FUNDAÇÃO GETULIO VARGAS

ESCOLA DE ECONOMIA DE SÃO PAULO

PEDRO MONTERO MATTOS

NOWCASTING BRAZILIAN GDP

SÃO PAULO

2017

PEDRO MONTERO MATTOS

NOWCASTING BRAZILIAN GDP

Dissertação apresentada à Escola de Economia de São Paulo como requisito à obtenção do título de Mestre em Economia.

Campo de conhecimento:

Economia, Econometria.

Orientador:

Prof. Dr. Ricardo Masini

SÃO PAULO

2017

Mattos, Pedro Montero.
Nowcasting Brazilian GDP / Pedro Montero Mattos. - 2017.
Digite aqui o total de folhas f.

Orientador: Ricardo Masini
Dissertação (MPFE) - Escola de Economia de São Paulo.

1. Previsão econômica - Brasil. 2. Brasil - Condições econômicas. 3. Modelos não-lineares (Estatística). 4. Produto interno bruto - Brasil. 5. Macroeconomia. I. Masini, Ricardo. II. Dissertação (MPFE) - Escola de Economia de São Paulo. III. Título.

CDU 330.55(81)

PEDRO MONTERO MATTOS

NOWCASTING BRAZILIAN GDP

Dissertação apresentada à Escola de Economia de São Paulo como requisito à obtenção do título de Mestre em Economia.

Campo de conhecimento:

Economia, Econometria.

Data de aprovação:

___/____/_____

Banca examinadora:

____

Prof. Dr. Ricardo Masini (Orientador)

EESP – FGV

____

Prof. Dr. Emerson Marçal

EESP – FGV

____

Prof. Dr. Gustavo Soares

INSPER

Este trabalho é dedicado à minha família, namorada e amigos que me apoiaram na jornada, aos meus educadores do curso e anteriores que me deram condições de confeccionar o presente trabalho e aos meus amigos de turma, sem os quais o curso não teria sido tão proveitoso.

Acknowledgements

Os agradecimentos principais são direcionados à Ricardo Masini, Gustavos Soares e Émerson Marçal, professores do mestrado que me orientaram e deram base para confeccionar o presente trabalho.

Agradecimentos especiais são direcionados à minha família, namorada e amigos. O apoio deles foi fundamental para a conclusão do curso. Em especial meu primo Daniel que também fez o curso e teve muita paciência com as minhas discussões.

“He remarks that, while the individual man is an insoluble puzzle„
in the aggregate he becomes a mathematical certainty.
You can, for example, never foretell what any one man will do,
but you can say with precision what an average number will be up to.
Individuals vary, but percentages remain constant. So says the statistician.
(The Sign of Four - Sherlock Holmes, by Arthur Conan Doyle, about Winwood Reade)

Resumo

Baseado em recentes pesquisas em métodos de Nowcasting, foi aplicada a estimação de modelos de fatores dinâmicos em um passo ao caso brasileiro. Esta metodologia lida com os problemas de frequências mistas, amostras recortadas, horizonte temporal e alta dimensão da amostra. Foram utilizadas as expectativas diárias do PIB publicadas pelo Banco Central como um benchmark do modelo. Não foram encontradas evidências que rejeitam a hipótese de igual poder preditivo, para circunstâncias econômicas não estressadas.

Palavras-chave: Nowcasting. Forecasting. Dynamic Factor Models

Abstract

Based on recent surveys on nowcasting methods, we apply the one-step estimation of dynamic factor models to the Brazilian case. Such methodology copes well with the problems of mixed-frequency series, ragged edges, timeliness and high dimensionality of data sets. We use the daily expectation published by the Brazilian Central Bank as a benchmark for our model and we do not find enough evidence to reject that both models have equal predictive accuracy, under non-distressed circumstances.

Keywords: Nowcasting. Forecasting. Dynamic Factor Models.

List of Figures

List of Tables

List of abbreviations and acronyms

EM	Expectation Maximization
GDP	Gross Domestic Product
RMSFE	Root-Mean-Squared-Forecast-Error
YoY	Year over year

List of symbols

Ω	Information set
ν	Denotes the date in which the last piece of data in a give information set was published
Y_t	Vector of possibly observable variables at time t
y_t	Scalar of a possibly observable variable at time t
y_t^M	Vector of a possibly observable monthly variables at time t
y_t	Scalar of the standardized GDP variable at time t
k	Indicates the frequency, in months, at which a given variable is published
N	Total number of possibly observable variables
μ	Vector of constants
Λ	Matrix of Factor Loadings
F_t	Vector of unobserved factors at time t
E_t	Vector of measurement equation's disturbances
Σ_E	Covariance matrix of measurement equation's disturbances
Φ	Matrix of unobservable factors' auto-regressive coefficients.
U_t	Vector of transition equation's disturbances
Σ_U	Covariance matrix of transition equation's disturbances
$L()$	Likelihood function
$E()$	Expectation operator
$\theta()$	Parameters of a given iteration
r	Number of unobserved factors

Contents

1 Introduction

Monetary policy, fiscal policy and investment decisions are made by major players in the economy using both real-time information about key macroeconomic variables and expectation about those same variables. Thus, the issue of efficiently using such information to form expectations about macroeconomic variables arises.

Nowcasting is a contraction of the terms *Now* and *Forecasting*, implying that it attempts to "forecast" or estimate the present. Giannone, Reichlin and Small (2008) define it as:

> *The prediction of the present, the very near future and the very recent past.*

It differs from forecasting in the sense that it focusses only on the next point in time to be forecast. According to Banbura et al. (2013) the current quarter is the only one in which models produce significant gains relative to naive constant growth models for the GDP, thus motivating the focus on nowcasting. Moreover, nowcasting deals with problems not usually dealt with by models attempting to generate long term forecasts, at least simultaneously. Among the problems that will be discussed here are ragged edges, mixed frequencies, timeliness and high dimensionality of data sets.

