High-Dimensional and Mixed-Frequency Models for Forecasting Year-on-Year Growth of Brazilian Agricultural GDP

André N. Maranhão [a]

Abstract

The objective of this study is to provide a forecasting exercise that demonstrates predictive gains compared to expert forecasts computed by the regulator. This study presents 21 different models addressing both high dimensionality and mixed frequencies for forecasting Brazilian Agricultural GDP. Using a high-dimensional dataset with 79 covariates, including mixed monthly and quarterly frequencies, we estimate and utilize 24 out-of-sample periods for forecasting exercises. All forecasts were conducted using rolling windows, with the out-of-sample period spanning from 2017 to 2022. Among the estimated models, the combination of models with the highest predictive ability showed an 80% predictive gain compared to the forecasts recorded in the Central Bank's Focus Bulletin.

JEL: C53, E27, Q14

Keywords: Forecasting, Mixed Frequency, High Dimension, Agricultural GDP

1. Introduction

Agriculture plays a fundamental role in shaping the world's Gross Domestic Product (GDP), being a strategic sector for the economic and social development of many countries. Through the production of food, fibers, and raw materials, agriculture provides sustenance for the population and meets the growing global demand for food. Additionally, the agricultural sector generates direct and indirect employment, driving economic growth and contributing to poverty and inequality reduction in various regions. Agriculture is also a significant driver of international trade, promoting economic integration between countries and contributing to a positive trade balance. Through innovation and the adoption of technologies, the agricultural sector has the potential to increase productivity, enhance efficiency, and mitigate environmental impacts, enabling sustainable and resilient production.

Brazilian agricultural GDP plays a central role in global agriculture. In addition to its global importance, Brazilian agricultural GDP has shown resilience even during recessions. Given this importance, the monitoring of agricultural activity has increased, making the use of econometric forecasts more relevant. The challenges of forecasting agricultural economic activity are even greater than those traditionally associated with GDP forecasting. On the other hand, the set of available covariates that can be used for this purpose has grown, with many of these new data sources having temporal frequencies that differ from the target

[a] Data Scientist at the Credit Directorate, Bank of Brazil. Professor at Fundação Getúlio Vargas (FGV EESP), São Paulo School of Economics. Email: andre.maranhao@fgv.br .

variable's frequency. This combination of a large number of covariates with different frequencies represents the environment known as high dimensionality and mixed frequencies. The objective of this article is to briefly present the various possibilities for formulating these models for forecasting agricultural GDP.

Using a dataset with 77 covariates, 75 of which are monthly, we aim to forecast the year-on-year growth of agricultural GDP (IBGE) to achieve predictive gains, particularly compared to the forecasts presented in the Focus Bulletin at the beginning of the calendar year. In Section 2, we detail the forecasting problems currently present in the Focus Bulletin and the alternatives that high-dimensional and mixed-frequency models can offer for this forecasting task. In Section 3, we provide a brief summary of the methodological approach and the literature on these approaches. In Section 4, we detail all the results of the estimated models and forecasts for each model, along with an analysis of predictive ability, concluding the study in Section 5.

2. Literature and the Problem of Forecasting Agricultural Economic Activity

The literature on economic forecasting closely follows the evolution brought about by the new era of data. The expansion of measurements of new variables, as well as the increase in computational capacity, has brought new questions and challenges to econometrics in this century. Despite all the technological advancements and even new theoretical developments in the field of econometrics, they have not yet been sufficient to overcome the old challenges of forecasting economic activity. Economic activity is inherently a complex and dynamic system, always affected by and affecting different covariates. On the other hand, monitoring the level of activity is of paramount importance for a wide range of professions and fields of activity. Although for different reasons, regulators and market agents have the task of generating forecasts of this economic activity.

Regulators, in addition to having their own forecasts of the level of activity, understand the great importance of capturing the expectations and uncertainties of agents regarding economic activity, price levels, etc. An important proxy for capturing expectations and uncertainties can be obtained through a forecasting competition. Although there are no methodological restrictions on how each institution elaborates its forecasts, the Central Bank of Brazil has been monitoring weekly forecasts of different economic variables and formalized a document called the Focus Bulletin. In this document, we can observe how expectations and uncertainties of a sample of market agents are forming, their difficulties with certain economic variables, and their anchoring relative to others. In this context, forecasts have been recorded in the Central Bank's databases since 2001 for a particularly challenging type of economic activity: agricultural economic activity.

The difficulty of forecasting GDP is related to its intrinsically multivariate nature and susceptibility to different shocks and structural breaks. In the context of agricultural activity, in addition to different types of shocks and structural breaks, the level of activity exhibits greater volatility and, therefore, greater uncertainty. Figures 1 and 2 illustrate the differences in the challenges of

estimating and forecasting these economic activities. In non-atypical years, GDP forecast errors are concentrated within the minimum and maximum forecast margins; however, agricultural GDP forecasts, even in typical years, have the median of the forecasts outside these limits.

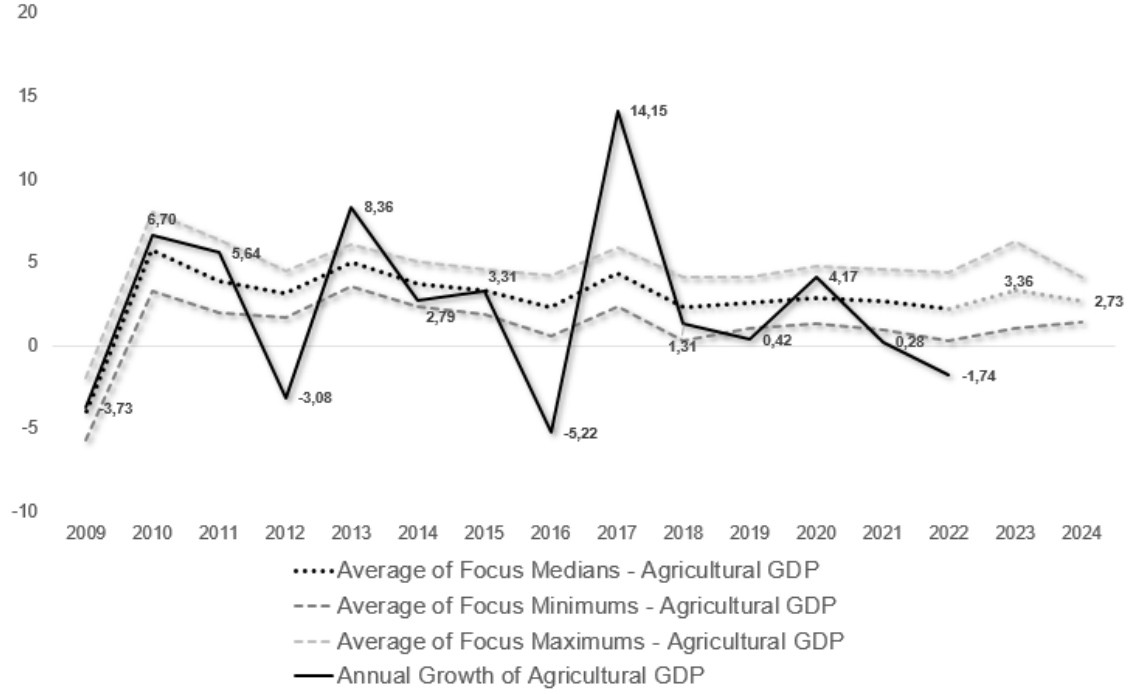

Figure 1: Focus Forecast of Agricultural GDP Growth

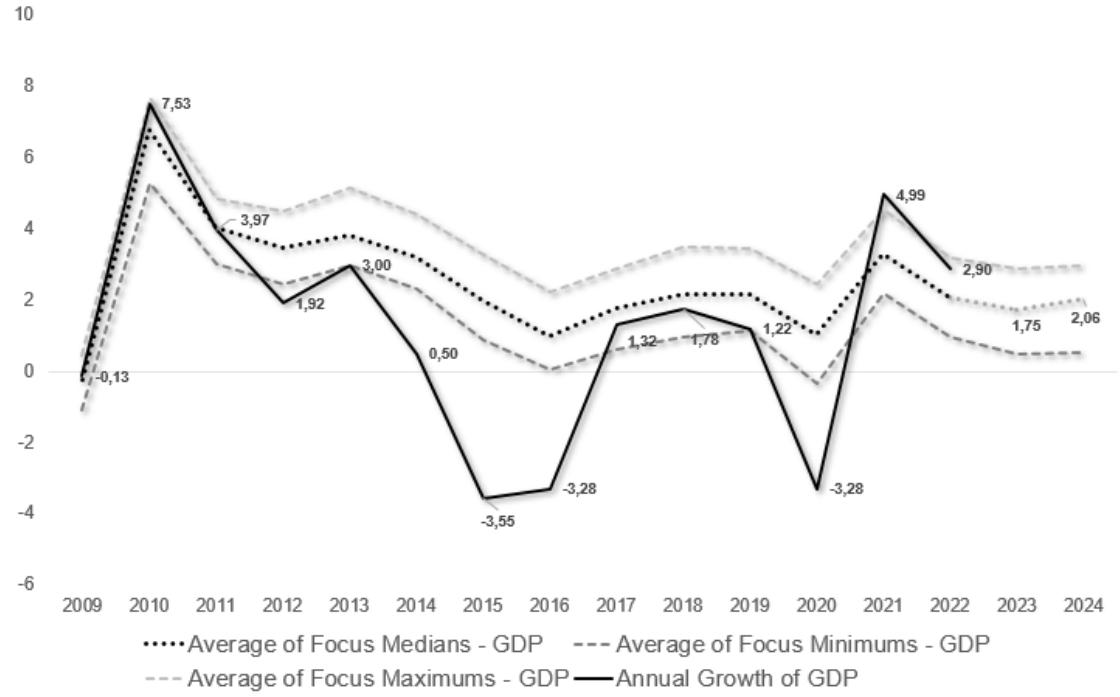

Figure 2: Focus GDP Growth Forecast

Uncertainty about agricultural economic activity is also greater compared to GDP uncertainty. This uncertainty can be measured by the standard deviations observed by the regulator and published. The levels of uncertainty in forecasts of agricultural economic activity are about 2.5 times the uncertainty associated with GDP forecasts, as can be seen in Figure 3. Another important characteristic of this set of forecasts is related to the fact that, as they are forecasts with weekly revisions, the longer the forecast horizon, the greater the uncertainty of these forecasts. When analyzing the predictive error in the annual growth forecasts made in January, we can observe the highest level of uncertainty among agents, which decreases as revisions are made and approach the end of the calendar year, with the interval between the minimum and maximum forecasts decreasing sharply.

