The Effects of Estimation Sample Size in Forecast Performance: The Case of Brazilian Industrial Production Index

Jordano Vieira Rocha*
Pedro L. Valls Pereira†

August 10, 2019

Abstract

Brazilian Industrial Production Index undergoes different methodological updates and periods of high inflation over time, which prompts researchers to avoid using too long industrial production series. We analyze how performance of different models in forecasting the Brazilian Industrial Production Index one-step ahead is influenced by the use of samples of different lengths. Relative performance of these models is also assessed. Results show that most models benefit from expanding the estimation sample beginning at least up to 1993:12. Autometrics lag selection with impulse dummy saturation forecasting performance is improved almost monotonically with sample size. For estimation starting in January 1975 and 1985, predictions from Autometrics with impulse dummy saturation and the average of forecasts are statistically more accurate than those from the benchmark AR model. However, the average of predictions performs better in the first half of the forecast horizon and Autometrics performs better in the second half.

Keywords: industrial production index, nonlinear methods, lag selection, dummy saturation, forecasting.
JEL Classification: C22, C52, C53.

1 Introduction

Macroeconomic data are an indispensable input to policy makers and private agents in capturing the economic situation of an entire country or region. They are intended to translate a complex set of heterogeneous variables into informative simple statistics. Since collecting information on all economic agents is unfeasible, macroeconomic data are usually originated from periodic surveys and a limited number of administrative records.

*IBGE
†Sao Paulo School of Economics - FGV and CEQEF -FGV

Given technological advances and changes in the structure of the economy, the methodology of such surveys are updated over time. Therefore, using long macroeconomic series may be problematic, specially in forecasting, since recent and old data may be of different qualities. Additionally, old data may contain information that is no longer relevant in forecasting future values (this is a characteristic not only of macroeconomic data, but of socioeconomic time series in general).

To investigate the effects of sample size for forecast performance, we analyze the Brazilian industrial production series. The main index of industrial production in Brazil is the Monthly Industrial Survey - Physical Production (PIM-PF), produced by the Brazilian Institute of Geography and Statistics (IBGE). This is one of the longest monthly macroeconomic time series for Brazil, with initial observations in 1975. It is widely used as a business cycle indicator, since the GDP is disclosed quarterly, while the PIM-PF is a monthly index.

Although, in its essence, the index is practically the same over the years — a weighted average of relative quantities, developed over a Laspeyres *quantum* index — this series has undergone methodological benchmark revisions over time, which consisted, basically, in updating the weights for products and activities, the surveyed set of products and respondents, the level of regional disaggregation, and the industry activity and product classifications, following revisions in international standards.

The last benchmark revision was made in March 2014, which produced revised values up to January 2002. This current series may be insufficiently large, particularly when a fraction of the sample is used for forecast evaluation. On the other hand, as stated before, larger series may contain irrelevant information, suffer from the different methodologies and from the presence of outliers, which may harm parameter estimation. Furthermore, since weights for activities composing the index are fixed for each methodological benchmark, periods of high inflation tend to have distorting influence over the series, possibly generating additional noise.

One objective of this chapter is to study the effect of the sample size selection in the forecast performance. Given the intrinsic volatile behavior of industrial production and the structural breaks that may arise also due to changes in the survey methodology, a second objective is to evaluate different linear and nonlinear models, comparing relative performance between them using different sample sizes. We consider only univariate models to rule out the influence of other variables in the analysis.

It is a standard assumption that economic time series are realizations of stochastic processes, represented by a Data Generating Process (DGP)[1]. The DGP (or at least a close approximation) is assumed to be known and any abnormal shift, departing from the standard linear dynamics, in the series must be captured by its form.

Nonlinear models use different techniques to capture these kind of break in a series over time. Some of the most known are composed by a set of different linear models that are alternated in time according to some rule of transition. It is the case of the Markov Switching Autoregressive Model (MSAR), which have a transition rule based on a probabilistic struc-

[1]Some authors consider that the DGP represents the joint density, not necessarily with constant parameters, of all variables characterizing an economy. Following the Theory of Reduction (Hendry, 1995), a subset of these variables can be represented by a Local Data Generating Process (LDGP). To properly parametrize any time-varying effect, an econometric model, usually different from the LDGP, must be a function with constant parameters. See, for example, Bontemps and Mizon (2003).

ture, and the Smooth Transition Autoregressive Model (STAR), with a changing rule based on a transition function that depends on the value of the variable of interest. STAR transitions tend to be more abrupt than those of MSAR, because, in the latter, the transitions are ruled by a Markov Process. This process is characterized by a transition matrix which give the probabilities of switching between regimes, and this characterizes the smoothness of transitions. Early examples of MSAR and STAR models application can be found in Engel (1994) and Teräsvirta and Anderson (1992), analyzing the the US dollar exchange rate and the quarter industrial production indexes from several countries, respectively.

Another way to capture nonlinearities, without explicitly modelling their behavior, is to automatically detect outliers and level shifts, using impulse and step dummies in the model. Santos et al. (2008) propose this approach, selecting the relevant dummies based on the Autometrics algorithm of variable selection (Doornik, 2009). Correctly modeled impulse and step dummies can improve parameter estimation and detect a change in the series' mean, modifying the conditional mean and avoiding repeatedly wrong forecasts, which would tend to be biased towards the previous unconditional mean. Hendry and Mizon (2011), for example, study the food expenditure in United States using this technique.

However, econometric models may still fail in the presence of unpredictable structural breaks (or changes in the underlying stochastic process). Such fact have been gaining more attention specially after the subprime financial crisis. Abrupt changes in the dynamic of several economic and financial series have caused a phenomenon called forecast failure (*e.g.*, Hendry, 2012), *i.e.*, persistent failures in the forecasts after a certain period.

Defining robust forecasting mechanisms, even in the presence of structural breaks of unknown form, becomes important in a world with abrupt, unpredictable events. Hendry (2006) suggests that to predict using the difference of a series (even a stationary one) would make its forecast robust. In the moment of a structural break occurrence, the forecast would be biased, since this change would be unpredictable by definition. In the following periods, however, the forecast would become unbiased, although at the cost of a higher forecast error variance in times of economic stability.

Besides the aforementioned methods, we analyze forecasts from models selecting lags of the dependent variable: the benchmark Autoregressive Model of order p (chosen by the Bayesian Information Criterion), against which all other forecasts will be compared, the Autometrics algorithm without any dummy saturation, and the Least Absolute Shrinkage and Selection Operator (LASSO) and two of its variants, the Adaptive LASSO (AdaLASSO) and the Weighted Lag Adaptive LASSO (WLAdaLASSO), well known regularization and variable selection methods (Tibshirani, 1996; Zou, 2006; Konzen and Ziegelmann, 2016). The simple average of forecasts from all models are also computed.

We perform Diebold and Mariano (1995) test to assess the relative forecast performance using different sample sizes in relation to the current smaller sample (starting in January 2002) for each method. We also test, for each sample, if these methods' forecasts are more accurate than the benchmark AR(p) model within each sample. Forecasts are made one-step ahead, re-estimating and re-specifying the models at each step. Results show that most models benefit from enlarging the estimation sample at least to 1993:12 and that a rolling window forecasting scheme with this same initial sample performs worse than its expanding window analogous, pointing that information in that period may remain relevant in forecasting future values.

In addition to this introduction, this chapter is structured as follows. Section 2 contains an overview of the PIM-PF and the nature of its methodological revisions, Section 3 provides a theoretical review of the considered models, Section 4 details the data used and the settings of the forecasting exercise performed, Section 5 presents the results, Section 6 contains the concluding remarks.

2 An Overview of the Monthly Industrial Survey and its Methodological Revisions

The Monthly Industrial Survey is part of an integrated system of business surveys[2], consisting of annual (or structural) and monthly (or conjunctural) surveys — the annual and monthly surveys of Mining and Manufacturing (PIA and PIM-PF), of Trade (PAC and PMC), and of Services (PAS and PMS) and the annual survey of Construction (PAIC). There are, additionally, satellite (or special) surveys, with specific themes, for example, the Innovation Survey (formerly Industrial Survey of Technological Innovation, or PINTEC).

This system is constructed over a unified business register, named Central Business Register (CEMPRE), which was built in the major revision of business surveys carried out in the mid 1990s. The CEMPRE consists of data and industry classifications of all formal enterprises and their local units, *i.e.*, all units registered in the Internal Revenue Service, identified by a unique National Register of Juridical Person (CNPJ) number. It is used to construct the samples of the annual surveys, which, in turn, serve as a basis to calculate the weights and extract the samples of monthly surveys.

Regarding the PIM-PF, the selection of products and respondents of the survey is based on a panel of units representing a percentage of the Industrial Transformation Value (VTI) or the Gross Value of Industrial Production (VBPI) on a base year. In the current benchmark, for example, in each region, only those activities representing 80% of the VTI are surveyed. Of those activities, only the products representing 80% of the VBPI are selected. The local units responsible for 70% of production for each product are surveyed. These fractions are relative to the base year of 2010.

The methodological revisions must be performed periodically to update the set of products and respondents surveyed, the activities analyzed, the industrial classifications, the weighting structure, based on more timely information from economic censuses or annual surveys, and the disagreggation of the index in terms of geographical regions. Seasonal adjustment procedures are also updated as new methodological techniques are developed and suggested by international standards.

The index *per se* is basically the same, simply a weighted average of relative quantities, developed over a Laspeyres *quantum* index:

[2]This section is based on IBGE's methodological reports (IBGE, 1991, 1996, 2004, 2015b), on information from PIM-PF monthly publications and IBGE's metadata website — available, respectively, at `https://metadados.ibge.gov.br` and at `https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=7228` — and on Góes (2005).

$$I_t^0 = \sum_{i=1}^{n} w_i^0 \left(\frac{q_i^t}{q_i^0} \right), \tag{1}$$

where I_t^0 is the index in time t relative to the base year 0, $w_i^0 = (VTI_i^0 / \sum_{i=1}^{n} VTI_i^0)$, and q_i^t/q_i^0 is the quantity[3] of product i in time t relative to the base period 0 (*e.g.*, 2012 in the current series)[4].

The first statistics of industrial physical production in Brazil were made in the early 1970s. Initially, the survey included 110 products and around 1000 respondents, using weights from the then recently developed Annual Industrial Survey of 1968. The first reformulation was made with the 1970 economic census publication, available in 1975, which included 600 products and 2500 respondents.

The following update in the series used weights from the 1978 PIA and information from the 1980 Industrial Census, including 736 products and 5000 respondents. With the 1985 Industrial Census (the last economic census to be carried out), the new PIM-PF structure comprised 944 products and 6200 respondents.

The economic statistics system described in the beginning of this section started to be implemented in the mid 1990s. Before this framework, the Industrial Census worked both as a register and provided information for the weighting structure and selection of product and respondent samples, with additional information from the Annual Surveys. The register based on the censuses was rapidly out of date, due to the lag which the census results were available, aggravated by the high rates of birth and deaths of enterprises. The CEMPRE, which is updated each year with administrative data and results from the most recent surveys, was developed to overcome those difficulties. Other important technical changes were the use of a new industrial classification system of activities, the National Classification of Economic Activities (CNAE — IBGE, 2003), and of products (PRODLIST-Indústria), and the redefinition of the statistical unit of investigation, from establishment to local unit[5].

These changes were incorporated in the PIM-PF only in 2004, covering data from January 2002, which used information from the 1998-2000 editions of the PIA, including 830 products and 3700 local units. The last methodological revision, made in 2015, covering data starting from January 2012, implemented the changes to make the survey compatible with second version of CNAE (IBGE, 2015a). It selected 944 products and 7800 local units.

---

[3]The relative quantities are actually calculated as $\frac{q_i^t}{q_i^0} = \left(\frac{q_i^1}{q_i^0}\right)^* . \left(\frac{q_i^2}{q_i^1}\right)^* \ldots \left(\frac{q_i^t}{q_i^{t-1}}\right)^*$, where $\left(\frac{q_i^t}{q_i^{t-1}}\right)^*$ is the relative quantity between months t and $t-1$ of product i, maintaining the same intersection of respondent and product panels from months $t-1$ and t. Note that $\left(\frac{q_i^t}{q_i^{t-1}}\right)^* \neq \frac{q_i^t}{q_i^{t-1}}$, since panels may be different in t and $t-1$. This adjustment is done to account for this fact, ensuring that relatives are calculated in compatible panels.

[4]In some editions, the weights are calculated over the Value Added instead of the Industrial Transformation Value, *i.e.*, $w_i^0 = (VA_i^0 / \sum_{i=1}^{n} VA_i^0)$, such as in the series that use weights from the 1985 Industrial Census.

[5]The establishment is defined as an enterprise or part of enterprise that, in a given location, independently engages predominantly in one economic activity. Thus one enterprise may have multiple establishments. The local unit is defined as a physical space where one or more economic activities are developed, corresponding to a unique CNPJ (14 digits) number. The use of establishment as statistical unit revealed problematic, since respondent had difficulties filling out questionnaires and it brought additional complexity in the collecting and processing of information. Furthermore, the harmonization between statistical and legal definition, together with a common industrial classification, made it easier to use administrative data in the register and surveys.

2.1 The Chaining of Series from Different Methodological Benchmarks

Each methodological update also changes the base year, *i.e.*, the year in which the average of the index is 100. This results in multiple series with different levels. For the sake of the series continuity, if we want to obtain a longer time length containing all available data, the entire period must be expressed in the same base.

IBGE's last technical report (IBGE, 2015b) describe how to link two consecutive series with different base periods. Although this is done to the series of all economic activities that compose the index, we made this procedure only to the General Index, since we do not work with disaggregated data.

Suppose we want to build a series $\{I_t\}_{t=1}^{K}$ of an index. Let $b_1 = (k_{b_1}^{Jan}, k_{b_1}^{Feb}, \ldots, k_{b_1}^{Dec})$ and $b_2 = (k_{b_2}^{Jan}, k_{b_2}^{Feb}, \ldots, k_{b_2}^{Dec})$ be the sets of months composing two different base years. Suppose that we have two different monthly series of the index expressed in the base years b_1 and b_2, $\{I_t^{b_1}\}_{t=1}^{k_1}$ and $\{I_t^{b_2}\}_{t=k_2}^{K}$, $k_2 < k_1$, and that $b_2 \subseteq (k_2, k_2+1, \ldots, k_1)$. This states that, to make the link between the series possible, both series must have observations in a given base year. The definition of base year implies that

$$\frac{\sum_{t\in b_1} I_t^{b_1}}{12} \equiv 100 \text{ and } \frac{\sum_{t\in b_2} I_t^{b_2}}{12} \equiv 100. \tag{2}$$

The index of base b_1 expressed in base b_2 is given by

$$I_t^{b_1|b_2} = I_t^{b_1}.\frac{\sum_{t\in b_2} I_t^{b_2}}{\sum_{t\in b_2} I_t^{b_1}}. \tag{3}$$

Note that

$$\frac{\sum_{t\in b_2} I_t^{b_1|b_2}}{12} = \frac{\sum_{t\in b_2} I_t^{b_1}}{12}.\left(\frac{\sum_{t\in b_2} I_t^{b_2}}{\sum_{t\in b_2} I_t^{b_1}}\right) = 100. \tag{4}$$

The entire series can be, thus, expressed in terms of b_2:

$$\{I_t\}_{t=1}^{K} \equiv \left(I_1^{b_1|b_2}, I_2^{b_1|b_2}, \ldots, I_{k_2-1}^{b_1|b_2}, I_{k_2}^{b_2}, I_{k_2+1}^{b_2}, \ldots, I_K^{b_2}\right). \tag{5}$$

3 Theoretical Review

This section briefly describes the models used in this paper to forecast the Brazilian Industrial Production Index series. In following sections, forecast performance of these techniques will be compared to the benchmark Autoregressive Model of order p. We consider nonlinear methods (the Markov Switching and Smooth Transition Autoregressive Models), an automatic lag selection and outlier detection method (Autometrics), a naive method robust to structural breaks (Double Difference Device), and Lasso-type penalty methods (LASSO, Adaptive LASSO and Weighted Lag Adaptive LASSO).