Relevance of nowcasting is given by the fact that key economic variables are published with significant delay (GDP and inflation rates being notable examples), but investment and policy decisions have to be made in real-time.

Policy institutions have dealt with the problem of nowcasting for some time and their approach has evolved from simpler bridge equations (Baffigi, Golinelli and Parigi (2004) to dynamic factor models (Banbura et al. (2013)). This paper applies the former approach to the Brazilian case.

This methodology has the advantage of incorporating real-time information to the nowcasting process, despite its (or lack thereof) timeliness. Thus coping with ragged edges, mixed frequency and high dimensionality issues in a proper way.

Opposed to most of the world's economies' general framework, the Central Bank of Brazil (BCB) provides a rare framework in which daily forecasts made by professional forecasters are published. The system is called Market Expectation System, in which economic players report their expectations for key macroeconomic variables.

This paper compares the nowcasting methodology to the expectations provided by the BCB in nowcasting the Brazilian GDP. We find that for non-distressed conditions there is no statistical difference between the nowcasts made by the model and reported by the BCB. Although, during political crisis, the model is not able to efficiently capture the economic conditions and anticipate surprises as efficiently as the expectations reported.

This paper is organized as follows. Section 2 discusses the problem of nowcasting and past approaches to it. Section 3 reviews the methodology used in the empirical part in depth. Section 4 reviews the data used and the transformations applied to it. Section 5 presents empirical results. Section 6 concludes.

Part I

Theoretical Basis

2 Literature and Methodology Review

Nowcasting is the process of obtaining expectations about variables of interest based on expading information sets. Adopting the notation proposed in Banbura et al. (2013), such an information set might be called Ω_ν, where ν denotes a particular data release, (and therefore Ω_ν contains that release and all past releases). Due to the nature of economic variables, ν is usually not equally spaced.

Additionally, for a variable y, y_t^k denotes the observation for period t for a variable collected at an interval of k periods, and $k = 1$ is the higher frequency among data, at which the model is defined. In the case where $k = 1$ the superscript is omitted for simplicity. The vector of N variables observed at different frequencies is denoted by $Y_t^{K_y} = (y_{t,1}^{k_1}, y_{t,2}^{k_2}, ..., y_{t,N}^{k_N})'$, and the vector that denotes the high-frequency (possibly unobserved) counterpart of those variables is denoted by $Y_t = (y_{t,1}, y_{t,2}, ..., y_{t,N})'$

Such an information set, might contain data collected at different frequencies and with different publication lags. That conjuncture leads to what is known as a "ragged edge" (see Giannone, Reichlin and Small (2008)). In order to use all available information, the model has to cope with the fact that, more often than not, a sample's end is different for the different time series. Also, this information set can be very large, leading to the issue known as "high dimensionality".

At first, Central Banks tackled the problem by the use of Bridge Equations, as defined by Clements and Galvão (2008) and Kuzin, Marcellino and Schumacher (2011). One application by the European Central Bank can be found in Baffigi, Golinelli and Parigi (2004). Bridge Equations are essentially linear regression models that estimate the GDP growth, aggregating monthly macroeconomic variables to quarterly frequency. Whenever there is no available data for one of the explanatory variables, ARIMA models are used to generate expectations about them. In case there is a wide variety of explanatory variables available, pooling is an option to make estimation possible, as described by Kitchen and Monaco (2003).

Another approach is known as MIDAS type equations. Defined in Clements and Galvão (2009) and Kuzin, Marcellino and Schumacher (2011), MIDAS type equations do not rely on hard temporal aggregation. Instead, the predictors are included in their original frequency, and the aggregation weights are endogenous in a data-driven fashion. Usually these weights are calculated by a parameterised polynomial as proposed by the previously mentioned authors or, alternatively, as proposed by Ghysels and Wright (2009).

Bridge Equations and MIDAS Equations are not able to properly handle wide data sets, mixed frequency series and do not take advantage of timeliness of data nor provide good estimations to deal with ragged edges.

Another approach to nowcasting is the estimation of a Mixed-Frequency VAR as defined in Giannone, Reichlin and Simonelli (2009) and Kuzin, Marcellino and Schumacher (2011). Through the Kalman Filter estimation, a Mixed-Frequency VAR is able to provide expectations for the missing points of data of lower frequency series, thus coping with the problem of different frequencies, timeliness and ragged edge. Moreover, those expectations are estimated in a one step fashion, fostering consistency. Applications can be found in the previously mentioned papers, and earlier in Zadrozny (1990) and Mittnik and Zadrozny (2004).

Lastly, since Mixed Frequency VARs cannot handle wide (or High Dimensional) data sets preserving its consistency characteristics, Factor Models estimation is in order. The framework provided in Giannone, Reichlin and Small (2008) and Evans (2005) consists of estimating Factor Models, in a state space representation, by the means of using a Kalman Filter. As stated by Banbura et al. (2013):

> *The framework in Giannone, Reichlin and Small (2008) is capable of handling a high-dimensional problem. That framework exploits the fact that some co-moving series can be captured by a few common factors. All the variables in the information set are assumed to be generated by a dynamic factor model (which copes with the 'curse of dimensionality').*

Dynamic factor models are fit for wide Macroeconomic data sets. Broad forecasting applications and theoretical background can be found in Sargent, Sims et al. (1977); Giannone, Reichlin and Sala (2005); Watson (2004) and Stock and Watson (2011).

The estimation procedure in the Giannone, Reichlin and Small (2008) follows a PCA-based two-step approach described by Doz, Giannone and Reichlin (2011), which consists in first estimating the state space parameters using only the balanced portion of the sample, and afterwards applying a Kalman Filter to the entire data set.