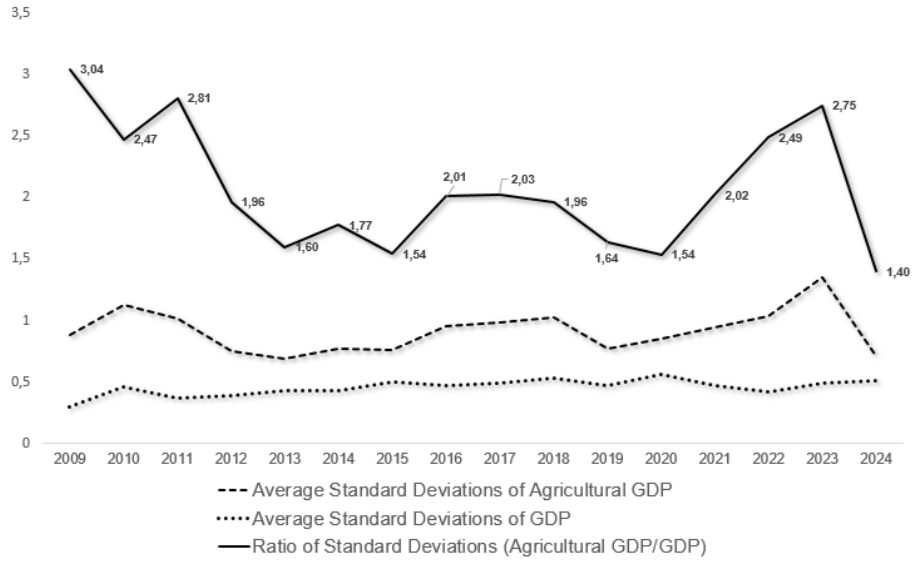

Figure 3: Uncertainty Focus of GDP and Agricultural GDP

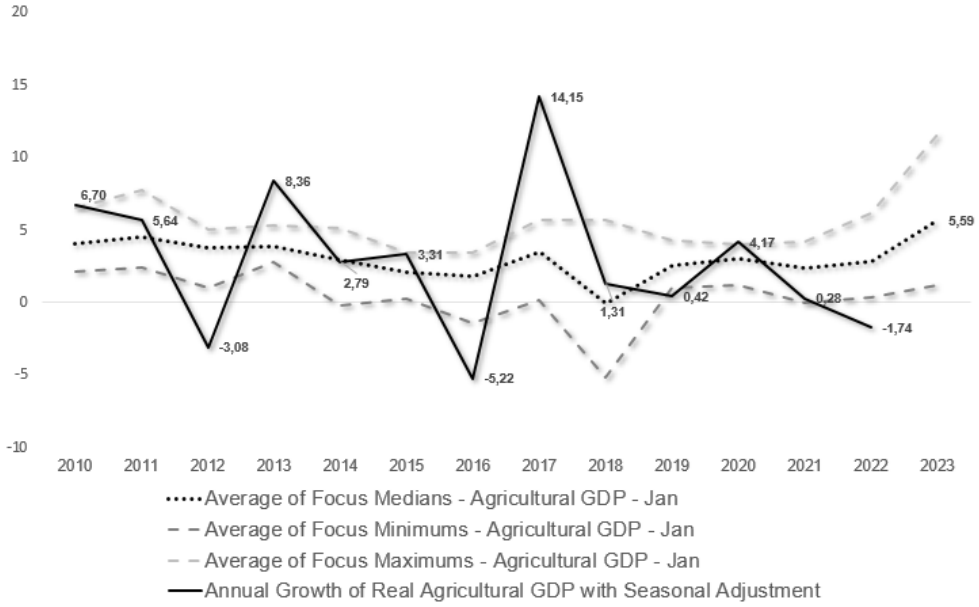

Figure 4: FOCUS Agricultural GDP Growth Forecast - January Forecasts

Thus, there is a belief that the largest predictive error in the Focus Bulletin occurs at the beginning of the calendar year, decreases when we observe the median of all forecasts, and reaches its minimum a short time before the official announcement of that period. This behavior is clearly observed in Table 1, which depicts the predictive errors of agents over time. The forecasts (median of the means) for January are always higher than the forecasts for the entire period, with December forecasts being lower for the calendar year.

Table 1 - Predictive Errors of the Focus Bulletin - GDP

Years	FOCUS Forecast - Average of GDP Medians	FOCUS Forecast - Average of GDP Medians - January	FOCUS Forecast - Average of Median GDP - December	Annual GDP Growth	Absolute Percentage Error - FOCUS Medians	Root Mean Square Error - Medians FOCUS	Absolute Percentage Error - FOCUS Medians - January	Root Mean Square Error - FOCUS Medians - January	Absolute Percentage Error - FOCUS Medians - Decembers	Root Mean Square Error - FOCUS Medians - Decembers
2001	3,48	4,00	2,90	1,39	2,06	2,33	2,57	2,90	1,49	1,67
2002	2,76	3,46	2,67	3,05	0,29	0,33	0,39	0,46	0,37	0,43
2003	2,87	3,33	2,53	1,14	1,71	2,01	2,16	2,54	1,38	1,62
2004	3,62	3,53	3,96	5,76	2,02	2,51	2,11	2,62	1,70	2,11
2005	3,58	3,73	3,38	3,20	0,37	0,47	0,52	0,66	0,18	0,23
2006	3,69	3,85	3,55	3,96	0,26	0,35	0,10	0,14	0,40	0,53
2007	3,79	3,60	3,85	6,07	2,15	3,04	2,33	3,29	2,10	2,96
2008	3,99	3,89	4,11	5,09	1,05	1,56	1,14	1,70	0,94	1,39
2009	2,96	3,32	2,67	-0,13	3,09	4,59	3,45	5,12	2,80	4,16
2010	4,53	4,22	4,92	7,53	2,79	4,45	3,08	4,92	2,43	3,88
2011	4,13	4,30	4,04	3,97	0,15	0,24	0,31	0,52	0,06	0,10
2012	3,85	4,08	3,58	1,92	1,90	3,21	2,12	3,59	1,62	2,75
2013	3,95	4,07	3,81	3,00	0,92	1,60	1,04	1,81	0,78	1,36
2014	3,37	3,64	3,01	0,50	2,85	5,00	3,12	5,47	2,49	4,36
2015	2,21	2,72	1,48	-3,55	5,97	10,09	6,49	10,97	5,21	8,81
2016	1,17	1,39	0,40	-3,28	4,59	7,51	4,83	7,89	3,80	6,21
2017	1,80	1,62	1,57	1,32	0,47	0,78	0,29	0,48	0,24	0,40
2018	2,23	2,27	2,13	1,78	0,44	0,74	0,47	0,80	0,34	0,57
2019	2,21	2,49	2,19	1,22	0,97	1,66	1,25	2,14	0,95	1,63
2020	1,23	2,45	1,01	-3,28	4,66	7,69	5,92	9,77	4,43	7,30
2021	3,24	2,72	3,40	4,99	1,67	2,89	2,16	3,74	1,51	2,61
2022	2,04	1,63	2,04	2,90	0,83	1,48	1,23	2,20	0,84	1,49
				Average	1,88	2,91	2,16	3,33	1,63	2,53

Note: The values for the MAPE and RMSE calculations were constructed from the GDP Series linked to the IBGE Quarterly National Accounts, considering the respective variations predicted in FOCUS in each reference year.

Source: IBGE, BACEN and own elaboration.

Tabela 2 - Predictive Errors of the Focus Bulletin - Agricultural GDP

Years	FOCUS Forecast - Median of Medians Agricultural GDP	FOCUS Forecast - Median of Medians Agricultural GDP - January	FOCUS Forecast - Median of Medians Agricultural GDP - December	Annual Agricultural GDP Growth	Absolute Percentage Error - Median FOCUS	Root Mean Square Error (RMSE) - Median FOCUS	Absolute Percentage Error - Median FOCUS - January	Root Mean Square Error (RMSE) - Median FOCUS - January	Absolute Percentage Error - Median FOCUS - December	Root Mean Square Error (RMSE) - Median FOCUS - December
2001	4,28	5,37	4,10	5,20	0,88	1,09	1,77	2,19	1,04	1,29
2002	3,54	3,33	3,04	8,02	4,14	5,53	4,23	5,64	4,61	6,15
2003	3,94	3,45	3,74	8,31	4,03	5,83	4,00	5,79	4,22	6,10
2004	4,22	3,97	4,21	2,00	2,18	3,21	1,92	2,84	2,17	3,20
2005	3,95	3,96	4,01	1,12	2,80	4,17	2,92	4,35	2,86	4,26
2006	3,85	4,07	3,91	4,64	0,75	1,18	0,61	0,95	0,69	1,08
2007	3,98	4,00	3,88	3,25	0,71	1,14	0,66	1,07	0,61	0,99
2008	4,26	3,93	4,08	5,77	1,43	2,43	1,78	3,04	1,60	2,72
2009	3,11	3,89	3,61	-3,73	7,11	11,66	8,09	13,27	7,62	12,51
2010	4,35	4,06	4,08	6,70	2,20	3,84	2,25	3,93	2,46	4,30
2011	4,18	4,30	4,28	5,64	1,38	2,56	1,22	2,25	1,29	2,38
2012	3,75	4,35	4,31	-3,08	7,05	12,62	7,62	13,65	7,63	13,66
2013	4,81	4,30	4,26	8,36	3,28	6,36	3,74	7,27	3,79	7,36
2014	3,85	4,31	4,12	2,79	1,03	2,06	0,99	1,97	1,29	2,58
2015	3,42	3,80	3,49	3,31	0,10	0,20	0,10	0,20	0,17	0,35
2016	2,62	3,41	3,04	-5,22	8,28	16,17	9,04	17,66	8,72	17,04
2017	4,11	3,34	3,38	14,15	8,79	19,62	9,67	21,57	9,44	21,05
2018	2,44	3,11	2,54	1,31	1,12	2,54	1,67	3,78	1,22	2,75
2019	2,66	3,00	2,91	0,42	2,23	5,07	2,59	5,88	2,49	5,64
2020	2,89	3,02	3,06	4,17	1,24	2,92	0,97	2,29	1,07	2,52
2021	2,92	3,16	2,97	0,28	2,64	6,25	2,71	6,43	2,68	6,36
2022	2,41	3,00	2,69	-1,74	4,22	9,84	1,77	4,13	4,51	10,51
				Average	2,98	5,54	3,13	5,75	3,18	5,92

Note: The values for the MAPE and RMSE calculations were derived from the chained Agro GDP series of the IBGE Quarterly National Accounts, considering the respective variations forecasted in FOCUS for each reference year.

Source: IBGE, BACEN, and own elaboration.

In this article, we investigate the technical and practical issues involved in using mixed-frequency data (quarterly and monthly, with the possibility of extending this approach to weekly and daily data) to forecast monthly and quarterly economic activity in a country. The analysis considers alternative high-frequency forecasting models for agricultural GDP growth, using indicators observable at different frequencies. The study focuses particularly on dynamic time series models involving latent factors and compares the forecasting performance of this approach with more commonly used data-intensive methods developed in applications in the United States and Europe—specifically, Mixed Data Sampling (MIDAS) regression and Current Quarter Modeling (CQM) with bridge equations. Although these alternatives are primarily data-intensive, dynamic latent factor modeling with mixed frequencies presents a parsimonious approach that depends on a much smaller dataset that needs to be updated regularly. However, it also faces additional methodological and computational complications, as mixed-frequency data are included in the analysis.

In the next section, we present a methodological summary of the models that will be detailed in the results section.

3. **Methodology**

In general, data analysts may encounter situations with a mixed-frequency dataset, which may include quarterly, monthly, weekly, and daily observations. In this chapter, the target variables are the growth rates of Brazilian agricultural GDP; these are available quarterly. On the other hand, all indicator variables are available monthly. Note that the forecasting procedures implemented here can be adapted to more general situations where the indicator variables come in even lower mixed frequencies.

The alternative forecasting models considered here can be labeled as "quarterly" or "monthly," according to the basic or underlying frequency explicitly modeled. For quarterly models, the observed quarterly values of the target variables are used directly, while the monthly observations for the indicators are aggregated over the quarter. For example, for stock variables, averages are calculated over the quarter, sums are used for flow variables, and growth rates are calculated from the aggregated series.

A monthly model, on the other hand, treats all data series (target or indicator) as generated at the highest frequency (monthly, in our case), but some of the data points are not observed. The variables observed at the low frequency (quarterly) are treated as having periodically missing or unobserved data points, available only at the end of the month of the quarter. The estimation procedures are then implemented to account for the presence of systematically missing observations. Note that an estimated monthly model would also provide forecasts of the target variables disaggregated at the high frequency.

3.1 **Quarterly Models**

The following quarterly models are covered in this study:

$$y_{tQ} \sim ARMA(p,q) \;\; ou \;\; VARMA(p,q) \qquad (1)$$

In this first class, we use univariate parametric models classically used in the Box-Jenkins methodology, as well as exponential smoothing models in state-space form with great flexibility for adaptations (Helske, J., 2018).

$$y_{tQ} \sim \left(ARMA(p,q), Z_{tQ}\right) \qquad (2)$$

Bridge equation models (expand the univariate benchmark by introducing indicator variables, possibly with lags, as additional explanatory variables).

$$y_{tQ} \sim (ARMA(p,q), PC(Z_{tQ})) \qquad (3)$$

Bridge equation models will be estimated using principal components as an alternative to handle high dimensionality. Current Quarterly Model (CQM)—high-frequency bridge modeling with updates of GDP projections and their components. Here, the objective is timely forecasting of agricultural GDP, typically available quarterly. Bridge equations are used, relating GDP components to observable quarterly and monthly indicator variables. Monthly observations are calculated over the quarter, with updates as more monthly observations become available. To forecast monthly and quarterly indicators, ARIMA models are used. If no indicator is available, an ARIMA model would be estimated for the GDP component itself. CQM with bridge equations for the United States has been extensively researched by Lawrence Klein—for example, in Klein and Sojo (1987, 1989), Klein and Park (1993, 1995), Klein and Ozmucur (2001, 2002, 2004, 2008), Mariano and Tse (2008), and Mariano and Ozmucur (2018) in Pauly (2018). Now, CQM models have been developed to update quarterly forecasts in several countries, such as Turkey (Ozmucur, 2009), Japan (Inada, 2005), Mexico (Coutino, 2005), Russia (Klein et al., 2003, 2005), and China (Klein and Mak, 2005).

