came from the LASSO model through the aggregated approach. However, the disaggregated method was able to show improvements in other high dimensional models such as the adaptive-LASSO, that ranked second in terms of showing the smallest forecasting errors, and the random forest (that still showed relatively good accuracy). Moreover, their results were very similar to the aggregated LASSO, meaning that there's probably no significant difference in which model presented the most accurate forecasts. The target factor model showed a considerably better performance than the factor model, but was less competitive than the other high dimensional models that were tested.

Ultimately, it's noteworthy that there remains plenty of ground to cover in this field. Other possibilities for this study would be testing out other models, such as the complete subset regression (CSR), which showed a good overall performance in other studies (Garcia, Medeiros and Vasconcelos, 2017). Additionally, analyzing the results for different forecast horizons, as well as the combination of different models could also pose as motivation for a following study. Lastly, one thing that we could've done to deepen this analysis would be implementing a forecasts comparison test (Diebold-Mariano, 1995) to define if there was any statistical difference in the performance of the best models.

6. BIBLIOGRAPHY

BREIMAN, L. **Random Forests**. University of California Statistics Department. Berkeley. 2001.

DIEBOLD, F. M. R. Comparing predictive accuracy. **Journal of Business & economic statistics**, v. 13, p. 253-265, 1995.

GARCIA, M. G. P.; MEDEIROS, M. C.; VASCONCELOS, G. F. R. Real-time inflation forecasting with high-dimensional models: The case of Brazil. **Elsevier, International Journal of Forecasting**, 2017.

MEDEIROS, M. C. et al. Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods. **Journal of Business & Economics Statistics**, p. 98-119, August 2019.

ONATSKI, A. Determining the Number of Factors from Empirical Distribution of Eigenvalues. **The Review of Economics and Statistics**, v. 92, n. 4, p. 1004-1016, nov. 2010.

PRINCE, D. D.; MARÇAL, E. F.; PEREIRA, P. L. V. Forecasting Industrial Production Using Its Aggregated and Disaggregated Series or a Combinantion of Both: Evidence from One Emerging Market Economy. **Econometrics 10 (2), 27**, 2022.

IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. PIM-PF methodology and sub-indexes weights. Available at: <https://www.ibge.gov.br/estatisticas/economicas/industria/9294-pesquisa-industrial-mensal-producao-fisica-brasil.html?=&t=downloads>. Access in 17 Aug. 2022.