Banbura and Modugno (2010) propose a one-step estimation based on a maximum likelihood procedure, for which Doz, Giannone and Reichlin (2012) have demonstrated consistency and robustness properties for large enough data sets. The algorithm used is named Expectation Maximization (EM), which allows to write the likelihood function in terms of observed and non observed data. Since, by the use of Monte Carlo simulations, Doz, Giannone and Reichlin (2012) established consistency and robustness to misspecification of the model for as few as five variables, the approach is fit for the present paper. Moreover, the EM algorithm is more efficient in small systems, when compared to the two-step approach and allows for imposing restriction on the parameters. In the author's view this approach represents the current state of art in the matter.

Banbura et al. (2013), applying the abovementioned approach, conclude that nowcasting models gains, compared to naive constant growth models are only relevant for the current quarter for the US GDP; That there is no significant difference between nowcasts and institutional forecasts; and that the exploitation of timely data leads to accuracy gains for the nowcast.

3 Methodology

Following Banbura et al. (2013), we estimate a dynamic factor model in a state space representation. A factor model takes advantage of the co-movements among the observable variables and thus copes well with the "curse of dimensionality". Moreover, its estimation by the means of the EM Algorithm plus the Kalman Filter deals elegantly with missing points of data, by producing consistent conditional expectation for those. The EM Algorithm is a Maximum Likelihood estimation procedure. The procedure's consistency is shown in Doz, Giannone and Reichlin (2012).

Factor models, generally, are specified by measurement equations, which links the observable variables to the unobservable factors; and transition equations, specifying the unobservable factors' dynamics.

Following Banbura et al. (2013) notation, the measurement equation is given by (3.1) and the transition equation is given by (3.2).

$$Y_t = \mu + \Lambda F_t + E_t, \qquad E_t \sim i.i.d.N(0, \Sigma_E), \tag{3.1}$$

$$F_t = \Phi(L) F_t + U_t, \qquad U_t \sim i.i.d.N(0, \Sigma_U), \tag{3.2}$$

Where Σ_E is assumed to be diagonal, but Doz, Giannone and Reichlin (2012) show that the model is robust to violations of this assumption for large sample sizes both across time and cross-section; and Φ is a lag function.

Formally, the nowcast of the GDP is the orthogonal projection of Y_t on the available information set Ω_ν.

Y_t consists of all the observable data set. We define our model on a monthly frequency. Since the GDP is a quarterly variable, the data set has structural missing values.

3.1 Incorporating the GDP

In order to account for the fact that the quarterly GDP is directly dependent on the three months that compose the quarter, we need to force constraints. To do so, following what is done in Banbura et al. (2013), Banbura and Modugno (2010), Kuzin, Marcellino and Schumacher (2011) and Bragoli, Metelli and Modugno (2014), we adopt the approximation defined in Mariano and Murasawa (2003).

They define that for log-differenced flow variables (applicable to the YoY GDP), in case y_t^k is a differenced version of z_t^k:

$$y_t^k = \log(z_t^k) - \log z_{t-k}^k = \log(\sum_{i=0}^{k-1} z_{t-i}) - \log(\sum_{i=k}^{2k-1} z_{t-i}) \approx \sum_{i=0}^{k-1} \log(z_{t-i}) - \sum_{i=k}^{2k-1} \log(z_{t-i})$$
$$= \sum_{i=0}^{2k-2} \omega_i^{k,f} y_{t-i} \qquad t = k, 2k, ...,$$

Where $y_t = \Delta \log(z_t)$; $\omega_i^{k,f} = i + 1$ for $i = 0, ..., k-1$; $\omega_i^{k,f} = 2k - i - 1$ for $i = k, ..., 2k-2$; $\omega_i^{k,f} = 0$ otherwise.

The approximation allows to keep the observational constraints stemming from the temporal aggregation linear. The quarterly GDP becomes:

$$y_t^3 = y_t + 2y_{t-1} + 3y_{t-2} + 2y_{t-3} + y_{t-4}$$

3.2 Temporal Aggregation

Since the model is defined at a monthly frequency and some data series are daily, the issue of temporal aggregation arises. The aggregation methology is described below.

Let z_t be the high frequency counterpart of z_t^k, where k is the frequency period and t is the reference date.

For stock variables we have:

$$z_t^k = z_t \qquad t = k, 2k, ...,$$

Whereas for flow variables we have:

$$z_t^k = \sum_{i=0}^{k-1} z_{t-i}, \qquad t = k, 2k, ...,$$

3.3 Number of Factors

We chose to use a single unobserved factor with a first-order auto-regressive form following what is done in Banbura et al. (2013) and Bragoli, Metelli and Modugno (2014). The latter have applied the Bai and Ng (2002) Information criteria and Akaike to select the number of factors and lags and came to the conclusion that one factor and one lag is optimal.

3.4 Estimation

Doz, Giannone and Reichlin (2012) show that large systems like (3.1)-(3.2) can be estimated by maximum likelihood. We adopt their methodology in which the Expectation Maximization (EM) Algorithm is used.

The algorithm basically iterates between two steps: (i) The expectation step in which the log-likelihood conditional expectation is calculated based on the data using the parameter estimates from the previous step; and (ii) maximizing the log-likelihood to obtain the parameter's new estimation. The procedure is repeated until convergence is achieved.

According to Banbura et al. (2013):

> *Maximum likelihood has a number of advantages compared to the principal components and the two-step procedure. First it is more efficient for small systems. Second, it allows to deal flexibly with missing observations. Third, it is possible to impose restrictions on the parameters. For example, Banbura and Modugno (2010) impose the restrictions on the loadings to reflect the temporal aggregation. Banbura, Giannone, and Reichlin (2011) introduce factors that are specific to groups of variables.*

Doz, Giannone and Reichlin (2012) show that the estimation is consistent and robust to omitted serial and cross-sectional correlation of the idiosyncratic components and non-normality.

Considering only one unobserved factor in the system (3.1)-(3.2) we would have our model parameters $\theta = (\mu, \Lambda, \Phi, \Sigma_E, \Sigma_U)$ where Σ_E is diagonal.