3.1 Markov Switching Autoregressive Model

This technique supposes that the DGP consists in different autoregressive models which alternate between them according to a process represented by a discrete latent variable $s_t = \{1, 2, \ldots, N\}$, in which each value represents a different regime. An additional assumption is that s_t follows a Markov process, *i.e.*, the probability of the realization of state j in the current period depends only of the realization of the state from the immediately previous period:

$$P(s_t = j|s_{t-1} = i, s_{t-2} = k, \ldots) = P(s_t = j|s_{t-1} = i) = p_{ij}. \tag{6}$$

The set of probabilities of transition p_{ij} can be represented, thus, by a transition matrix P:

$$P = \begin{bmatrix} p_{11} & p_{21} & \cdots & p_{N1} \\ p_{12} & p_{22} & \cdots & p_{N2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1N} & p_{2N} & \cdots & p_{NN} \end{bmatrix}. \tag{7}$$

Suppose, for simplicity and without loss of generality, an AR(1) process in which the intercept (c_{s_t}) and autoregressive values (ϕ_{s_t}) change according to each regime. We can write this process as:

$$y_t = c_{s_t} + \phi_{s_t} y_{t-1} + \varepsilon_t, \tag{8}$$

where we assume that $\varepsilon_t \sim NI(0, \sigma^2_{s_t})$, *i.e.*, the variance of the error term also depends on s_t.

Observe that, even if we know all the involved parameters, we cannot be sure of which regime the process is in each period. Let θ be the vector containing all the parameters in the model. We can infer the state by the probability that the process is in regime j conditional to the realization of the series values up to t, denoted $P(s_t = j|y_t, \theta)$. Following the law of conditional probabilities:

$$P(s_t = j|y_t, \theta) = \frac{p(y_t, s_t = j, \theta)}{f(y_t, \theta)} = \frac{\pi_j f(y_t|s_t = j, \theta)}{f(y_t, \theta)}. \tag{9}$$

In the expression above, $\pi_j = P(s_t = j, \theta)$ is the unconditional probability of s_t assuming the value j, $f(y_t|s_t = j, \theta)$ is the conditional density, assumed normal, that is:

$$f(y_t|s_t = j, \theta) = \frac{1}{\sqrt{2\pi}\sigma_{s_t}} exp\Big\{\frac{-(y_t - c_{s_t} - \phi_{s_t} y_{t-1})^2}{2\sigma^2_{s_t}}\Big\} \tag{10}$$

and the unconditional density is given by:

$$f(y_t, \theta) = \sum_{j=1}^{N} p(y_t, s_t = j, \theta) = \sum_{j=1}^{N} \frac{\pi_j}{\sqrt{2\pi}\sigma_{s_t}} exp\Big\{\frac{-(y_t - c_{s_t} - \phi_{s_t} y_{t-1})^2}{2\sigma^2_{s_t}}\Big\}. \tag{11}$$

Stacking the conditional densities $f(y_t|s_t = j, \theta)$, the conditional probabilities $P(s_t = j|y_t, \theta)$ and the forecasts $P(s_{t+1} = j|y_t, \theta)$ in vectors $(N \times 1)$, respectively, η_t, $\hat{\xi}_{t|t}$ e $\hat{\xi}_{t+1|t}$, *i.e.*,

$$\eta_t = \begin{bmatrix} f(y_t|s_t = 1, y_{t-1}; \theta) \\ f(y_t|s_t = 2, y_{t-1}; \theta) \\ \vdots \\ f(y_t|s_t = N, y_{t-1}; \theta) \end{bmatrix} \tag{12}$$

$$\hat{\xi}_{t|t} = \begin{bmatrix} P(s_t = 1|y_t, \theta) \\ P(s_t = 2|y_t, \theta) \\ \vdots \\ P(s_t = N|y_t, \theta) \end{bmatrix} \tag{13}$$

$$\hat{\xi}_{t+1|t} = \begin{bmatrix} P(s_{t+1} = 1|y_t, \theta) \\ P(s_{t+1} = 2|y_t, \theta) \\ \vdots \\ P(s_{t+1} = N|y_t, \theta) \end{bmatrix}. \tag{14}$$

Hamilton (1994) shows that the optimal inference for the regimes is given by the following recursion:

$$\hat{\xi}_{t|t} = \frac{(\hat{\xi}_{t|t-1} \odot \eta_t)}{1'(\hat{\xi}_{t|t-1} \odot \eta_t)} \tag{15}$$

$$\hat{\xi}_{t+1|t} = P\hat{\xi}_{t|t}, \tag{16}$$

where P is defined by (7), $\odot$ represents element-by-element multiplication and 1 is a $(N \times 1)$ vector of 1's.

The likelihood maximization problem is given, thus, by:

$$max\ \mathcal{L}(\theta) = \sum_{t=1}^{T} f(y_t|y_{t-1}, \theta) = \sum_{t=1}^{T} 1'(\hat{\xi}_{t|t-1} \odot \eta_t) \tag{17}$$

$$s.t.\ (p_{i1} + p_{i2} + \cdots + p_{iN}) = 1, p_{ij} \geq 0, \forall j, i. \tag{18}$$

Expression (15) is similar to (9), and (17) similar to (11).

Given the values of θ_0 and $\hat{\xi}_{1|0}$ (*e.g.* $\hat{\xi}_{1|0} = \rho$, with $\rho = 1/N$), the recursion in (15)-(16) returns the values of $\hat{\xi}_{t|t}$. Note that, for each t, this method utilizes only the realized values in the current and in previous periods $(t, t-1, t-2, \dots)$ in order to estimate $P(s_t = j|y_{t-1}, \theta)$. However, we possess additional information to increase this estimate's accuracy, namely, the remainder of the sample $(t+1, t+2, \dots, T)$. For this, we have to use a smoothed inference mechanism.

Let us denote, generalizing the previous notation, the probability of being in a regime conditional to the realized values from the series and to the parameters $P(s_t = j|y_\tau, \theta)$. For $t < \tau$, we will have a smoothed inference of regime t. Kim (1994) develops a recursion analogous to (16):

$$\hat{\xi}_{t|T} = \hat{\xi}_{t|t} \odot \left\{ P' \left[\hat{\xi}_{t+1|T} (\div) \hat{\xi}_{t+1|t} \right] \right\}, \tag{19}$$

where $(\div)$ means term-by-term division.

The iteration algorithm starts in $\hat{\xi}_{T|T}$, which, in turn, is calculated from (15)-(16), passing through $\hat{\xi}_{T-1|T}, \hat{\xi}_{T-2|T}, \ldots, \hat{\xi}_{1|T}$. Once having $\hat{\xi}_{t|t}$ and $\hat{\xi}_{t+1|t}$ known, the full vector of maximum likelihood estimates $\hat{\theta}$ is obtained solving (17)-(18), proceeding recursively until there is convergence to a θ^*. Restricting $p_{ij} \geq 0$ and $p_{i1} + p_{i2} + \cdots + p_{iN} = 1$ for all i, j, and supposing $\hat{\xi_{1|0}}$ is fixed and not related to the other parameters, Hamilton (1990) demonstrate the the transition probabilities estimators p_{ij} are given by:

$$\hat{p}_{ij} = \frac{\sum_{t=2}^{T} P(s_{t=j}, s_{t-1} = i | y_T, \hat{\theta})}{\sum_{t=2}^{T} P(s_{t-1} = i | y_T, \hat{\theta})}, \tag{20}$$

that is the sum of the probabilities of regime i being followed by regime j divided by the sum of the probabilities of being in regime i.

Finally, to calculate the optimal forecast one-step ahead, suppose we know which will be the regime in period $t+1$, so the prediction will be obtained by $h_{jt} = \mathbb{E}(y_{t+1} | s_{t+1} = j, \mathcal{Y}_t; \theta) = \int y_{t+1} f(y_{t+1} | s_{t+1} = j, \mathcal{Y}_t; \theta) dy_{t+1}$, where $\mathcal{Y}_t$ is a vector containing all y observations through date t. The unconditional expectation will be, thus, given by:

$$\begin{aligned}
\mathbb{E}(y_{t+1} | x_{t+1}, \mathcal{Y}_t; \theta) &= \int y_{t+1} f(y_{t+1} | x_{t+1}, \mathcal{Y}_t; \theta) dy_{t+1} \\
&= \int y_{t+1} \Big\{ \sum_{j=1}^{N} p(y_{t+1}, s_{t+1} = j | x_{t+1}, \mathcal{Y}_t; \theta) \Big\} dy_{t+1} \\
&= \int y_{t+1} \Big\{ \sum_{j=1}^{N} [f(y_{t+1} | s_{t+1} = j, x_{t+1}, \mathcal{Y}_t; \theta) P(s_{t+1} = j | x_{t+1}, \mathcal{Y}_t; \theta)] \Big\} dy_{t+1} \\
&= \sum_{j=1}^{N} P(s_{t+1} = j | x_{t+1}, \mathcal{Y}_t; \theta) \int y_{t+1} f(y_{t+1} | s_{t+1} = j, x_{t+1}, \mathcal{Y}_t; \theta) dy_{t+1} \\
&= \sum_{j=1}^{N} P(s_{t+1} = j | \mathcal{Y}_t; \theta) \mathbb{E}(y_{t+1} | s_{t+1} = j, x_{t+1}, \mathcal{Y}_t; \theta).
\end{aligned} \tag{21}$$

If we stack h_{jt} in a vector $(N \times 1)$ h_t, then:

$$\mathbb{E}(y_{t+1} | \mathcal{Y}_t; \theta) = h_t' \hat{\xi}_{t+1|t}. \tag{22}$$

Although the Markov Chain admits a linear representation, the optimal forecast one-step ahead is a nonlinear function of the of the observed variables, for $\xi_{t|t}$ depends on $\mathcal{Y}_t$ in a nonlinear way. As pointed by Hamilton (1994), this implies that, if, inside a regime, an outlier (an observation little likely to be generated in this regime) is observed, the framework analyzed here has high probability of inferring that a change of state has occurred.

3.2 Smooth Transition Autoregressive Model

As it is the case with MSAR models, the STAR models consist in two or more regimes following different linear autoregressive processes. The transition dynamics from one regime to another is, though, different. The change between states is given by a transition function as described below.

Suppose a stochastic process represented by the following DGP:

$$y_t = \pi_{10} + \pi_1' z_t + (\pi_{20} + \pi_2' z_t) F(y_{t-d}) + \epsilon_t. \tag{23}$$

$F(y_{t-d})$ is called the transition function, it depends on the model's dependent variable, y_t, lagged by the parameter d, that will be estimated. The vector $z_t = (y_{t-1}, y_{t-2}, \ldots, y_{t-p})$ contains the lagged series, (π_1, π_2) are parameter vectors with correspondent dimensions and (π_{10}, π_{20}) are scalars. The error term is assumed normal, independent and identically distributed, *i.e.*, $\epsilon_t \sim NI(0, \sigma^2)$. It is important to stress that ϵ_t is symmetric, so that rejections of the null hypothesis from linearity tests would come from the model parametrization, and not from the error term.

The F function is limited between 0 and 1. In case it is an indicator function $F(y_{t-d}) = I(y_{t-d} > c)$[6], the model is denoted *Self Exciting Threshold Autoregressive* (SETAR), in which, when y_{t-d} is greater than the threshold c, the regime is changed instantaneously. To establish a smooth transition between the regimes, we can use some continuous function between 0 and 1, so that the parameters of (23) vary, as function of y_{t-d}, between π_{10} and $(\pi_{10} + \pi_{20})$ for the intercept, and between π_1 and $(\pi_1 + \pi_2)$, for the autoregressive parameters. These values correspond to the two extreme regimes ($F = 0$ and $F = 1$, respectively).

There are two functions used predominantly in the literature. The first is the logistic function:

$$F(y_{t-d}) = (1 + exp[-\gamma(y_{t-d} - c)])^{-1}. \tag{24}$$

When the model uses the function above, it is called *Logistic Smooth Transition Autoregressive* (LSTAR). The inclination parameter γ measures the speed of transition from one regime to another, and the location parameter c is the threshold. Besides, when $\gamma \to \infty$, the model approximates a SETAR model.

LSTAR model has been used to model macroeconomic series with asymmetric behavior. Öcal and Osborn (2000) and Teräsvirta and Anderson (1992) use it to model industrial production series and Skalin and Teräsvirta (2002) for unemployment series.

The second function is the exponential function:

$$F(y_{t-d}) = 1 - exp(-\gamma(y_{t-d} - c)^2). \tag{25}$$

Using this function, we denote the model Exponential Smooth Transition Autoregressive (ESTAR). The interpretation given to the parameters γ and c is the same as in LSTAR. When $\gamma \to \infty$, however, F tends to $I(y_{t-d} = c)$.

[6]We will use, in this work, only two regimes, corresponding in theory to periods of expansion and periods of recession, but we could generalize for more regimes. In this case, the model would have, in the indicator function example, the form $y_t = \sum_{j=1}^{r} \alpha_j' z_t I(c_{j-1} < y_{t-d} < c_j) + \epsilon_t$, with c_j representing each *threshold*.

The ESTAR model is normally used in series with symmetric behavior. In this case, we would have one regime when y_{t-d} gets near the threshold and another when it gets far. An example of application is Arango and Gonzalez (2001), who models the Colombian inflation acceleration.

To verify if there are gains in formulating a nonlinear specification, we should do a linearity test. Besides, we will need this test to estimate the parameter d. There is, however, a caveat in performing this test, for there are many ways of defining linearity in (23)-(24)/(25).

When we test $H_{01} : \gamma = 0$ in (23)-(24), the model will be identified only under the alternative hypothesis, because π_{20} and π_2 will be able to assume any value. If, on the other side, we test $H_{01} : \pi_{20}, \pi_2 = 0$, γ and c will be able to assume any value. A similar reasoning applies to (23)-(25).

To contour this problem, Teräsvirta (1994), based on Davies (1977) and on Luukkonen et al. (1988), elaborates a test from a Taylor expansion around $\gamma = 0$, leading to the following auxiliary regression:

$$\hat{u}_t = \hat{w}'_{1t}\tilde{\beta}_1 + \sum_{j=1}^{p} \beta_{2j} y_{t-j} y_{t-d} + \sum_{j=1}^{p} \beta_{3j} y_{t-j} y_{t-d}^2 + \sum_{j=1}^{p} \beta_{4j} y_{t-j} y_{t-d}^3 + v'_t, \tag{26}$$

where $\hat{u}_t$ are the residuals from the (linear) AR(p) model estimation, $\hat{w}'_{1t} = -(1, z'_t)'$ and $\tilde{\beta}_1 = (\beta_{10}, \beta'_1)'$. The null hypothesis of linearity is:

$$H'_0 : \beta_{2j} = \beta_{3j} = \beta_{4j} = 0, \ j = 1, \ldots, p. \tag{27}$$

Each β_{ij} is a combination of the parameters from (23), different, thus, for (24) and (25). Comparing these combinations, (26) will have different formats and Teräsvirta (1994) elaborates successive F-tests which can distinguish between on function and another. They are[7]:

1. $H'_{01} : \beta_4 = 0$.
2. $H'_{02} : \beta_3 = 0 | \beta_4 = 0$.
3. $H'_{03} : \beta_2 = 0 | \beta_3 = \beta_4 = 0$.

If H'_{01} is rejected, we have evidence in favor of the LSTAR model, otherwise, we can use the ESTAR model. If H'_{10} and H'_{20} are not rejected, we also have evidence in favor of LSTAR, the hypothesis rejection, however, is not informative. If the true model is a LSTAR, H'_{03} is generally rejected. If H'_{03} is not rejected and H'_{02} is rejected, we choose the ESTAR model. If H'_{03} and H'_{02} are not rejected, but H'_{01} rejected, LSTAR model is selected. The only inconclusive case is when H'_{01} and H'_{02} are rejected. In this situation, we test:

$$H''_{02} : \beta_3 = 0 | \beta_2 = \beta_4 = 0. \tag{28}$$

However, if H'_{02} is rejected, H''_{02} should be rejected even more strongly.

[7]The notation $\beta_i = 0 | \beta_j = 0$ means that the test $\beta_i = 0$ is made conditional on the non-rejection of the previous test $\beta_j = 0$. Thus, H'_{02} is made conditional on the non-rejection of H'_{01}, and H'_{03} is made conditional on the non-rejection of H'_{02} and of H'_{01}.

Parameter estimation can be done by Nonlinear Least Squares, that, under the assumption of normality, is equivalent to the estimation of Conditional Maximum Likelihood. Under some assumptions and regularity conditions (Wooldridge, 1994), estimators are consistent and asymptotically normal. The estimation procedure is thus carried out with standard numerical algorithms (*e.g.* Newton-Raphson and BFGS).

This work's objective is to make one-step ahead forecasts, re-specifying the model at each period. It is, therefore, impracticable and counter-intuitive to decide, at each step, if the DGP comes from a STAR model with logistic function or with an exponential function. Impracticable for the fact that the test explained above for the most appropriate function contains inconclusive situations, making it difficult to construct an algorithm for real-time automatic forecast. Counter-intuitive because the change in the function at each step would imply that, given the information at each new observation, the entire series generating process would be revalued, being governed alternatively by a function characterizing asymmetric series and by a function characterizing symmetric series.

