Igor da Silva Carvalho

Nowcasting CPI using online retail prices: forecasting combination of dynamic factor models

Rio de Janeiro

2020

Igor da Silva Carvalho

Nowcasting CPI using online retail prices: forecasting combination of dynamic factor models

Dissertation presented to the Escola de Matemática Aplicada (EMAp-FGV) in partial fulfillment of the requirements for the degree of Mestre em Modelagem Matemática.

Fundação Getúlio Vargas – FGV

Escola de Matemática Aplicada – EMAp

Supervisor: Eduardo Fonseca Mendes

Rio de Janeiro

2020

Dados Internacionais de Catalogação na Publicação (CIP)
Ficha catalográfica elaborada pelo Sistema de Bibliotecas/FGV

Carvalho, Igor da Silva
Nowcasting CPI using online retail prices : forecasting combination of dynamic factor models / Igor da Silva Carvalho. – 2020.
47 f.

Dissertação (mestrado) - Fundação Getulio Vargas, Escola de Matemática Aplicada.
Orientador: Eduardo Fonseca Mendes.
Inclui bibliografia.

1. Inflação – Brasil – Previsão – Modelos matemáticos. 2. Preços – Previsão – Modelos matemáticos. 3. Comércio eletrônico. I. Mendes, Eduardo Fonseca. II. Fundação Getulio Vargas. Escola de Matemática Aplicada. III. Título.

CDD – 338.520184

Elaborada por Márcia Nunes Bacha – CRB-7/4403

IGOR DA SILVA CARVALHO

“NOWCASTING CPI USING ONLINE RETAIL PRICES: FORECASTING COMBINATION OF DYNAMIC FACTOR MODELS’’.

Dissertação apresentado(a) ao Curso de Mestrado em Modelagem Matemática do(a) Escola de Matemática Aplicada para obtenção do grau de Mestre(a) em Modelagem Matemática.

Data da defesa: 06/08/20

ASSINATURA DOS MEMBROS DA BANCA EXAMINADORA

Presidente da Comissão Examinadora: Profº **Eduardo Fonseca Mendes**

Eduardo Fonseca Mendes
Orientador

Pedro Guilherme Costa Ferreira
Membro

Marcelo Cunha Medeiros
Membro

Nos termos da Lei nº 13.979 de 06/02/20 - DOU nº 27 de 07/02/20 e Portaria MEC nº 544 de 16/06/20 - DOU nº 114 de 17/06/20 que dispõem sobre a suspensão temporária das atividades acadêmicas presenciais e a utilização de recursos tecnológicos face ao COVID-19, as apresentações das defesas de Tese e Dissertação, de forma excepcional, serão realizadas de forma remota e síncrona, incluindo-se nessa modalidade membros da banca e discente.

César Leopoldo Camacho Manco
Diretor

Antonio de Araujo Freitas Junior
Pró-Reitor de Ensino, Pesquisa e Pós-Graduação FGV

Antonio Freitas, PhD
Pró-Reitor de Ensino, Pesquisa e Pós-Graduação
Fundação Getúlio Vargas

Instrução Normativa nº 01/19, de 09/07/19 - Pró-Reitoria FGV
Em caso de participação de Membro(s) da Banca Examinadora de forma não-presencial*, o Presidente da Comissão Examinadora assinará o documento como representante legal, delegado por esta I.N.
*Skype, Videoconferência, Apps de vídeo etc

Acknowledgements

First, I thank my parents Marlene Maria da Silva Carvalho and Jocemar da Silva Carvalho for all support and incentive during my academic journey.

I thank my advisor Eduardo Fonseca Mendes for the partnership and guidance throughout the elaboration of this work. I thank Marcelo Cunha Medeiros and all team of Digital Lab for the advices, opportunities and learning for the last four years. I also thank Pedro Guilherme Ferreira for the pertinent comments on this work.

Last but not least, I thank my fiancée Bruna Scoralick Sousa Lisboa for the inspiration, patience and care for the last seven years, and my nephew and godson Miguel Vicente Carvalho Martins that arrived recently in our family and makes our days funnier.