If some initial estimate of the parameters $\theta(0)$ is given, the algorithm would proceed as follows:

$$E - step : \qquad L(\theta, \theta(j)) = E_{\theta(j)}[l(Y, F; \theta)|\Omega_\nu],$$

$$M - step : \qquad \theta(j+1) = argmax_\theta L(\theta, \theta(j)),$$

The EM Algorithm's parameters $\theta(0)$ are initialized to the values obtained from regressing factors obtained by PCA on the actual data. Missing values are replaced with zeros. That should only speed up the EM convergence and should have no effect on final values.

Further details on the EM algorithm can be found in appendix A.

3.5 State-Space Representation

To represent the (3.1)-(3.2) system in a state space form, we adopt the following specifications:

$$\begin{bmatrix} y_t^M \\ y_t^{GDP} \end{bmatrix} = \begin{bmatrix} \Lambda_M & 0 & 0 & 0 & 0 & I_n & 0 & 0 & 0 & 0 & 0 \\ \Lambda_Q & 2\Lambda_Q & 3\Lambda_Q & 2\Lambda_Q & \Lambda_Q & 0 & 1 & 2 & 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} f_t \\ f_{t-1} \\ f_{t-2} \\ f_{t-3} \\ f_{t-4} \\ \epsilon_t^M \\ \epsilon_t^Q \\ \epsilon_{t-1}^Q \\ \epsilon_{t-2}^Q \\ \epsilon_{t-3}^Q \\ \epsilon_{t-4}^Q \end{bmatrix} + \begin{bmatrix} \xi_t^M \\ \xi_t^Q \end{bmatrix}$$

$$\begin{bmatrix} f_t \\ f_{t-1} \\ f_{t-2} \\ f_{t-3} \\ f_{t-4} \\ \epsilon_t^M \\ \epsilon_t^Q \\ \epsilon_{t-1}^Q \\ \epsilon_{t-2}^Q \\ \epsilon_{t-3}^Q \\ \epsilon_{t-4}^Q \end{bmatrix} = \begin{bmatrix} A_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ I_r & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & I_r & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I_r & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I_r & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \alpha_M & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \alpha_Q & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} f_{t-1} \\ f_{t-2} \\ f_{t-3} \\ f_{t-4} \\ f_{t-5} \\ \epsilon_{t-1}^M \\ \epsilon_{t-1}^Q \\ \epsilon_{t-2}^Q \\ \epsilon_{t-3}^Q \\ \epsilon_{t-4}^Q \\ \epsilon_{t-5}^Q \end{bmatrix} + \begin{bmatrix} u_t \\ 0 \\ 0 \\ 0 \\ 0 \\ e_t^M \\ e_t^Q \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Where Λ_M denotes the factor loadings for the monthly data (y_t^M) and Λ_Q denotes the factor loading for the GDP data y_t^{GDP}; $\alpha_M = diag(\alpha_1, ...\alpha_{n_M})$ are the auto regressive coefficient of the idiosyncratic component of monthly data; n_M is the number of monthly series used in the model; and α_Q is a scalar auto regressive coefficient of ϵ_t^Q; r is the number of factors; and ξ_t^M and ξ_t^Q have fixed variances. The specified form accounts for the aggregation described in section 3.1.

3.6 Software

The econometric software used was developed by the author using Python open-source programming language. The widely-known third-party packages *pandas* and *numpy* were used for data manipulation.

Part II

Empirical Application

4 Data

We gathered 21 time series, comprised of annual, quarterly, monthly and daily series. Among them, macroeconomic variables, e.g. the GDP itself; indexes, e.g. the Consumer Price Index (IPCA); and financial variables, e.g. Crude Oil Prices. The data was gathered from *Bloomberg*.

The series are demeaned and standardized to unitary variance. Also, due to estimation methods' limitations, we had to apply logarithmic and differentiation transformations to induce stationarity. These transformations, the sources, the publication delay, and the series types, are described in table 1.

Due to the impossibility of obtaining actual publication dates, we use a stylized publication calendar, as discussed in section 4.2.

Table 1 – Data Summary

	Name	**Frequency**	**Delay**	**Bloomberg Code**	**Type**	**log**	**diff**
1	GDP YOY %	Quarterly	60	BZGDYOY% Index	Flow	FALSE	FALSE
2	Markit PMI Manufacturing SA	Monthly	14	MPMIBRMA Index	Stock	TRUE	TRUE
3	Housing Starts	Monthly	50	BZREELHT Index	Stock	TRUE	TRUE
4	Consumer Price Index, All Urban Consumers	Monthly	10	BZPIIPCM Index	Flow	FALSE	FALSE
5	Merchandise Exports	Monthly	2	BZEXTOT$ Index	Stock	TRUE	TRUE
6	Merchandise Imports	Monthly	2	BZTBBALY INDEX	Stock	TRUE	TRUE
7	Philadelphia Fed Survey, General Business Conditions	Monthly	13	BZBXPBCM INDEX	Stock	TRUE	TRUE
8	Retail and Food Services Sales	Monthly	12	BZRTFBSA INDEX	Stock	TRUE	TRUE
9	Conference Board Consumer Confidence	Monthly	29	BZFGCCSA INDEX	Stock	TRUE	TRUE
10	S&P 500 Index	Daily	1	IBOV INDEX	Stock	TRUE	TRUE
11	Crude Oil, West Texas Intermediate (WTI)	Daily	1	CL1 COMDTY	Stock	TRUE	TRUE
12	10-Year Treasury Constant Maturity Rate	Daily	1	BCSWLPD CURNCY	Stock	TRUE	TRUE
13	3-Month Treasury Bill, Secondary Market Rate	Daily	1	BCSWFPD CURNCY	Stock	TRUE	TRUE
14	Trade Weighted Exchange Index, Major Currencies	Monthly	17	BZMOTRFB INDEX	Flow	FALSE	FALSE
15	Economic Activity Index	Monthly	45	BZEASA INDEX	Stock	TRUE	TRUE
16	Retail Trade: Volume	Monthly	25	OEBRD003 INDEX	Stock	TRUE	TRUE
17	Real Industrial Production SA	Monthly	21	BZIPTLSA Index	Stock	TRUE	TRUE
18	Brazil Formal Employment SA	Monthly	38	BFOETTSA Index	Stock	TRUE	TRUE
19	YoY CPI IPCA	Monthly	8	BZPIIPCY Index	Stock	TRUE	TRUE
20	Brazil Retail Sales Volume MoM SA	Monthly	25	BZRTRETM Index	Flow	FALSE	FALSE
21	Brazil Government Registered Job Creation Total SA	Monthly	41	BZJCTOTS Index	Flow	FALSE	FALSE