3.2 **Monthly Models**

The following monthly models are addressed in this study:

Monthly VAR using averages or cubic splines to "fill in the blanks"—i.e., estimate the missing monthly observations. Mixed-Frequency Vector Autoregressive (MF-VAR):

$$y_{tm} \sim (VAR(p), PC(Z_{tm})) \qquad (4)$$

This is a state-space model formulation, and Kalman filtering methods can be used to estimate the model and compute forecasts at the highest frequency—for example, see Harvey (1989).

Typical bridge equation modeling relates a quarterly variable to 3-month averages of monthly variables. This implicitly imposes a restriction on the coefficients for the months of the quarter and, consequently, introduces asymptotic biases and inefficiencies—Ghysels, 2013. In contrast, MIDAS estimates a monthly regression of GDP on monthly (and possibly quarterly) indicators using parsimonious distributed lags to represent the lack of observations. The initial reference is Ghysels et al. (2004), with initial applications in finance, now also used to forecast macroeconomic time series.

Since its introduction, this modeling approach has been widely used in the mixed-frequency forecasting literature and has been enhanced with numerous variations—as described, for example, in Ghysels (2016a,b); Ghysels et al. (2007); and Ghysels and Marcellino (2018). For implementation, MIDAS applies a more parsimonious parameterization of distributed lag structures to model the relationship between GDP and current and lagged indicators at the monthly frequency, so that the basic model can be expressed as:

$$y_{tm} \sim DL(Z_{tm}) + \epsilon_{tm} \qquad (5)$$

Finally, we use the MIDAS-DFM Model:

$$y_{tm} \sim DL(f(Z_{tm})_{t-k}) + \epsilon_{tm} \qquad (6)$$

The underlying philosophy is that macroeconomic fluctuations are driven by a small number of shocks or common factors and an idiosyncratic component peculiar to each economic time series. The seminal articles are Sargent and Sims (1977) and Stock and Watson (1989). Earlier work (e.g., Stock and Watson) develops single-factor models to construct composite indexes of economic activity based on a handful of coincident indicators. More recent studies use the model to extract unobserved common factors from a large collection of observable indicator variables. More recently, the approach has been revived for forecasting purposes in the United States and larger European countries—Foroni and Marcellino (2012, 2013).

Another (related) application dealt with the combination of mixed frequencies in the construction of composite indexes—for example, Mariano and Murasawa (2003), Aruoba et al. (2009). The estimated MIDAS-DFM factor model, properly validated, can also be used to forecast macroeconomic variables of interest at the highest frequency, for example, Liu and Hall (2001), Mariano and Murasawa (2010).

In summary, the underlying model consists of two parts. The first explains the dynamics of the target and indicator variables depending on their own lags, unobservable common factor(s), and possibly observable exogenous variables. The second part explains the behavior of the latent common factor(s) in terms of their own joint dynamics and possibly interactions with observable indicators. The system may also have other observable exogenous variables that serve as indicators for the latent common factors. A similar modeling approach is used in Mariano and Murasawa (2003, 2010) in the construction of an enhanced coincident economic index for the United States using mixed frequencies

(quarterly and monthly), as well as in Aruoba et al. (2009) in the construction of a "real-time" business conditions index (daily) for the United States, using four indicators (quarterly, monthly, weekly, daily). To make the analysis implementable, we have to deal with the two confounding complications of missing data observations as well as unobserved common factors. One solution is to derive from the underlying model a state-space model formulation with measurement and state equations involving only fully observed variables, latent state variables, predetermined variables, and measurement and transition shocks.

The "missing" observations need to be factored into the construction of the observation matrices in the state-space formulation, and it is necessary to distinguish the treatment of stock and flow variables. Moreover, the linear state-space formulation is only an approximation of the true relationship—nonlinear filtering procedures, typically through stochastic simulations, would be needed to obtain an exact solution; but linear approximations may be sufficient.

Details on the specific expressions for the variables and parameters in the measurement and state equations depend on the mixed frequencies present in the model. And they become more complicated and more computationally intensive as higher and higher frequencies are involved.

Kalman filtering procedures can be applied to re-estimate unknown parameters in this state-space formulation and perform signal extraction to compute estimates of the latent factor. This Kalman filtering approach needs to be adapted for special complicated features of the high-dimensional and mixed-frequency problem. In particular, the use of mixed-frequency data for the indicators introduces missing data in the "measured" variables. In addition, additional attention is needed, and other complications in the calculations arise when dealing with indicators that are flow variables.

Details for formulating the "observable" state-space model are in Harvey (1989), Mariano and Murasawa (2003, 2010), and Aruoba et al. (2009). For both monthly and quarterly models, we can assume that the functional form with the target variable can take on nonlinear characteristics. Recently, the use of machine learning models has become popular to deal with, among other things, this functional nonlinearity. Among many algorithms used, Random Forest (RF) has stood out:

Random Forest (RF; Breiman, 2001) is an ensemble method that uses a large number of decision trees. The underlying idea is to build a large number of uncorrelated trees. Then, by averaging the predictions over several noisy trees, the variance of the aggregate prediction is reduced. And since the trees can also have relatively low bias, the aggregate prediction can exhibit both low variance and low bias. The key in this technique is the low correlation between the trees: this is guaranteed by (i) growing each tree on a bootstrap subsample of the initial dataset, and (ii) restricting the number of variables considered at each node—only a random subset of variables is allowed, forcing an even lower correlation between the trees. This technique is increasingly used in economic forecasting (Soybilgen and Yazgan, 2021; Medeiros et al., 2021).

3.3 **Covariate Pre-Selection Methods**

When forecasting with a high-dimensional dataset, the literature generally concludes that factor models are significantly more accurate when selecting fewer, but more informative predictors (Bai and Ng, 2008). On a more theoretical level, Boivin and Ng (2006) show that larger datasets lead to poorer forecasting performance when idiosyncratic errors are cross-correlated or when variables with higher predictive power are dominated.

The underlying idea of pre-selection is to rank the regressors x_{it} based on a measure of their predictive power relative to the target variable (or goal). In this study, we consider three techniques from the literature:

1. The "Sure Independence Screening" (SIS) of Fan and Lv (2008): regressors are ranked based on their marginal correlation with the target predictor. Fan and Lv (2008) provide a theoretical basis for their approach, demonstrating that it has the sure screening property that "all important variables survive after applying a screening relative to the target variable in this procedure with probability tending to 1". This approach has been used for short-term forecasting in Ferrara and Simoni (2019) or Proietti and Giovannelli (2021).
2. Based on the t-statistic: each regressor x_{it} is ranked based on the absolute value of the t-statistic associated with its coefficient estimates in a univariate regression of x_{it} on the target variable y_t. The univariate regression also includes four lags of the dependent variable to control for endogenous dynamics. Although originating from genetic studies (Bair et al., 2006), this technique has been applied to economics, for example, in Jurado et al. (2015).
3. Least Angle Regression (LARS) as in Bai and Ng (2008): while the two methods above are based on univariate relationships of regressors with the target variable, this considers the presence of other predictors. LARS (Efron et al., 2004) is an iterative forward selection algorithm. Starting with no predictors, it adds the predictor x_i most correlated with the target variable y and then moves the coefficient β_i in the direction of its least squares estimate so that the correlation of x_i with the residual ($y - \beta_i x_i$) decreases. The procedure continues until another predictor x_j has a similar correlation with $y - \beta_i x_i$ as x_i. At this point, x_j is added to the active set, and the procedure continues moving both coefficients β_i and β_j equi-angularly in the direction of their least squares estimates, until another predictor x_k has as much correlation with the residual (now $y - \beta_i x_i - \beta_j x_j$). This approach has been used in short-term forecasting, such as in Schumacher (2010), Bulligan et al. (2015), or Falagiardia and Sousa (2015).

3.4 **Factor Extraction Methods**

The econometric framework for dealing with high dimensionality, in general, is based on a factor model. Formally, we assume that the pre-selected dataset X_t can be represented by a factor structure with an r-dimensional factor vector Ft, a loading matrix Λ, and an idiosyncratic component ξt of the common factors:

$$X_{tm} = \Lambda * Ft + \xi t \quad (7)$$

Following the canonical structure of Stock and Watson (2002), static factors are extracted via Principal Component Analysis (PCA). PCA assumes that Ft and ξt are independent and identically distributed (i.i.d.). The factors can be estimated via maximum likelihood and are consistent estimators, provided that the factors are generalized and the idiosyncratic dependence and cross-correlation in ξt are weak.

Exploring all these models and their combinations in detail would make this article excessively long. Therefore, we present in Table 3 the models that will be estimated in detail in the results section.

Table 3 - Models for High-Dimensional and Mixed-Frequency Forecasting

MODELS	DESCRIPTION	Target Variable
M1	ARIMA(1,0,1)	Interannual Variation of Agricultural GDP
M2	Exponential Smoothing in State-Space Models with Idiosyncratic Shocks	Agricultural GDP at Level
M3	Exponential Smoothing in State-Space Models with Idiosyncratic Shocks	Interannual Variation of Agricultural GDP
M4	Generalized Exponential Smoothing in State-Space Models	Agricultural GDP at Level
M5	Generalized Exponential Smoothing in State-Space Models	Interannual Variation of Agricultural GDP
M6	Bridge Equation Model - VAR(5) Structure - Integration of Expectations	Interannual Variation of Agricultural GDP
M7	Bridge Equation Model - VAR(4) Structure - Integration of Expectations	Interannual Variation of Agricultural GDP - 1st Difference
M8	Bridge Equation Model - FA-VAR(2) Structure - With SIS Method	Interannual Variation of Agricultural GDP - 1st Difference
M9	Bridge Equation Model - FA-VAR(3) Structure - With SIS Method	Interannual Variation of Agricultural GDP
M10	Bridge Equation Model - FA-VAR(5) Structure - With t-Test Method	Interannual Variation of Agricultural GDP - 1st Difference
M11	Bridge Equation Model - FA-VAR(5) Structure - With t-Test Method	Interannual Variation of Agricultural GDP
M12	Bridge Equation Model - FA-VAR(5) Structure - With LARS Method	Interannual Variation of Agricultural GDP
M13	MIDAS-DFM - SIS Method	Interannual Variation of Agricultural GDP - 1st Difference
M14	MIDAS-DFM - t-Test Method	Interannual Variation of Agricultural GDP - 1st Difference
M15	MIDAS-DFM - LARS Method	Interannual Variation of Agricultural GDP - 1st Difference
M16	RF - LARS Method	Interannual Variation of Agricultural GDP - 1st Difference
M17	RF - MIDAS - LARS Method	Interannual Variation of Agricultural GDP - 1st Difference
M18	MIDAS-DFM	Interannual Variation of Agricultural GDP - 1st Difference

Source: Own elaboration

4.1 **Data Description and Treatment**

The construction of the dataset used in this study considered the possibility of a wide range of time series with potential impact on agricultural economic activity. Table 9 in Appendix details a total of 79 variables for the study. With the objective of forecasting the year-on-year growth of agricultural GDP, we consider two response variables: the chained series from IBGE without seasonal adjustment, from which we can obtain the official annual growth numbers of agricultural GDP through the moving average, as well as the year-on-year variation itself. With the exception of 19 time series (in Table 9 in the appendix, variables 5 to 23), all are the result of the first difference of the natural logarithm, which represents a continuous approximation of the marginal variation. This treatment is common in time series, with the aim of reducing the chance of non-stationarity and attenuating the presence of structural breaks. The remaining 19 are already represented in units suitable for modeling. The dataset used was based on the covariates most cited in articles on forecasting agricultural activities, and includes variables ranging from climatic variables and climate shocks, such as El Niño, to variables directly related to agricultural activity, such as the food and beverage industries, commodity prices, agricultural exports, agricultural credit, food inflation components, and expectations and uncertainties from the Focus Bulletin.

We can observe in Figure 5 that there is a trend of growth in agricultural GDP over its history. However, this growth has erratic variations with shocks that include deep troughs and large peaks of growth. A natural suspicion is that part of these abrupt variations may be related to shocks in the implicit prices in the composition of GDP; however, the implicit deflator of agricultural GDP makes it clear that the sources of these large variations may have other origins. This greater variability in agricultural GDP growth makes its forecasting a greater and more challenging task.