Since, usually, industrial production tends to have an asymmetric behavior, *i.e.*, its growth dynamics is different from its recession dynamics, we opted to use only the logistic function in the specification of the STAR model.

Castle and Hendry (2014) investigate, through Monte Carlo simulations, the LSTAR models forecast and estimation efficiency for different magnitudes in the transition probability from one regime to another and differences between the intercepts of each regime. They conclude that, for very high γ's, estimation becomes imprecise. Besides, there must be a large number of observations in each regime. For this to happen in small samples, it would be necessary a large probability of change between regimes. On the other side, the constant change of regimes can imply that the series is better modeled by a linear autoregressive process. These problems tend to be solved in reasonably large samples.

Concerning the forecast results, precision, measured by the Root Mean Squared Forecast Error (RMSFE), of a correctly specified LSTAR with one lag improves considerably in large samples and when the difference in the mean of each regime is not very large nor very small (the cited authors analyze changes of 1, 2, and 5 standard-deviations in the intercept).

From equation (23), the one-step ahead forecast, conditional to the information up to time t ($\mathcal{I}_t$), is given by:

$$\begin{aligned}\mathbb{E}[y_{t+1}|\mathcal{I}_t] &= \mathbb{E}[\pi_{10} + \pi_1' z_{t+1} + (\pi_{20} + \pi_2' z_{t+1})F(y_{t+1-d}) + \epsilon_{t+1}|\mathcal{I}_t] \\ &= \pi_{10} + \pi_1' z_{t+1} + (\pi_{20} + \pi_2' z_{t+1})F(y_{t+1-d}),\end{aligned} \tag{29}$$

where the terms z_{t+1} and $F(y_{t+1-d})$ are values already realized in period t, because $z_{t+1} = (y_t, y_{t-1}, \ldots, y_{t-p-1})$ and $d \geq 1$.

More-steps ahead forecasts are more complex, because obtaining $\mathbb{E}(y_{t+2}|\mathcal{I}_t)$ is more difficult and would have to be done by numerical integration. This work, however, makes only one-step ahead predictions, which dismisses a deeper explanation of the subject[8].

[8]See Lundbergh and Teräsvirta (2005) for more detailed reference in forecasting more than one steps ahead in Smooth Transition Autoregressive models.

3.3 Autometrics

Autometrics has its origin in the so-called London School of Economics methodology, whose econometric technique is based on a general-to-specific modeling. In this approach, the researcher chooses a reasonably large number of variables which have a theoretical relation with some dependent variable and, from this general model, he performs specification and significance tests to gradually reduce the set of relevant variables.

This methodology opposes to the so called specific-to-general, that is more frequently used by econometrics practitioners. It consists in starting with simple models containing few variables and, according to diagnostic tests results, adding other variables or modifying the used specification.

The algorithm called Autometrics is presented in Doornik (2009). This type of algorithm, however, has its origin in the works of Lovell (1983) and Hoover and Perez (1999). Later, Hendry and Krolzig (1999, 2005) improve previous works, arriving at a structure similar to Autometrics, which has been continuously improved since then.

Hoover and Perez (1999) consider four main characteristics in their selection mechanism. First, the concept of General Unrestricted Model (GUM), that is a linear regression model involving a large number of variables relating with the utilized dependent variable. Second, the use of multiples search paths, where the first path begins with the elimination of the variable with the largest p-value in the significance test, the second path with the elimination of the variable with the second largest p-value and so on until the tenth variable. Third, at each step of each path, it is performed a joint significance test including all the eliminated variables so far in relation to the GUM. Fourth, at each step of each path, a battery of specification tests are also performed.

In Hendry and Krolzig (1999, 2005), the authors improve the existing algorithm, creating the software called *PcGets* (from general-to-specific). Differently for the previous mechanism, more paths are followed eliminating sets of variables, and there is a pre-selection of lags before the reduction begins. Besides, each path's "terminal" — when it is not possible to remove any more variable without failing in some test — are united in a set of variables which will form the new GUM. The algorithm is finished when there is convergence, *i.e.*, the current GUM equals the previous GUM.

In the end, we will have many candidates which cannot be reduced. Among them, the final candidate is chosen by the minimum Bayesian information criterion (BIC) in Hendry and Krolzig (1999, 2005) and by the best fitted model in Hoover and Perez (1999).

3.3.1 Autometrics' Selection Algorithm

In Doornik (2009), starting with the GUM, the selection mechanism seeks to consider the models with all possible variable combinations. It utilizes a *tree representation* (Figure 1), in which, each letter represents one variable and each node a model to be estimated.

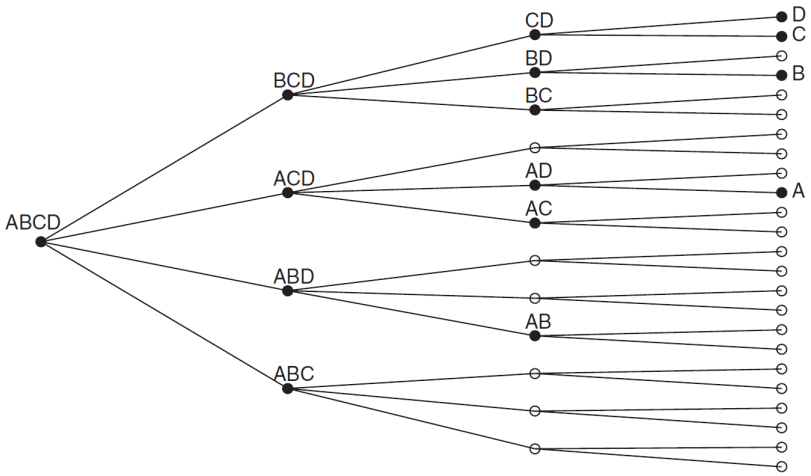

Figure 1: Tree for Possible Models in Autometrics.

The algorithm runs through the tree, following the nodes from left to right and from top to bottom, so that, in the Figure 1 example, it would pass through the models ABCD, BCD, CD, D, C, BD and so on. In each node, the variables are ordered in crescent order of significance, that is, in the GUM, A is the least significant variable and D is the most significant. It is possible that, in the following node, the order will be changed. For example, if C is less significant than B, the ordering would be given by CBD and the variable to be eliminated would be C.

If all the nodes from the tree were evaluated, the computational cost would be very large (for n regressors, $2^n - 1$ models would be estimated), so a number of strategies to jump the irrelevant nodes are used. These strategies are classified by Doornik (2009) as Pruning, Bunching, Chopping and Model Contrast.

Pruning consists in ignoring the nodes derived from a node invalidated by a diagnostic test or by backtesting (a joint significance F-test with relation to the initial GUM). If, *e.g.*, starting from BCD, variable B cannot be eliminated, models CD, D and C are not estimated. The p-value used is denoted p_a.

Bunching stems from the idea that it is possible, instead of eliminating the variables individually based in a t-test, eliminating groups of variables based in a F-test with relation to the current node. The size of the group to be tested is defined by $p_b = max\{\frac{1}{2}p_a^{1/2}, p_a^{3/4}\}$[9], the p-value of the joint significance test. The F-tests significance level for which the group is eliminated, however, is the same as Pruning's, *i.e.*, p_a. Once the group of variables is eliminated, we get to the correspondent following node, avoiding estimation of many models. In the example of Figure 1, if we are at BCD and are able to effectively eliminate jointly B and C, the next model to be estimated is that containing only D, ignoring CD. If, in group, the variables cannot be withdrawn, we test the removal of groups with less variables, until a valid reduction is achieved.

Chopping means ignoring models, in a given ramification from a node, containing some variable or group of variables if they are of very little significance. For example, if we are, again, in BCD and BC presents a F-test with a greater p-value than a certain p_c, the algorithm

[9]See Doornik (2009) for an explanation about this value.

will estimate, in this ramification of the tree, only model D and will go next to node ACD. The standard p-value used by Autometrics is $p_c = p_b$.

Model Contrasts consist in the elimination or shortening of paths in a certain ramification of the tree if it contains some terminal already found previously, so that it jumps to the nodes possessing possible terminals not yet estimated. As an example, suppose that, in the above Figure, we had selected D as a terminal in the first ramification. In the next step, ACD, the following node, AD, would originate two possible models: A and D. As D was already selected, the algorithm tests directly a reduction of ACD to A, speeding the calculations.

There are two types of Model Contrast: Union Contrast and Terminal Contrast. In the first, used when the current GUM is different from the previous GUM, the program considers, to jump, nodes that lead to any model different from the terminal union. In the second, used when the current GUM equals the previous, the jumped nodes takes into account paths that lead to terminals different from each one of the previously individually found terminals.

Several specification tests are then performed. However, differently from the algorithms cited earlier, in Autometrics they are performed only when the terminals are attained. If some test fail in some terminal, the program follows the path backwards from the node, making specification tests until some valid terminal is found. The tests performed are: residuals normality (Jarque and Bera, 1980), second order residual autocorrelation (χ^2 test, Godfrey, 1978; Breusch and Pagan, 1980), autocorrelated conditional heteroscedasticity (ARCH) to second order (Engle, 1982) and that of in-sample stability (Chow, 1960).

In each iteration of the GUM, Autometrics divides the search in two stages. In the first, it ignores paths whose nodes contain terminals. In the second, it follows the root paths which contain terminals, using Model Contrast to reach definitive terminals more quickly.

Furthermore, before initializing the algorithm *per se*, there is the option of performing a presearch to eliminate sets of variables or lags. The general model after the presearch is used as the GUM in the algorithm. Several options of lag and variable reduction are available. It is out of the scope of this chapter to explain the details of the presearch procedures. For a description of the different options, see Doornik (2009).

Finally, when all the tree paths are covered, the final terminals where specification tests failed are discarded and, after the GUM convergence, *i.e.*, when the previous GUM is equal to the current GUM, the terminal with the minimum BIC is chosen.

3.3.2 Dummy Saturation

A way to detect discrepant observations (outliers) in the series and other possible nonlinearities is the use of impulse dummy variables — which assume value 1 for observation t and 0 for the others — for the periods they occur. Proceeding according to the *Gets* methodology, we would have one dummy variable for each observation, besides other exogenous variables which could possibly affect the dependent variable. This framework is denoted as a Dummy Saturation model.

Trying to estimate the GUM this way is obviously impracticable, because there are more variables than observations. Autometrics, however, utilizes a technique to reduct these models, initially proposed by Hendry and Krolzig (2005) and applied to the Impulse Indicator Saturation (IIS) context by Santos et al. (2008).

In case no variables beside the dummies are included, the first step is to apply the reduc-

tion mechanism described above only with the T/2 first dummies, so that the resulting model will contain a subset of them. After this, the same is done with only the T/2 last dummies. Thus, we will obtain two models containing subsets of the initial dummies, the union of these models will generate a new GUM. Applying once more the reduction mechanism, we obtain the model containing the relevant dummies.

A generalization of this method can be applied when step dummies — which assume value 0 up to observation $t-1$ and 1 in the following observations — and other exogenous variables are included in the GUM. In this case, the algorithm introduced by Hendry and Johansen (2012) assumes $N = \sum_{j=1}^{N} n_j$ regressors so that $N > T$ and $n_j < T$ for all j, that is, the N regressors are partitioned in portions smaller than the sample size.

Once the best fitted model is found, with relevant variables and dummies, the forecast is made in the standard way, *i.e.*, the model's n-steps ahead mathematical expectation.

3.4 Double Difference Device

Consider a variable y_t that we want to predict[10]. Suppose $y_t \sim D_{y_t}(y_t|Y_{t-1}, \theta)$, where $\theta \in \Theta \subseteq \mathbb{R}^k$ and $Y_{t-1} = (y_1, \ldots, y_{t-1})$. For a sample of size T, a forecast h step ahead is a combination of the sample values from 1 through T, *i.e.*, $\hat{y}_{T+h|T} = f_h(Y_T)$. It can be proven (for example in Clements and Hendry, 1998) that $\overline{y}_{T+h|T} = \mathbb{E}[y_{T+h}|Y_T]$ is the unbiased predictor which minimizes the Mean Square Forecast Error (MSFE) and holds the minimum variance within the unbiased predictor class.

Suppose y_t follows the linear autoregressive process below:

$$y_t = \mu + \rho y_{t-1} + \gamma z_t + \epsilon_t, \text{ with } \epsilon_t \sim IN(0, \sigma_\epsilon^2) \text{ e } |\rho| < 1, \tag{30}$$

where $\{z_t\}$ is an exogenous variable, forecast error's mean and variance are, respectively, $\mathbb{E}[y_{T+1} - \overline{y}_{T+1|T}] = 0$ and $V[(y_{T+1} - \overline{y}_{T+1|T})] = \sigma_\epsilon^2$.

Consider $\mathbb{E}[y_t] = \theta$ and $z_t \sim N(\kappa, 1)$. We can write the mean of y_t as:

$$\mathbb{E}[y_t] = \theta = \frac{\mu + \gamma\kappa}{1 - \rho}. \tag{31}$$

If $\mu = \kappa = 0$, a break in ρ will not affect the mean of y_t. However, in case that we have one of these parameters different from zero, a change in the autoregressive parameter will imply a change in the mean. Besides this, if more than one of these parameters suffer a break, the forecast failure can be even more severe. For example, if the autoregressive parameter and the intercept, ρ and μ, are, respectively, 0.8 and 8, and suffer a break, becoming 0.6 and 6, the unconditional expectation shifts from $\theta = 40$ to $\theta^* = 15$.

Let us rewrite the model (30) as:

[10]This exposition is based in Hendry (2012), and simplifies the more general theory with multiple variables in Hendry (2006).

$$\begin{aligned}\Delta y_t &= \mu + (\rho-1)y_{t-1} + \gamma z_t + \epsilon_t + \left((\rho-1)\frac{\mu+\gamma\kappa}{1-\rho} - (\rho-1)\frac{\mu+\gamma\kappa}{1-\rho}\right) \\ &= \mu - \mu - \gamma\kappa + (\rho-1)(y_{t-1}-\theta) + \gamma z_t + \epsilon_t \\ &= (\rho-1)(y_{t-1}-\theta) + \gamma(z_t-\kappa) + \epsilon_t.\end{aligned}$$

Rearranging the terms, i steps ahead values can be written as:

$$\Delta y_{T+i} = \omega + \alpha(y_{T+i-1} - \theta) + \gamma z_{T+i} + \epsilon_{T+i}, \tag{32}$$

where $\omega = -\gamma\kappa$ and $\alpha = \rho - 1$.

This formulation is denoted Equilibrium Correction Model (EqCM), where the equilibrium is given by θ. Forecasts will tend to return to θ independently of the behavior of the data. Changes in this equilibrium, which occur mainly by shifts in the level, consist in the main factor which imply forecast failure.

A naive predictor is the called Double Difference Device (DDD). Its intuition comes from the fact that most economic variables do not continuously accelerate and, hence, its second differences have unconditional expectation $\mathbb{E}[\Delta^2 y_t] = 0$, suggesting the following DDD estimator:

$$\widetilde{\Delta y}_{T+1|T} = \Delta y_T, \tag{33}$$

where $\widetilde{\Delta y}_{T+1|T}$ denotes a one-step ahead forecast for Δy_T given the information in time T.

Because the forecast does not utilizes any parameter, structural breaks do not persistently influence it, and the estimator is unbiased.

Let us modify equation (32) so that the DGP does not contain exogenous variables besides the dependent variable lags (*i.e.* $\gamma = 0$):

$$\Delta y_{T+i} = \alpha(y_{T+i-1} - \theta) + \epsilon_{T+i}. \tag{34}$$

Suppose there is a break in the parameters, so that the DGP becomes:

$$\Delta y_{T+i} = \alpha^*(y_{T+i-1} - \theta^*) + \epsilon_{T+i}. \tag{35}$$

EqCM forecast error is given by:

$$\Delta y_{T+i} - \widehat{\Delta y}_{T+i|T+i-1} = \Delta y_{T+i} - \hat{\alpha}(y_{T+i-1} - \hat{\theta}) + \epsilon_{T+i} = w_{T+i}, \tag{36}$$

where the hat over the parameters means that they were estimated on the EqCM formulation.