4.1 The BCB Survey

In 1999 in order to foster transparency and information access the Brazilian Central Bank created a system called Market Expectation System in which daily expectations of key macroeconomic variables are provided.

The system is comprised of a web interface in which economic players inform their expectation about those variables. Those players are financial institutions, consulting firms and universities which are required to have a specialized team to produce those macroeconomic forecasts. According to Bragoli, Metelli and Modugno (2014) even though the process through which these institutions produce their forecasts is not clear, it is reasonable to assume that these predictions are not entirely model based.

At 5 pm everyday, the expectations are consolidated and the data's average, median, standard deviations, minimum and maximum values are published. For the purpose of benchmarking our model we use the median, following what is done in Bragoli, Metelli and Modugno (2014).

The daily series of expectations about the GDP is used to benchmark our model.

4.2 *Pseudo* Real-Time Calendar

Due to the impossibility of obtaining actual publication dates, we use a stylized publication calendar following what is done in Banbura et al. (2013), Giannone, Reichlin and Small (2008), Giannone, Reichlin and Simonelli (2009) and others. This *pseudo* real-time calendar is created using the publication delays in the arbitrarily chosen year of 2013. Few studies for the US market were conceived with an actual real-time calendar, including Camacho and Perez-Quiros (2010), Lahiri and Monokroussos (2013), Liebermann (2012) and Siliverstovs (2012).

We stress here that if the stylized calendar is a monotonic transformation of the real time calendar, no impact is expected. I.e., if the order of releases is respected, the information set expands in the same fashion, thus producing numerically identical results for equivalent vintages.

However, it is of interest to produce a realistic enough calendar since that (i) fosters readability of results in a familiar measure, i.e, actual days; and (ii) allows for precise comparison with other nowcasts (such as institutional), since for a valid comparison we need a precise estimation of the available information set at certain points in time.

5 Results

In order to evaluate the model's performance we first report a historical reconstruction of the quarterly GDP from 2001:Q3 to 2016:Q4 in figure 1.

Figure 1 was built using the original GDP year-over-year variation series, the original BCB Expectation series and the reconstructed model's nowcast series. The nowcast series needs reconstruction because all variables are demeaned and standardized to unitary variance previous to model estimation, thus yielding standardized results. We discretionarily chose the nowcasts made 30 days after the end of the reference period to plot.

From the chart we can note that the three series are quite similar, except for the model's nowcasts during the most of 2016, which seems to be quite distant from both the true values and the expectations' values. We believe this is expected since the model does not use any kind of news-related data and most of the GDP anticipation during that year was done based on political issues and scandals.

In figure 2 we plot the RMSFE (Root-mean-squared-forecast-error) for both the model's and expectations's nowcasts. The x-axis is the difference (in days) from the nowcasting date and the end of the reference period for a given quarter. Thus, negative numbers represent nowcasts made before each quarter was over.

We can heuristically note from figure 2 that the BCB's nowcasts are more accurate than the model's for any periods. In order to statistically measure that difference in performance we use the Diebold-Mariano test of equal forecasting accuracy. We conducted

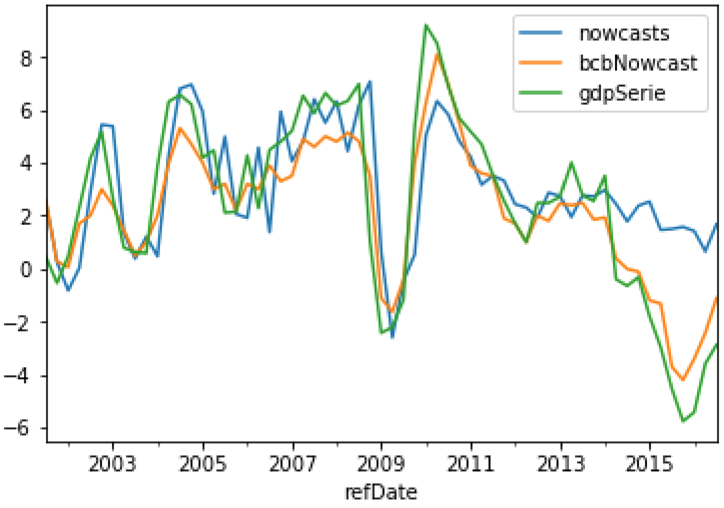

Figure 1 – GDP Reconstruction with the Nowcasting model and BCB Expectations from 30 after the end of the reference period