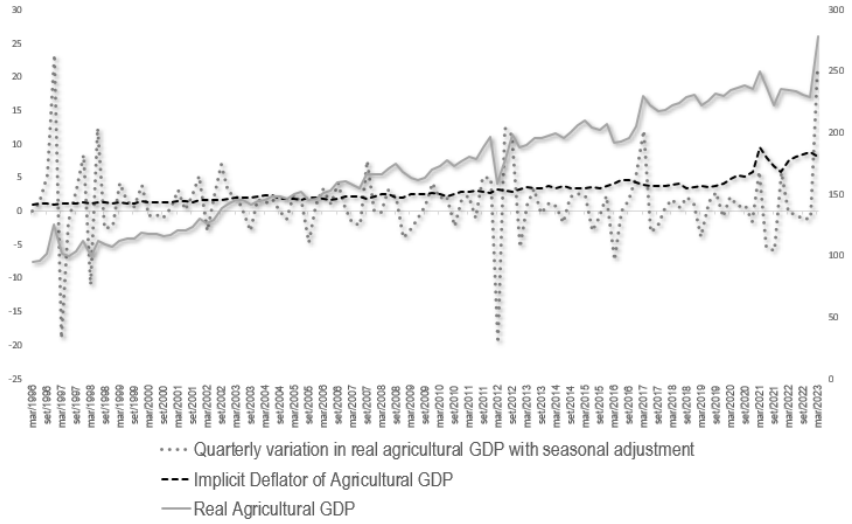

Figure 5: Dynamics of Variation in Quarterly Agricultural GDP Growth and the Implicit Deflator

As is common in the forecasting literature, the first approach will be the adjustment of univariate time series models, with the aim of identifying the data-generating process of the series, and a subsequent evolution to multivariate models, in search of identifying a more informative set of time series that bring predictive gains. For all cases, we consider forecasts with 4 steps ahead within a rolling window of 60 quarters (or 180 months for monthly models), covering the period from the 1st quarter of 2002 to the 4th quarter of 2016, resulting in a forecasting exercise with 24 quarters up to the last quarter of 2002.

4.2 **Univariate Models for Forecasting Annual Agricultural GDP**

To identify the data-generating process of the series, the Box and Jenkins approach is traditionally used for univariate cases. A decisive step is the identification of the presence of a unit root, which can render econometric efforts entirely spurious. Therefore, we present the results of the ADF tests for both the level series of agricultural GDP used to calculate the year-on-year variation, as well as the first difference of this variation.

Table 4 - ADF Test

Agricultural GDP at Level

Lag	No Drift and No Trend	With Drift without Trend	With Drift and with Trend
1	0,99	0,58	0,32
2	0,99	0,63	0,01
3	0,99	0,67	0,02
4	0,99	0,67	0,01

Variation - Agro GDP.

Lag	No Drift and No Trend	With Drift without Trend	With Drift and with Trend
1	0,01	0,03	0,08
2	0,01	0,01	0,01
3	0,01	0,01	0,01
4	0,01	0,01	0,01

Margin variation - Agricultural GDP 1st Difference

Lag	No Drift and No Trend	With Drift without Trend	With Drift and with Trend
1	0,01	0,01	0,01
2	0,01	0,01	0,01
3	0,01	0,01	0,01
4	0,01	0,01	0,01

Source: Own elaboration

Note: The body of the Table reports the P-Values of the ADF Test

The results considering four lags indicate the presence of a unit root in agricultural GDP in levels, with the exceptions being in the presence of a data-generating process of the series with drift and deterministic trend for higher lags. Analyzing the year-on-year variation, we have evidence of stationarity under certain conditions, but there may still be a unit root if the data-generating process of the series has drift and deterministic trend considering one lag. Finally, the first difference of the year-on-year variation constitutes the series in which we have the greatest evidence of stationarity. With these results, we test different specifications for use in the models that will be submitted to predictive exercises, with the predictive ability for each of these conditions (series in levels, year-on-year variation, and its first difference) being determinants for the production of good forecasts.

The next step is the identification of the dependence structure of the series through the Autocorrelation, Partial Autocorrelation, and Cross-Correlation functions. This identification has its own statistical tests that consider different AR and MA orders and their respective information criteria values (AIC, BIC, HQ, etc.). The automatic procedure suggests an order (1,1,1) for agricultural GDP in levels and (1,0,1) for its year-on-year variation[^3]. As detailed in the methodology section, parametric and non-parametric models were estimated for the construction of forecasts for univariate models. We present in Figure 6 the results of the univariate parametric ARIMA models and the exponential smoothing models in State-Space with idiosyncratic shocks as described in Table 3:

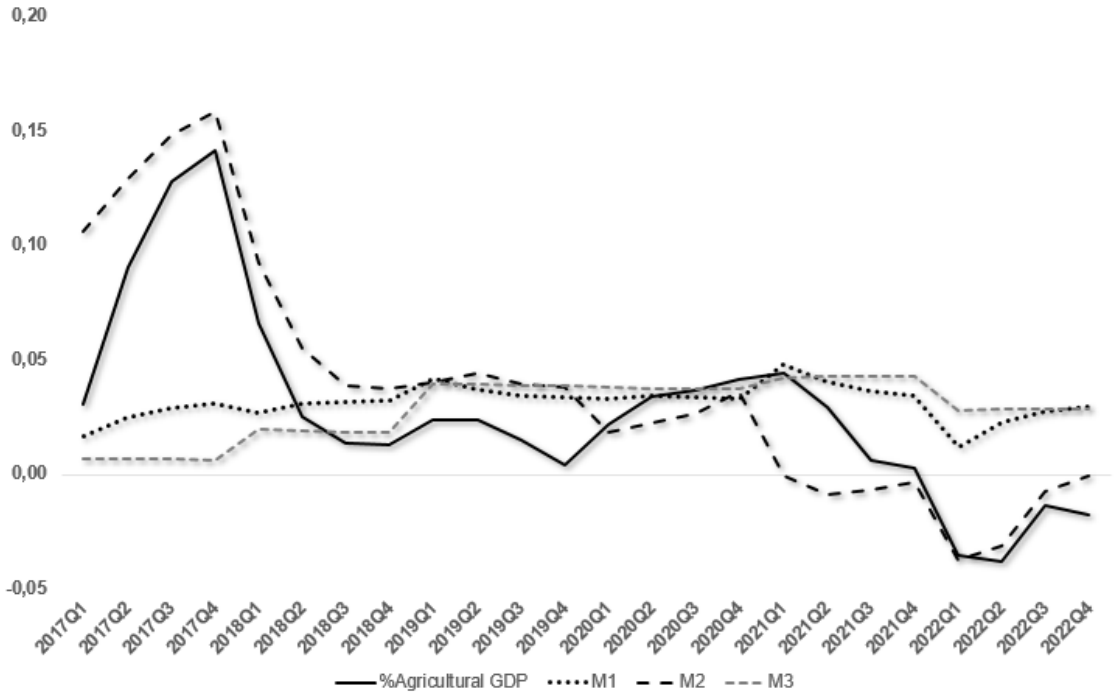

Figure 6: Predictions of Parametric and Nonparametric Models

The results for the parametric models suggest that the ARIMA(1,0,1)[1] model considering the year-on-year variation has a worse fit compared to the non-parametric models with state-space specification; however, the M2 model with agricultural GDP in levels captures movements at the beginning of the out-of-sample forecasting exercise not captured by the M1 and M3 models, which are forecasting the year-on-year variation. Next, we present in Figure 7 the results of the generalized non-parametric models estimated in State-Space form via the Kalman filter.

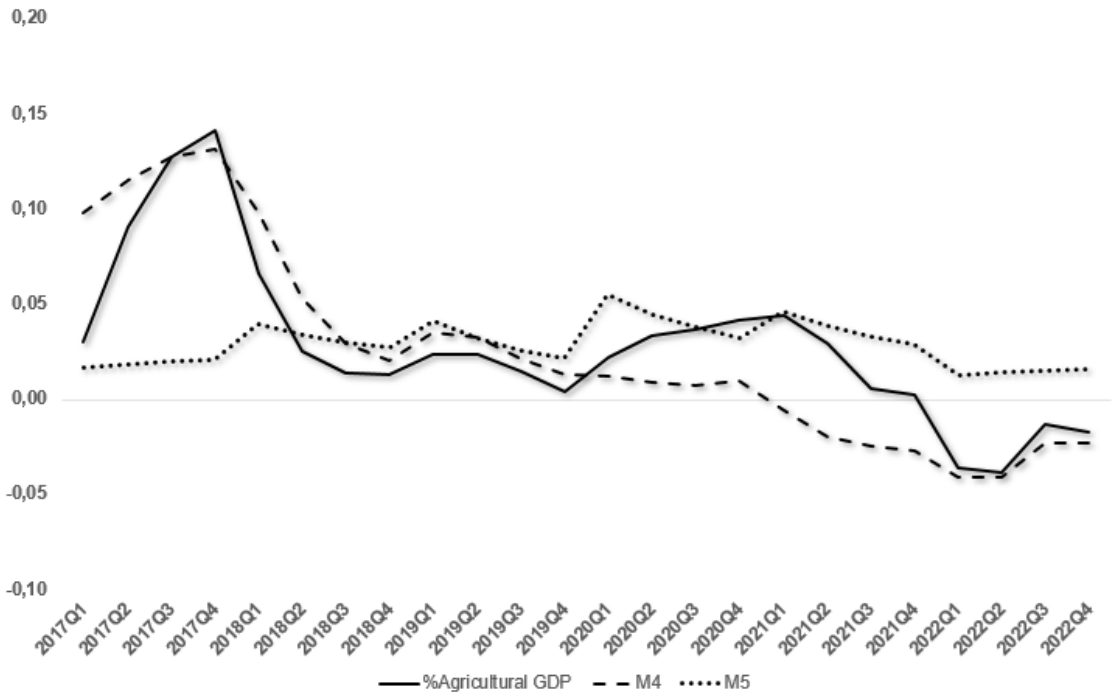

Figure 7: Predictions of Generalized Nonparametric Models

Considering this class of models, the forecasts with agricultural GDP in levels (M4) presented a better result. In the next section, we will begin the use of multivariate models with different specifications.

4.3 **Temporal Interaction of Expectations, Uncertainties, and Agricultural Production**

The dynamics between expectations, uncertainties, and production is an important step for identifying predictive ability, as the expectations measured by the Focus Bulletin forecasts and their uncertainty may have simultaneity impacts with producers' decisions. To highlight this aspect, we use a VAR model

[1] Several specifications were tested, both parametric and non-parametric models, as well as the other models that will be discussed, however we present in the article the results that have the best predictive performance given the objective of the predictive exercise.

considering this endogeneity, and for this purpose, we select the order of temporal dependence in this system with 3 significant lags[2]:

Considering the balance between dependence structure and parsimonious model, we choose the suggestion of the HQ information criterion, with order 3. Next, we identify a univariate Granger causality[3] simultaneously between expectations, uncertainties, and agricultural production, which validates the endogenous hypothesis of these variables. We follow these results for estimation and forecasting in a rolling window of the VAR(3) with these three time series. The highlight of this model is the fact that only the combination of these variables results in an R^2 of these combinations above 60% (0.73;0.66 and 0.62), reinforcing the hypothesis of endogeneity of these variables. The predictive results:

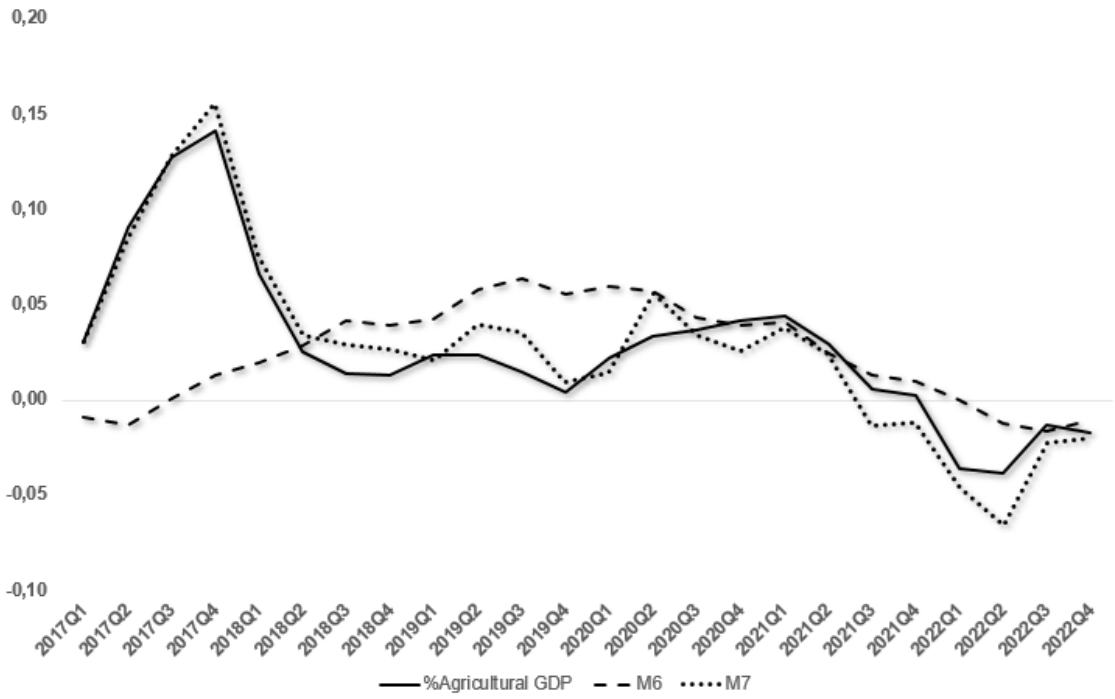

Figure 8: Interaction of Expectations and Agricultural Production

The out-of-sample results suggest that although there is a simultaneity shock between the variables[4], the multivariate dynamics are centered on the first difference of the year-on-year variation, with results that follow the movements of agricultural GDP. The results of multivariate Granger causality show that variations in agricultural GDP cause, in the Granger sense, uncertainties and expectations.

