Universidade de Brasília

Faculdade de Economia, Administração e Contabilidade

Departamento de Economia

Previsão da inflação de alimentos no domicílio usando dados meterológicos

Júlia Regina Scotti

Brasília

2019

Júlia Regina Scotti

Previsão da inflação de alimentos no domicílio usando dados meterológicos

Dissertação apresentada ao Programa de Mestrado em Economia da Universidade de Brasília como requisito à obtenção do título de Mestre em Ciências Econômicas.

Universidade de Brasília

Faculdade de Economia, Administração e Contabilidade

Departamento de Economia

Orientador: Prof. Marina Delmondes de Carvalho Rossi, PhD

Brasília

2019

Júlia Regina Scotti
Previsão da inflação de alimentos no domicílio usando dados meterológicos/
Júlia Regina Scotti. – Brasília, 2019-
45 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Marina Delmondes de Carvalho Rossi, PhD

Dissertação (Mestrado) – Universidade de Brasília
Faculdade de Economia, Administração e Contabilidade
Departamento de Economia, 2019.

1. Economia - Dissertação. 2. Previsão de inflação. 3. Dados meteorológicos I. Rossi, Marina Delmondes de Carvalho. II. Universidade de Brasília. III. Departamento de Economia. IV. Título

CDU 02:141:005.7

Resumo

A inflação de alimentos é o componente mais imprevisível da inflação ao consumidor e o que mais afeta a vida das pessoas, especialmente os pobres. Uma das razões para a volatilidade da inflação de alimentos é sua dependência das variações climáticas. Nos países em desenvolvimento, todos esses três fatos são mais pronunciados: primeiro, os preços dos alimentos são mais voláteis devido à maior prevalência de alimentos frescos sobre processados; Em segundo lugar, os alimentos frescos são mais suscetíveis às variações climáticas, não só porque são frescos, mas também porque a produção é geralmente menos tecnológica do que a dos alimentos processados; Finalmente, nos países em desenvolvimento, a alimentação no domicílio representa uma parcela maior do já limitado orçamento familiar. Isto tem duas consequências significativas, o peso da inflação de alimentos no domicílio na inflação cheia é maior nos países em desenvolvimento, e é mais difícil para as pessoas para acomodar aumentos dos preços dos alimentos.

Portanto, nós pesquisamos se os dados meteorológicos podem ajudar a prever a inflação de alimentos no domicílio. Estudamos o caso do Brasil, um país em desenvolvimento com ricos dados meteorológicos diários históricos e com um índice de inflação mensal confiável publicado 24 vezes por ano. Usamos dados de 2001 a 2018. Como método, usamos o lasso e a Random Forest porque eles lidam bem com modelos de alta dimensionalidade. Todas as nossas estimativas são pseudo-fora-da-amostra. Nossos resultados mostram que os dados meteorológicos melhoram as previsões quando comparados ao benchmark para cada um dos horizontes considerados de 1 a 7 meses. Em média ao longo dos horizontes, a razão para o benchmark da raiz quadada do erro médio quadrático (RMSE) no conjunto de holdout foi de 0,70 para o lasso e de 0,73 para Random Forest. Além disso, usamos nossas projeções de inflação de alimentos no domicílio para prever a inflação cheia. Descobrimos que, novamente, poderíamos melhorar as previsões para todos os sete horizontes com os dois modelos. No conjunto de holdout, o índice médio de RMSE para o benchmark para o lasso foi de 0,87, e para o Random Forest, 0,91. Nossos resultados sugerem que os dados meteorológicos podem melhorar substancialmente as previsões de inflação nos países em desenvolvimento.

Palavras-chave: Previsão de inflação, inflação de alimentos, dados metereológicos, lasso, random forests.

Abstract

Food inflation is the most unpredictable component of CPI and the one that mostly affects people's lives, especially the poor. One of the reasons for food inflation volatility is its dependence on weather variations. In developing countries, all these three facts are more pronounced: First, food prices are more volatile due to the higher prevalence of fresh over processed foods; Secondly, fresh foods are more susceptible to weather variations not only because they are fresh, but also because production is usually less technological than that of processed foods; Finally, in developing countries, food-at-home represents a higher share of the already limited household budget. This has two significant consequences, the weight of food-at-home inflation is larger in developing countries CPI, and it is more difficult for people to accommodate food price increases.

Therefore, we research whether meteorological data can help forecast food-at-home inflation. We study the case of Brazil, a developing country with rich historical daily weather data and with a reliable monthly inflation index published 24 times per year. We retrieved data for the last 18 years. As method, we use the lasso and Random Forest because they handle well high dimensional models. All our estimations are pseudo-out-of-sample, and we use, as benchmark, a direct estimated AR with lag order selected by BIC. Using a validation set, we choose to use climate normals averaged over the last 30 and 90 days. Our results show that the weather data improve forecasts when compared to the benchmark for each of the 1- to 7-month horizons considered. On average over the horizons, the ratio to benchmark of the Root Mean Squared Error (RMSE) in the holdout set was 0.70 for the lasso model and 0.73 for the Random Forest model. Additionally, we use our forecasts of food-at-home inflation to forecast headline inflation. We find that, again, we could improve forecasts for all seven horizons with both models. In the holdout set, the average RMSE ratio to benchmark for the lasso was 0.87, and for the Random Forest, 0.91. Our findings suggest that weather data might substantially improve inflation forecasts in developing countries.

Keywords: Food inflation, inflation forecasting, metereological data, lasso, random forest.

List of figures

List of tables

List of abbreviations

AIC	Akaike Information Criterion
AR	Auto-Regressive
BCB	Brazilian Central Bank
BIC	Bayesian Information Criterion
CPI	Consumer Price Index
INMET	Brazilian Institute of Metereology
IPCA	Extended National Consumer Price Index
LASSO	Least Absolute Shrinkage and Selection Operator
RMSE	Root Mean Squared Error
OLS	Ordinary Least Squares
POF	Consumer Expenditure Survey
RF	Random Forest
WMO	World Metereological Organization

Contents

1 Introduction

Food-at-home is the most unpredictable component of consumer price inflation and yet it is the one that mostly affects people's lives, especially the poor: the poorer the person, the larger the participation of food in the home budget (ENGEL, 1857) and at the same time, the smaller is the space to accommodate price increases. One of the reasons food-at-home is unpredictable is that it is highly dictated by the weather[1].

We use real-time weather data to forecast food-at-home inflation up to 7 months in advance. Then we use these forecasts to forecast full inflation and we show that, compared to well-constructed benchmarks, predictions using our forecasts of the food-at-home component reduce the mean squared error of forecasts of full inflation, on average for all 7 horizons, by 13%.

Inflation is hard to forecast: When only out-of-sample performance is considered, it is very hard to beat the best univariate model using any multivariate inflation forecasting model (STOCK; WATSON, 2008). Yet, it is also hard to ignore information that should influence inflation when forecasting. Subjective inflation forecasts seem to outperform model-based forecasts (FAUST; WRIGHT, 2013), sometimes by a wide margin, suggesting that the problem with inflation forecasting is not that the process of inflation itself is unpredictable, but that the determinants or the channels or interactions have not been well established.

The weather is the standard explanation for disturbances in food-at-home inflation in Brazil. The Inflation Report, published quarterly by the Brazilian Central Bank, regularly uses the weather to explain this component of Brazilian CPI (BCB, 2016; BANK, 2017a; BANK, 2017b). However, these documents usually cite the low weather predictability to justify not fully exploring the topic, as in Bank (2017a). We posit that weather influences food-at-home inflation with different lags and consequently past weather can help forecast future food-at-home inflation.