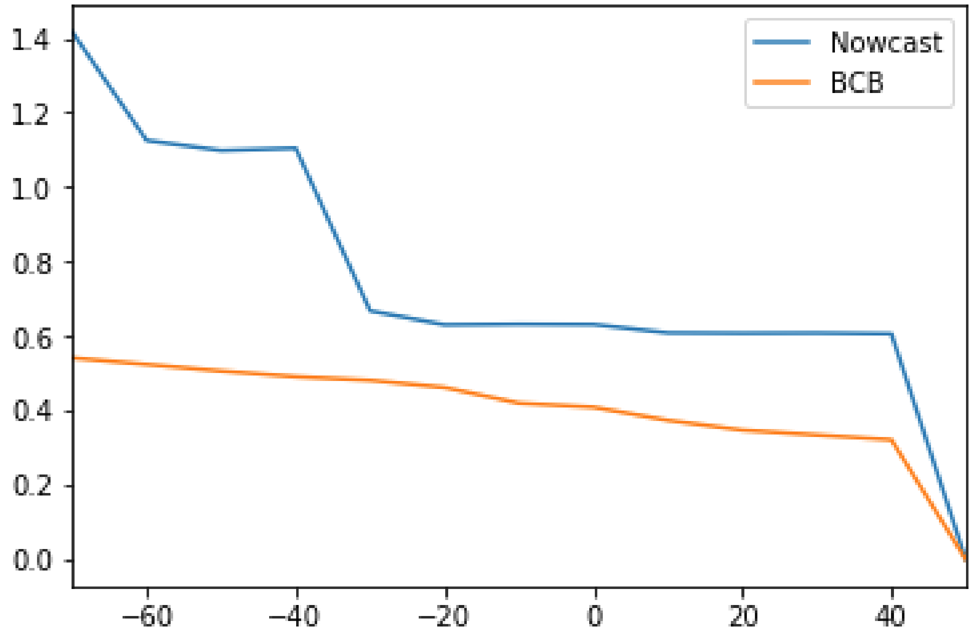

Figure 2 – RMSFE for the model and the BCB Expectations' nowcasts

Table 2 – Diebold-Mariano test of equal forecasting accuracy for the full sample

	Nowcast date				
	-60	-30	0	30	59
DM p-value	**0.0185****	0.2171	0.1472	**0.0968***	**0.0645***

Notes. The table reports the p-values for the DM statistic conducted for each series of forecasts. One, two and three starts indicate significance at 10,5 and 1 percent levels, respectively.

the test for nowcasts made 60 and 30 days before the reference quarter is over; at the very last day of the quarter; and for 30 and 59 days after the reference quarter is over. 59 days represent the day before the true value is published. The results are reported in table 2.

Table 2 reassures our previous conclusions drawn from the chart in figure 2 and indicates that the nowcasts based on the expectations have significantly better performance when compared to the model's, specifically for the nowcasts made in -60, 30 and 59 days from the end of the reference quarter, for which the test showed statistical difference.

As briefly mentioned before, we can see in figure 1 that our nowcasting model performs specially worse for the most of 2016. During this period, Brazil went through a sequence of critical political events including the indictment of several congressmen and the president's impeachment. Since that sort of political issues are not expected to be instantly reflected on most of the macroeconomic variables we chose for the model, we

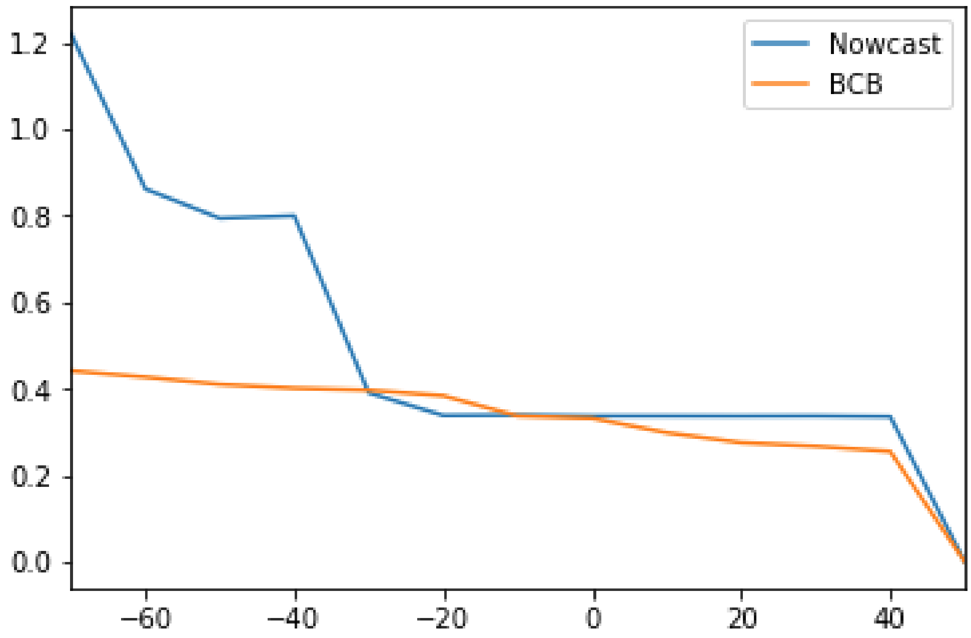

Figure 3 – RMSFE for the model and the BCB Expectations nowcasts

Table 3 – Diebold-Mariano test of equal forecasting accuracy for the sub-sample

	Nowcast date				
	-60	-30	0	30	59
DM p-value	**0.0438****	0.9468	0.9456	0.2971	0.1256

Notes. The table reports the p-values for the DM statistic conducted for each series of forecasts. One, two and three starts indicate significance at 10,5 and 1 percent levels, respectively. The sub-sample is comprised of data and nowcasts ranging from 2001:Q3 to 2015:Q4.

believe it is worth to observe the RMSFE chart and the Diebold-Mariano test made using a subsample of data up to 2015, thus disregarding the troubled year of 2016, in order to assess the model's performance when the country is not under deep political crisis. The results are reported in figure 3 and in table 3.

Figure 3 demonstrates that a great portion of the model's excess RMSFE disappears when we look strictly at the sub-sample that excludes the year of 2016. Table 3 corroborates with that observation by not rejecting the null hypothesis that the model's and the BCB's nowcasts have equal predictive accuracy, except for the earliest nowcasts, which could have showed up by chance.

From the developed results we draw the conclusion that the model and the agents' expectations reported to the BCB are equally good nowcasters, given that the country is not under distressed political conditions.