Replacing the in-sample estimated values for the pseudo-true in-sample population values, where $\mathbb{E}[\hat{\theta}] = \theta_p$, we can reduce the forecast error variance without damage to the general analysis. Thus, we have:

$$\begin{aligned}\mathbb{E}[w_{T+i}|y_{T+i-1}] &= (\alpha^*\theta^* - \alpha_p\theta_p) + (\alpha^* - \alpha_p)y_{T+i-1} \\ V_{T+i}[w_{T+i}|y_{T+i-1}] &= \sigma_\epsilon^2.\end{aligned} \tag{37}$$

With respect to the DDD ($\widetilde{\Delta y}_{T+i|T+i-1} = \Delta y_{T+i-1}$), for $i > 1$, we have:

$$\Delta y_{T+i-1} = \alpha^*(y_{T+i-2} - \theta^*) + \epsilon_{T+i-1}. \tag{38}$$

The respective forecast error is calculated as:

$$\begin{aligned} \Delta y_{T+i} - \widetilde{\Delta y}_{T+i|T+i-1} = u_{T+i} &= \alpha^*(y_{T+i-1} - \theta^*) + \epsilon_{T+i} \\ &\quad - [\alpha^*(y_{T+i-2} - \theta^*) + \epsilon_{T+i-1}] \qquad (39) \\ \Rightarrow u_{T+i} &= \alpha^* \Delta y_{T+i-1} + \Delta\epsilon_{T+i}. \qquad (40) \end{aligned}$$

On the long term, values are replaced by their unconditional expectations, *i.e.*:

$$\mathbb{E}[u_{T+i}] = \alpha^* \mathbb{E}[\Delta y_{T+i-1}] + \mathbb{E}[\Delta\epsilon_{T+i}] = 0 \tag{41}$$

$$\begin{aligned} V[u_{T+i}] &= V[\alpha^* \Delta y_{T+i-1}] + V[\Delta\epsilon_{T+i}] \\ &= \alpha^{*2} V[\Delta y_{T+i-1}] + 2\sigma_\epsilon^2. \qquad (42) \end{aligned}$$

As it is evident from (40) and (42), the Double Differencing mechanism adds some noise sources by the extra differentiation of $\alpha^* y_{T-i+1}$ and of ϵ_{T+i}. This extra source of noise, however, can be of lower dimension than those from the traditional autoregressive model when there is presence of structural breaks in the series' unconditional mean.

To illustrate, suppose that the shift occurs only in μ, so that θ changes and α remains constant. We will have:

$$\begin{aligned} w_{T+i} &= -\alpha(\theta^* - \theta) + \epsilon_{T+i} \\ \mathbb{E}[w_{T+i}] &= -\alpha(\theta^* - \theta) \\ V[w_{T+i}] &= \sigma_\epsilon^2 \\ u_{T+i} &= \alpha \Delta y_{T+i-1} + \Delta\epsilon_{T+i} \\ \mathbb{E}[u_{T+i}] &= 0 \\ V[u_{T+i}] &= V[\alpha \Delta y_{T+i-1}] + V[\Delta\epsilon_{T+i}]. \end{aligned}$$

The MSFE is approximately:

$$M[w_{T+i}] = \alpha^2(\theta^* - \theta)^2 + \sigma_\epsilon^2. \tag{43}$$

Comparing with the DDD MSFE:

$$M[u_{T+i}] = 2\sigma_\epsilon^2 \left(1 + \frac{\alpha^2}{2 + \alpha}\right). \tag{44}$$

Assuming $\rho = 0.8$, $\nabla\mu^* = \mu^* - \mu = 0.2$, and $\sigma_\epsilon = 0.06$, we have $\alpha = -0.2$ and $\nabla\theta^* = \theta^* - \theta = 1$. With these values, $M[w_{T+i}]$ is approximately 6-fold larger than $M[u_{T+i}]$.

3.5 LASSO and Derived Methods

Developed by Tibshirani (1996), the Least Absolute Shrinkage and Selection Operator (LASSO) is based on a class of estimation procedures known as shrinkage methods, which shrink Ordinary Least Squares (OLS) coefficients estimators towards zero. This approach aims to reduce the variance in the bias-variance trade-off, introducing some bias in the estimators.

Consider the data $(\mathbf{x^i}, y_i)$, $i = 1, \ldots, N$, where $\mathbf{x^i} = x_{i1}, x_{i2}, \ldots, x_{ip}$ are the regressors and y_i the dependent variable, and assume that the $y_i s$ are conditionally independent given the $x_{ij} s$. The LASSO estimator solves the following problem:

$$\underset{\alpha,\beta}{min}\left\{ \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\}, \text{ subject to } \sum_j |\beta_j| \leq t, \tag{45}$$

where $t \geq 0$ is a tuning parameter. The independent variables are standardized (*i.e.*, $\sum_i x_{ij}/N = 0$ and $\sum_i x_{ij}^2/N = 1$) to rule out problems of differences in their scales. For all t, the solution for α is given by $\hat{\alpha} = \overline{y}$, so that (45) can be solved omitting the intercept.

The problem can be expressed in its equivalent Lagrangian form:

$$\underset{\alpha,\beta}{min}\left\{ \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\} + \lambda \sum_j |\beta_j|. \tag{46}$$

For $\lambda = 0$ (for a sufficiently large t), the restriction in (46) is nonbinding and $\hat{\beta}_{LASSO} = \hat{\beta}$, the OLS estimator. Higher values of λ impose the estimators a penalty, shrinking them towards zero. Depending on the value of λ, some of the LASSO coefficients are set to be exactly equal to zero, performing, thus, a method of variable selection.

To better understand this characteristic of the LASSO, consider an earlier example of regularization method introduced by Hoerl and Kennard (1970), the Ridge Regression estimator, which solves the following problem (also in its Lagrangian form):

$$\underset{\alpha,\beta}{min}\left\{ \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\} + \lambda \sum_j \beta_j^2, \tag{47}$$

with the same set-up of the LASSO.

The difference between the two methods lies in the form of the penalty: the LASSO uses the L^1-norm penalty and the Ridge Regression uses the L^2-norm penalty.

Due to the L^1-norm used in LASSO, the solution to (46) is nonlinear in y_i, for which there is no closed form expression. For the Ridge Regression, the solution takes the form $\hat{\beta}_{RIDGE} = (X'X + I\lambda)^{-1}X'y$. Note that, differently from the LASSO, $\hat{\beta}_{RIDGE}$ will generally have all its components different from zero. To see this, consider Figure 2 below.

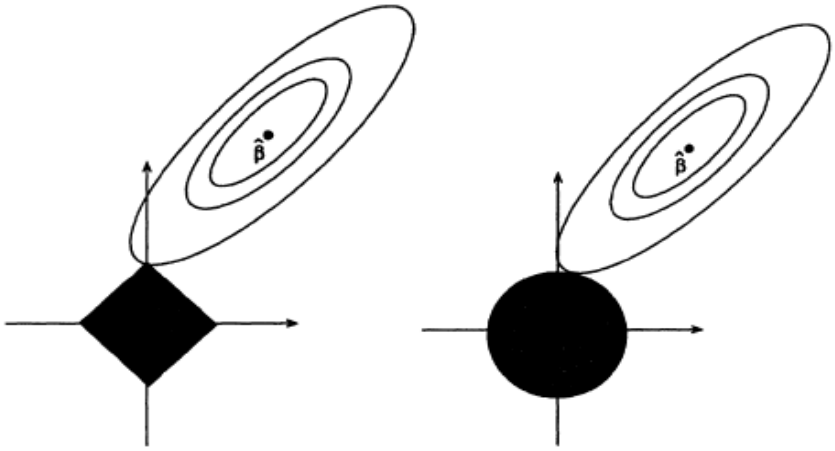

Figure 2: Estimation picture for LASSO (left) and Ridge Regression (right). Source: Tibshirani (1996).

In a two parameter space, the OLS estimator is represented by $\hat{\beta}$, the ellipses are contour plots for the same sum of squared residuals, and the black area represents the possible values under the constraint for some t. The corners on the restriction area make the LASSO estimator, in the left, to fix one of the coefficients exactly at zero. The Ridge Regression estimator, in the right, shrinks the OLS estimator towards zero, but does not, as a rule, estimate exactly zero coefficients.

Up to this point, our analysis treated λ as given. In practice, we have to choose it according to the data. Probably, the most common approach to estimate λ is by cross-validation, which aims to minimize an estimate of the expected prediction error[11]. However, cross-validation may present difficulties in a time-series framework. Zou et al. (2007), Zhang et al. (2010), Wang et al. (2007) argue that using the Bayesian Information Criterion (BIC) is a good alternative to cross-validation in selecting λ. We follow Medeiros and Mendes (2016), Medeiros and Vasconcelos (2016) and Konzen and Ziegelmann (2016), for example, in using the BIC to select the tuning parameter.

3.5.1 AdaLASSO

Fan and Li (2001) defined a procedure satisfying the oracle properties as one that, asymptotically, selects the correct subset of variables with nonzero coefficients and has an optimal estimation rate. Zou (2006) states that if, for some λ, the LASSO has an optimal estimation rate, then it does not satisfy variable selection consistency. The author also argues that, even relaxing an optimal estimation rate, variable selection is consistent only under a nontrivial condition (similar conclusions were made by Meinshausen and Bühlmann, 2006 and Zhao and Yu, 2006).

Zou (2006) proposed, thus, a slight modification to the LASSO, the Adaptive LASSO (AdaLASSO) estimator, which solves the following problem:

[11] See Hastie et al. (2009) for a more detailed explanation of the cross-validation method for model selection and parameter tuning.

$$\min_{\alpha,\beta}\left\{\sum_{i=1}^{N}\left(y_i-\alpha-\sum_j\beta_j x_{ij}\right)^2\right\}+\lambda\sum_j\hat{w}_j|\beta_j|, \tag{48}$$

where $\hat{w}_j=|\hat{\beta}_{FSTEP}|^{-\gamma}$, $\gamma>0$, and $\hat{\beta}_{FSTEP}$ is a first step estimator of β. One can use, for example, the OLS estimator, the Ridge Regression estimator, or the LASSO estimator (as in Garcia et al. (2017)). We opt to use $\hat{\beta}_{FSTEP}=\hat{\beta}_{RIDGE}$ and $\gamma=1$.

The AdaLASSO assigns different penalties to each variable, with higher penalties to variables whose coefficients are closer to zero in the first step estimation. Zou (2006) proves that the AdaLASSO is an oracle procedure.

3.5.2 WLAdaLASSO

Another variant of the LASSO we use is the Weighted Lag Adaptive LASSO (WLAdaLASSO), proposed by Konzen and Ziegelmann (2016), based on the work by Park and Sakaori (2013). This method is similar to the AdaLASSO, and it comes from the observation that, in a time-series autoregression framework, more distant lags tend to have less influence in forecasting the dependent variable, imposing on them higher penalties. The estimator solves:

$$\min_{\alpha,\beta}\left\{\sum_{i=1}^{N}\left(y_i-\alpha-\sum_j\beta_j x_{ij}\right)^2\right\}+\lambda\sum_j\hat{w}_j^{WL}|\beta_j|, \tag{49}$$

where $\hat{w}_j^{WL}=(|\hat{\beta}_{RIDGE}|e^{-\alpha l})^{-\gamma}$, l is the lag order, and $\gamma>0$, $\alpha\geq 0$ are tuning parameters. As in AdaLASSO, we set $\gamma=1$. To select α, for a given λ, we estimate the model considering a grid $(0,0.5,1,\ldots,10)$ for α and choose that with the lowest BIC. The parameter λ is then selected considering the model with the lowest BIC (see Konzen and Ziegelmann, 2016 and Prince and Marçal, 2018).

Monte Carlo simulations performed by Konzen and Ziegelmann (2016) pointed that the WLAdaLASSO was superior to both LASSO and AdaLASSO in variable selection, parameter estimation and forecasting, particularly when the candidate variables included a high number of lags and predictors presented stronger linear dependence.

4 Empirical Framework

The objective of this work is twofold. The first is to test if there is any gain in using extended series of Brazilian Industrial Production, the second is to assess if nonlinear univariate methods have better forecasting performance than the autoregressive model.

This section details the PIM-PF time-series available and describes the forecasting exercise we perform. It also details the method to test whether forecasts accuracy are statistically different between models.

4.1 PIM-PF Data

There are, as mentioned in the previous chapter, three different series available in IBGE's System of Automatic Data Retrieval (SIDRA): 1985:01-2004:01, with base year 1991, 1991:01-2014:02, with base year 2002, and 2002:01-2018:12, with base year 2012. There is an extended version of the first series available in the Time Series Management System (SGS) of Brazilian Central Bank (BCB), ranging from 1975:01 to 2004:01, where the 1985:01-2004:01 values are identical to those from SIDRA.

Figure 3 plots the three series in one graph. Note that there is considerable overlap between their covered time span. This is because the closed series published by IBGE already contain chained values from the previous one.

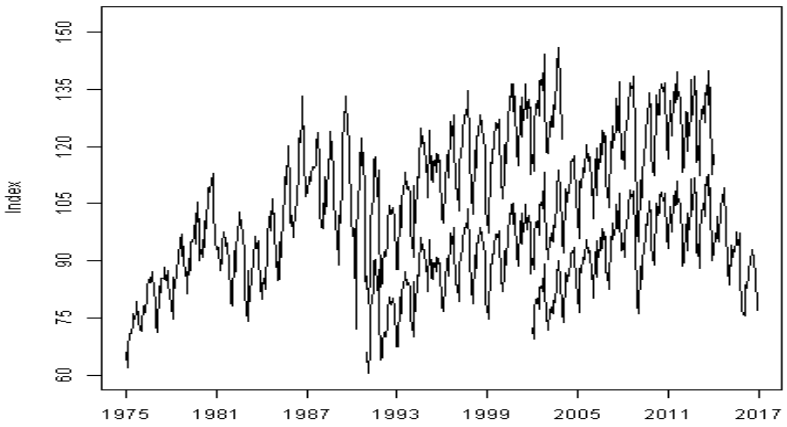

Figure 3: PIM-PF Series Expressed in Different Base Years.

One reason the entire chained series is not always published is that there are difficulties in the chaining components disaggregated by activities, due to changes in the classification system and different list of selected products in each revision. The chaining of the aggregated series is, on the other hand, straightforward and plausible. Figure 4 contains the chained series, constructed as described in (3) and (5), of the General Industrial Production *quantum* for whole sample.

The presence of non-stationarity and seasonality in the series is easily noticed. To deal with these characteristics, the log of the series is differenced with relation to the same month in the previous year. This corrects both problems if the series is integrated of order one and has constant seasonality[12]. Figure 5 shows a graph with the treated series.

For nonlinear models with different regimes, specification becomes more robust when data shows well defined different regimes. In STAR models, for example, γ estimation becomes difficult if the probability of crossing the threshold c is large, *i.e.*, the regime are not very well

[12]The presence of stochastic seasonality, however, may not be eliminated with this transformation. In fact, even standard seasonal adjustment methods may fail to correctly account for stochastic seasonality. The use of Autometrics with dummy saturation, however, can improve the modelling of non-deterministic seasonality.

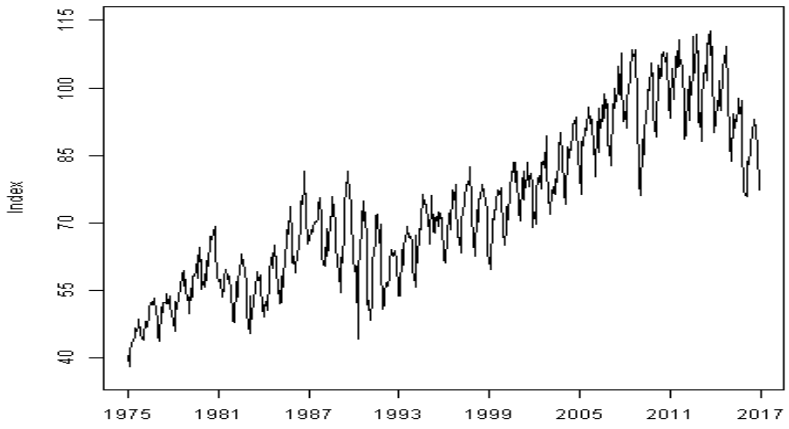

Figure 4: Chained PIM-PF Series. Mean 2012=100

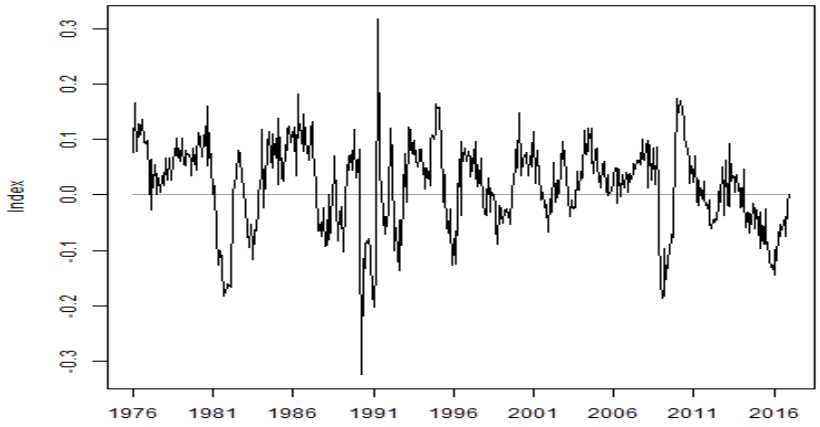

Figure 5: Log General Industrial Production - Annual Difference

defined (Castle and Hendry, 2014; Granger and Teräsvirta, 1993). The Markov Switching models also adapt better when there are well defined regimes.