The influence of weather variability in commodity crops yields is well established (see, for example, Iizumi and Ramankutty (2015), Ray et al. (2015)). Less known is the impact of weather variability in non-commodity produce. But because non-commodity agriculture is less technology intensive and the producers tend to be less sophisticated, the influence of the weather is arguably greater in non-commodity products.

In Brazil, 34% of vegetables and 30% of fruits harvested are lost before commer-

[1] "Among the determinants of the forecast errors, an important part refers to weather events, which support the commonsense view that most of the uncertainty about the future evolution of food prices arises from low weather predictability, especially over longer horizon" (BANK, 2017a)

cialization (CENCI; SOARES; JúNIOR, 1997). A considerable share of these losses are caused, directly or indirectly, by less than ideal environmental conditions. Pre-harvest, the development of the crop is highly dependent on environmental conditions, in particular, temperature (SPAGNOL et al., 2017). During harvesting, excessive rains can not only destroy the products, but also prevent reaping, diminishing the quality of the late-harvested product and encouraging fast harvests, which cause the products to be incorrectly handled, leading to further damage to fruits and vegetables. Finally, the crop must then be stored. The ideal storage is a temperature and humidity controled warehouse, rarely seen in the developing world.

Transportation is also impacted by weather conditions, as usually trucks are not refrigerated in Brazi. In addition, with poor infrastructure, the shipping time is greatly dependent on the weather conditions, and these products are highly perishable.

Note that the mechanisms by which weather conditions influence food production are less pronounced in developed countries: with more technology and resources the dependence on the weather is mitigated in every step of the chain.

In developing countries, as a result of the limitedt household budgets, the share of food-at-home in total CPI is higher than in developed countries. In addition to the higher share, food prices are usually more volatile in developing countries due to the higher prevalence of fresh over processed foods (GÓMEZ; GONZÁLEZ; MELO, 2012).Therefore, improvements in food-at-home inflation forecasting can have a significant impact on the full inflation forecast.

Although monetary policy should not respond to temporary shocks – as is usually the case of unexpected food inflation, and specifically, food inflation caused by changes in weather patterns – large changes in food prices may affect inflation expectations, and through this channel, increase the persistence of the shock.

Brazil is a country of relatively poor people[2] that has just beaten hyperinflation in 1994. This has two major consequences. The first is that Brazilian contracts and bonds are usually pegged to some price index. The other is that, since people are poor and food-at-home takes a considerable share of their household budget, the food-at-home component of inflation has a substantial impact on their lives.

Therefore, considering the importance of improving the food-at-home inflation forecasts and the possibility that using weather data could potentially help in this objective, we undertake an empirical study that aims to use this data to predict future food-at-home inflation, and consequently of inflation itself.

As method, we explore the newly mainstream machine learning methods of $l1$ penalized regression (lasso and lasso variations) and random forest (Random Forest, as

[2] Income per capita was US$9159 in 2018 (FUND, 2018).

described by Breiman (2001) and variations). We chose these methods because both of them can handle high dimensional models, linear and nonlinear, respectively, and have shown to deliver forecast improvements over traditional approaches [3].

For food-at-home inflation we use this component of IPCA (Brazilian official CPI), and also, for training, the IPCA-15 series[4] We retrieve the historical weather data from the Brazilian Institute of Meteorology (INMET). We find that models using the weather data improved the forecasts upon the benchmark for all 7 horizons. On average over the horizons, the ratio to benchmark of the RMSE in the holdout set was 0.70 for the lasso model, 0.73 for the random forest model, and 0.65 for the ensemble model.

We then use these food-at-home inflation forecasts to forecast full inflation. We recreate a series of inflation without food-at-home, and pseudo-out-of-sample forecast this inflation-ex-food-at-home using an AR, BIC selected lag order, model. We then calculate our forecasts for full inflation as the forecasts for inflation-ex-food-at-home plus 0.16 [5] times the estimates for food-at-home inflation we previously obtained with each model. We are able to improve upon the benchmark for all seven horizons. On average over the horizons, the ratio to benchmark of the RMSE is 0.87 for the lasso model, 0.91 for the Random Forest model and 0.88 for the ensemble model.

Therefore our research contributes to the literature in two ways. First, we show that meteorological data improves food-at-home inflation forecasts dramatically from a standard benchmark up to seven months in advance. Secondly, we show that these forecasts can significantly improve Brazilian inflation (IPCA) forecasts.

Following this introduction, this paper is organized as follows. We briefly review the literature in section 2. In section 3 we describe our dataset and in section 4 describe our selection of benchmarks. Section 5 describes the methods used, and, finally, our results are shown in section 6.

[3] Smeekes and Wijler (2018) show this in general for macroeconomic forecasting using lasso, Garcia, Medeiros and Vasconcelos (2017) show this for real-time Brazilian inflation forecasting using lasso and Medeiros et al. (2018) show the robustness of random forests for US macroeconomic data.

[4] Which, with a slightly different methodology, measures inflation from the 15th of the previous month until the 14th of the reference month.

[5] The average weight of the food-at-home component of inflation.

2 Review of Literature

The literature on inflation forecasting is vast (GARCIA; MEDEIROS; VASCONCELOS, 2017). Before 2000, the focus of the literature was in finding the drivers of inflation, as in Phillips curve based models (STOCK; WATSON, 2008). However, since the publication of Atkeson et al. (2001), showing that none of the Phillips curve models for US inflation published before that time were able to improve upon a four-quarter random walk benchmark when considering the full sample period [1], which, at that time was 1984-1990, the focus turned to three questions: (1) Can we improve the random walk model benchmark? Stock and Watson (2008) showed that the answer is yes if we use their unobserved component stochastic volatility model, UC-SV. (2) If we are willing to forgo generalization ability, do Phillips curve models perform better than univariate benchmarks for specific ranges of time, that is, for particular business cycles? Again, the answer seems to be yes (see Stock and Watson (2007)); and (3): Can machine learning methods improve inflation forecasts? Garcia, Medeiros and Vasconcelos (2017) and Medeiros et al. (2018) show that yes, at least for US and Brazilian inflation, improvements can be made for most horizons using lasso or Random Forests.

In the last ten years, two branches of research stand out. The first focuses on the use of novel time series models and machine learning techniques to forecast inflation (see Ülke, Sahin and Subasi (2018), Smith and Maneesoonthorn (2018), Medeiros et al. (2018), Garcia, Medeiros and Vasconcelos (2017)). And the second reviews older literature aiming to consolidate their results (see Faust and Wright (2013), Stock and Watson (2008)).

There has also been much interest in using alternative datasets to help improve inflation forecasts. These alternative datasets are usually daily, or even higher frequency data, which brings about the issue of mixed frequency estimation. Breitung and Roling (2015) comprehensibly studies the forecasting of monthly inflation using daily indicators. Some examples of alternative datasets are Google trends (LI et al., 2015), commodity prices (CHEN; TURNOVSKY; ZIVOT, 2011; BREITUNG; ROLING, 2015), and raw materials and energy prices (MODUGNO, 2011).

Another branch of research is nowcasting inflation, the most famous example being the Billion Prices Project, which calculates, for example, *Inflacion verdadera Argentina* using online prices in that country and *Inflacion verdadera Venezuela* which uses crowdsourcing via a mobile app to estimate the inflation in that country [2].

[1] "It is difficult to make comparisons across papers in this literature because the papers use different sample periods, different inflation series, and different benchmarks models, and the quantitative results in the literature are curiously dependent upon these details" (STOCK; WATSON, 2008).

[2] Argentina and Venezuela do not have a reliable official CPI.