APPENDICES

Appendix A – All variables used and their release lags

Variable Description	Source	Release Lags*
Activity		
General industry index (PIM-PF)	IBGE (Brazilian Institute of Geography and Statistics)	2
Manufacturing industries index (PIM-PF)	IBGE (Brazilian Institute of Geography and Statistics)	2
Mining & Quarrying industries index (PIM-PF)	IBGE (Brazilian Institute of Geography and Statistics)	2
Industry sector sentiment	FGV Ibre (Fundação Getúlio Vargas)	1
Retail sector sentiment	FGV Ibre (Fundação Getúlio Vargas)	1
Consumer sentiment	FGV Ibre (Fundação Getúlio Vargas)	1
Business sentiment	FGV Ibre (Fundação Getúlio Vargas)	1
Usage of industrial capacity level (NUCI)	FGV Ibre (Fundação Getúlio Vargas)	1
Economic uncertainty index (IIE-Br)	FGV Ibre (Fundação Getúlio Vargas)	1
Steel production	IAB (Instituto Aço Brasil)	1
Laminated steel production	IAB (Instituto Aço Brasil)	1
Corrugated cardboard production	Empapel (Associação Brasileira de Embalagens em Papel)	1
Total vehicles production	ANFAVEA (Associação Nacional de Frabricantes de Veículos Automores)	1
Credit for corporations	Brazilian Central Bank	2
Oil production	Agência Nacional de Petróleo	2
Gas production	Agência Nacional de Petróleo	2
Manufacturing industries sales	CNI (Confederação Nacional da Indústria)	2
Usage of industrial capacity level (NUCI)	CNI (Confederação Nacional da Indústria)	2
Vehicles sales	Fenabrave (Federação Nacional da Distribuição de Veículos Automotores)	1
Manufacturing industries activity indicator	FIESP (Federação das Indústrias do Estado de São Paulo)	1
Manufacturing industries market sentiment indicator	FIESP (Federação das Indústrias do Estado de São Paulo)	1
Manufacturing industries sales indicator	FIESP (Federação das Indústrias do Estado de São Paulo)	1
Manufacturing industries inventory indicator	FIESP (Federação das Indústrias do Estado de São Paulo)	1
Manufacturing industries investment indicator	FIESP (Federação das Indústrias do Estado de São Paulo)	1
Mobility indicator for heavy vehicles (buses and trucks)	ABCR (Associação Brasileira de Captadores de Recursos)	1
Energy		
Energy consumption for the industry sector	EPE (Empresa de Pesquisa Energética)	2
General energy consumption	ONS (Operador Nacional do Sistema Elétrico)	0
Volume levels in water reservoirs	ONS (Operador Nacional do Sistema Elétrico)	0
Labor market		
Real expanded wage bill (PNAD + Tesouro Nacional)	IBGE (Brazilian Institute of Geography and Statistics) + Tesouro Nacional	2
Employed population in the industry sector	IBGE (Brazilian Institute of Geography and Statistics)	2
Employed population in the industry sector	CAGED - Ministério do Trabalho	2
Employed population in the mining & quarrying industries sector	CAGED - Ministério do Trabalho	2
Employed population in the manufacturing industries sector	CAGED - Ministério do Trabalho	2
Unemployment rate - interpolated series with PNAD and PME	IBGE (Brazilian Institute of Geography and Statistics)	2
Employment indicator for the manufacturing industries sector	CNI (Confederação Nacional da Indústria)	2
Hours worked indicator for the manufacturing industries sector	CNI (Confederação Nacional da Indústria)	2
Real wage bill for employees in the manufacturing industries sector	CNI (Confederação Nacional da Indústria)	2
Real productivity for employees in the manufacturing industries sector	CNI (Confederação Nacional da Indústria)	2
Employment indicator for the manufacturing industries sector	FIESP (Federação das Indústrias do Estado de São Paulo)	1
Prices		
Consumer's price index (IPCA)	IBGE (Brazilian Institute of Geography and Statistics)	1
Producer's price index for the general industry (IPP)	IBGE (Brazilian Institute of Geography and Statistics)	2
Producer's price index for the manufacturing industries (IPP)	IBGE (Brazilian Institute of Geography and Statistics)	2
Producer's price index for the mining & quarrying industries (IPP)	IBGE (Brazilian Institute of Geography and Statistics)	2
Retail prices index	FGV Ibre (Fundação Getúlio Vargas)	1
Trade balance		
Total imports in quantity	Funcex (Fundação Centro de Estudos do Comércio Exterior)	2
Capital goods imports in quantity	Funcex (Fundação Centro de Estudos do Comércio Exterior)	2
Durable goods imports in quantity	Funcex (Fundação Centro de Estudos do Comércio Exterior)	2
Non-durable goods imports in quantity	Funcex (Fundação Centro de Estudos do Comércio Exterior)	2
Intermediate goods imports in quantity	Funcex (Fundação Centro de Estudos do Comércio Exterior)	2
Monetary policy & exchange rate		
Real exchange rate index	Brazilian Central Bank	1
Real interest rate (Selic)	Brazilian Central Bank	1

*Release lags refers to the number of months in which the release date lags the reference date

Appendix B – Proportion that each variable was selected by shrinkage models: General industry index