4.2 Description of the Forecasting Exercise

This work evaluates the performance of forecasts using samples of different lengths from the Brazilian Industrial Production Index. Given the volatile and asymmetric nature of the investigated variable, it is also of interest to compare forecasts accrued from nonlinear methods with those of a standard benchmark methods. Therefore, we make two forecasting exercises. One compares, for each method, forecasts using chained series, extended to different starting periods, with forecasts using the latest PIM-PF series available, beginning in 2002:01. The other compares, for each sample, forecasts of different methods with those of AR(p).

Forecasts are made one-step ahead, re-specifying and re-estimating the models at each step. The AR and STAR models order p is selected by the BIC. To specify the MSAR model, at each forecast step, we estimate 27 different models (1 to 13 lags with switching autoregressive parameters, 1 to 13 lags with fixed autoregressive parameters, and one only with intercept, all with two regimes and including switching intercept and variance) selecting that with the lowest BIC. LASSO, AdaLASSO, WLAdaLASSO and Autometrics algorithms are reapplied at every step. For all methods, a maximum lag order of 13 is considered. In Autometrics, we use a standard p-value of 0.01 for significance tests and saturations of impulse and step indicators.

Since we used a twelve month difference to treat the data, we will suppose $\mathbb{E}[\Delta\Delta_{12}y_t] = 0$ instead of $\mathbb{E}[\Delta^2 y_t] = 0$ in the double difference device and denote it DD12, *i.e.*, the industrial production growth with respect to the same month of the previous year do not continuously accelerate. This modication is straightforward and the interpretation is not impaired.

We analyze the forecast horizon starting in January 2007 and ending in December 2016, comprising a total of 120 months, in which forecasts are made one-step ahead. We also analyze the forecast performance in both halves of the forecast horizon, *i.e.*, 2007:01-2011:12 and 2012:01-2016:12, in which the first half is characterized by abrupt changes in the series, reflecting the effects of the subprime crisis in Brazilian industrial production and its subsequent recovery, and the second by a less abrupt, but persistent, decline in the industrial activity level.

Results are obtained using the chained series starting in 1975:01, 1985:01 and 1991:01. We also consider forecasts made with a 121 months rolling window and with an extending window with the same initial size, *i.e.*, starting in 1993:12[13]. Note that these starting periods coincide with those from the closed series obtained from SIDRA and BCB. Using these samples may help identify if more distant methodological benchmarks contain useful information for forecasting recent values.

4.3 Forecast Comparison (Diebold and Mariano Test)

To compare the forecasts, we used the Diebold and Mariano (1995) test, which utilizes the forecast errors from two different models to evaluate if one forecast is statistically superior the other.

For the test to be valid, it suffices that assumptions about the forecast errors are valid and it is not necessary to make hypotheses about the models being tested, so that it is possible to compare even predictions that do not come from models.

Let e_{it} be the forecast error from model i in period t and $g(e_{it})$ some loss function. In this study, we analyze the results for both $g(e_{it}) = e_{it}^2$, the quadratic loss function, and $g(e_{it}) = |e_{it}|$, the absolute loss function. The key hypotheses are actually made about the difference between the loss function associated with the forecast errors, $d_t = g(e_{it}) - g(e_{jt})$. It is assumed that d_t is covariance-stationary:

[13]These are the starting months of the level series, the annual difference of the index logarithm drops 12 observations from each sample.

$$\begin{cases} \mathbb{E}(d_t) = \mu, \forall t \\ cov(d_t, d_{t-\tau}) = \varphi(\tau), \forall t \\ V(d_t) = \sigma_d^2 < \infty \end{cases}$$

The hypothesis of equal predictive capacity is equivalent to $\mathbb{E}(d_t) = 0$. In this case, the test statistics is:

$$DM = \frac{\overline{d}}{\hat{\sigma}_{\overline{d}}} \xrightarrow{d} N(0,1) \tag{50}$$

where $\overline{d} = \frac{1}{T}\sum_{t=1}^{T}(g(e_{it}) - g(e_{jt}))$ and $\hat{\sigma}_{\overline{d}}$ is a consistent estimator for the standard deviation of $\overline{d}$.

As d_t has frequently some autocorrelation, mainly due to imperfect predictions, DM statistic must be calculated with a robust $\hat{\sigma}_{\overline{d}}$. Diebold and Mariano (1995) suggest $\hat{\sigma}_{\overline{d}} = \sqrt{2\pi \hat{f}_d(0)/T}$, where $\hat{f}_d(0)$ is a consistent estimator for the spectrum of the loss function differential at frequency zero. $f_d(0) = \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty}\varphi(\tau)$.

One caveat must be made concerning this method. Its objective is to test if different predictions from diverse sources are statistically different and, for this, it uses as primitive the forecast error loss function differential. When we compare two models, saving a fraction of the sample to calculate the loss function, what is used as primitives for the Diebold and Mariano test are, ultimately, the estimated parameters of the model, which originate the forecasts and, thus, the forecast errors. In doing so, part of the sample is lost, increasing the estimation uncertainty and jeopardizing efficiency in small samples.

Diebold (2015) argues that there are consolidated methods using the whole sample which are more efficient in model comparison. In a Bayesian context, the best model selection criterion resumes to that with the largest marginal likelihood function, that, as shown in Schwarz (1978), corresponds asymptotically to the model with smallest BIC. In fact, in the Gaussian linear regression context, the Bayesian information criterion can be written as $BIC = T^{(k/T)}MSFE$, where $MSFE = \frac{\sum_{t=1}^{T} e_t^2}{T}$, that is, the BIC is an estimate of the out-of-sample MSFE.

Comparing different models, however, is not the only focus of this work. Other two important objectives are: (i) test if using longer samples, with series produced using different methodological benchmarks, is superior to using only the last available sample; (ii) assess how one-step ahead forecasts for different methods behave in different periods and different estimation samples. Diebold (2015) criticizes the indiscriminate use of the Diebold and Mariano test to compare models, but argues that it may be useful to, among other things, flag for structural breaks and to assess predictive performance over different business cycles. Furthermore, at each forecast step a new model is specified and estimated, and estimation using different sample sizes are performed. Therefore, Diebold and Mariano test arrive as an appropriate test, since it uses only the series of loss function differentials between two models.

5 Results

Tables 1 and 2 present the results of the Root Mean Squared Forecast Error (RMSFE) and Mean Absolute Forecast Error (MAFE) for each sample and model. We also present, in Table 3, the Mean Forecast Error (MFE), which gives an idea about the bias of forecasts.

Table 1: Root Mean Squared Forecast Error

2007:01 - 2016:12											
	AR	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	DD12	AUTO	AutoIIS	AutoSIIS	AVG
2002:01	4.08	4.70	4.25	4.20	4.18	4.28	4.17*	4.29	4.57	4.37	**4.01**
121-RW	3.88	4.31	4.08	4.02	3.89	3.89	4.17*	3.95	**3.77**	4.09	**3.77**
1993:12	3.76*	3.96*	3.96*	3.82	3.79	3.79	4.17*	3.81	3.83	3.82*	**3.65***
1991:01	3.88	4.35	3.97	3.85	3.76	3.76	4.17*	3.72	3.84	3.88	**3.71**
1985:01	3.99	4.27	4.05	3.94	3.80	3.87	4.17*	3.74	**3.60**	4.01	3.76
1975:01	3.93	4.04	3.97	3.81*	3.75*	3.75*	4.17*	3.64*	**3.57***	4.23	3.72
2007:01 - 2011:12											
	AR	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	DD12	AUTO	AutoIIS	AutoSIIS	AVG
2002:01	4.60	5.46	4.64	4.60	4.65	4.75	**4.30***	4.92	5.44	5.02	4.46
121-RW	4.43	4.70	4.55	4.37	4.26	4.26	4.30*	4.34	4.25	4.58	**4.13**
1993:12	4.25*	4.49	4.32	4.13	4.14	4.14	4.30*	4.32	4.45	4.20*	**4.02***
1991:01	4.38	5.10	4.30	4.15	**4.08**	**4.08**	4.30*	4.11	4.45	4.25	4.09
1985:01	4.37	4.65	4.38	4.12	3.99	4.05	4.30*	**3.95***	4.12*	4.54	4.05
1975:01	4.29	4.45*	4.27*	4.05*	**3.98***	**3.98***	4.30*	4.04	4.12*	4.77	4.04
2012:01 - 2016:12											
	AR	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	DD12	AUTO	AutoIIS	AutoSIIS	AVG
2002:01	**3.48**	3.79	3.81	3.75	3.65	3.75	4.03*	3.55	3.50	3.61	3.50
121-RW	3.24	3.89	3.54*	3.63	3.49	3.48	4.03*	3.51	**3.23**	3.54	3.38
1993:12	3.20*	3.34*	3.56	3.49*	3.41	3.41	4.03*	3.22	**3.08**	3.40	3.24*
1991:01	3.30	3.44	3.60	3.51	3.40*	3.40*	4.03*	3.29	**3.12**	3.47	3.30
1985:01	3.58	3.84	3.69	3.74	3.61	3.67	4.03*	3.53	**3.00**	3.39*	3.45
1975:01	3.53	3.58	3.63	3.56	3.50	3.50	4.03*	3.18*	**2.92***	3.62	3.35

Notes: Values are multiplied by 100. The first column denotes the initial observation of the estimation sample, 121-RW denotes a 121 months rolling window forecasting scheme. Lowest absolute values for each method (column) are marked by *, boldface numbers mark lowest absolute values for each sample (row).

Figure 6 plots RMSFE, MAFE and MFE for all models in the first, second and third lines, respectively. Results are shown for the whole forecast horizon (first column) and both of its halves (second and third columns). The dashed line links the median of plotted values (*i.e.*, median of RMSFE, MAFE and MFE values). Point forecast results and the value of RMSFE, MAFE and MFE medians can be found in the appendix.

We first focus on the difference between the results for different forecast periods. A first fact we may note is that forecasts for 2012:01-2016:12 are more accurate than for 2007:01-2011:12. In fact, the highest RMSFE for all models and samples considering only the second half of the forecast horizon, that corresponding to the double difference device (DD12), is only slightly higher than the lowest RMSFE for all models and samples considering only the first half, that of Autometrics selection without any dummy saturation (AUTO) with the sample starting in 1985:01 (4.05 for DD12 and 3.95 for AUTO).

The MAFE results are less discrepant between the different forecast horizons, although the median of MAFEs (the dashed lines in Figure 6) are also lower in the second half of forecast horizon for each sample. This arises from the fact that squared forecast error loss function punishes more heavily forecasts which are more distant from the actual realization

Table 2: Mean Absolute Forecast Error

2007:01 - 2016:12											
	AR	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	DD12	AUTO	AutoIIS	AutoSIIS	AVG
2002:01	**2.93**	3.29	3.27	3.15	3.05	3.16	3.22*	3.14	3.26	3.43	3.01
121-RW	**2.88**	3.19	3.02	3.13	3.06	3.04	3.22*	3.07	2.92	3.06	2.91
1993:12	**2.75***	2.86*	2.99*	3.02	3.00	3.00	3.22*	2.93	2.90	2.91*	2.79*
1991:01	2.87	3.02	3.04	2.98*	2.90*	2.90*	3.22*	2.88	2.85	3.03	**2.83**
1985:01	3.09	3.25	3.13	3.11	3.04	3.08	3.22*	2.95	**2.74**	3.09	2.92
1975:01	3.00	3.05	3.05	3.00	2.96	2.96	3.22*	2.86*	**2.67***	3.28	2.88
2007:01 - 2011:12											
	AR	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	DD12	AUTO	AutoIIS	AutoSIIS	AVG
2002:01	**3.10**	3.56	3.40	3.16	3.12	3.25	3.28*	3.39	3.78	3.88	3.14
121-RW	3.14	3.38	3.16	3.20	3.16	3.16	3.28*	3.21	3.28	3.27	**2.99**
1993:12	2.90*	2.98*	3.08*	3.06	3.09	3.09	3.28*	3.15	3.21	2.94*	**2.83***
1991:01	3.01	3.22	3.14	2.97*	2.91*	2.91*	3.28*	2.95*	3.10	3.15	**2.88**
1985:01	3.30	3.37	3.23	3.08	3.03	3.07	3.28*	2.97	**2.93**	3.32	2.96
1975:01	3.16	3.22	3.14	3.00	2.97	2.97	3.28*	3.03	**2.89***	3.57	2.96
2012:01 - 2016:12											
	AR	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	DD12	AUTO	AutoIIS	AutoSIIS	AVG
2002:01	2.77	3.02	3.13	3.13	2.99	3.06	3.16*	2.90	**2.74**	2.99	2.88
121-RW	2.63	3.00	2.88*	3.06	2.96	2.91	3.16*	2.92	**2.56**	2.85	2.82
1993:12	2.59*	2.73*	2.90	2.98*	2.91	2.91	3.16*	2.72	**2.58**	2.88	2.74
1991:01	2.73	2.82	2.95	3.00	2.88*	2.88*	3.16*	2.80	**2.59**	2.91	2.77*
1985:01	2.89	3.12	3.03	3.13	3.05	3.09	3.16*	2.93	**2.54**	2.85*	2.87
1975:01	2.83	2.89	2.97	3.00	2.95	2.95	3.16*	2.68*	**2.45***	3.00	2.80

Notes: see notes from Table 1.

Table 3: Mean Forecast Error

2007:01 - 2016:12											
	AR	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	DD12	AUTO	AutoIIS	AutoSIIS	AVG
2002:01	0.22	0.85	0.43	1.21	1.10	1.02	**0.00***	0.32	0.67	0.62	0.64
121-RW	0.16	0.75	0.43	0.88	0.80	0.71	**0.00***	0.44	0.31	0.15	0.50
1993:12	0.14	0.39*	0.40*	0.92	0.88	0.88	**0.00***	0.61	0.55	0.23	0.50
1991:01	0.24	0.72	0.48	0.97	0.88	0.88	**0.00***	0.72	0.40	0.10*	0.54
1985:01	0.14	0.65	0.59	0.74*	0.65*	0.71	**0.00***	0.24*	0.26	0.40	0.44
1975:01	0.13*	0.45	0.52	0.76	0.67	0.67*	**0.00***	0.32	0.13*	0.19	0.38*
2007:01 - 2011:12											
	AR	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	DD12	AUTO	AutoIIS	AutoSIIS	AVG
2002:01	-0.13*	0.43	0.13	0.82	0.79	0.74	**0.02***	-0.32	0.76	-0.10*	0.31
121-RW	-0.25	0.24	0.22	0.23	0.24	0.24	**0.02***	-0.19	0.16	-0.63	0.06
1993:12	-0.30	-0.19	-0.02	0.08	0.16	0.16	0.02*	-0.06*	**-0.01***	-0.85	-0.10
1991:01	-0.30	0.34	**0.00***	0.24	0.23	0.23	0.02*	0.10	-0.01*	-0.95	-0.01*
1985:01	-0.52	0.31	-0.07	-0.31	-0.23	-0.22	**0.02***	-0.52	-0.25	-0.34	-0.21
1975:01	-0.39	0.13*	-0.04	-0.04*	-0.05*	-0.05*	**0.02***	-0.23	-0.21	-1.00	-0.19
2012m01 - 2016m12											
	AR	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	DD12	AUTO	AutoIIS	AutoSIIS	AVG
2002:01	0.57	1.26	0.73	1.61	1.40	1.30	**-0.02***	0.97	0.59	1.33	0.97
121-RW	0.56*	1.26	0.64*	1.53*	1.37*	1.18*	**-0.02***	1.07	0.45*	0.93*	0.94*
1993:12	0.58	0.97	0.83	1.75	1.59	1.59	**-0.02***	1.29	1.10	1.31	1.10
1991:01	0.77	1.10	0.95	1.69	1.54	1.54	**-0.02***	1.34	0.82	1.15	1.09
1985:01	0.81	0.98	1.24	1.79	1.54	1.64	**-0.02***	1.01	0.77	1.13	1.09
1975:01	0.65	0.77*	1.08	1.56	1.38	1.38	**-0.02***	0.88*	0.47	1.38	0.95

Notes: see notes from Table 1.