[2] We always use a maximum lag of 12 quarters to identify the order of the VAR(p) models that will follow.
[3] Detailed results of the Granger Causality tests are described in the Appendix
[4] The results of the Portmanteau and ARCH tests show that the model eliminated the serial autocorrelation structure.

4.4 Covariate Pre-Selection Method in High Dimension, Dimensionality Reduction, and Bridge Equation Models

The bridge equation model is a model in which monthly data are quarterlyized and used for different other econometric approaches. Once the data are quarterlyized, we proceed to deal with the issue of high dimensionality of the data.

As presented in Section 3.3 and 3.4, we will use three different approaches for variable pre-selection, and subsequently, the extraction of factors from these selected covariates. This approach becomes necessary in the context of high dimensionality, as, for our out-of-sample forecasting exercise with 24 quarters, there would be 60 quarters of model adjustment for a set of 79 possible covariates.

In Table 4, we present the results of the selection of the three methods. We observe that the LARS method selected 57 covariates, the method based on the t-test selected only 17 covariates, while the SIS method selected 24 covariates. These results, although they may serve for more descriptive analysis of agricultural GDP, allow us to identify an ideal number of factors for these pre-selected datasets, indicating a significantly smaller number of factors compared to the initial set of covariates. However, it is worth noting that each pre-selection method has its implicit hypotheses, which generate different sets of information that may contribute more or less to the forecasting of our target variable.

Table 4 - Pre-Selection of Variables

Variable Index	LARS Selection	Selection t-test	SIS Selection	Variable Index	LARS Selection	Selection t-test	SIS Selection	Variable Index	LARS Selection	Selection t-test	SIS Selection
3	PIB_EUA			52	PRECO_MILHO	PRECO_MILHO	PRECO_MILHO	28	CARCACA_FRANGO		
5	PIB_CHINA			54	PRECO_TRIGO			33	QTD_LEITE		QTD_LEITE
6	VAR_FOCUS_PIB_AGRO		VAR_FOCUS_PIB_AGRO	55	PRECO_SOJA	PRECO_SOJA		35	IMPORTACOES_CHINA		IMPORTACOES_CHINA
24	QTD_COURO		QTD_COURO	60	IND_FAO_CARNE	IND_FAO_CARNE		38	INFL_ALIMENTOS_EUA		INFL_ALIMENTOS_EUA
25	QTD_BOVINO_ABATIDO		QTD_BOVINO_ABATIDO	63	IND_FAO_OLEOS	IND_FAO_OLEOS		40	ENERGIA_ONS	ENERGIA_ONS	
29	QTD_SUINO_ABATIDO			64	DOLAR	DOLAR		42	ABCR_PESADOS	ABCR_PESADOS	ABCR_PESADOS
30	CARCACA_SUINO			65	CRED_AGROP			43	LIC_VEIC_NOVOS		
31	QTD_GALINHAS			58	IBC_BR		IBC_BR	53	PRECO_SUCO_LAR	PRECO_SUCO_LAR	PRECO_SUCO_LAR
32	QTD_OVOS			67	ICC_FERCOMERCIO		ICC_FERCOMERCIO	56	PRECO_FAR_SOJA	PRECO_FAR_SOJA	PRECO_FAR_SOJA
34	QTD_LEITE_IND		QTD_LEITE_IND	69	PRECO_BOI_CEPEA	PRECO_BOI_CEPEA		57	PRECO_OLEO_SOJA	PRECO_OLEO_SOJA	
36	INFL_ALIMENTOS_EUROPA	INFL_ALIMENTOS_EUROPA		72	EXP_AGROP			58	IC_BR_BACEN	IC_BR_BACEN	
37	IMPORTACOES_EUROPA		IMPORTACOES_EUROPA	74	PMC_SUP_ALIMENTO			59	IND_FAO_COM		IND_FAO_COM
41	PROD_MAQ_AGRO		PROD_MAQ_AGRO	77	PIM_BEBIDAS			61	IND_FAO_LAT		IND_FAO_LAT
44	CONSUMO_ABRAS			78	PIM_FUMO			62	IND_FAO_CEREAIS	IND_FAO_CEREAIS	IND_FAO_CEREAIS
46	PRECO_ALGODAO			79	PIM_TEXTIL		PIM_TEXTIL	68	PRECO_ALGODAO_CEPEA		
47	PRECO_ARROZ	PRECO_ARROZ		4	PIB_EUROPA			71	PRECO_SOJA_CEPEA		PRECO_SOJA_CEPEA
48	PRECO_CACAU	PRECO_CACAU		19	TEMP_NORTE			73	PMC_RESTRITO		
50	PRECO_BOI			26	CARCACA_BOVINA		CARCACA_BOVINA	75	PIM		
51	PRECO_LEITE			27	QTD_FRANGO_ABATIDO			76	PIM_ALIMENTOS	PIM_ALIMENTOS	PIM_ALIMENTOS

Source: Own elaboration

Table 5 - Factor Extraction

Method	Number of Variables	Quantity of Factors
LARS	57	8
t Test	17	5
SIS	24	8

Source: Own elaboration

Note: In the SIS method, we use a truncation with correlations greater than 0.1 in modulus, as a criterion.

When we consider a general dimensionality reduction via principal components, in search of a representative variance structure, at least 5 components were necessary to explain 72% of the original variability of the data. When we estimate a model with these components, we have an adjusted R^2 above 45% in any of the pre-selection methods, which is higher than the result produced only with the multivariate model with the selected variables. These results indicate that we do not gain from the use of general components, and therefore, we will continue to use the specific components resulting from the individual selection of covariates.

4.5 **Estimating and Forecasting in High Dimension and Mixed Frequency**

The specific components allow us to include a new class of model that assumes endogeneity between the estimated principal components and the target variable (year-on-year variation of agricultural GDP). The results of the order selection test for these models traditionally suggest low orders, even if some higher orders offer better forecasts. For each estimated model, different autoregressive orders were tested in the VAR(p) model and decided by information criteria, and only then were the models submitted for predictive use.

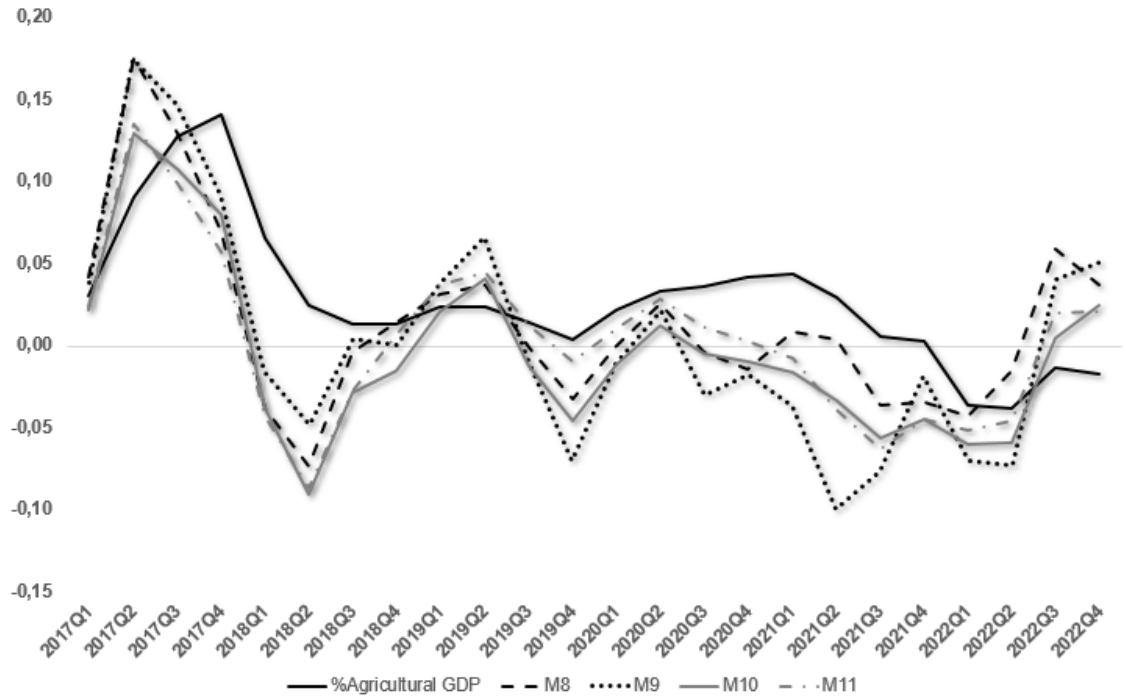

Figure 9: Predictions from Mixed-Frequency Factor Models

The results of the models in Figure 9 show that the dynamics obtained from the SIS and t-statistic-based selection methods have movements, almost throughout the test sample, synchronized, indifferent, when we consider the year-on-year variation or its first difference.

An alternative widely used in the literature is the MIDAS models, which incorporate dynamics of different data frequencies. Initially, we estimate the specific principal components from the monthly data. Once the base is adjusted to compose the two frequencies, monthly and quarterly, we select the order of a VAR structure. We present in Figure 10 the results of the forecasts:

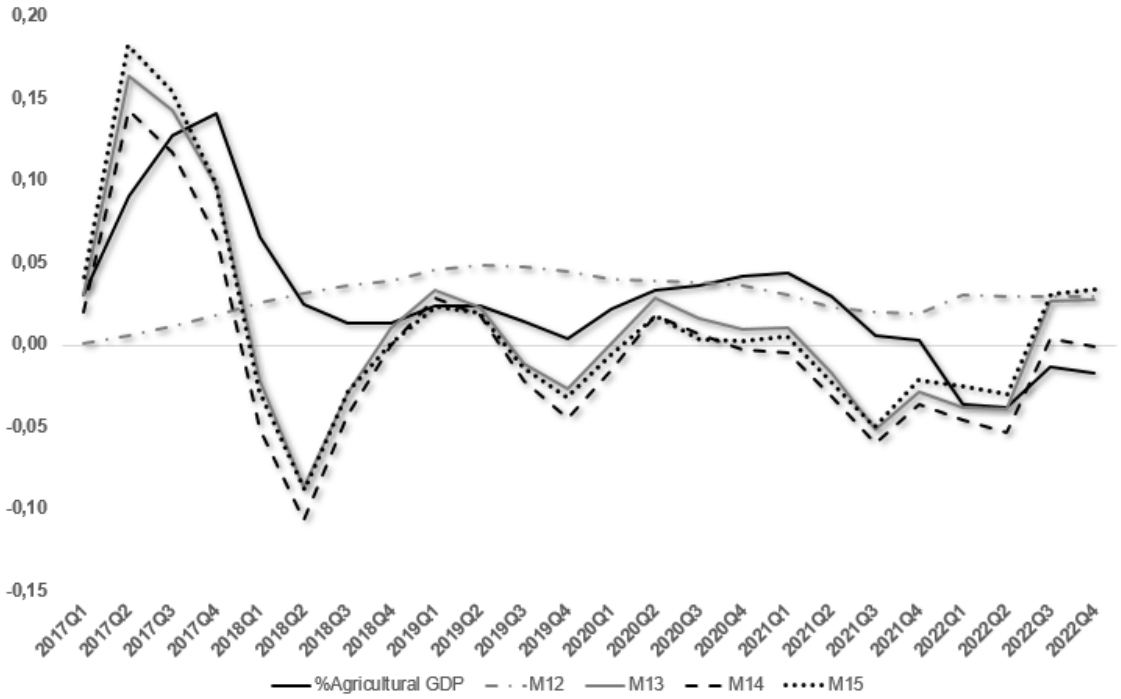

Figure 10: Factorial and Dynamic Factorial Models with Mixed Frequency Predictions

The results show that, unlike the models with factors extracted by the SIS and t-test methods, the factors generated by the LARS selection method are not suitable for the year-on-year variation. The DFM-MIDAS models, regardless of the pre-selection method, have predictive movements that follow the movements of agricultural GDP, but with large mismatches in magnitude for some periods.