	Random Forest	Lasso	Factor Models	Principal Component	OLS
2004	1.3	29.5	2.6	15.0	26.9
2005	1.2	33.1	2.1	12.8	28.3
2006	1.3	35.5	1.8	9.3	27.5
2007	1.0	37.9	1.4	7.7	27.6
2008	1.1	38.0	1.4	6.5	29.6
2009	1.1	38.1	1.6	6.8	29.3
2010	1.3	45.3	1.5	6.8	27.8
2011	1.7	59.5	1.8	6.3	27.4
2012	2.3	63.0	1.6	6.4	28.4
2013	3.3	58.3	1.8	6.1	29.0
2014	4.0	63.7	1.3	6.0	33.3
2015	5.9	56.4	1.3	6.0	35.8
2016	7.7	60.3	1.4	5.5	37.9
2017	11.3	63.9	1.5	5.8	37.3
2018	13.6	56.6	1.1	5.4	35.3
2019	12.5	59.0	1.0	5.0	35.5

Table 1 – Google trends results for selected methods of forecasting

Concomitantly, since historical weather data has been made available in several countries, researchers have been studying how temperature, precipitation, and windstorms influence economic outcomes. Dell, Jones and Olken (2014) review studies that demonstrate impacts on agricultural output, industrial output, labor productivity, energy demand, health, conflict, and economic growth. They conclude that the findings provide rigorous econometric evidence that weather has manifold effects on economic activity and that developing countries appear particularly vulnerable to detrimental weather effects. They also show that this effect appears to be largely driven by the impact in agriculture.

Unusual weather is routinely cited as a factor in explaining unexpected fluctuations in economic activity (WILSON, 2017) and this relationship has been extensively studied (see, for example, Bloesch and Gourio (2015) and Sandqvist and Siliverstovs (2018)).

Similiarly to our work, Brown and Kshirsagar (2015) focused on short-term weather changes influencing the price of food by modelling the impact of weather disturbances on local food affordability. They examined 554 local markets in 51 developing countries and found that almost 20% of local market prices were affected by weather disturbances.

When Stock and Watson (2006) surveyed methods for forecasting with many predictors, they did not mention machine learning methods, including the lasso. Since then, research using these methods for forecasting economic variables has increased year after year [3]. Table 1 shows Google trends data (averaged over each year) reflecting the interest in each method. While the interest for Principal Components and Factor Models was reduced to roughly one-third of the interest at the beginning of the 2000s, the interest for lasso was doubled and that of random forests increased ten times.

Other interesting examples of these methods forecasting economic phenomena are Lohrmann and Luukka (2019), who use random forest to build a classification model for predicting the open-to-close returns of the S&P500, and Smeekes and Wijler (2018), who

[3] For a review of research using lasso and random forests, see Medeiros et al. (2018).

find that lasso-type penalized regression techniques to macroeconomic forecasting with high dimensional datasets are more robust to misspecification than factor models and deliver forecast improvements over traditional approaches.

3 Dataset

Our sample consists of 18 years (2001 to 2018). We used years 2001 to 2014 for the training set. The validation set consists of year 2015 and 2016, and the holdout set comprises 2017 and 2018.

3.1 Food-at-home Inflation

For food-CPI, we use the food-at-home component of the Brazilian consumer price index (IPCA), which is the official inflation measure in Brazil. IPCA is published monthly around the 7th day of the following month. Additionally, 15 days later, the Brazilian Institute of Geography and Statistics (IBGE) publishes IPCA-15, which uses the same methodology of IPCA (except it considers fewer metropolitan areas), but shifted 15 days.

For the training step, we build a series with both IPCA and IPCA-15 interspersed. Note that as two subsequent data points of these series share 15 days, they are not independent. We use this enhanced series for training our algorithms.

Food-at-home inflation is more volatile than inflation itself. Actually, with the exception of the components which, by law, suffer price adjustments only once a year (as formal education and private healthcare plans), food-at-home is the most volatile component of Brazilian inflation, with a standard deviation, for the entire sample, almost three times higher than inflation itself (1.06 and 0.39). The higher volatility of food-at-home inflation in comparison to full inflation is apparent in Figure 1.

The weight of each component in the inflation index changes over time according to price changes (maintaining the quantities of the original basket). In Brazil, the basket is established using the Consumer Expenditure Survey (POF) and the current weight is ca. 16% (see Figure 2).

3.2 Weather data

We retrieve weather data from the Brazilian Metereological Institute (INMET), which collects data from 256 ground weather stations around the country. The data consists of daily maximum temperature, minimum temperature, average humidity, hours of sunshine and millimeters of precipitation for each station.

We posit that the influence of the weather in food prices comes not from the absolute value of the meteorological variables, but from the difference to the typical value these variables take. Therefore, and inspired by Boldin and Wright (2015), we weather-adjusted

Figure 1 – Full and food-at-home inflation volatility

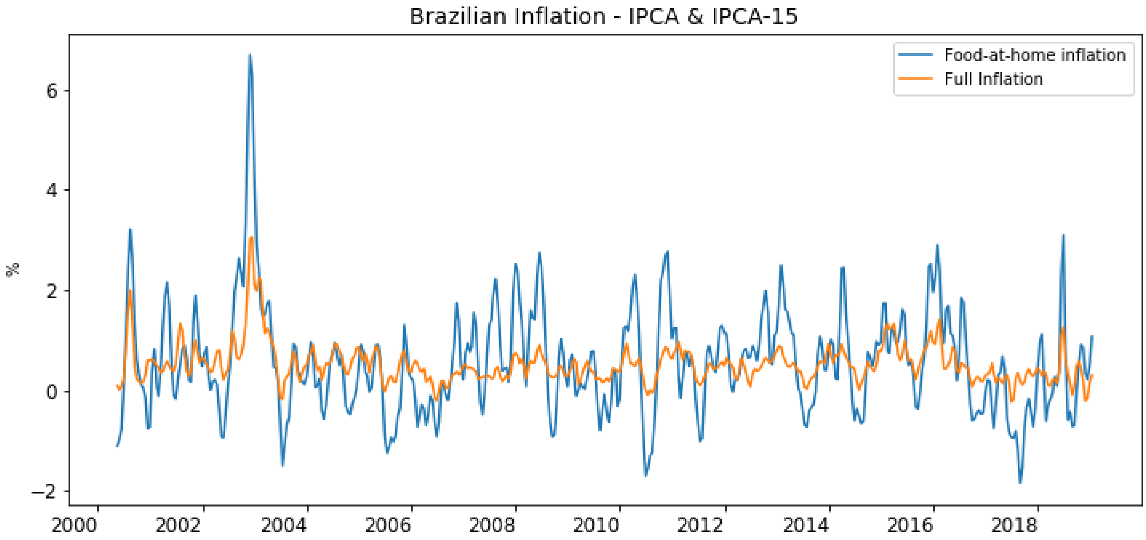
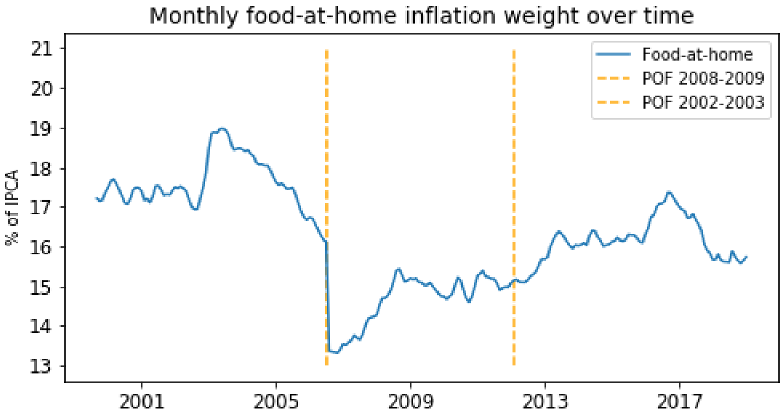


Figure 2 – Weight of food-at-home inflation in the inflation index



the weather data by subtracting the average for that day over the past years, if the day belongs to the validation or hold-out sample; and over the entire training sample, if the day belongs to the training sample. This methodology is in line with the World Meteorological Organization guidelines for the calculation of climate normals (ORGANIZATION, 2017).