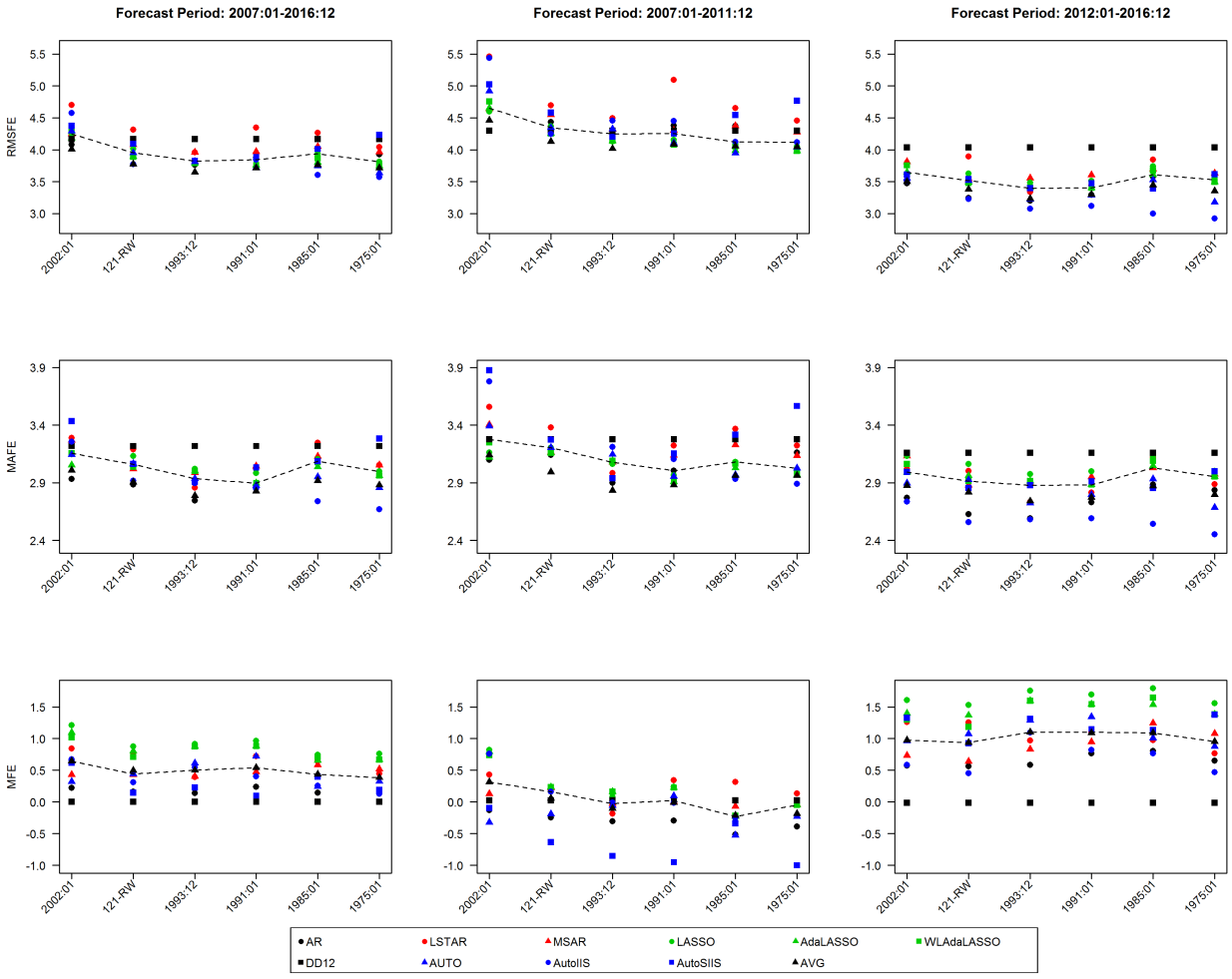

Figure 6: RMSFE, MAFE and MFE results for each estimation sample. The dashed line connects the median of these values.

than the absolute forecast error. The abrupt changes in 2007:01-2011:12 causes most models to, even at one-step ahead forecasts, miss the actual value by a relatively large amount, specially in turning points.

The MFE results in the forecast horizon's second half, however, show an upward bias, except for DD12 (which is consistent with theory, as seen in section 3.4), presenting a near zero forecast bias, whereas, in the first half, models present upward and downward biases, with median MFE nearer zero. Although forecasts are less precise in 2007:01-2012:12, the PIM-PF annual difference plunge and soar in this period causes them to deviate up and down from actual values in similar magnitude. Note that, as expected, forecasts from LASSO and derived methods tend to have larger bias than from the other models.

Consider now how forecast accuracy changes as we enlarge the initial estimation sample. The median of RMSFE and MAFE values, for both halves of the forecast horizon, tends to become lower if the sample is extended at least to 1993:12, and higher (or at least in the same level) if it is extended further, although forecasts using the sample starting in 1985:01 tend to be less precise than using that starting in 1975:01. In 2012:01-2016:12, variability between

the RMSFE (and MAFE) of different models increases for the larger estimation samples. In 2007:01-2011:12, variability of accuracy from the different models is the greatest using the sample starting in January 2002.

These differences, as we extend the estimation sample, in the median forecast performance of the models are, however, small. The median RMSFE for the whole forecast horizon, for example, changes from 4.22, using the sample starting in January 2002, to 3.82, using the expanding sample (*i.e.*, not using rolling windows) starting in December 1993.

Considering the relative performance of the models, reported in Tables 1 and 2, for the whole forecast horizon, the most accurate model depends on which estimation sample is used. Autometrics with Impulse Indicator Saturation (AutoIIS) using the sample starting in 1975:01 gives the best forecast accuracy, for both RMSFE and MAFE. For smaller samples, however, its accuracy tends to deteriorate, being inferior to either the average of forecasts from all models (AVG) or the benchmark AR.

For the first half of the forecast horizon, a period with abrupt shifts in the series, AVG delivers the best results for the 121 months rolling window estimation scheme, and for the samples starting in 1993:12 and 1991:01. For the remaining samples, the result depends on what loss function we analyze. DD12 returns the lowest RMSFE and AR the lowest MAFE using the sample starting in 2002:01. Using the sample starting in 1985:01, the minimum RMSFE is attained by Autometrics without any dummy saturation (AUTO), and, starting in 1975:01, by both AdaLASSO and WLAdaLASSO models. AutoIIS returns the lowest MAFE for both of these samples.

For the second half of the forecast horizon, a period with relatively less volatility than the first half, AutoIIS performs better than all other models, except when using the smallest sample, in which the benchmark AR gives a slightly lower RMSFE.

5.1 Diebold and Mariano Tests

Up to this point, we reported the Root Mean Squared Forecast Error and Mean Absolute Forecast Error to measure predictive performance for different models and estimation samples. However, to assess statistical differences in forecast accuracy, we perform, as described in 4.3, the Diebold and Mariano test of equal predictive performance. The analysis is separated in two subsections. The first tests, for each model, if forecasts using different samples are more accurate than the those using the latest series available, *i.e.*, that starting in January 2002. The second tests if, for each sample, any particular model have better predictive performance than the benchmark Autoregressive model of order p. Augmented Dickey-Fuller tests of the loss function differentials used in each performed DM tests are presented in the Appendix.

5.1.1 Comparing Forecasts Using Different Samples

Tables 4 and 5 report Diebold and Mariano (DM) tests with quadratic and absolute loss functions. The null hypothesis is that of equal predictive accuracy between forecasts using a particular sample and the sample starting in January 2002. The one-sided alternative hypothesis is that forecasts using the sample starting in 2002 is less accurate than those using

the other sample. We do not report the results for DD12, since its forecasts are invariant to the sample size.

DM test results clearly shows that the predictions from AR and LSTAR are improved if we use the expanding sample starting in 1993:12 instead of 2002:01, a result that is robust to the loss function used and fraction of the forecast horizon analyzed. The improvement from using the remaining estimation samples, however, is less clear cut — the null is not rejected in the majority of the cases, *i.e.*, depending on which loss function and forecast horizon is being considered.

Using larger samples with the Markov Switching Autoregressive model (MSAR) also tends to produce more accurate forecasts, particularly considering the second half of the forecast horizon. For the first half, however, the DM test using the absolute loss function rejects the null for the samples starting in 1991:01 and 1993:12, but fails to reject the null for all samples using the quadratic loss function.

Results for Lasso related methods are also not robust to the loss function considered. Although the DM test null is rejected for a number of different settings using the quadratic loss function, only WLAdaLASSO predictions with estimation sample beginning in January 1991 is statistically more accurate than those using the latest PIM-PF series at a 5% level of significance with an absolute value loss function of forecast errors.

Considering models applying the Autometrics algorithm, AutoIIS, in particular, shows robust results pointing to improvements in forecast performance from using larger samples — estimation using larger samples tends to more strongly reject the null (note, from Tables 1 and 2, that RMSFE and MAFE tend to decrease with the sample size). Autometrics without any dummy saturation and with both step and impulse indicator saturation (AUTO and AutoSIIS, respectively) produce more erratic results, although one can clearly benefit from using larger samples in a number of settings.

Using the sample starting in 1991:01 and 1993:12 also tends to improve results for the average of forecasts (AVG) in relation to using the sample starting in January 2002.

5.1.2 Comparing Forecasts Using Different Methods

Tables 6 and 7 report DM tests with quadratic and absolute loss functions, respectively. The null hypothesis is that of equal predictive accuracy between forecasts from each method and the benchmark AR model. The one-sided alternative hypothesis is that AR forecasts are less accurate than those produced by other method.

In general, when using the 121 month rolling window estimation scheme and estimation samples smaller than that starting in 1985:01, evaluated methods do not produce more accurate forecasts than those from AR. With a 5% level of significance, the one exception is the average of forecasts from all models considering only the first half of the forecast horizon, using the series starting in January 1991 and the quadratic loss function in the DM test.

Consider the results for estimation using the samples starting in January 1975 and 1985. For the first half of the forecast horizon, the DM test, using both loss functions, points that AVG predictions are more accurate than those from AR. For AutoIIS, the null from the test is rejected only when the quadratic loss function is used to construct the loss function differential series.

In the second half of the forecast horizon, the test points that AutoIIS forecasts are

Table 4: Diebold Mariano Test Comparing Samples: $g(e_{it}) = e_{it}^2$

2007:01 - 2016:12										
	AR	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	AUTO	AutoIIS	AutoSIIS	AVG
121-RW	0.03**	0.19	0.15	0.19	0.10*	0.04**	0.05**	0.02**	0.06*	0.05**
1993:12	0.00**	0.05**	0.05**	0.04**	0.05**	0.02**	0.02**	0.01**	0.00**	0.01**
1991:01	0.05**	0.17	0.07*	0.05**	0.03**	0.01**	0.01**	0.01**	0.02**	0.03**
1985:01	0.31	0.15	0.12	0.19	0.12	0.09*	0.07*	0.01**	0.05**	0.12
1975:01	0.18	0.06*	0.07*	0.10*	0.08*	0.04**	0.03**	0.01**	0.28	0.09*
2007:01 - 2011:12										
	AR	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	AUTO	AutoIIS	AutoSIIS	AVG
121-RW	0.13	0.15	0.37	0.27	0.16	0.11	0.05**	0.03**	0.05**	0.10*
1993:12	0.03**	0.10*	0.14	0.11	0.10*	0.06*	0.06*	0.02**	0.01**	0.06*
1991:01	0.12	0.27	0.15	0.12	0.08*	0.05**	0.03**	0.03**	0.03**	0.08*
1985:01	0.21	0.13	0.18	0.19	0.13	0.11	0.07*	0.02**	0.10*	0.14
1975:01	0.12	0.08*	0.14	0.15	0.12	0.08*	0.07*	0.03**	0.27	0.14
2012:01 - 2016:12										
	AR	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	AUTO	AutoIIS	AutoSIIS	AVG
121-RW	0.07*	0.60	0.02**	0.02**	0.05**	0.01**	0.39	0.08*	0.37	0.07*
1993:12	0.04**	0.03**	0.03**	0.03**	0.06*	0.04**	0.07*	0.04**	0.13	0.02**
1991:01	0.12	0.08*	0.05**	0.01**	0.03**	0.02**	0.10*	0.02**	0.20	0.04**
1985:01	0.74	0.57	0.20	0.47	0.39	0.34	0.44	0.02**	0.11	0.31
1975:01	0.63	0.21	0.09*	0.07*	0.11	0.07*	0.02**	0.01**	0.51	0.06*

Notes: Diebold-Mariano test p-values. $g(e_{it})$ specifies the loss function used in the test. The first column denotes the initial observation of the estimation sample, 121-RW denotes a 121 months rolling window forecasting scheme. Alternative hypothesis is that forecasts from the same method using sample starting in 2002:01 are less accurate than using the sample given by the respective line. Significance at 10% and 5% levels are marked by * and **, respectively.

Table 5: Diebold Mariano Test Comparing Samples: $g(e_{it}) = |e_{it}|$

2007:01 - 2016:12										
	AR	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	AUTO	AutoIIS	AutoSIIS	AVG
121-RW	0.26	0.32	0.01**	0.45	0.52	0.18	0.26	0.01**	0.02**	0.07*
1993:12	0.01**	0.01**	0.00**	0.15	0.35	0.14	0.08*	0.01**	0.00**	0.00**
1991:01	0.24	0.06*	0.01**	0.09*	0.13	0.04**	0.04**	0.01**	0.02**	0.01**
1985:01	0.93	0.42	0.10*	0.40	0.46	0.33	0.15	0.00**	0.04**	0.19
1975:01	0.75	0.11	0.02**	0.16	0.29	0.13	0.04**	0.00**	0.24	0.10*
2007:01 - 2011:12										
	AR	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	AUTO	AutoIIS	AutoSIIS	AVG
121-RW	0.63	0.31	0.08*	0.58	0.57	0.35	0.19	0.04**	0.02**	0.12
1993:12	0.09*	0.04**	0.02**	0.32	0.45	0.25	0.17	0.01**	0.01**	0.02**
1991:01	0.28	0.13	0.05**	0.18	0.20	0.10*	0.06**	0.01**	0.01**	0.03**
1985:01	0.88	0.29	0.15	0.39	0.39	0.28	0.11	0.01**	0.06*	0.16
1975:01	0.67	0.14	0.06*	0.27	0.31	0.17	0.12	0.01**	0.22	0.16
2012:01 - 2016:12										
	AR	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	AUTO	AutoIIS	AutoSIIS	AVG
121-RW	0.07*	0.47	0.02**	0.19	0.38	0.07*	0.58	0.07*	0.22	0.14
1993:12	0.03**	0.05**	0.03**	0.13	0.31	0.18	0.12	0.16	0.24	0.06*
1991:01	0.36	0.15	0.06*	0.11	0.20	0.10*	0.24	0.16	0.31	0.09*
1985:01	0.80	0.66	0.23	0.50	0.65	0.56	0.59	0.12	0.22	0.49
1975:01	0.69	0.29	0.10*	0.14	0.39	0.25	0.06*	0.03**	0.51	0.19

Notes: see notes from Table 4.

Table 6: Diebold Mariano Test Comparing Methods: $g(e_{it}) = e_{it}^2$

	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	DD12	AUTO	AutoIIS	AutoSIIS	AVG
2007:01 - 2016:12										
2002:01	0.95	0.92	0.67	0.64	0.77	0.63	0.80	0.92	0.90	0.32
121-RW	0.97	0.97	0.79	0.52	0.50	0.85	0.64	0.26	0.84	0.19
1993:12	0.98	0.98	0.62	0.56	0.56	0.95	0.62	0.65	0.61	0.14
1991:01	0.94	0.78	0.43	0.25	0.25	0.88	0.22	0.41	0.51	0.07*
1985:01	1.00	0.82	0.36	0.15	0.23	0.81	0.08*	0.01**	0.54	0.01**
1975:01	0.97	0.83	0.21	0.13	0.13	0.90	0.04**	0.01**	0.91	0.01**
2007:01 - 2011:12										
	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	DD12	AUTO	AutoIIS	AutoSIIS	AVG
2002:01	0.91	0.58	0.49	0.54	0.63	0.26	0.78	0.92	0.89	0.29
121-RW	0.83	0.79	0.40	0.24	0.24	0.38	0.39	0.26	0.67	0.07*
1993:12	0.93	0.71	0.34	0.36	0.36	0.55	0.62	0.79	0.44	0.08*
1991:01	0.92	0.34	0.21	0.16	0.16	0.42	0.20	0.63	0.35	0.05**
1985:01	0.95	0.51	0.13	0.09*	0.11	0.41	0.07*	0.14	0.76	0.02**
1975:01	0.94	0.43	0.13	0.10*	0.10*	0.52	0.15	0.20	0.92	0.04**
2012:01 - 2016:12										
	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	DD12	AUTO	AutoIIS	AutoSIIS	AVG
2002:01	0.97	0.98	0.93	0.83	0.95	0.99	0.67	0.55	0.69	0.58
121-RW	0.95	0.98	0.99	0.94	0.96	1.00	0.95	0.45	0.89	0.92
1993:12	0.96	1.00	0.92	0.83	0.83	1.00	0.54	0.27	0.80	0.61
1991:01	0.88	0.98	0.92	0.75	0.75	0.99	0.49	0.12	0.83	0.52
1985:01	0.99	0.98	0.79	0.56	0.67	0.98	0.41	0.02**	0.22	0.16
1975:01	0.92	0.97	0.56	0.44	0.44	0.99	0.08*	0.01**	0.63	0.09*

Notes: Diebold-Mariano test p-values. $g(e_{it})$ specifies the loss function used in the test. The first column denotes the initial observation of the estimation sample, 121-RW denotes a 121 months rolling window forecasting scheme. Alternative hypothesis is that forecasts from the AR model are less accurate than those from the model given by the respective column using the same sample. Significance at 10% and 5% levels are marked by * and **, respectively.