Dynamic factor models offer an alternative approach to dimensionality reduction. The proposal involves estimating factors instead of principal components, aiming to achieve a representative covariance structure, which is then used to dynamically compose forecasts of the target variable. A limitation of this approach is the appropriate determination of the number of factors to be used. Although parametric tests exist for this purpose, in practice, the percentage of explained variance often guides the decision on the number of factors to include. To incorporate higher-frequency data, the factors are estimated using the selected monthly variables. Once the number of factors is determined, we seek to identify the lag order to make these factors dynamic. Finally, using bridge equation models, we estimate the MIDAS-DFM model. This latter model is compared with the forecasts generated by machine learning models, considering both bridge equation specifications (quarterly data) and MIDAS (monthly data). The results are presented in Graph 11.

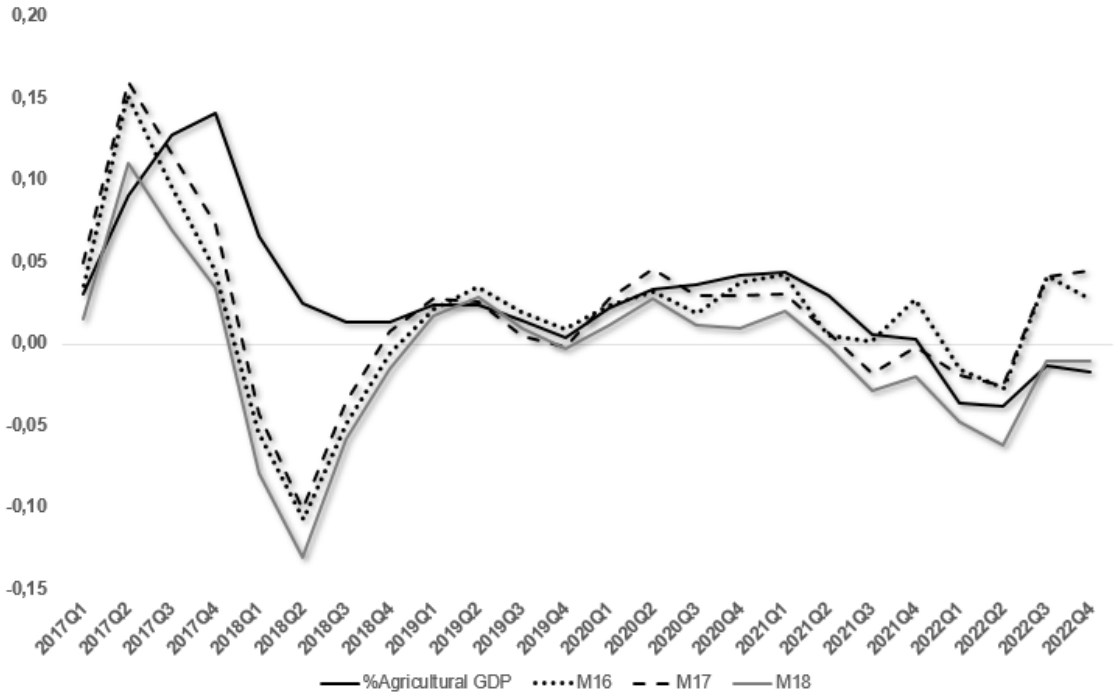

Figure 11: Machine Learning and Factorial Dynamic Model Predictions

Except for the beginning of the out-of-sample forecast, the three models quickly adjust to track the movements of agricultural GDP. The response of these models to a shock is faster, which, in some situations, can be extremely useful for other predictive exercises.

4.6 **Exploratory Analysis of the Predictive Ability of the Models**

A final exercise is the comparison of the predictive ability of the presented models. This final section, the exploratory analysis, highlights important aspects that relate all the methodologies presented in this study. Table 6 summarizes the two main predictive error measures, MAPE and RMSE. The results vary across models, with some showing good MAPE values but not necessarily the best RMSE results. These findings reinforce the need for inferential methods to analyze the predictive ability of the models tested in this study. Although specific models performed as well as or better than the Focus Bulletin forecasts—particularly for the January predictions, when uncertainty is higher—exploratory evidence suggests the possibility of a model combination with superior performance.

Table 6 - Exploratory Analysis of the Predictive Ability of the Models

MODELS	MAPE	RMSE
M1	3,12	0,18
M2	**2,16**	**0,07**
M3	3,55	0,25
M4	**2,08**	**0,07**
M5	**2,98**	**0,18**
M6	3,41	0,25
M7	**1,06**	**0,02**
M8	3,70	0,23
M9	4,87	0,33
M10	4,17	0,25
M11	3,73	0,24
M12	3,65	0,24
M13	3,26	**0,19**
M14	4,02	0,27
M15	3,77	0,22
M16	3,17	0,24
M17	**3,01**	0,20
M18	3,57	0,30
Focus - Medians	**3,58**	**0,22**
Focus - Medians -Januarys	**4,00**	**0,26**
Focus - Medians - Decembers	**3,79**	**0,25**

Source: Own elaboration

Note: The body of the Table reports MAPE in percentage terms

4.7 **Predictive Ability Test of the Models and Model Ensemble**

To assess the predictive ability among the models, we use the Diebold-Mariano test for predictive accuracy. The results are presented in Table 7.

Table 7 - Diebold-Mariano Test for Predictive Ability

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18
M1		**0,076**	0,985	**0,078**	0,574	0,922	**0,009**	0,721	0,923	0,757	0,739	0,993	0,536	0,796	0,683	0,721	0,603	0,815
M2			0,957	0,455	0,900	0,951	**0,020**	0,982	0,998	0,988	0,978	0,963	0,955	0,979	0,979	0,946	0,933	0,948
M3				**0,045**	**0,002**	0,416	**0,010**	0,401	0,729	0,472	0,436	0,214	0,278	0,547	0,393	0,452	0,341	0,629
M4					0,898	0,946	**0,014**	0,984	0,999	0,991	0,983	0,960	0,959	0,982	0,981	0,947	0,934	0,949
M5						0,961	**0,020**	0,688	0,901	0,728	0,710	0,999	0,520	0,773	0,657	0,700	0,584	0,802
M6							**0,014**	0,428	0,745	0,496	0,463	0,425	0,300	0,568	0,416	0,476	0,365	0,645
M7								0,998	1,000	0,999	0,997	0,993	0,995	0,995	0,997	0,985	0,984	0,978
M8									0,898	0,664	0,583	0,554	**0,108**	0,788	0,441	0,595	0,267	0,792
M9										0,129	0,113	0,225	**0,033**	0,248	**0,078**	0,202	**0,098**	0,418
M10											0,336	0,483	**0,021**	0,814	0,268	0,453	0,150	0,768
M11												0,517	**0,081**	0,878	0,375	0,532	0,208	0,821
M12													0,305	0,592	0,433	0,496	0,375	0,665
M13														0,979	0,992	0,847	0,714	0,901
M14															0,149	0,208	**0,038**	0,722
M15																0,623	0,294	0,807
M16																	**0,093**	0,909
M17																		0,926

Source: Own elaboration

Note: P-value of the Diebold-Mariano test for the one-sided predictive ability test

The inferential results suggest that models M2, M4, M5, M7, M13, M15, and M17 exhibit superior predictive ability compared to the other models. To evaluate these findings, we propose three different ensemble approaches:

1. The average of these models' predictions;
2. A composition of these models' forecasts at each predictive window, such as a regression of their predictions;
3. A variation of approach 2, but considering only the models that were statistically significant.

The ensemble results are presented in Table 8. Both exploratory and inferential evidence were confirmed, as even the simple average of the models demonstrated superior predictive performance compared to the Focus Bulletin. The second ensemble approach yielded the best predictive results, achieving an 80% improvement in MAPE and a fivefold improvement in RMSE.

Table 8 - Model Ensemble and Predictive Ability

Combination of Forecasts	MAPE	RMSE
Average with Best Models	1,68	0,04
Ensemble with Best Models	**0,71**	**0,01**
Ensemble with Best Significant Est. Models	1,02	0,02
Focus - Medians	**3,58**	**0,22**
Focus - Medians -Januarys	**4,00**	**0,26**
Focus - Medians - Decembers	**3,79**	**0,25**

Source: Own elaboration

Note: The body of the Table reports MAPE in percentage terms

The results become even clearer when analyzing Figure 12. The combination of the best models captures, with each model incorporating its own assumptions, key characteristics of agricultural GDP movements. These findings suggest that any predictive gains will depend both on high dimensionality and on the information set derived from high-frequency data.

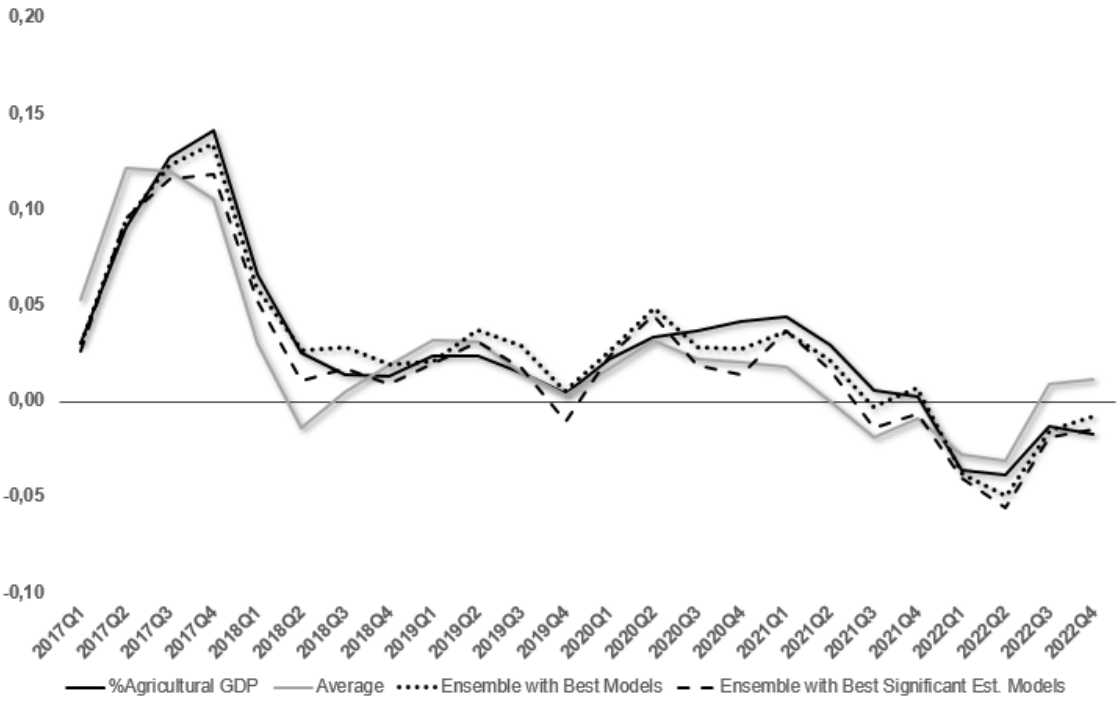

Figura 12: Predictions from Ensemble of Predictive Models

5. **Conclusion**

Forecasting economic activity variables has proven to be a challenging task, requiring complex analyses and models due to the wide range of factors that can influence these variables. Specifically, agricultural activity presents a unique set of forecasting difficulties, given its strong dependence on climatic variables, government policies, national and international commodity prices, and market fluctuations. Structural breaks caused by uncertainty related to weather conditions, such as rainfall and temperature, are not uncommon and represent a critical factor for agricultural production. Additionally, government policies—such as agricultural subsidies, environmental regulations, and trade agreements—can significantly influence agricultural production and, consequently, impact economic forecasts. Finally, market fluctuations, including supply and demand dynamics, commodity prices, and changes in consumer preferences, also play a crucial role in the complexity of predicting agricultural activity. Therefore, forecasting agricultural GDP is even more challenging than forecasting overall GDP. On the other hand, the increasing availability of covariates may provide a way to address this complexity.