6 Conclusion

Since Brazil provides a rare framework of expectations in which the Brazilian Central Bank publishes daily agents' expectations for the GDP, we aimed at verifying whether the Banbura and Modugno (2010) methodology could outperform the agents' expectations in nowcasting the GDP.

We found statistical results showing us that the model does not outperform, but is at least as good as the agents' expectations for nowcasting the GDP, except for non-economical crisis periods.

Even though for economic agents it is easier to simply gather the published expectations than developing a nowcasting model, we believe it is safe to say that a nowcasting model is less costly than building an economic team to generate projections.

That said, if we take those results as a general valid assessment of the model for macroeconomic variables, we can still produce good nowcasts for macroeconomic variables of which daily expectations are not available. Routine examples are Industrial Production, Unemployment, Retail Sales, Commodities production and several others.

Since those variables do not have published daily expectations we do not have a benchmark to compare the model to. Being that so, accepting the validity of the results found using this model to nowcast the GDP can be extrapolated to other economic variables, allowing us to state that the model's nowcasts of macroeconomic variables would be as good as the economic agents' unpublished expectations.

Moreover, the model is also applicable for other economies in which daily expectations are not available.

Yet, we believe further research should focus on the lack of performance during political crisis. That issue can be addressed through the inclusion of more predictor variables, specially soft data; or through the natural course of time, since our limited sample is comprised of few years and few crisis. It is worth noting that if the model is replicated for economies, with longer datasets and more variety of soft data available, it is expected to cope with this problem, as extensively reported by the literature.

Moreover we believe that future research should also focus on the generation of soft data for economies in which it is not institutionally available. More specifically, using frameworks like Natural Language Processing models for datasets like Newspapers data, Facebook data, Twitter data or Google Tools might be substantially beneficial for the model's performance, specially during crisis.

Bibliography

BAFFIGI, A.; GOLINELLI, R.; PARIGI, G. Bridge models to forecast the euro area gdp. *International Journal of Forecasting*, Elsevier, v. 20, n. 3, p. 447–460, 2004. Citado 2 vezes nas páginas 14 and 17.

BAI, J.; NG, S. Determining the number of factors in approximate factor models. *Econometrica*, Wiley Online Library, v. 70, n. 1, p. 191–221, 2002. Citado na página 21.

BANBURA; MODUGNO. Maximum likelihood estimation of large factor model on datasets with arbitrary pattern of missing data. *Working Paper*, European Central Bank, v. 164, n. 1, p. 188–205, 2010. Citado 3 vezes nas páginas 18, 20, and 31.

BANBURA, M. et al. Now-casting and the real-time data flow. ECB working paper, 2013. Citado 9 vezes nas páginas 14, 17, 18, 19, 20, 21, 22, 27, and 35.

BRAGOLI, D.; METELLI, L.; MODUGNO, M. The importance of updating: Evidence from a brazilian nowcasting model. 2014. Citado 4 vezes nas páginas 20, 21, 26, and 27.

CAMACHO, M.; PEREZ-QUIROS, G. Introducing the euro-sting: Short-term indicator of euro area growth. *Journal of Applied Econometrics*, Wiley Online Library, v. 25, n. 4, p. 663–694, 2010. Citado na página 27.

CLEMENTS, M. P.; GALVÃO, A. B. Macroeconomic forecasting with mixed-frequency data: Forecasting output growth in the united states. *Journal of Business & Economic Statistics*, Taylor & Francis, v. 26, n. 4, p. 546–554, 2008. Citado na página 17.

CLEMENTS, M. P.; GALVÃO, A. B. Forecasting us output growth using leading indicators: An appraisal using midas models. *Journal of Applied Econometrics*, Wiley Online Library, v. 24, n. 7, p. 1187–1206, 2009. Citado na página 17.

DOZ, C.; GIANNONE, D.; REICHLIN, L. A two-step estimator for large approximate dynamic factor models based on kalman filtering. *Journal of Econometrics*, Elsevier, v. 164, n. 1, p. 188–205, 2011. Citado na página 18.

DOZ, C.; GIANNONE, D.; REICHLIN, L. A quasi–maximum likelihood approach for large, approximate dynamic factor models. *Review of economics and statistics*, MIT Press, v. 94, n. 4, p. 1014–1024, 2012. Citado 3 vezes nas páginas 18, 20, and 22.

EVANS, M. D. *Where are we now? real-time estimates of the macro economy.* [S.l.], 2005. Citado na página 18.

GHYSELS, E.; WRIGHT, J. H. Forecasting professional forecasters. *Journal of Business & Economic Statistics*, Taylor & Francis, v. 27, n. 4, p. 504–516, 2009. Citado na página 17.

GIANNONE, D.; REICHLIN, L.; SALA, L. Monetary policy in real time. In: *NBER Macroeconomics Annual 2004, Volume 19.* [S.l.]: MIT Press, 2005. p. 161–224. Citado na página 18.

GIANNONE, D.; REICHLIN, L.; SIMONELLI, S. Nowcasting euro area economic activity in real time: the role of confidence indicators. *National Institute Economic Review*, SAGE Publications, v. 210, n. 1, p. 90–97, 2009. Citado 2 vezes nas páginas 18 and 27.

GIANNONE, D.; REICHLIN, L.; SMALL, D. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, Elsevier, v. 55, n. 4, p. 665–676, 2008. Citado 4 vezes nas páginas 14, 17, 18, and 27.

KITCHEN, J.; MONACO, R. Real-time forecasting in practice: The us treasury staff's real-time gdp forecast system. 2003. Citado na página 17.

KUZIN, V.; MARCELLINO, M.; SCHUMACHER, C. Midas vs. mixed-frequency var: Nowcasting gdp in the euro area. *International Journal of Forecasting*, Elsevier, v. 27, n. 2, p. 529–542, 2011. Citado 3 vezes nas páginas 17, 18, and 20.

LAHIRI, K.; MONOKROUSSOS, G. Nowcasting us gdp: The role of ism business surveys. *International Journal of Forecasting*, Elsevier, v. 29, n. 4, p. 644–658, 2013. Citado na página 27.