Our dataset contains many missing values. In order not to discard series containing a few missing values, we fill the missing values of the daily weather-adjusted weather series with zeros. The rationale is that we assume that there was no variation from the climate normal of that day.

To match the frequency of the weather data and the food-at-home inflation series, we build moving averages series (15, 30, 45, 60, 90 days moving averages) and resample these series with the food-at-home inflation frequency.

4 Benchmarks

Food-at-home inflation has not been, to our knowledge, extensively investigated in the literature. Therefore, we study the best univariate models to pseudo-out-of-sample forecast h-periods ahead for the specific case of Brazilian food-at-home inflation and during the period that ranges from the beginning of 2016 to the end of 2017. We then use this model as a benchmark when evaluating our forecasts using weather data.

A question that arises when aiming to estimate the best univariate linear model is whether to use the iterated ("plug in") method, in which case the multistep forecasts are obtained from plugging in forecasts to obtain the h-period ahead forecast, or the direct method, in which the model itself is estimated h-periods ahead.

The earlier theoretical literature on this problem tended to favor the direct method. Bhansali (1996) derived asymptotic lower bounds for the h-step mean squared error (MSE) of prediction for the direct and iterated methods. His results show that the bound of the direct method is always smaller than that of the iterated method and that the iterated bound is not attainable. However, he notes that these results are asymptotic, and with a finite time, the variance term for the direct method can be expected to be larger than that for the plug-in method and that, for an ARMA model, the difference between their bias terms need not necessarily be large.

Therefore, the choice involves a trade-off between bias and variance: the iterated method produces more efficient parameter estimates than the direct method, but it is prone to bias if the model is misspecified (MARCELLINO; STOCK; WATSON, 2006).

Marcellino, Stock and Watson (2006) then undertake a large scale empirical comparison of iterated vs. direct forecasts using data on 170 U.S. macroeconomic time series. They consider whether the iterated or direct forecasts are more accurate on average and whether the distribution of pseudo-out-of-sample MSE for direct forecasts is statistically and substantially below that of iterated forecasts, as suggested by previous literature. They find that iterated forecasts tend to have lower sample MSE than direct forecasts, particularly if the lag length in the one period ahead is selected by AIC. Furthermore, they find that these improvements are modest and that the relative performance of iterated forecasts improve as the horizon lengthens.

We calculate pseudo-out-of-sample RMSE for iterated and direct forecasts for recursively estimated data-dependent selected and fixed order AR models for Brazilian food-at-home inflation. We found that the best model was direct forecast with lag order selected by BIC and therefore we use this model as a benchmark both for food-at-home and full inflation.

5 Methods

As forecasting methods, we use the lasso and Random Forests because both are particularly suited for high dimensional models.

5.1 Lasso

The lasso (least absolute shrinkage and selection operator), first described by Tibshirani (1996) tries to select the relevant variables using a norm-1 penalty on the standard loss function. Because the penalty is norm-1, most of the coefficients will be zero, and the operator effectively selects variables.

If X_t is the matrix with all the possible variables as columns, β is a row vector with the linear coefficients, many possibly zero, and $y_{t+h} = \beta X_t^{'} + \epsilon_{t+h}$, then, the lasso operator is:

$$\hat{\beta} = argmin \sum_{t=1}^{T} (y_{t+h} - \beta' x_t)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

and λ is the weight of the penalty term. If λ is zero then the lasso reduces to standard OLS. If λ is too high, all the entries in $\hat{\beta}$ will be zero.

The choice of λ is usually made using out-of-sample cross-validation, that is, the sample is separated into train and test set. For each value of λ, the train set is used to calculate $\hat{\beta}$, and these values are used to predict on the test set. Some metric is evaluated on the test set and this procedure, including the separation of test and train sets, is repeated increasing λ until the evaluation metric starts to deteriorate. The best λ is then selected.

Here, because we have a time series and cannot use the future to choose parameters to predict the future, we use only the training set to determine all hyperparameters, including the λ.

Recently, Garcia, Medeiros and Vasconcelos (2017) and Medeiros, Vasconcelos and Freitas (2016) showed that, from a collection of over 100 macroeconomic series, the lasso selected variables model outperforms benchmarks for some horizons, in the case of Brazilian and US inflation, respectively.

5.2 Random Forest

Random forests form a family of methods that consist in building an ensemble (forest) of decision trees grown from a randomized variant of the tree induction algorithm (LOUPPE, 2014). The most famous of which is the Random Forest (here, with capitals) algorithm described by Breiman (2001). Random Forest is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (BREIMAN, 2001).

Decision trees are non-parametric and can model arbitrarily complex relations between inputs and outputs and intrinsically implement feature selection. However, they can perfectly overfit (in which case every data point is classified as an individual leaf). Growing a forest of trees, and letting them vote for the most popular class (if in a classification problem) or averaging the forecasts (if in a regression problem), overcomes the overfitting obstacle while maintaining the advantages of decision trees.

In a recent working paper about the benefits of machine learning for forecasting economic data, Medeiros et al. (2018) advocate that the machine learning algorithm that deserves more attention is the random forest, both because of its variable selection ability and the potential to identify nonlinearities between macroeconomic variables. In their working paper, they forecast US CPI using hundreds of variables from the FRED-MD, a monthly database of 130 macroeconomic variables from the Federal Reserve of St Louis. They find that random forest performs better for several horizons and both in the 1990s and 2000s.

Previously, the same authors published a similar exercise but using Brazilian inflation (and therefore Brazilian macroeconomic variables, including forecasts by specialists from the FOCUS report [1]) and obtained slightly different results: they found that shrinkage (lasso) and complete subset regression (CSR) perform very well (GARCIA; MEDEIROS; VASCONCELOS, 2017). Random Forest, in this empirical exercise, performed better than RW and AR benchmarks, but worse than lasso, CSR, and Factor models.

5.3 Ensemble (Forecast combination)

Forecast combination is the combination of two or more individual forecasts from a panel of forecasts to produce a single, pooled forecast. Combining methods typically outperform individual forecasts, often by a wide margin, and while the forecasts can be combined linearly with weights estimated by OLS or with time-varying parameter weights estimated using the Kalman filter, simple combining methods – the mean, trimmed mean, or median – often perform as well as or better than sophisticated regression models

[1] Brazilian Central Bank publishes a weekly market expectations report, they call it the FOCUS report.

(STOCK; WATSON, 2006).

The machine learning literature calls the same process as bagging (short for bootstrap aggregating) (GOODFELLOW et al., 2016). They describe it as training different models separately, then having all the models vote on the output for test examples.

The idea is that, if the errors of the models are perfectly correlated, the model averaging does not help at all, but does not hurt. However, if the errors are uncorrelated, the expected error of the ensemble (combined forecast) is smaller than any of its members. This is also the idea behind random forests.

5.4 Estimation

For each technique (lasso or Random Forest), we evaluate a large number of combination of input features in the validation set, including and excluding different lags of IPCA and its components, and including or excluding different weather series and its lags. For example, we would test using the average of the last 30 days and the average of the previous 60 days.

Also, because the lasso is a linear model and we hypothesize that the influence of the weather on prices is not linear, but proportional to the deviation of the current weather to the climate normal of that day, we also considered the absolute value of our weather-adjusted weather series as a potential input to our models.

We called the rolling average series by its rolling window size, that is “90” means the series with the averages of the last 90 days, sampled at the inflation frequency. Moreover, “abs90” represents the same, but the averages are calculated over the absolute weather series. When we include the lagged values of one of these series, we added in parentheses the lag included. So that “abs30(t-2)” represents the absolute values of the weather series averaged over the last 30 days, sampled with the inflation series frequency and lagged by two periods.