This study presented a broad range of econometric methodologies for forecasting agricultural GDP in an environment characterized by high dimensionality (where the number of covariates approaches the number of time-series observations of the target variable) and mixed frequencies. The proposed solutions helped elucidate characteristics that link expectations and uncertainty to agricultural production. The study also identified time series with the most

significant contributions to explaining agricultural GDP variability. Different dimensionality reduction approaches were employed, allowing for the construction of models that combine monthly and quarterly dynamics.

The results compared the predictive performance of univariate models, including simpler structures such as ARIMA models and more complex approaches such as State-Space models. When considering individual models, the results for the out-of-sample period and rolling window forecasts indicated the superiority of model ensembles. Three different statistical variable selection methods proved to be valuable mechanisms for handling high dimensionality and incorporating high-frequency versions of models. The study introduced a final ensemble model with a MAPE of 0.71%, compared to the Focus Bulletin's MAPE of 4% for January forecasts. This represents a predictive gain of more than 80% over the more uncertain forecasts made by the Focus Bulletin regarding annual agricultural GDP growth. Finally, to the best of our knowledge, this study represents a pioneering effort in Brazil in applying these methodologies within a high-dimensional and mixed-frequency data context.

References

Aruoba, S.B., Diebold, F.X., Scotti, C., ADS, 2009. Real-time measurement of business conditions. J. Bus. Econ. Stat. 27 (4), 417–427.

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006). "Prediction by supervised principal components", Journal of the American Statistical Association, 101(473), pp. 119–137

Bai, J., and Ng, S. (2008). "Forecasting economic time series using targeted predictors", Journal of Econometrics, 146(2), pp. 304–317

Bennett, M.J., Hugen, D.L., 2016. Financial Analytics With R, Building a Laptop Laboratory for Data Science. UK, Cambridge.

Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. J. Bus. Econ. Stat. 13, 253–265.

Boivin, J. and Ng, S. (2006). "Are more data always better for factor analysis", Journal of Econometrics, 132, pp. 169–194

Breiman, L. (2001). "Random Forests", Machine Learning, 45(1), pp. 5–32

Bulligan, G., Marcellino, M., and Venditti, F. (2015). "Forecasting economic activity with targeted predictors", International Journal of Forecasting, 31(1), pp. 188–206

Coutino, A., 2005. A High-Frequency Model for Mexico. Project LINK web-site, http://www.chass.utoronto.ca/LINK>.

Doz, C., Giannone, D., Reichlin, L., 2011. A two-step estimator for large approximate dynamic factor models based on Kalman filtering. J. Econ. 164 (1), 188–205.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). "Least angle regression", Annals of Statistics, 32(2), pp. 407–499

Falagiarda, M., and Sousa, J. (2015). “Forecasting euro area inflation using targeted predictors: is money coming back?”, European Central Bank Working Paper Series, No 2015

Fan, J., and Lv, J. (2008). “Sure independence screening for ultrahigh dimensional feature space”, Journal of the Royal Statistical Society Series B, 70(5), pp. 849–911

Ferrara, L., and Marsilli, C. (2019). “Nowcasting global economic growth: A factor-augmented mixed-frequency approach”, The World Economy, 42(3), pp. 846–875

Foroni, C., Marcellino, M., 2012. A Comparison of Mixed Frequency Approaches for Modelling Euro Area Macroeconomic Variables. Economics Working Papers ECO 2012/07, European University Institute.

Foroni, C., Marcellino, M., 2013. A Survey of Econometric Methods for Mixed Frequency Data. Economics Working Papers ECO 2013/02, European University Institute. 224 PART II Macro Econometrics

Ghysels, E., 2013. Matlab Toolbox for Mixed Sampling Frequency Data Analysis Using MIDAS Regression Models. Version 5. May, 2013.

Ghysels, E., 2016a. MIDAS Matlab Toolbox. Version 2.1.

Ghysels, E., 2016b. Macroeconomics and the reality of mixed-frequency data. J. Econ. 193, 294–314.

Ghysels, E., Marcellino, M., 2018. Applied Economic Forecasting Using Time Series Methods.Oxford University Press.

Ghysels, E., Santa-Clara, P., Valkanov, R., 2004. The MIDAS Touch: Mixed Data Sampling Regression Models. Working paper, Chapel Hill, NC.

Ghysels, E., Sinko, A., Valkanov, R., 2007. MIDAS regressions: further results and new directions. Econ. Rev. 26 (1), 53–90.

Ghysels, E., Kvedaras, V., Zemlys, V., 2016. Mixed frequency data sampling regression models: the R package midasr. J. Stat. Softw. 72 (4), 1–35, https://www.jstatsoft.org/article/view/ v072i04.

Gilbert, P., Meijer, E., 2015. Package ‘tsfa’—Time Series Factor Analysis. May 1, 2015, https://cran.r-project.org/web/packages/tsfa/tsfa.pdf.

Harvey, A.C., 1989. Forecasting, Structural Time Series Models and the Kalman Filter.Cambridge University Press, Cambridge.

Heiss, F., 2016. Using R for Introductory Econometrics. Germany.

Helske, J., 2018. Package ‘KFAS’—Kalman Filter and Smoother for Exponential Family State Space Models. September 19, 2018, https://cran.r-project.org/web/packages/KFAS/KFAS.pdf.

Holmes, E., Ward, E., Scheuerell, M.D., Wills, K., 2018. Package ‘MARSS’—Multivariate Autoregressive State-Space Modeling. November 2, 2018, https://cran.r-project.org/web/packages/ MARSS/MARSS.pdf.

Hyndman, R.J., Athanasopoulos, G., 2018. Forecasting Principles and Practice, second ed.Otexts, Online, Open-Access Textbooks.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M.,Petropoulos, F., Razbash, S., Wang, E., Yasmeen, F., R Core Team, Ihaka, R., Reid, D., Shaub, D., Tang, Y., Zhou, Z., 2019. Package 'Forecast'—Forecasting Functions for Time Series and Linear Models. January 18, 2019, https://cran.r-project.org/web/packages/forecast/forecast.pdf.

Inada, Y., 2005. A High-Frequency Model for Japan. Project LINK web-site, http://www.chass.utoronto.ca/LINK>.

Jurado, K., Ludvigson, S., and Ng, S. (2015). "Measuring Uncertainty", American Economic Review, 105(3), pp. 1177–1216

Klein, L.R., Mak, W., 2005. University of Pennsylvania Current Quarter Model of the Chinese Economy. Forecast Summary. Project LINK web-site, http://www.chass.utoronto.ca/LINK.

Klein, L.R., Ozmucur, S., 2001. The use of surveys in macroeconomic forecasting. In: Welfe, W.(Ed.), Macromodels 2001. University of Lodz, Poland.

Klein, L.R., Ozmucur, S., 2002. Some possibilities for indicator analysis in economic forecasting.In: Project LINK Fall Meeting, University of Bologna, October 2002.

Klein, L.R., Ozmucur, S., 2004. University of Pennsylvania Current Quarterly Model of the United States Economy Forecast Summary. Project LINK website, http:www.chass.utoronto.ca/LINK.

Klein, L.R., Ozmucur, S., 2008. University of Pennsylvania high frequency macroeconomic modeling. In: Mariano, R.S., Tse, Y.-K. (Eds.), Econometric Forecasting and High-Frequency Data Analysis. Lecture Notes Series, vol. 13. World Scientific Publishers, Singapore, pp. 53–91. Institute for Mathematical Sciences, National University of Singapore. 2008. Earlier version presented at the High Frequency Modeling Conference, Singapore Management University (SMU), Singapore, May 7–8, 2004.

Klein, L.R., Park, J.Y., 1993. Economic forecasting at high-frequency intervals. J. Forecast.12, 301–319.

Klein, L.R., Park, J.Y., 1995. The University of Pennsylvania model for high-frequency economic Forecasting. In: Economic and Financial Modelling, vol. 2, pp. 95–146. Autumn 1995.

Klein, L.R., Sojo, E., 1987. Combinations of high and low frequency data in macroeconometric models. In: Paper Presented at the Session "Can Economic Forecasting Be Improved?" American Economic Association. Chicago, Illinois. December.

Klein, L.R., Sojo, E., 1989. Combinations of high and low frequency data in macroeconometric models. In: Klein, L.R., Marquez, J. (Eds.), Economics in Theory and Practice: An Eclectic Approach. Kluwer Academic Publishers, pp. 3–16.

Klein, L.R., Eskin, V., Roudoi, A., 2003. Empirical regularities in the Russian economy.In: Project LINK Spring Meeting. United Nations, New York, April 23–25, 2003.

Klein, L.R., Eskin, V., Roudoi, A., 2005. University of Pennsylvania and Global Insight Current Quarter Model of the Russian Economy. Forecast Summary. Project LINK web-site, http:// www.chass.utoronto.ca/LINK>.

Kvedaras, V., Zemlys, V., 2016. Package 'midasr'—Mixed Data Sampling Regression, August 16, 2016. CRAN (Comprehensive R Archive Network), https://cran.r-project.org/web/ packages/midasr/midasr.pdf.

Liu, H., Hall, S.G., 2001. Creating high frequency national accounts with state-space modelling: a Monte Carlo experiment. J. Forecast. 20, 441–449.

Luethi, D., Erb, P., Otziger, S., 2018. Package 'FKF'—Fast Kalman Filter. July 20, 2018, https://cran.r-project.org/web/packages/FKF/FKF.pdf.

Mariano, R.S., 2002. Testing forecast accuracy. In: Clements, M.P., Hendry, D.F. (Eds.), A Companion to Economic Forecasting. Blackwell, Oxford.

Mariano, R.S., Murasawa, Y., 2003. A new coincident index of business cycles based on monthly and quarterly series. J. Appl. Economet. 18 (4), 427–443.

Mariano, R.S., Murasawa, Y., 2010. A coincident index, common factors, and monthly real GDP. Oxf. Bull. Econ. Stat. 72 (1), 27–46.

Mariano, R.S., Ozmucur, S., 2018. High-mixed-frequency forecasting models for GDP and inflation. Ch. 1, In: Pauly, P. (Ed.), Global Economic Modeling—A Volume in Honor of Lawrence Klein. World Scientific Publishing Co. Pt. Ltd, pp. 2–29.

Mariano, R.S., Preve, D., 2012. Statistical tests for multiple forecast comparison. J. Econ. 169 (1),123–130.

Mariano, R.S., Tse, Y.-K. (Eds.), 2008. Econometric Forecasting and High-Frequency Data Analysis, In: Lecture Notes Series, vol. 13. World Scientific Publishers, Singapore. Institute for Mathematical Sciences, National University of Singapore.

Ozmucur, S., 2009. Current quarter model for Turkey. In: Klein, L.R. (Ed.), The Making of National Economic Forecasts. Edward Elgar Publishing Ltd., Cheltenham, UK/Northampton, MA, USA, pp. 245–264, 2009, Chapter 9.

Medeiros, M., Vasconcelos, G., Veiga, A., and Zilberman, E. (2021). "Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods", Journal of Business & Economic Statistics, 39(1), pp. 98–119

Pauly, P. (Ed.), 2018. Global Economic Modeling—A Volume in Honor of Lawrence Klein.World Scientific Publishing Co. Pt. Ltd.

Petris, G., 2018. Package 'dlm'—Bayesian and Likelihood Analysis of Dynamic Linear Models.June 13, 2018, https://cran.r-project.org/web/packages/dlm/dlm.pdf.

Proietti, T., and Giovannelli, A. (2021). "Nowcasting monthly GDP with big data: A model averaging approach", Journal of the Royal Statistical Society: Series A, 184(2), pp. 683–706

Pfaff, B., Stigler, M., 2018. Package 'vars'—VAR Modelling. August 6, 2018, https://cran.r-project.org/web/packages/vars/vars.pdf.