General Industry Index					
Variable Description	**Source**	**Number of times selected by the LASSO**	**LASSO Proportion**	**Number of times selected by the adaptive-LASSO**	**adaptive-LASSO Proportion**
Mobility indicator for heavy vehicles (buses and trucks) - 1st lag	ABCR (Associação Brasileira de Captadores de Recursos)	120	100,0%	99	82,5%
Gas production	Agência Nacional de Petróleo	113	94,2%	48	40,0%
Hours worked indicator for the manufacturing industries sector	CNI (Confederação Nacional da Indústria)	110	91,7%	110	91,7%
Usage of industrial capacity level (NUCI)	CNI (Confederação Nacional da Indústria)	110	91,7%	106	88,3%
Oil production	Agência Nacional de Petróleo	106	88,3%	94	78,3%
Manufacturing industries sales	CNI (Confederação Nacional da Indústria)	106	88,3%	29	24,2%
Intermediate goods imports in quantity - 1st lag	Funcex (Fundação Centro de Estudos do Comércio Exterior)	96	80,0%	13	10,8%
Total vehicles production - 1st lag	ANFAVEA (Associação Nacional de Frabricantes de Veículos Automores)	94	78,3%	29	24,2%
Corrugated cardboard production - 1st lag	Empapel (Associação Brasileira de Embalagens em Papel)	86	71,7%	61	50,8%
Consumer sentiment - 1st lag	FGV Ibre (Fundação Getúlio Vargas)	72	60,0%	22	18,3%
Real expanded wage bill (PNAD + Tesouro Nacional) - 2nd lag	IBGE (Brazilian Institute of Geography and Statistics) + Tesouro Nacional	70	58,3%	38	31,7%
Steel production - 1st lag	IAB (Instituto Aço Brasil)	67	55,8%	0	0,0%
Volume levels in water reservoirs	ONS (Operador Nacional do Sistema Elétrico)	66	55,0%	29	24,2%
Consumer's price index (IPCA) - 1st lag	IBGE (Brazilian Institute of Geography and Statistics)	62	51,7%	62	51,7%
Unemployment rate - interpolated series with PNAD and PME - 2nd lag	IBGE (Brazilian Institute of Geography and Statistics)	56	46,7%	54	45,0%
Real productivity for employees in the manufacturing industries sector - 2nd lag	CNI (Confederação Nacional da Indústria)	55	45,8%	2	1,7%
Consumer sentiment - 2nd lag	FGV Ibre (Fundação Getúlio Vargas)	49	40,8%	13	10,8%
Steel production	IAB (Instituto Aço Brasil)	43	35,8%	0	0,0%
Manufacturing industries sales indicator - 1st lag	FIESP (Federação das Indústrias do Estado de São Paulo)	34	28,3%	0	0,0%
Real exchange rate index	Brazilian Central Bank	32	26,7%	0	0,0%
Oil production - 1st lag	Agência Nacional de Petróleo	32	26,7%	11	9,2%
Usage of industrial capacity level (NUCI)	FGV Ibre (Fundação Getúlio Vargas)	27	22,5%	23	19,2%
Economic uncertainty index (IIE-Br) - 2nd lag	FGV Ibre (Fundação Getúlio Vargas)	24	20,0%	0	0,0%
Laminated steel production - 1st lag	IAB (Instituto Aço Brasil)	22	18,3%	0	0,0%
Business sentiment - 2nd lag	FGV Ibre (Fundação Getúlio Vargas)	21	17,5%	6	5,0%
Durable goods imports in quantity - 1st lag	Funcex (Fundação Centro de Estudos do Comércio Exterior)	19	15,8%	0	0,0%
Energy consumption for the industry sector	EPE (Empresa de Pesquisa Energética)	18	15,0%	0	0,0%
Retail sector sentiment	FGV Ibre (Fundação Getúlio Vargas)	17	14,2%	0	0,0%
Real wage bill for employees in the manufacturing industries sector - 2nd lag	CNI (Confederação Nacional da Indústria)	17	14,2%	0	0,0%
Industry sector Sentiment	FGV Ibre (Fundação Getúlio Vargas)	13	10,8%	0	0,0%