In all the models we also included three lags (45 days) of Brazilian food-at-home inflation. We noticed that these features(regressors) are selected only for the first and second forecasting horizons.

5.4.1 Training, validation and holdout sets

We divided our sample as follows:

- Trainning set: From January, 15th, 2001 to December, 31st, 2014
- Validation set: From January, 31st 2015 to December, 31st, 2016

- Holdout set: From January, 31st 2017 to December, 31st, 2018

All calculations using the training set use the entire training set and use IPCA-15 in addition to IPCA. For example, when calculating the averages to weather-adjust the weather, we use the average over all the training set. The hyperparameter of the shrinkage of the lasso, λ, is also calculated using the full training sample (and chosen using the BIC criterion).

We used the validation set to determine the combination of features (regressors) and the holdout set to evaluate the performance of our chosen model in unseen data.

5.4.2 Evaluation Metrics and choice of model

To evaluate our results we use the root mean squared error (RMSE). We test several models and calculate their respective RMSE. We chose the model with the smallest cumulative RMSE in the validation set over the 7 horizons. We then use this model, and only this one, to forecast in the holdout set.

Theoretically, as both lasso and Random Forest implement feature selection and can handle high dimensional models, we could use all possible input features in one estimation. However, that did not work. If we kept the number of iterations small, the models became unstable and would give highly different results in each run. If we increased the number of iterations, then the time necessary to run the models was impractical.

6 Results

We now summarize the results. Table 2 shows the performance over the validation set for the models with smallest cumulative RMSE. Recall that the h-step is 30 days, therefore our forecasting horizon covers seven months.

Considering the cumulative RMSE of the different models, we chose the direct forecast benchmark, the abs90+abs30+30+90 model for the lasso, and the 30+90(t-2) model for the random forest. We also included the ensemble model that uses the simple mean of the predictions from the chosen lasso and Random Forest models. The ratio to benchmark of the RMSE over the validation set can be seen in Table 3. On average, the lasso models outperform the random forest models, and the ensemble model outperforms both.

The results for the holdout set are shown in Tables 4 and 5. As was the case for the validation set, the ensemble model far outperforms both lasso and Random Forest models and the lasso model outperforms the Random Forest model. Compared to the benchmark, all models have smaller forecasting error for all horizons except the first.

Both lasso and Random Forest models did generalize well, as they did not lose much performance in the holdout set when compared to the validation set. Another possible reason to the relative worse performance in the holdout set is that in the validation set time frame, the majority peaks of food-at-home inflation were caused by weather disturbances (BANK, 2017b; BANK, 2017c; BANK, 2017a; BCB, 2016), while the peak of food-at-home inflation during the holdout set was caused by the truck driver strike in May 2018 (BANK, 2018).

The ensemble model is far superior to each of the models, in both validation and holdout sets, suggesting that each model is extracting different, possibly uncorrelated, information from the data.

In Figures 3 and 5 we show the forecasts for each method and each horizon on the validation and holdout sets, respectively. In Figures 4 and 6 we show the forecasted

h (months ahead)	1	2	3	4	5	6	7	Average
AR(BIC) iterated	774.3	836.8	958.7	952.4	959.9	955.7	956.2	913.4
AR(BIC) direct	761.9	844.8	964.4	965.3	959.8	962.7	948.6	915.4
Lasso	504.3	593.8	695.2	628.3	674.4	648.7	703.0	635.4
RF	521.4	670.0	660.2	690.7	692.6	753.6	650.4	662.7
Ensemble	473.5	541.6	594.4	600.9	602.7	630.0	623.0	580.9

Table 2 – 1000xRMSE for validation set

h (months ahead)	1	2	3	4	5	6	7	Average
Lasso	0.56	0.69	0.50	0.67	0.56	0.56	0.56	0.59
RF	0.55	0.73	0.80	0.78	0.80	0.76	0.76	0.74
Ensemble	0.45	0.63	0.57	0.64	0.61	0.60	0.63	0.59

Table 3 – Ratio to benchmark for validation set

h (months ahead)	1	2	3	4	5	6	7	Average
AR(BIC) iterated	1067.5	1147.6	1110.8	1131.4	1126.7	1134.1	1145.2	1123.3
AR(BIC) direct	1099.5	1148.9	1104.5	1121.7	1140.1	1151.0	1166.3	1133.1
Lasso	596.6	793.1	556.6	754.3	626.8	638.9	646.3	659.0
RF	589.6	841.4	884.4	884.8	900.7	867.5	865.3	833.4
Ensemble	529.6	756.0	627.6	743.2	692.7	681.6	676.2	672.4

Table 4 – 1000xRMSE for holdout set

h (months ahead)	1	2	3	4	5	6	7	Average
Lasso	0.65	0.71	0.73	0.66	0.70	0.68	0.74	0.70
RF	0.67	0.80	0.69	0.73	0.72	0.79	0.68	0.73
Ensemble	0.58	0.70	0.66	0.65	0.60	0.67	0.65	0.65

Table 5 – Ratio to benchmark for holdout set

trajectory for selected dates.

Figure 3 – Forecasts in validation set

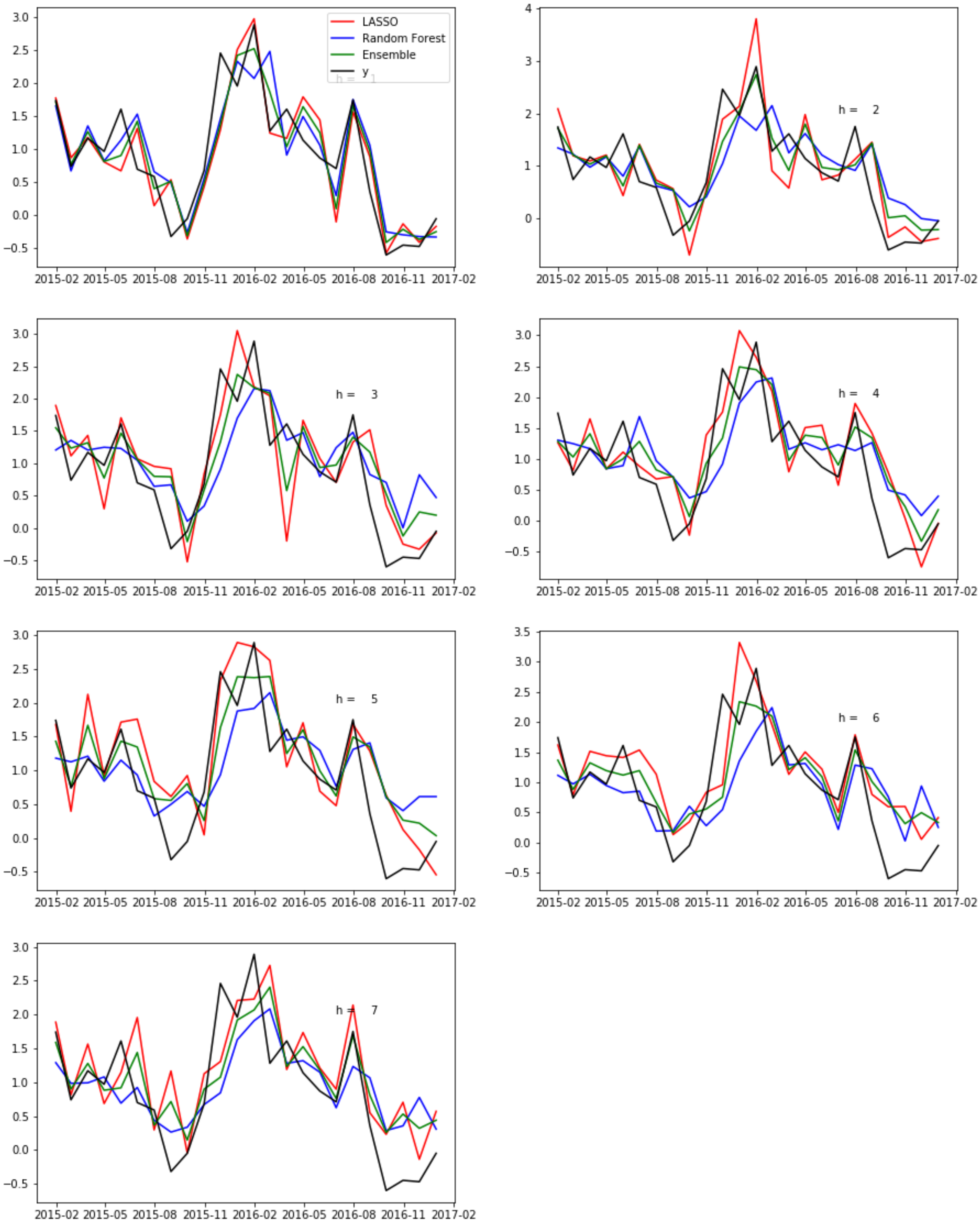
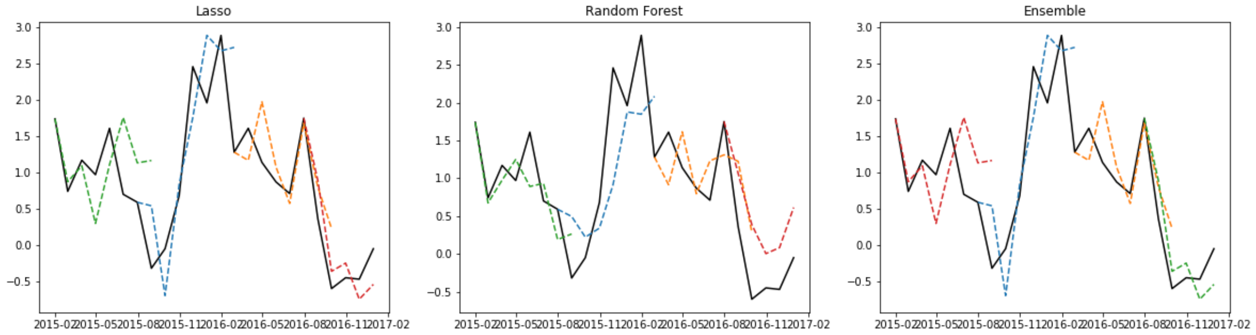