Reinhart, A., 2017. Package 'pdfetch'. October 15, 2017, https://cran.r-project.org/web/packages/pdfetch/pdfetch.pdf.

Revelle, W., 2019. Package 'psych'—Procedures for Psychological, Psychometric, and Personality Research. January 13, 2019. https://cran.r-project.org/web/packages/psych/psych.pdf.

Sargent, T., Sims, C., 1977. Business cycle modeling without pretending to have too much a priori economic theory. In: New Methods in Business Cycle Research. Federal Reserve Bank of Minneapolis.

Schumacher, C. (2010). “Factor forecasting using international targeted predictors: The case of German GDP”, Economics Letters, 107(2), pp. 95–98

Shumway, R.H., Stoffer, D.S., 2017. Time Series Analysis and Its Applications, With R Examples, fourth ed. Springer International Publishing, Switzerland.

Soybilgen, B., and Yazgan, E. (2021). “Nowcasting US GDP Using Tree-Based Ensemble Models and Dynamic Factors”, Computational Economics, 57, pp. 387–417

Stock, J., and Watson, M. (2002). “Forecasting using principal components from a large number of predictors”, Journal of the American Statistical Association, 97(460), pp. 1167–1179

Stock, J.H., Watson, M.W., 1989. New indexes of coincident and leading economic indicators. In: Blanchard, O.J., Fischer, S. (Eds.), NBER Macroeconomics Annual, vol. 4. MIT Press,Cambridge, Massachusetts, pp. 351–409.

Stoffer, D., 2017. Package ‘astsa’—Applied Statistical Time Series Analysis. December 15, 2017,https://cran.r-project.org/web/packages/astsa/astsa.pdf.

Vinod, H.D., 2011. Hands-on Intermediate Econometrics Using R, Templates for Extending Dozens of Practical Examples. World Scientific Publishing Co., Singapore

Wickham, H., Bryan, J., Kalicinski, M., Valery, K., Leitienne, C., Colbert, B., Hoerl, D.,

Miller, E., 2018. Package ‘Readxl’—Read Excel Files. December 20, 2018, https://cran.r-project.org/web/packages/readxl/readxl.pdf

6. **Appendix**

6.1 **Description of the time series used**

Table - High-Dimensional Database

Variable index	Variable Name	Variable	Source	Periodicity
1	PIB_AGROPECIARIO	Quarterly Agricultural GDP Chained Series - Moving Average	IBGE (Brazilian Institute of Statistics and Geography)	Quarterly
2	PIB_AGRO_P	Quarterly Agricultural GDP - Chained Series - Moving Average - Interannual variation	IBGE	Quarterly
3	PIB_EUA	U.S. Gross Domestic Product, constant values - year-on-year change	Bureau of Economic Analysis	Quarterly
4	PIB_EUROPA	GDP Eurozone - Demand side - Year-on-year change	Eurostat	Quarterly
5	PIB_CHINA	China's GDP - % change compared to the same period last year	National Bureau of Statistics of China	Quarterly
6	VAR_FOCUS_PIB_AGRO	Monthly Forecast of Annual Agricultural GDP - Average of Medians - Focus Bulletin	BACEN (Brazilian Central Bank)	Monthly

7	DP_FOCUS_PIB_AGRO	Standard Deviation of Annual Agricultural GDP Forecasts	BACEN	Monthly
8	INDICE_ELNINO	Oceanic Niño Index (ONI)	NOAA (Estados Unidos)	Monthly
9	INDICE_OSCILACAO_CLI	Southern Oscillation Index (SOI)	BOM (Austrália)	Monthly
10	INFL_ALIMENTOS_CHINA	Chinese Food Inflation	National Bureau of Economic Statistics	Monthly
11	IPCA_ALIMENTO	IPCA - Food	IBGE	Monthly
12	PRECIP_BRASIL	Precipitation in mm (average) - Brazil	INMET	Monthly
13	PRECIP_NORTE	Precipitation in mm (average) - North	INMET	Monthly
14	PRECIP_NORDESTE	Precipitation in mm (average) - Northeast	INMET	Monthly
15	PRECIP_SUD	Precipitation in mm (average) - Southeast	INMET	Monthly
16	PRECIP_SUL	Precipitation in mm (average) - South	INMET	Monthly
17	PRECIP_CEO	Precipitation in mm (average) - Central-West	INMET	Monthly
18	TEMP_BRASIL	Temperature in C (average) - Brazil	INMET	Monthly
19	TEMP_NORTE	Temperature in C (average) - North	INMET	Monthly
20	TEMP_NORDESTE	Temperature in C (average) - Northeast	INMET	Monthly
21	TEMP_SUD	Temperature in C (average) - Southeast	INMET	Monthly
22	TEMP_SUL	Temperature in C (average) - South	INMET	Monthly
23	TEMP_CEO	Temperature in C (average) - Central-West	INMET	Monthly
24	QTD_COURO	Total quantity of whole raw cowhide purchased and received from third parties for tanning (Units)	IBGE	Monthly
25	QTD_BOVINO_ABATIDO	Total quantity of cattle carcasses slaughtered in Brazil	IBGE	Monthly
26	CARCACA_BOVINA	Total weight of carcasses of cattle slaughtered in Brazil	IBGE	Monthly
27	QTD_FRANGO_ABATIDO	Total quantity of carcasses of slaughtered chickens in Brazil	IBGE	Monthly
28	CARCACA_FRANGO	Total weight of carcasses of slaughtered chickens in Brazil	IBGE	Monthly
29	QTD_SUINO_ABATIDO	Total quantity of pig carcasses slaughtered in Brazil	IBGE	Monthly
30	CARCACA_SUINO	Total weight of carcasses of slaughtered pigs in Brazil	IBGE	Monthly
31	QTD_GALINHAS	Number of laying hens (Heads)	IBGE	Monthly
32	QTD_OVOS	Quantity of eggs produced (Thousand dozen)	IBGE	Monthly
33	QTD_LEITE	Quantity of raw milk, chilled or not, purchased (Thousand liters)	IBGE	Monthly

34	QTD_LEITE_IND	Quantity of raw milk, chilled or not, industrialized (Thousand liters)	IBGE	Monthly
35	IMPORTACOES_CHINA	Chinese imports	Customs General Administration PRC	Monthly
36	INFL_ALIMENTOS_EUROPA	Food Inflation Europe	Eurostat, European Central Bank, European Commission.	Monthly
37	IMPORTACOES_EUROPA	European imports	Eurostat, European Central Bank, European Commission.	Monthly
38	INFL_ALIMENTOS_EUA	American Food Inflation	Bureau Economic Statistics	Monthly
39	IMPORTACAO_BENS_EUA	Imports of American Goods	U. S. Census Bureau	Monthly
40	ENERGIA_ONS	Average Electric Power Load Mw Average	National Electric System Operator (ONS) - IPDO	Monthly
41	PROD_MAQ_AGRO	Agricultural and Road Machinery in Units	National Association of Automotive Vehicle Manufacturers (ANFAVEA)	Monthly
42	ABCR_PESADOS	ABCR Activity Index - Heavy	Brazilian Association of Highway Concessionaires (ABCR)	Monthly
43	LIC_VEIC_NOVOS	Licensing of New Vehicles in units	National Federation of Motor Vehicle Distribution (FENABRAVE)	Monthly
44	CONSUMO_ABRAS	Consumption in Brazilian Households - Index Number Jan/2001=100 and Percentage	Brazilian Association of Supermarkets (ABRAS)	Monthly
45	PRECO_ACUCAR	Commodities Spot Prices - End of Monthly Period - Sugar	NYMEX	Monthly
46	PRECO_ALGODAO	Commodities Spot Prices - End of Monthly Period - Cotton	ICE	Monthly
47	PRECO_ARROZ	Commodities Spot Prices - End of Monthly Period - Rice	Chicago	Monthly
48	PRECO_CACAU	Commodities Spot Prices - End of Monthly Period - Cocoa	ICE	Monthly
49	PRECO_CAFE	Commodities Spot Prices - End of Monthly Period - Arabia Coffee	ICE	Monthly
50	PRECO_BOI	Commodities Spot Prices - End of Monthly Period - Live Cattle	CME	Monthly
51	PRECO_LEITE	Commodities Spot Prices - End of Monthly Period - Milk	CME	Monthly
52	PRECO_MILHO	Commodities Spot Prices - End of Monthly Period - Corn	Chicago	Monthly
53	PRECO_SUCO_LAR	Commodities Spot Prices - End of Monthly Period - Orange Juice	ICE	Monthly
54	PRECO_TRIGO	Commodities Spot Prices - End of Monthly Period - Wheat	Chicago	Monthly
55	PRECO_SOJA	Commodities Spot Prices - End of Monthly Period - Soybeans	Chicago	Monthly

56	PRECO_FAR_SOJA	Commodities Spot Prices - End of Monthly Period - Soybean Oil	Chicago	Monthly
57	PRECO_OLEO_SOJA	Commodities Spot Prices - End of Monthly Period - Soybean Meal	Chicago	Monthly
58	IC_BR_BACEN	Brazil Commodities Index (IC-Br)	BACEN	Monthly
59	IND_FAO_COM	International Agricultural Commodity Quotations Index - FAO	Food and Agriculture Organization of United Nations	Monthly
60	IND_FAO_CARNE	International Agricultural Commodity Quotations Index - FAO - Meat	Food and Agriculture Organization of United Nations	Monthly
61	IND_FAO_LAT	International Agricultural Commodity Quotations Index - FAO - Dairy	Food and Agriculture Organization of United Nations	Monthly
62	IND_FAO_CEREAIS	International Agricultural Commodity Quotations Index - FAO - Cereals	Food and Agriculture Organization of United Nations	Monthly
63	IND_FAO_OLEOS	International Agricultural Commodity Price Index - FAO - Oils and Fats	BACEN	Monthly
64	DOLAR	Exchange Rates - End of Monthly Period - Ptax sale	BACEN	Monthly
65	CRED_AGROP	Credit Balance by Economic Activity - Agriculture (Backcast of the series using the monthly median indicator of the Agricultural GDP from the Focus Bulletin)	BACEN	Monthly
66	IBC_BR	Central Bank Economic Activity Index (IBC-Br)	Federation of Commerce of the State of São Paulo (Fecomércio)	Monthly
67	ICC_FERCOMERCIO	Consumer Confidence Index	Esalq	Monthly
68	PRECO_ALGODAO_CEPEA	Esalq Agricultural Commodities - Monthly Average - Cotton	Esalq	Monthly
69	PRECO_BOI_CEPEA	Esalq Agricultural Commodities - Monthly Average - Beef	Esalq	Monthly
70	PRECO_CAFE_CEPEA	Esalq Agricultural Commodities - Monthly Average - Coffee		Monthly
71	PRECO_SOJA_CEPEA	Esalq Agricultural Commodities - Monthly Average - Soy	Esalq	Monthly
72	EXP_AGROP	Agricultural Exports - US$ FOB	Secex/MDIC	Monthly
73	PMC_RESTRITO	Retail Sales Volume Index - PMC Volume	IBGE	Monthly
74	PMC_SUP_ALIMENTO	Retail Sales Volume Index - PMC - Volume - Hypermarkets, Supermarkets, Food Products, Beverages and Tobacco	IBGE	Monthly
75	PIM	PIM-PF - Industrial Sections and Activities - (Backcast of the series using the monthly median indicator of the Agricultural GDP from the Focus Bulletin)	IBGE	Monthly

76	PIM_ALIMENTOS	PIM-PF - Food - Industrial Sections and Activities - (Backcast of the series using the monthly median indicator of the Agricultural GDP from the Focus Bulletin)	IBGE	Monthly
77	PIM_BEBIDAS	PIM-PF - Beverages - Industrial Sections and Activities - (Backcast of the series using the monthly median indicator of the Agricultural GDP from the Focus Bulletin)	IBGE	Monthly
78	PIM_FUMO	PIM-PF - Tobacco - Industrial Sections and Activities - (Backcast of the series using the monthly median indicator of the Agricultural GDP from the Focus Bulletin)	IBGE	Monthly
79	PIM_TEXTIL	PIM-PF - Textile - Industrial Sections and Activities - (Backcast of the series using the monthly median indicator of the Agricultural GDP from the Focus Bulletin)	IBGE	Monthly

Source: Own elaboration