Table 7: Diebold Mariano Test Comparing Methods: $g(e_{it}) = |e_{it}|$

	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	DD12	AUTO	AutoIIS	AutoSIIS	AVG
2007:01 - 2016:12										
2002:01	0.99	1.00	0.89	0.76	0.91	0.95	0.90	0.95	0.99	0.79
121-RW	0.98	0.94	0.98	0.95	0.94	0.97	0.93	0.62	0.85	0.61
1993:12	0.98	0.99	0.98	0.96	0.96	1.00	0.92	0.87	0.82	0.70
1991:01	0.89	0.97	0.82	0.58	0.58	0.98	0.52	0.42	0.85	0.29
1985:01	0.95	0.77	0.54	0.34	0.45	0.77	0.15	0.01**	0.48	0.01**
1975:01	0.86	0.93	0.50	0.37	0.37	0.92	0.13	0.01**	0.93	0.04**
2007:01 - 2011:12										
	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	DD12	AUTO	AutoIIS	AutoSIIS	AVG
2002:01	0.96	0.96	0.58	0.53	0.69	0.75	0.85	0.97	0.99	0.60
121-RW	0.90	0.58	0.63	0.54	0.54	0.70	0.64	0.79	0.68	0.10*
1993:12	0.83	0.90	0.78	0.81	0.81	0.94	0.90	0.92	0.55	0.28
1991:01	0.84	0.83	0.42	0.31	0.31	0.86	0.40	0.72	0.71	0.15
1985:01	0.66	0.15	0.08*	0.06*	0.09*	0.46	0.04**	0.02**	0.53	0.00**
1975:01	0.72	0.29	0.11	0.11	0.11	0.69	0.20	0.05**	0.90	0.01**
2012:01 - 2016:12										
	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	DD12	AUTO	AutoIIS	AutoSIIS	AVG
2002:01	0.94	0.98	0.98	0.90	0.98	0.96	0.79	0.43	0.81	0.86
121-RW	0.96	0.96	1.00	0.99	0.99	0.98	0.97	0.29	0.85	0.97
1993:12	0.97	0.99	0.98	0.96	0.96	0.99	0.77	0.47	0.92	0.92
1991:01	0.78	0.96	0.96	0.83	0.83	0.96	0.65	0.16	0.85	0.68
1985:01	1.00	0.99	0.92	0.79	0.86	0.89	0.59	0.06*	0.44	0.46
1975:01	0.91	1.00	0.85	0.76	0.76	0.94	0.23	0.03**	0.77	0.36

Notes: see notes from Table 6.

statistically more accurate than AR forecasts for both loss functions, with 5% significance level for the largest sample, whereas the DM test for AVG forecasts is unable to reject the null using the quadratic loss function.

For the complete forecast horizon, 2007:01-2016:12, DM test null is rejected for AutoIIS and AVG for both loss functions[14].

5.2 Additional Comments

As a general result, all models tend to benefit, in terms of forecast accuracy, from using chained older PIM-PF series rather than the latest series within the current methodological benchmark, at least up to a point.

Forecasts performed using an expanding estimation sample beginning in 1993:12 show better results than those using a sample beginning in 1991:01. This may be due to the fact that the 1991-1992 period is extremely volatile, with values ranging from -20% to +30% in industrial production 12 month percentage variation, adding noise to the estimation of the latter sample.

Additionally, when estimation applying Autometrics is done with larger series, specially with impulse dummy saturation, starting in January 1975 and 1985, forecast accuracy is improved, whereas performance from other models generally worsens. In fact, AutoIIS performance improves almost monotonically with estimation sample size.

Our results point, however, that Autometrics with impulse dummy saturation may not be the best choice when forecasting in abnormally volatile periods (or in the presence of possible structural breaks), as the 2007:01-2011:12 forecast horizon in our study. Despite still performing relatively well, it fails to reject the Diebold and Mariano test null (with quadratic loss function) of equal predictive ability in relation to those from the AR model. Although forecasts from all models deteriorate in such cases, AdaLASSO/WLAdaLASSO, Autometrics without dummy saturation, and the simple average of all forecasts return equal or better results than Autometrics with impulse indicator saturation (in terms of RMSFE) in this chapter's empirical exercise, being possible alternative candidates.

6 Concluding Remarks

In this chapter, we analyzed how performance in forecasting the Brazilian Industrial Production Index is influenced by the use of samples of different lengths. The complete series, starting in January 1975, contains methodological updates and periods of high inflation, which could cause the weighting structure of the index to be inefficient. Additionally, long series may contain information that is no longer relevant in forecasting future values. On the other hand, using only the current series, starting in January 2002, may harm parameter estimation due to small sample size.

Annually differenced series were used to avoid problems of non-stationarity and seasonality. We separate the forecast horizon in two, 2007:01-2011:12 and 2012:01-2016:12. The first

[14]Test statistic p-value for AVG using the largest sample, starting in 1975:01, is actually 0.07, so the null would only be rejected at a 10% significance level.

is characterized by a great volatility in industrial production, caused by the subprime crisis. The second shows a less abrupt, yet persistent, decrease in industrial activity.

We assessed one-step ahead predictions from different models, re-estimating and re-specifying the models at each step. Besides the benchmark Autoregressive Model of order p and a naive method robust to structural breaks, we analyzed forecasts from nonlinear and data reduction techniques, and also from the simple average from all models. The relative performance between each model and the benchmark autoregressive model is also assessed for each sample.

Results show that Autometrics lag selection with impulse dummy saturation forecasting performance is improved almost monotonically with sample size — the minimum RMSFE and MAFE results, in the 2007:01-2016:12 forecast horizon, are achieved using such method with the largest series available, starting in 1975. The majority of remaining models benefit from expanding the estimation sample beginning at least up to 1993:12. Further enlarging the series may deteriorate forecast performance.

Forecast performance of AR model is superior to all other methods using the estimation sample beginning in January 1991 and smaller samples, and also using a 121 months rolling window (only the average of all forecasts is statistically more accurate for a 5% significance level in a Diebold Mariano test using a quadratic loss function). For estimation starting in January 1975 and 1985, predictions from Autometrics with impulse dummy saturation and the average of forecasts are statistically more accurate than those from AR. However, the average of predictions performs better in the first half of the forecast horizon and Autometrics performs better in the second half.

This hints that, controlling for outliers, industrial production dynamics stays relatively stable, allowing one to benefit from large samples. Using large samples with standard models, however, may impair forecast performance owing to estimation error rather than because of heterogeneity in industrial production index methodology.

References

Arango, L. and Gonzalez, A. (2001). Some evidence of smooth transition nonlinearity in Colombian inflation. *Applied Economics*, 33(2):155–162.

Bontemps, C. and Mizon, G. E. (2003). Congruence and encompassing. In Stigum, B., editor, *Econometrics and the Philosophy of Economics*, pages 354–378. Princeton University Press, New Jersey.

Breusch, T. S. and Pagan, A. R. (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *The Review of Economic Studies*, 47(1):239–253.

Castle, J. L. and Hendry, D. F. (2014). Semi-automatic nonlinear model selection. In Haldrup, N., Meitz, M., and Saikkonen, P., editors, *Essays in Nonlinear Time Series Econometrics*, chapter 7, pages 163–197. Oxford University Press, Oxford.

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28(3):591–605.

Clements, M. and Hendry, D. (1998). *Forecasting Economic Time Series*. Cambridge University Press, Cambridge.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 64(2):247–254.

Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold–mariano tests. *Journal of Business & Economic Statistics*, 33(1):1–1.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Doornik, J. A. (2009). Autometrics. In Castle, J. L. and Shephard, N., editors, *The Methodology and Practice of Econometrics*, pages 88–121. Oxford University Press, Oxford.

Engel, C. (1994). Can the Markov switching model forecast exchange rates? *Journal of International Economics*, 36(1):151–165.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4):987–1007.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Garcia, M. G., Medeiros, M. C., and Vasconcelos, G. F. (2017). Real-time inflation forecasting with high-dimensional models: The case of Brazil. *International Journal of Forecasting*, 33(3):679–693.

Godfrey, L. G. (1978). Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables. *Econometrica*, 46(6):1303–1310.

Góes, M. C. (2005). Transition to an integrated system of business surveys — the Brazilian case. In *International Workshop on Economic Census, Beijing, China*. Available at <https://unstats.un.org/unsd/newsletter/unsd_workshops/country/Brazil_Transition%20to%20an%20integrated%20system%20of%20business%20surveys.pdf>.

Granger, C. W. and Teräsvirta, T. (1993). *Modelling Non-Linear Economic Relationships*. Oxford University Press, Oxford.

Hamilton, J. D. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics*, 45(1):39–70.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, New Jersey.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Hendry, D. F. (1995). *Dynamic Econometrics*. Oxford University Press, Oxford.

Hendry, D. F. (2006). Robustifying forecasts from equilibrium-correction systems. *Journal of Econometrics*, 135(1):399–426.

Hendry, D. F. (2012). Mathematical models and economic forecasting: Some uses and misuses of mathematics in economics. In Dieks, D., Gonzalez, W. J., Hartmann, S., Stöltzner, M., and Weber, M., editors, *Probabilities, Laws, and Structures*, pages 319–335. Springer, New York.

Hendry, D. F. and Johansen, S. (2012). Model discovery and Trygve Haavelmo's legacy. Discussion Paper 598, University of Oxford, Oxford.

Hendry, D. F. and Krolzig, H.-M. (1999). Improving on 'Data mining reconsidered' by K.D. Hoover and S.J. Perez. *The Econometrics Journal*, 2(2):202–219.

Hendry, D. F. and Krolzig, H.-M. (2005). The properties of automatic GETS modelling. *The Economic Journal*, 115(502):C32–C61.

Hendry, D. F. and Mizon, G. E. (2011). Econometric modelling of time series with outlying observations. *Journal of Time Series Econometrics*, 3(1). Article 1.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hoover, K. D. and Perez, S. J. (1999). Data mining reconsidered: encompassing and the general-to-specific approach to specification search. *The Econometrics Journal*, 2(2):167–191.

IBGE (1991). *Indicadores Conjunturais da Indústria: Produção, Emprego e Salário*, volume 11 of *Relatórios Metodológicos*. Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro, 1 edition.

IBGE (1996). *Indicadores Conjunturais da Indústria: Produção, Emprego e Salário*, volume 11 of *Relatórios Metodológicos*. Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro, 2 edition.

IBGE (2003). *Classificação Nacional de Atividades Econômicas - CNAE: Versão 1.0*. Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro.

IBGE (2004). *Indicadores Conjunturais da Indústria: Produção*, volume 31 of *Relatórios Metodológicos*. Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro, 1 edition.

IBGE (2015a). *Classificação Nacional de Atividades Econômicas - CNAE: Versão 2.0*. Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro, 2 edition.

IBGE (2015b). *Indicadores Conjunturais da Indústria: Produção*, volume 31 of *Relatórios Metodológicos*. Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro, 2 edition.

Jarque, C. M. and Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6(3):255–259.

Kim, C.-J. (1994). Dynamic linear models with Markov-switching. *Journal of Econometrics*, 60(1):1–22.

Konzen, E. and Ziegelmann, F. A. (2016). Lasso-type penalties for covariate selection and forecasting in time series. *Journal of Forecasting*, 35(7):592–612.

Lovell, M. C. (1983). Data mining. *The Review of Economics and Statistics*, 65(1):1–12.

Lundbergh, S. and Teräsvirta, T. (2005). Forecasting with smooth transition autoregressive models. In Clements, M. P. and Hendry, D. F., editors, *A Companion to Economic Forecasting*, pages 485–509. Wiley-Blackwell, New Jersey.

Luukkonen, R., Saikkonen, P., and Teräsvirta, T. (1988). Testing linearity against smooth transition autoregressive models. *Biometrika*, 75(3):491–499.

Medeiros, M. C. and Mendes, E. F. (2016). L1-regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors. *Journal of Econometrics*, 191(1):255–271.

Medeiros, M. C. and Vasconcelos, G. F. (2016). Forecasting macroeconomic variables in data-rich environments. *Economics Letters*, 138:50–52.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Öcal, N. and Osborn, D. R. (2000). Business cycle non-linearities in UK consumption and production. *Journal of Applied Econometrics*, 15(1):27–43.

Park, H. and Sakaori, F. (2013). Lag weighted lasso for time series model. *Computational Statistics*, 28(2):493–504.

Prince, D. d. and Marçal, E. F. (2018). Selection of lags and univariate models to forecast industrial production in an emerging country: is disagreggation useful? In *46º Encontro Nacional de Economia - ANPEC, Rio de Janeiro, Brazil*.

Santos, C., Hendry, D. F., and Johansen, S. (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics*, 23(2):317–335.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Skalin, J. and Teräsvirta, T. (2002). Modeling asymmetries and moving equilibria in unemployment rates. *Macroeconomic Dynamics*, 6(02):202–241.

Teräsvirta, T. (1994). Specification, estimation, and evaluation of Smooth Transition Autoregressive Models. *Journal of the American Statistical Association*, 89(425):208–218.

Teräsvirta, T. and Anderson, H. M. (1992). Characterizing nonlinearities in business cycles using smooth transition autoregressive models. *Journal of Applied Econometrics*, 7(S1):S119–S136.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Wang, H., Li, G., and Tsai, C.-L. (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(1):63–78.

Wooldridge, J. M. (1994). Estimation and inference for dependent processes. In Engle, R. F. and McFadden, D. L., editors, *Handbook of Econometrics*, volume 4, pages 2639–2738. Elsevier, Amsterdam.

Zhang, Y., Li, R., and Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105(489):312–323.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(Nov):2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H., Hastie, T., Tibshirani, R., et al. (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192.

Appendix

Unit Root Tests for Loss Function Differentials

Tables A1-A4 show the p-value of Augmented Dickey-Fuller tests for unit roots for the loss function differential series used in the Diebold and Mariano tests presented in Tables 4, 5, 6 and 7 in Section 5, comparing sample and methods with quadratic and absolute loss functions.

Table A1: Augmented Dickey-Fuller Test Comparing Samples: $g(e_{it}) = e_{it}^2$.

	2007:01-2016:12									
	AR	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	AUTO	AutoIIS	AutoSIIS	AVG
121-RW	0.03	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02
1993:12	0.01	0.03	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02
1991:01	0.01	0.01	0.01	0.01	0.01	0.01	0.04	0.01	0.01	0.01
1985:01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02
1975:01	0.01	0.04	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Notes: ADF tests include constant and trend. $g(e_{it})$ specifies the loss function used in the DM test. The first column denotes the initial observation of the estimation sample, 121-RW denotes a 121 months rolling window forecasting scheme. Minimum p-value reported is 0.01.

Table A2: Augmented Dickey-Fuller Test Comparing Samples: $g(e_{it}) = |e_{it}|$.

	2007:01-2016:12									
	AR	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	AUTO	AutoIIS	AutoSIIS	AVG
121-RW	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
1993:12	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
1991:01	0.01	0.01	0.01	0.01	0.01	0.01	0.10	0.01	0.01	0.01
1985:01	0.03	0.01	0.04	0.01	0.01	0.01	0.04	0.01	0.01	0.01
1975:01	0.02	0.01	0.01	0.01	0.01	0.01	0.04	0.01	0.01	0.01

Notes: see notes from Table A1.

Table A3: Augmented Dickey-Fuller Test Comparing Methods: $g(e_{it}) = e_{it}^2$.

	2007:01-2016:12									
	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	DD12	AUTO	AutoIIS	AutoSIIS	AVG
2002:01	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
121-RW	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
1993:12	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
1991:01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
1985:01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
1975:01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Notes: see notes from Table A1.