Producer's price index for the manufacturing industries (IPP) - 1st lag	IBGE (Brazilian Institute of Geography and Statistics)	12	10,0%	2	1,7%
Retail prices index - 1st lag	FGV Ibre (Fundação Getúlio Vargas)	12	10,0%	0	0,0%
Retail prices index	FGV Ibre (Fundação Getúlio Vargas)	11	9,2%	2	1,7%
Non-durable goods imports in quantity - 1st lag	Funcex (Fundação Centro de Estudos do Comércio Exterior)	11	9,2%	0	0,0%
Manufacturing industries market sentiment indicator - 1st lag	FIESP (Federação das Indústrias do Estado de São Paulo)	11	9,2%	0	0,0%
Industry sector Sentiment - 2nd lag	FGV Ibre (Fundação Getúlio Vargas)	11	9,2%	6	5,0%
Consumer's price index (IPCA)	IBGE (Brazilian Institute of Geography and Statistics)	10	8,3%	8	6,7%
Real interest rate (Selic) - 1st lag	Brazilian Central Bank	10	8,3%	10	8,3%
Employed population in the mining & quarrying industries sector - 2nd lag	CAGED - Ministério do Trabalho	8	6,7%	0	0,0%
Manufacturing industries inventory indicator	FIESP (Federação das Indústrias do Estado de São Paulo)	6	5,0%	0	0,0%
General energy consumption - 1st lag	ONS (Operador Nacional do Sistema Elétrico)	5	4,2%	0	0,0%
Manufacturing industries inventory indicator - 2nd lag	FIESP (Federação das Indústrias do Estado de São Paulo)	5	4,2%	0	0,0%
Employed population in the mining & quarrying industries sector - 1st lag	CAGED - Ministério do Trabalho	4	3,3%	0	0,0%
Manufacturing industries activity indicator - 1st lag	FIESP (Federação das Indústrias do Estado de São Paulo)	4	3,3%	0	0,0%
Employed population in the industry sector - 2nd lag	CAGED - Ministério do Trabalho	4	3,3%	0	0,0%
Gas production - 2nd lag	Agência Nacional de Petróleo	4	3,3%	0	0,0%
Real expanded wage bill (PNAD + Tesouro Nacional)	IBGE (Brazilian Institute of Geography and Statistics) + Tesouro Nacional	3	2,5%	0	0,0%
Employed population in the mining & quarrying industries sector	CAGED - Ministério do Trabalho	3	2,5%	2	1,7%
Usage of industrial capacity level (NUCI) - 1st lag	FGV Ibre (Fundação Getúlio Vargas)	3	2,5%	2	1,7%
Employment indicator for the manufacturing industries sector - 1st lag	FIESP (Federação das Indústrias do Estado de São Paulo)	3	2,5%	0	0,0%
Mobility indicator for heavy vehicles (buses and trucks)	ABCR (Associação Brasileira de Captadores de Recursos)	2	1,7%	0	0,0%
Employed population in the industry sector	CAGED - Ministério do Trabalho	2	1,7%	0	0,0%
Employed population in the manufacturing industries sector	CAGED - Ministério do Trabalho	2	1,7%	0	0,0%
Retail sector sentiment - 2nd lag	FGV Ibre (Fundação Getúlio Vargas)	2	1,7%	0	0,0%
Employed population in the manufacturing industries sector - 2nd lag	CAGED - Ministério do Trabalho	2	1,7%	0	0,0%
Manufacturing industries market sentiment indicator - 2nd lag	FIESP (Federação das Indústrias do Estado de São Paulo)	2	1,7%	0	0,0%
Manufacturing industries investment indicator - 2nd lag	FIESP (Federação das Indústrias do Estado de São Paulo)	2	1,7%	0	0,0%
Producer's price index for the manufacturing industries (IPP) - 2nd lag	IBGE (Brazilian Institute of Geography and Statistics)	2	1,7%	0	0,0%
Vehicles sales - 2nd lag	Fenabrave (Federação Nacional da Distribuição de Veículos Automotores)	0	0,0%	2	1,7%