LIEBERMANN, J. Real-time forecasting in a data-rich environment. 2012. Citado na página 27.

MARIANO, R. S.; MURASAWA, Y. A new coincident index of business cycles based on monthly and quarterly series. *Journal of applied Econometrics*, Wiley Online Library, v. 18, n. 4, p. 427–443, 2003. Citado na página 20.

MITTNIK, S.; ZADROZNY, P. A. Forecasting quarterly german gdp at monthly intervals using monthly ifo business conditions data. CESifo Working Paper Series, 2004. Citado na página 18.

SARGENT, T. J.; SIMS, C. A. et al. Business cycle modeling without pretending to have too much a priori economic theory. *New methods in business cycle research*, Federal Reserve Bank of Minneapolis: Minneapolis, v. 1, p. 145–168, 1977. Citado na página 18.

SILIVERSTOVS, B. Keeping a finger on the pulse of the economy: Nowcasting swiss gdp in real-time squared. KOF Working paper, 2012. Citado na página 27.

STOCK, J. H.; WATSON, M. W. Dynamic factor models. *Oxford handbook of economic forecasting*, Oxford University Press Oxford, v. 1, p. 35–59, 2011. Citado na página 18.

WATSON, M. W. Comment on giannone, reichlin, and sala. *NBER Macroeconomics Annual*, MIT Press, v. 2004, p. 216–221, 2004. Citado na página 18.

ZADROZNY, P. A. *Estimating a multivariate ARMA model with mixed frequency data: an application to forecasting US GNP at monthly intervals.* [S.l.]: Federal Reserve Bank of Atlanta, 1990. v. 90. Citado na página 18.

Appendix

APPENDIX A – EM Algorithm

Appendix mostly taken from Banbura et al. (2013).

$$E - step: \qquad L(\theta,\theta(j)) = E_{\theta(j)}[l(Y,F;\theta)|\Omega_\nu], \tag{A.1}$$

$$M - step: \qquad \theta(j+1) = argmax_\theta L(\theta,\theta(j)), \tag{A.2}$$

Let T_ν be the index of the most recent observation.

The new parameter estimates in the M-step can be obtained in two steps, first $\Lambda(j+1)$ and $\Phi(j+1)$ are given by:

$$\Lambda(j+1) = (\sum_{t=1}^{T_\nu} E_{\theta(j)}[Y_t F_t^t|\Omega_\nu])(\sum_{t=1}^{T_\nu} E_{\theta(j)}[F_t F_t^t|\Omega_\nu])^{-1} \tag{A.3}$$

$$\Phi(j+1) = (\sum_{t=1}^{T_\nu} E_{\theta(j)}[F_t F_{t-1}^t|\Omega_\nu])(\sum_{t=1}^{T_\nu} E_{\theta(j)}[F_{t-1} F_{t-1}^t|\Omega_\nu])^{-1} \tag{A.4}$$

Second, given the new estimates of Λ and Φ, the covariance matrices can be obtained as follows:

$$\begin{aligned}\Sigma_E(j+1) = diag(\frac{1}{T_\nu}\sum_{t=1}^{T_\nu} E_{\theta(j)}[(Y_t - \Lambda(j+1)F_t)(Y_t - \Lambda(j+1)F_t)^t|\Omega_\nu]) = \\ diag(\frac{1}{T_\nu}\sum_{t=1}^{T_\nu} E_{\theta(j)}[Y_t Y_t^t|\Omega_\nu] - \Lambda(j+1)E_{\theta(j)}[F_t Y_t^t|\Omega_\nu]))\end{aligned} \tag{A.5}$$

and

$$\Sigma_U(j+1) = \frac{1}{T}(E_{\theta(j)}[F_t F_t^t|\Omega_\nu]) - \Phi(j+1)E_{\theta(j)}[F_{t-1}F_t^t|\Omega_\nu]) \tag{A.6}$$

see Watson and Engle (1983) and Shumway and Stoffer (1982). If Y_t did not contain missing observations, we would have that:

$$E_{\theta(j)}[Y_t Y_t^t|\Omega_\nu] = Y_t Y_t^t and E_{\theta(j)}[Y_t F_t^t|\Omega_\nu] = Y_t E_{\theta(j)}[F_t^t|\Omega_\nu], \tag{A.7}$$

which can be plugged to the formulas above. The expectations $E_{\theta(j)}[F_t F_t^t|\Omega_\nu]$, $E_{\theta(j)}[F_t F_{t-1}^t|\Omega_\nu]$ and $E_{\theta(j)}[F_t|\Omega_\nu]$ can be obtained via the Kalman filter and smoother. When Y_t contains missing observations equations A.3 and A.5 become:

$$vec(\Lambda(j+1)) = (\sum_{t=1}^{T_\nu} E_{\theta(j)}[F_t F_t^t|\Omega_\nu] \otimes S_t)^{-1} vec(\sum_{t=1}^{T_\nu} S_t Y_t E_{\theta(j)}[F_t^t|\Omega_\nu]) \tag{A.8}$$

and

$$\begin{aligned}\Sigma_E(j+1) = diag(\frac{1}{T_\nu}\sum_{t=1}^{T_\mu}(S_t Y_t Y_t^t S_t^t - S_t Y_t E_{\theta(j)} \Lambda(j+1)^t S_t - S_t \Lambda(j+1) E_{\theta(j)}[F_t|\Omega_\nu] Y_t^t S_t \\ + S_t \Lambda(j+1) E_{\theta(j)}[F_t F_t^t|\Omega_\nu] \Lambda(j+1)^t S_t + (I_N - S_t)\Sigma_E(j)(I_N - S_t)))\end{aligned} \tag{A.9}$$

where S_t is a selection matrix, i.e. it is a diagonal matrix with ones corresponding to the non-missing observations in Y_t and zeros otherwise.