Figure 4 – Forecasts trajectory for selected dates in validation set



Lasso

Random Forest

Ensemble

Figure 5 – Forecasts in holdout set

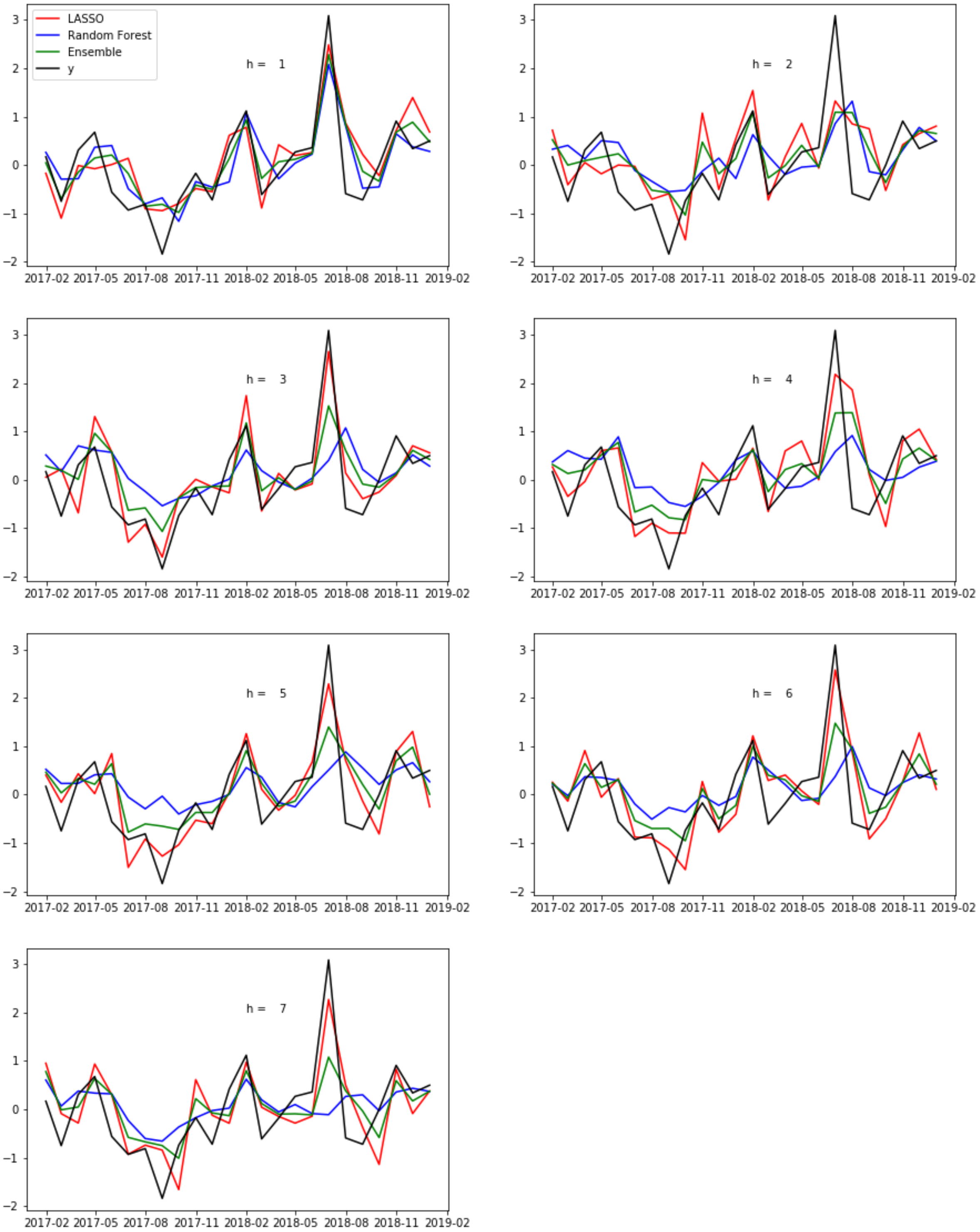


Figure 6 – Forecasts trajectory for selected dates in holdout set

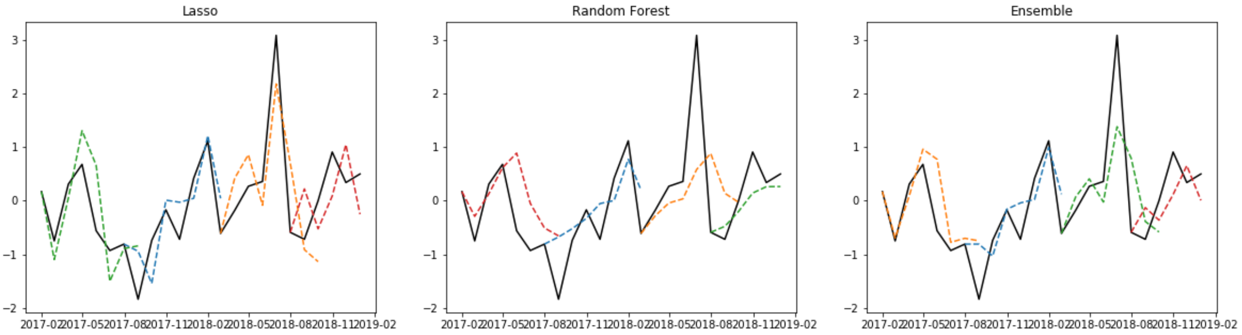
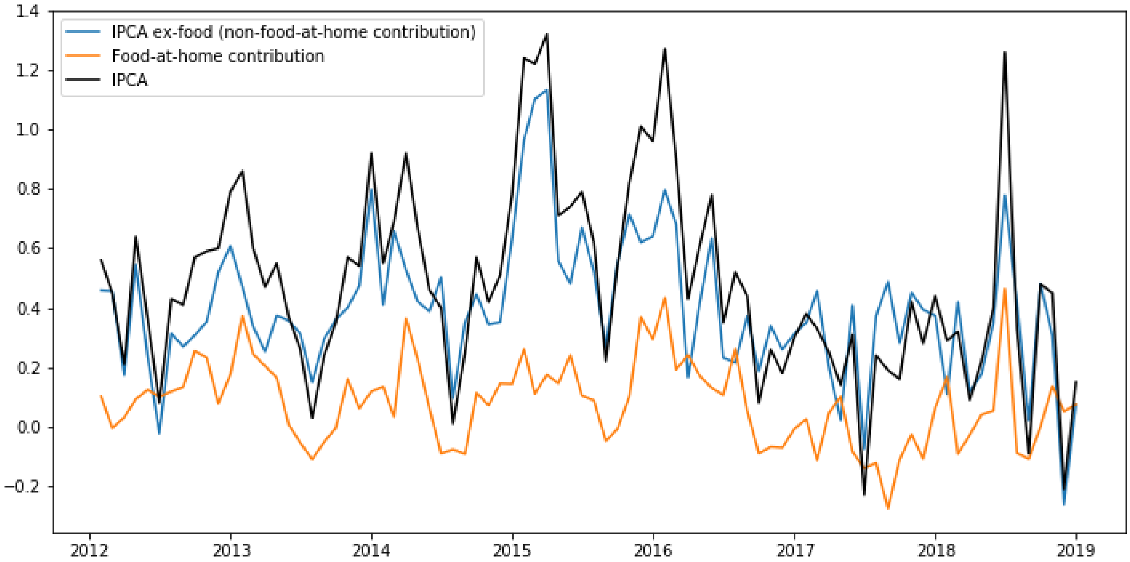


Figure 7 – Food-at-home inflation as driver of volatility in IPCA



6.1 Forecasting IPCA

We now turn to forecasting the full Brazilian consumer price (IPCA) using our results for its subgroup food-at-home. Recall that the food-at-home subgroup is the most unpredictable component of the inflation, so it is reasonable to expect that gains in the forecast of this component, even though its weight in the full index is only ca. 16%, can have a considerable impact on the forecasts of the whole index.

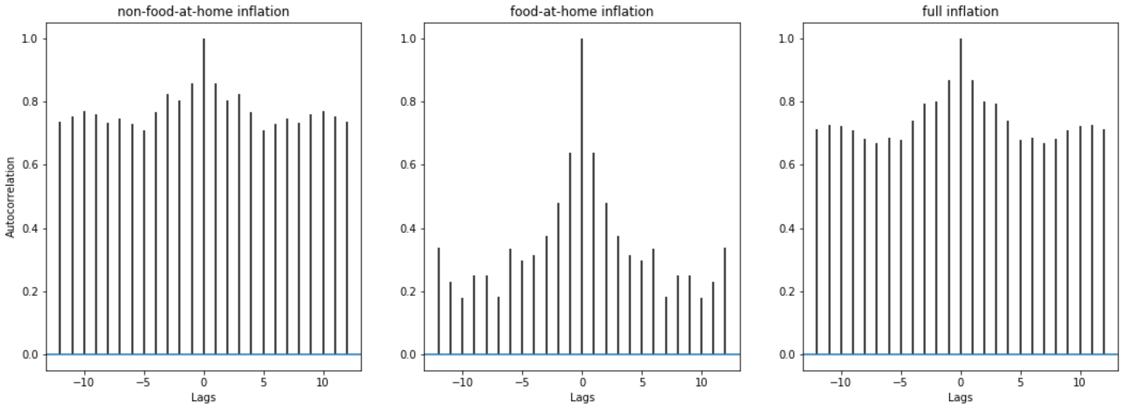
In Figure 7 we show inflation decomposed in the food-at-home and non-food-at-home components (adding the orange and blue lines results in the black line). Note that a good share of full inflation dynamics comes from the food-at-home component.

If we consider the series since 2016, the standard deviation of food-at-home inflation is 0.94, while it is 0.27 for non-food-inflation and 0.32 for the full index (the correlation between food-at-home and non-food-at-home inflation is 0.36). To further motivate the potential of food-at-home inflation forecasts to improve full inflation forecasts can the seen in Figure 8 that shows each of these series autocorrelation.

We use a naive model for full IPCA. We calculate “IPCA ex-food-at-home”, that is, we multiply each component, except food-at-home, by its weight in the full index and sum this up. We do not divide by the food-at-home weight, therefore, the weights do not sum to one. We call this series IPCA ex-food-at-home. We calculate this series beginning in 2012.

We forecast IPCA ex-food-at-home using, as before, an AR with BIC selected lag

Figure 8 – Autocorrelation of components of inflation



h (months ahead)	1	2	3	4	5	6	7	Average