*LASSO and adaptive-LASSO proportions are calculated by (number of times selected by model / total windows number)

Appendix C – Proportion that each variable was selected by shrinkage models: Manufacturing industries index

Manufacturing Industries Index					
Variable Description	**Source**	**Number of times selected by the LASSO**	**LASSO Proportion**	**Number of times selected by the adaptive-LASSO**	**adaptive-LASSO Proportion**
Mobility indicator for heavy vehicles (buses and trucks) - 1st lag	ABCR (Associação Brasileira de Captadores de Recursos)	120	100,0%	120	100,0%
Total vehicles production - 1st lag	ANFAVEA (Associação Nacional de Frabricantes de Veículos Automores)	120	100,0%	33	27,5%
Hours worked indicator for the manufacturing industries sector	CNI (Confederação Nacional da Indústria)	110	91,7%	110	91,7%
Usage of industrial capacity level (NUCI)	CNI (Confederação Nacional da Indústria)	110	91,7%	85	70,8%
Gas production	Agência Nacional de Petróleo	106	88,3%	44	36,7%
Intermediate goods imports in quantity - 1st lag	Funcex (Fundação Centro de Estudos do Comércio Exterior)	103	85,8%	10	8,3%
Manufacturing industries sales	CNI (Confederação Nacional da Indústria)	86	71,7%	3	2,5%
Corrugated cardboard production - 1st lag	Empapel (Associação Brasileira de Embalagens em Papel)	86	71,7%	62	51,7%
Oil production	Agência Nacional de Petróleo	83	69,2%	79	65,8%
Real expanded wage bill (PNAD + Tesouro Nacional) - 2nd lag	IBGE (Brazilian Institute of Geography and Statistics) + Tesouro Nacional	72	60,0%	47	39,2%
Steel production	IAB (Instituto Aço Brasil)	65	54,2%	6	5,0%
Consumer sentiment - 1st lag	FGV Ibre (Fundação Getúlio Vargas)	65	54,2%	37	30,8%
Steel production - 1st lag	IAB (Instituto Aço Brasil)	62	51,7%	0	0,0%
Unemployment rate - interpolated series with PNAD and PME - 2nd lag	IBGE (Brazilian Institute of Geography and Statistics)	57	47,5%	49	40,8%
Volume levels in water reservoirs	ONS (Operador Nacional do Sistema Elétrico)	54	45,0%	3	2,5%
Consumer's price index (IPCA) - 1st lag	IBGE (Brazilian Institute of Geography and Statistics)	50	41,7%	48	40,0%
Consumer sentiment - 2nd lag	FGV Ibre (Fundação Getúlio Vargas)	50	41,7%	12	10,0%
Laminated steel production - 1st lag	IAB (Instituto Aço Brasil)	37	30,8%	0	0,0%
Usage of industrial capacity level (NUCI)	FGV Ibre (Fundação Getúlio Vargas)	32	26,7%	28	23,3%
Oil production - 1st lag	Agência Nacional de Petróleo	28	23,3%	12	10,0%
Manufacturing industries sales indicator - 1st lag	FIESP (Federação das Indústrias do Estado de São Paulo)	26	21,7%	0	0,0%
Real productivity for employees in the manufacturing industries sector - 2nd lag	CNI (Confederação Nacional da Indústria)	25	20,8%	0	0,0%
Real expanded wage bill (PNAD + Tesouro Nacional)	IBGE (Brazilian Institute of Geography and Statistics) + Tesouro Nacional	19	15,8%	0	0,0%
Mobility indicator for heavy vehicles (buses and trucks)	ABCR (Associação Brasileira de Captadores de Recursos)	17	14,2%	12	10,0%
Manufacturing industries market sentiment indicator - 1st lag	FIESP (Federação das Indústrias do Estado de São Paulo)	16	13,3%	0	0,0%
Retail prices index - 1st lag	FGV Ibre (Fundação Getúlio Vargas)	15	12,5%	4	3,3%
Business sentiment - 2nd lag	FGV Ibre (Fundação Getúlio Vargas)	14	11,7%	9	7,5%