Table A4: Augmented Dickey-Fuller Test Comparing Methods: $g(e_{it}) = |e_{it}|$.

	2007:01-2016:12									
	LSTAR	MSAR	LASSO	AdaLASSO	WLAdaLASSO	DD12	AUTO	AutoIIS	AutoSIIS	AVG
2002:01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
121-RW	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
1993:12	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
1991:01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
1985:01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
1975:01	0.03	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Notes: see notes from Table A1.

Plots of the Forecasts

Figures A1-A6 plot the point forecasts of all models for the different estimation samples considered. The average of all forecasts is represented by the red lines.

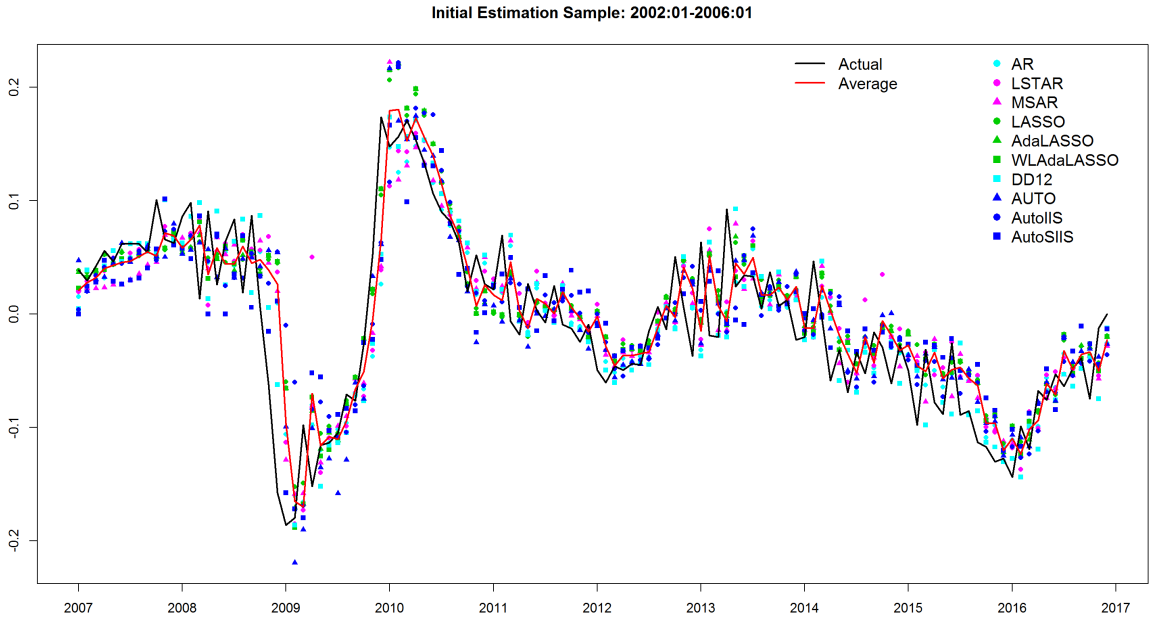

Figure A1: Point forecasts for all models for the estimation samples starting in 2002:01.

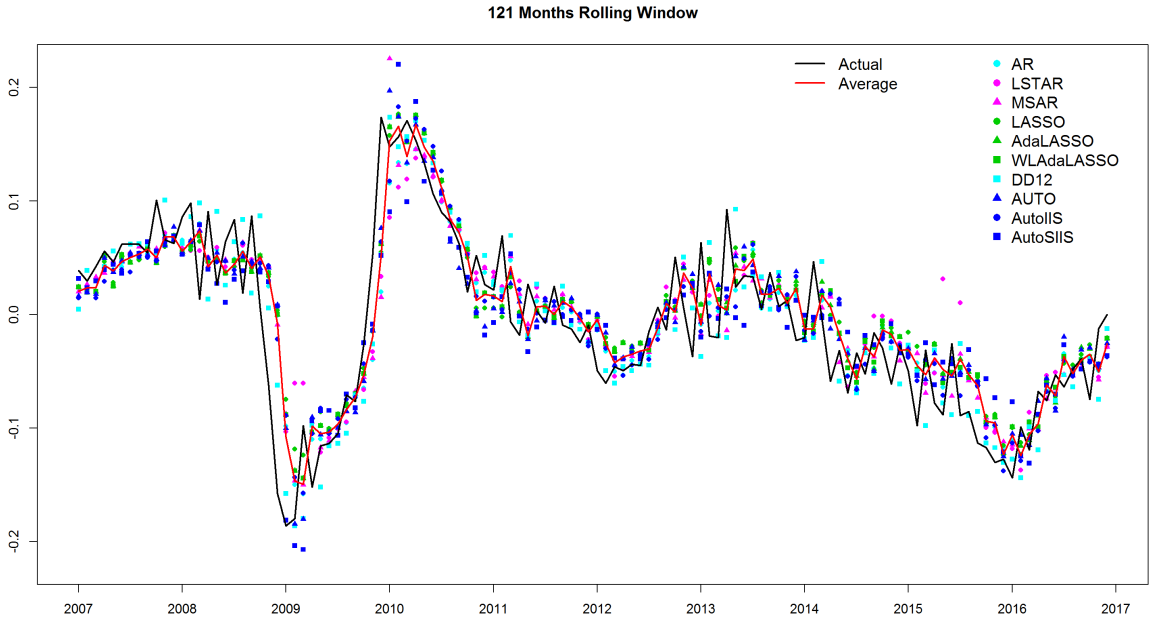

Figure A2: Point forecasts for all models for the 121 months rolling window forecasting scheme.

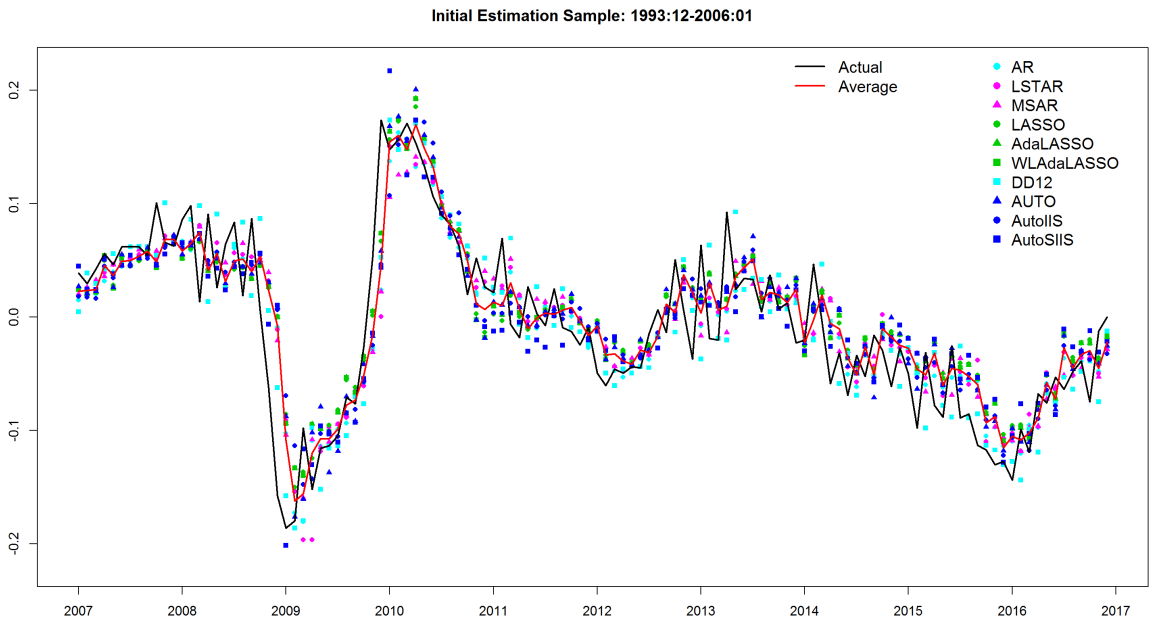

Figure A3: Point forecasts for all models for the estimation samples starting in 1993:12.

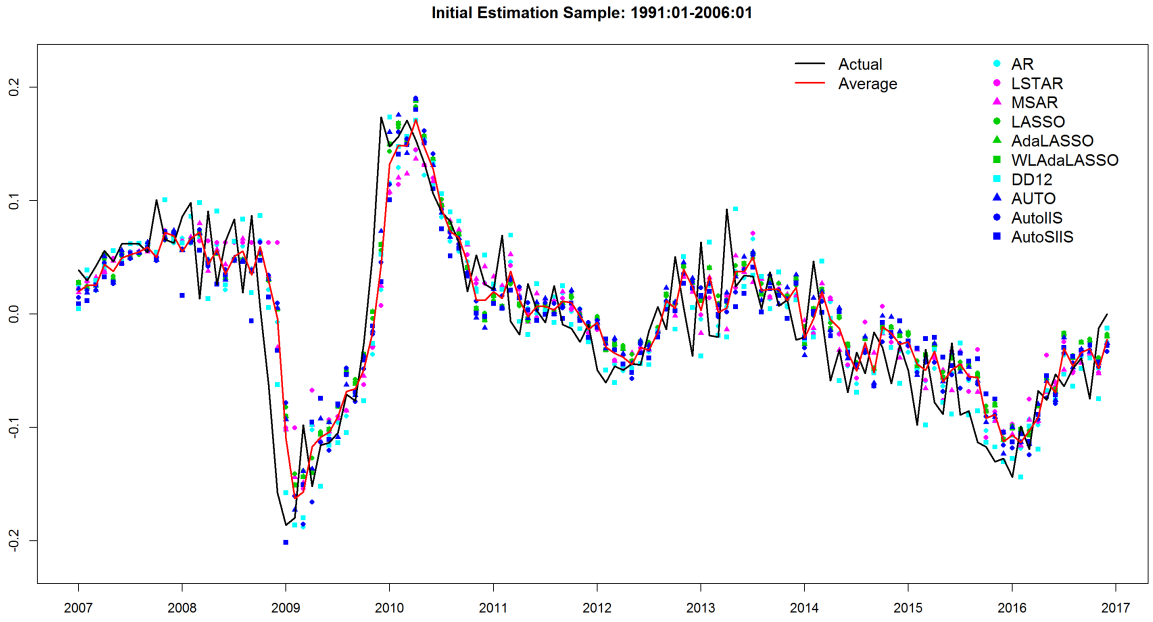

Figure A4: Point forecasts for all models for the estimation samples starting in 1991:01.

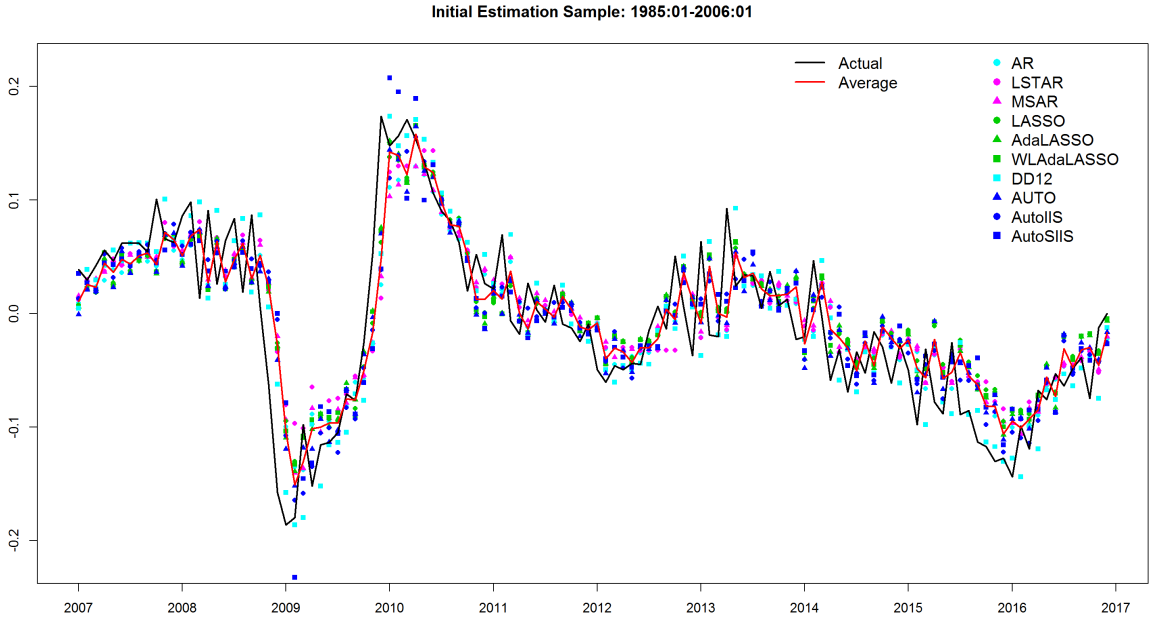

Figure A5: Point forecasts for all models for the estimation samples starting in 1985:01.

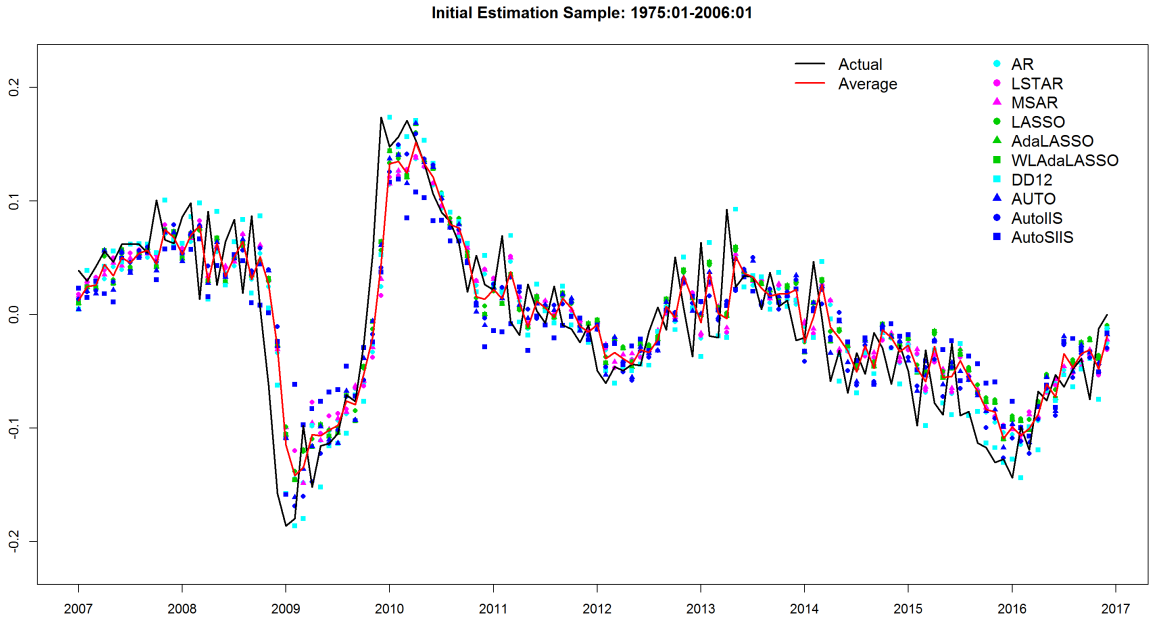

Figure A6: Point forecasts for all models for the estimation samples starting in 1975:01.

RMSFE, MAFE and MFE Results

Table A5 provides the medians plotted in Figure 6.

Table A5: Median Values of RMSFE, MAFE and MFE by Estimation Sample

	2007:01 - 2016:12			2007:01 - 2011:12			2012:01 - 2016:12		
	RMSFE	MAFE	MFE	RMSFE	MAFE	MFE	RMSFE	MAFE	MFE
2002:01	4.25	3.16	0.64	4.65	3.28	0.31	3.65	2.99	0.97
121-RW	3.95	3.06	0.44	4.34	3.20	0.16	3.51	2.91	0.94*
1993:12	3.82	2.93	0.50	4.25	3.08	-0.02*	3.40*	2.88*	1.10
1991:01	3.85	2.90*	0.54	4.25	3.01*	0.02*	3.40*	2.88*	1.10
1985:01	3.94	3.09	0.44	4.12	3.08	-0.23	3.61	3.03	1.09
1975:01	3.81*	3.00	0.38*	4.12*	3.03	-0.05	3.53	2.95	0.95

Notes: Median values for RMSFE, MAFE and MFE in Tables 1, 2 and 3. The first column denotes the initial observation of the estimation sample, 121-RW denotes a 121 months rolling window forecasting scheme. The first row marks the forecast horizon evaluated. Lowest absolute values in each column are marked by *.