AR(BIC) iterated	248.4	284.3	286.1	328.9	347.7	333.9	332.3	308.8
Lasso	162.4	207.6	223.2	245.9	252.3	262.5	269.6	231.9
RF	202.4	260.4	268.4	269.2	297.1	289.3	278.3	266.4
Ensemble	179.0	226.3	241.2	255.0	271.8	273.7	271.1	245.5

Table 6 – Forecast 1000xRMSE for Brazilian consumer price index (IPCA) (2015 to 2016)

h (months ahead)	1	2	3	4	5	6	7	Average
AR(BIC) iterated	248.4	284.3	286.1	328.9	347.7	333.9	332.3	308.8
Lasso	162.4	207.6	223.2	245.9	252.3	262.5	269.6	231.9
RF	202.4	260.4	268.4	269.2	297.1	289.3	278.3	266.4
Ensemble	179.0	226.3	241.2	255.0	271.8	273.7	271.1	245.5

Table 7 – Forecast 1000xRMSE for Brazilian consumer price index (IPCA) (2017 to 2018)

order, estimated pseudo-out-of-sample. Finally, we build our forecasts for IPCA adding the IPCA ex-food-at-home forecasts to our forecasts for food-at-home inflation multiplied by 0.16 (the typical weight of food-at-home).

All models outperformed the benchmark for all 7 horizons. However, the lasso outperformed the ensemble model when considering the average RMSE over the 7 horizons.

The results are summarized in Tables 6, 8, 9 and 7 and show that, as expected, the improvements in the food-at-home inflation component can have striking effects on the forecasts of full inflation, especially for longer horizons.

h (months ahead)	1	2	3	4	5	6	7	Average
Lasso	0.65	0.73	0.78	0.75	0.73	0.79	0.81	0.75
RF	0.81	0.92	0.94	0.82	0.85	0.87	0.84	0.86
Ensemble	0.72	0.8	0.84	0.78	0.78	0.82	0.82	0.79

Table 8 – Ratio to benchmark for Brazilian consumer price index (IPCA) (2015 to 2016)

h (months ahead)	1	2	3	4	5	6	7	Average
Lasso	0.88	0.89	0.83	0.99	0.87	0.86	0.79	0.87
RF	0.78	0.84	0.95	0.95	0.94	0.95	0.99	0.91
Ensemble	0.83	0.86	0.88	0.95	0.9	0.9	0.88	0.88

Table 9 – Ratio to benchmark for Brazilian consumer price index (IPCA) (2017 to 2018)

Bibliography

ATKESON, A. E. O. et al. Are phillips curves useful for forecasting inflation?[*]. *Federal Reserve bank of Minneapolis quarterly review*, Federal Reserve Bank of Minneapolis, v. 25, n. 1, p. 2–2, 2001. Cited on page 19.