Producer's price index for the manufacturing industries (IPP) - 1st lag	IBGE (Brazilian Institute of Geography and Statistics)	12	10,0%	5	4,2%
Industry sector Sentiment	FGV Ibre (Fundação Getúlio Vargas)	11	9,2%	6	5,0%
Real exchange rate index	Brazilian Central Bank	10	8,3%	0	0,0%
Industry sector Sentiment - 2nd lag	FGV Ibre (Fundação Getúlio Vargas)	10	8,3%	4	3,3%
Manufacturing industries sales indicator - 2nd lag	FIESP (Federação das Indústrias do Estado de São Paulo)	9	7,5%	0	0,0%
Manufacturing industries inventory indicator	FIESP (Federação das Indústrias do Estado de São Paulo)	8	6,7%	0	0,0%
Employment indicator for the manufacturing industries sector	FIESP (Federação das Indústrias do Estado de São Paulo)	8	6,7%	0	0,0%
Consumer's price index (IPCA)	IBGE (Brazilian Institute of Geography and Statistics)	7	5,8%	0	0,0%
Real interest rate (Selic) - 1st lag	Brazilian Central Bank	7	5,8%	11	9,2%
Employed population in the industry sector - 2nd lag	CAGED - Ministério do Trabalho	7	5,8%	0	0,0%
Energy consumption for the industry sector	EPE (Empresa de Pesquisa Energética)	6	5,0%	0	0,0%
Non-durable goods imports in quantity - 1st lag	Funcex (Fundação Centro de Estudos do Comércio Exterior)	6	5,0%	0	0,0%
Real wage bill for employees in the manufacturing industries sector - 2nd lag	CNI (Confederação Nacional da Indústria)	6	5,0%	0	0,0%
Retail sector sentiment	FGV Ibre (Fundação Getúlio Vargas)	5	4,2%	0	0,0%
Employed population in the mining & quarrying industries sector	CAGED - Ministério do Trabalho	5	4,2%	5	4,2%
Manufacturing industries activity indicator - 1st lag	FIESP (Federação das Indústrias do Estado de São Paulo)	5	4,2%	0	0,0%
Economic uncertainty index (IIE-Br) - 2nd lag	FGV Ibre (Fundação Getúlio Vargas)	4	3,3%	0	0,0%
Manufacturing industries market sentiment indicator - 2nd lag	FIESP (Federação das Indústrias do Estado de São Paulo)	4	3,3%	0	0,0%
Corrugated cardboard production	Empapel (Associação Brasileira de Embalagens em Papel)	3	2,5%	0	0,0%
Usage of industrial capacity level (NUCI) - 1st lag	FGV Ibre (Fundação Getúlio Vargas)	3	2,5%	0	0,0%
Gas production - 2nd lag	Agência Nacional de Petróleo	3	2,5%	0	0,0%
Manufacturing industries activity indicator - 2nd lag	FIESP (Federação das Indústrias do Estado de São Paulo)	3	2,5%	0	0,0%
Manufacturing industries inventory indicator - 2nd lag	FIESP (Federação das Indústrias do Estado de São Paulo)	3	2,5%	0	0,0%
Real interest rate (Selic)	Brazilian Central Bank	2	1,7%	2	1,7%
Durable goods imports in quantity	Funcex (Fundação Centro de Estudos do Comércio Exterior)	2	1,7%	0	0,0%
Manufacturing industries investment indicator	FIESP (Federação das Indústrias do Estado de São Paulo)	2	1,7%	0	0,0%
Retail prices index	FGV Ibre (Fundação Getúlio Vargas)	2	1,7%	0	0,0%
Economic uncertainty index (IIE-Br) - 1st lag	FGV Ibre (Fundação Getúlio Vargas)	2	1,7%	0	0,0%
Employed population in the mining & quarrying industries sector - 2nd lag	CAGED - Ministério do Trabalho	2	1,7%	0	0,0%
Total imports in quantity - 2nd lag	Funcex (Fundação Centro de Estudos do Comércio Exterior)	2	1,7%	0	0,0%
Employment indicator for the manufacturing industries sector - 1st lag	CNI (Confederação Nacional da Indústria)	0	0,0%	7	5,8%