BANK, B. B. C. *Effects of food price shocks on the IPCA*. [S.l.: s.n.], 2017. Cited 2 times on pages 15 and 33.

BANK, B. B. C. *Inflation Report, chapter 1.3*. [S.l.: s.n.], 2017. Cited 2 times on pages 15 and 33.

BANK, B. B. C. *Inflation Report, chapter 1.3*. [S.l.: s.n.], 2017. Cited on page 33.

BANK, B. B. C. *Effects on consumer inflation of the temporary halt in the transportation sector*. [S.l.: s.n.], 2018. Cited on page 33.

BCB, B. C. do B. *Evolução recente da inflação de alimentos*. [S.l.], 2016. Cited 2 times on pages 15 and 33.

BHANSALI, R. J. Asymptotically efficient autoregressive model selection for multistep prediction. *Annals of the Institute of Statistical Mathematics*, Springer, v. 48, n. 3, p. 577–602, 1996. Cited on page 27.

BLOESCH, J.; GOURIO, F. The effect of winter weather on us economic activity. *Economic Perspectives*, v. 39, n. 1, 2015. Cited on page 20.

BOLDIN, M.; WRIGHT, J. H. Weather-adjusting economic data. *Brookings Papers on Economic Activity*, Brookings Institution Press, v. 2015, n. 2, p. 227–278, 2015. Cited on page 23.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Cited 2 times on pages 17 and 30.

BREITUNG, J.; ROLING, C. Forecasting inflation rates using daily data: A nonparametric midas approach. *Journal of Forecasting*, Wiley Online Library, v. 34, n. 7, p. 588–603, 2015. Cited on page 19.

BROWN, M. E.; KSHIRSAGAR, V. Weather and international price shocks on food prices in the developing world. *Global environmental change*, Elsevier, v. 35, p. 31–40, 2015. Cited on page 20.

CENCI, S. A.; SOARES, A. G.; JúNIOR, M. F. Manual de perdas pós-colheita em frutos e hortaliças. 1997. Cited on page 16.

CHEN, Y.-c.; TURNOVSKY, S. J.; ZIVOT, E. Forecasting inflation using commodity price aggregates. 2011. Cited on page 19.

DELL, M.; JONES, B. F.; OLKEN, B. A. What do we learn from the weather? the new climate-economy literature. *Journal of Economic Literature*, v. 52, n. 3, p. 740–98, 2014. Cited on page 20.

ENGEL, E. Die productions-und consumtionsverhältnisse des königreichs sachsen. *Zeitschrift des Statistischen Bureaus des Königlich Sächsischen Ministeriums des Innern*, v. 8, p. 1–54, 1857. Cited on page 15.

FAUST, J.; WRIGHT, J. H. Forecasting inflation. In: *Handbook of economic forecasting*. [S.l.]: Elsevier, 2013. v. 2, p. 2–56. Cited 2 times on pages 15 and 19.

FUND, I. M. World economic outlook databse. October 2018. Cited on page 16.

GARCIA, M. G.; MEDEIROS, M. C.; VASCONCELOS, G. F. Real-time inflation forecasting with high-dimensional models: The case of brazil. *International Journal of Forecasting*, Elsevier, v. 33, n. 3, p. 679–693, 2017. Cited 4 times on pages 17, 19, 29, and 30.

GÓMEZ, M. I.; GONZÁLEZ, E. R.; MELO, L. F. Forecasting food inflation in developing countries with inflation targeting regimes. *American Journal of agricultural economics*, Oxford University Press, v. 94, n. 1, p. 153–173, 2012. Cited on page 16.

GOODFELLOW, I. et al. *Deep learning*. [S.l.]: MIT press Cambridge, 2016. Cited on page 31.

IIZUMI, T.; RAMANKUTTY, N. How do weather and climate influence cropping area and intensity? *Global Food Security*, Elsevier, v. 4, p. 46–50, 2015. Cited on page 15.

LI, X. et al. A midas modelling framework for chinese inflation index forecast incorporating google search data. *Electronic Commerce Research and Applications*, Elsevier, v. 14, n. 2, p. 112–125, 2015. Cited on page 19.

LOHRMANN, C.; LUUKKA, P. Classification of intraday s&p500 returns with a random forest. *International Journal of Forecasting*, Elsevier, v. 35, n. 1, p. 390–407, 2019. Cited on page 20.

LOUPPE, G. Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*, 2014. Cited on page 30.

MARCELLINO, M.; STOCK, J. H.; WATSON, M. W. A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series. *Journal of econometrics*, Elsevier, v. 135, n. 1-2, p. 499–526, 2006. Cited on page 27.

MEDEIROS, M. C.; VASCONCELOS, G.; FREITAS, E. Forecasting brazilian inflation with high-dimensional models. *Brazilian Review of Econometrics*, v. 36, n. 2, p. 223–254, 2016. Cited on page 29.

MEDEIROS, M. C. et al. Forecasting inflation in a data-rich environment: The benefits of machine learning methods. 2018. Cited 4 times on pages 17, 19, 20, and 30.

MODUGNO, M. Nowcasting inflation using high frequency data. 2011. Cited on page 19.

ORGANIZATION, W. W. M. *Guidelines on the Calculation of Climete Normals*. [S.l.]: Secretariat of the World Meteorological Organization, 2017. Cited on page 25.

RAY, D. K. et al. Climate variation explains a third of global crop yield variability. *Nature communications*, Nature Publishing Group, v. 6, p. 5989, 2015. Cited on page 15.

SANDQVIST, A. P.; SILIVERSTOVS, B. Is it good to be bad or bad to be good?: Assessing the aggregate impact of abnormal weather on consumer spending. *KOF Studies*, ETH Zurich, v. 443, 2018. Cited on page 20.

SMEEKES, S.; WIJLER, E. Macroeconomic forecasting using penalized regression methods. *International journal of forecasting*, Elsevier, v. 34, n. 3, p. 408–430, 2018. Cited 2 times on pages 17 and 20.

SMITH, M. S.; MANEESOONTHORN, W. Inversion copulas from nonlinear state space models with an application to inflation forecasting. *International Journal of Forecasting*, Elsevier, v. 34, n. 3, p. 389–407, 2018. Cited on page 19.

SPAGNOL, W. A. et al. Redução de perdas nas cadeias de frutas e hortaliças pela análise da vida útil dinâmica. *Brazilian Journal of Food Technology*, 2017. Cited on page 16.

STOCK, J. H.; WATSON, M. W. Forecasting with many predictors. *Handbook of economic forecasting*, Elsevier, v. 1, p. 515–554, 2006. Cited 2 times on pages 20 and 31.

STOCK, J. H.; WATSON, M. W. Why has us inflation become harder to forecast? *Journal of Money, Credit and banking*, Wiley Online Library, v. 39, p. 3–33, 2007. Cited on page 19.

STOCK, J. H.; WATSON, M. W. *Phillips curve inflation forecasts.* [S.l.], 2008. Cited 2 times on pages 15 and 19.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996. Cited on page 29.

ÜLKE, V.; SAHIN, A.; SUBASI, A. A comparison of time series and machine learning models for inflation forecasting: empirical evidence from the usa. *Neural Computing and Applications*, Springer, v. 30, n. 5, p. 1519–1527, 2018. Cited on page 19.

WILSON, D. J. The impact of weather on local employment: Using big data on small places. In: FEDERAL RESERVE BANK OF SAN FRANCISCO. [S.l.], 2017. Cited on page 20.