*LASSO and adaptive-LASSO proportions are calculated by (number of times selected by model / total windows number)

Appendix D – Proportion that each variable was selected by shrinkage models: Mining & quarrying industries index

Mining & Quarrying Industries Index					
Variable Description	**Source**	**Number of times selected by the LASSO**	**LASSO Proportion**	**Number of times selected by the adaptive-LASSO**	**adaptive-LASSO Proportion**
Oil production	Agência Nacional de Petróleo	120	100,0%	120	100,0%
Gas production	Agência Nacional de Petróleo	115	95,8%	66	55,0%
Durable goods imports in quantity - 1st lag	Funcex (Fundação Centro de Estudos do Comércio Exterior)	107	89,2%	85	70,8%
Intermediate goods imports in quantity - 1st lag	Funcex (Fundação Centro de Estudos do Comércio Exterior)	85	70,8%	26	21,7%
Unemployment rate - interpolated series with PNAD and PME - 2nd lag	IBGE (Brazilian Institute of Geography and Statistics)	83	69,2%	83	69,2%
Volume levels in water reservoirs	ONS (Operador Nacional do Sistema Elétrico)	65	54,2%	39	32,5%
Manufacturing industries sales	CNI (Confederação Nacional da Indústria)	60	50,0%	58	48,3%
Industry sector Sentiment - 2nd lag	FGV Ibre (Fundação Getúlio Vargas)	47	39,2%	19	15,8%
Real expanded wage bill (PNAD + Tesouro Nacional) - 2nd lag	IBGE (Brazilian Institute of Geography and Statistics) + Tesouro Nacional	40	33,3%	36	30,0%
Steel production - 1st lag	IAB (Instituto Aço Brasil)	39	32,5%	35	29,2%
Energy consumption for the industry sector	EPE (Empresa de Pesquisa Energética)	32	26,7%	30	25,0%
Hours worked indicator for the manufacturing industries sector - 2nd lag	CNI (Confederação Nacional da Indústria)	32	26,7%	0	0,0%
Total vehicles production - 1st lag	ANFAVEA (Associação Nacional de Frabricantes de Veículos Automores)	30	25,0%	13	10,8%
Business sentiment - 2nd lag	FGV Ibre (Fundação Getúlio Vargas)	27	22,5%	23	19,2%
Oil production - 2nd lag	Agência Nacional de Petróleo	20	16,7%	25	20,8%
Real wage bill for employees in the manufacturing industries sector - 2nd lag	CNI (Confederação Nacional da Indústria)	16	13,3%	0	0,0%
Consumer's price index (IPCA)	IBGE (Brazilian Institute of Geography and Statistics)	13	10,8%	13	10,8%
Manufacturing industries inventory indicator	FIESP (Federação das Indústrias do Estado de São Paulo)	12	10,0%	3	2,5%
Laminated steel production - 1st lag	IAB (Instituto Aço Brasil)	12	10,0%	8	6,7%
Employed population in the mining & quarrying industries sector - 1st lag	CAGED - Ministério do Trabalho	12	10,0%	3	2,5%
Capital goods imports in quantity - 1st lag	Funcex (Fundação Centro de Estudos do Comércio Exterior)	12	10,0%	0	0,0%
General industry index (PIM-PF) - 2nd lag	IBGE (Brazilian Institute of Geography and Statistics)	11	9,2%	2	1,7%
Employed population in the mining & quarrying industries sector	CAGED - Ministério do Trabalho	10	8,3%	2	1,7%
Mobility indicator for heavy vehicles (buses and trucks) - 1st lag	ABCR (Associação Brasileira de Captadores de Recursos)	10	8,3%	10	8,3%
Employment indicator for the manufacturing industries sector - 1st lag	CNI (Confederação Nacional da Indústria)	9	7,5%	7	5,8%
Hours worked indicator for the manufacturing industries sector	CNI (Confederação Nacional da Indústria)	8	6,7%	8	6,7%
Vehicles sales - 1st lag	Fenabrave (Federação Nacional da Distribuição de Veículos Automotores)	7	5,8%	0	0,0%
Real productivity for employees in the manufacturing industries sector - 2nd lag	CNI (Confederação Nacional da Indústria)	7	5,8%	0	0,0%
General industry index (PIM-PF) - 1st lag	IBGE (Brazilian Institute of Geography and Statistics)	6	5,0%	2	1,7%
Consumer's price index (IPCA) - 1st lag	IBGE (Brazilian Institute of Geography and Statistics)	4	3,3%	4	3,3%
Consumer sentiment - 1st lag	FGV Ibre (Fundação Getúlio Vargas)	3	2,5%	2	1,7%
Volume levels in water reservoirs - 1st lag	ONS (Operador Nacional do Sistema Elétrico)	3	2,5%	0	0,0%
Gas production - 2nd lag	Agência Nacional de Petróleo	3	2,5%	3	2,5%
Steel production	IAB (Instituto Aço Brasil)	2	1,7%	0	0,0%
Employed population in the manufacturing industries sector	CAGED - Ministério do Trabalho	2	1,7%	0	0,0%
Usage of industrial capacity level (NUCI)	CNI (Confederação Nacional da Indústria)	2	1,7%	0	0,0%
Gas production - 1st lag	Agência Nacional de Petróleo	2	1,7%	0	0,0%
General energy consumption - 2nd lag	ONS (Operador Nacional do Sistema Elétrico)	2	1,7%	3	2,5%
Capital goods imports in quantity - 2nd lag	Funcex (Fundação Centro de Estudos do Comércio Exterior)	2	1,7%	0	0,0%

*LASSO and adaptive-LASSO proportions are calculated by (number of times selected by model / total windows number)