"All models are wrong,
but some are useful."
(George E. P. Box)

Abstract

In this work we use daily online retail prices and financial data to construct nowcasts of consumer price index (CPI) for the current month, which we refer simply as inflation. We first adapt the dynamic factor model in (GIANNONE; REICHLIN; SMALL, 2008) to our setting and integrate it into a *nowcasting combination* algorithm. We also propose an extension to the nowcasting equation to take into account previous inflation level and nowcasting error. We compare our results with a base model and Focus survey of professional forecasters. The base model is constructed from a simple factor regression using individual categories monthly price change calculated in (CAVALLO, 2013), whereas the Focus survey aggregate daily inflation forecasts provided by a large number of institutions to the Brazilian Central Bank. We consider the Focus survey output as the *golden standard*. Our results indicate that (1) online retail prices are not enough to capture the whole dynamics of CPI; and (2) the factor regression has to be augmented with recent inflation information to be competitive.

Key-words: Nowcasting. Inflation. E-commerce prices.

Resumo

Neste trabalho utilizamos preços diários do varejo online e variáveis financeiras para computar *nowcasts* da inflação brasileira. Nós adaptamos o modelo de fatores dinâmicos proposto em (GIANNONE; REICHLIN; SMALL, 2008) ao nosso contexto e o integramos a um algoritmo de combinação de *nowcasts*. Propomos também um modelo estendido em que, além de fatores, utilizamos um componente auto-regressivo e média móvel para modelar e prever a inflação corrente. Para fins de avaliação, comparamos nossos resultados com um modelo base e com as previsões disponibilizadas pelo Focus survey. As previsões oriundas do modelo base são construídas a partir de fatores extraídos das séries de inflação a nível de categoria de produtos, estas últimas calculadas tal como em (CAVALLO, 2013). O Focus survey, por sua vez, agrega previsões diárias fornecidas por diversas instituições como bancos, consultorias e corretoras, e, por isso, consideramos estes resultados um *padrão ouro* no processo de avaliação. Nossos exercícios empíricos indicam que (1) preços do varejo online não são suficientes para capturar toda a dinâmica da inflação brasileira, e (2) componentes auto-regressivos devem ser utilizados para melhorar a performance dos *nowcasts* fora da amostra.

Palavras-chave: Nowcasting. Inflação. Preços do comércio eletrônico.

List of Figures

List of Tables

Contents

1 Introduction

1.1 Background and objectives

Central Banks around the world often declare price stability as the main goal of their monetary policy. The Consumer Price Index (CPI)[1] is the quantitative measure of price stability. Market practitioners, on the other hand, devote effort to develop up to date inflation forecast as to adjust investment strategies and expectations. For both government and market, tracking real time inflation is essential for timely decision making. Nevertheless, CPI, as well as other key statistics, are often released with delay.

Nowcasting models are used to track official statistics, such as inflation and GDP, along the current month while is has not been released. (MODUGNO, 2013), (KNOTEK; ZAMAN, 2017) and (FUNKE; MEHROTRA; YU, 2015) explore variables released intra-monthly, such as CPI sub-indices and surveys of fuel price, to compute nowcasts of inflation for Euro Area, United States and China. However, online retail market expands year after year making available prices for large number of products. It provides a cheap and accessible source of information to monitor official price index.

The main goal of this work is to exploit prices from online retailers to nowcast brazilian inflation, measured by relative changes of Extended National Consumer Price Index, also denoted IPCA in Brazil. We use the dataset collected and provided by *The Billion Price Project*, which contain daily price series for thousands of products scraped from the website of a leader retail company in Brazil.

We use a combination of dynamical factor models and forecasting combination to compute our nowcasts. We first adapt (GIANNONE; REICHLIN; SMALL, 2008) methodology to our framework and expand their linear nowcasting model to incorporate recent inflation information using an ARMA(1, 1) term. Then, we construct datasets with a random subsets of products to construct individual nowcasts. The main advantages of our approach are that we can handle very large number of daily series seamlessly and scale our method as number of scraped products increase. The data can be collected from websites in an automated way. Results indicate that there are some information yet to be captured by data and that augmenting the factor model with local dynamics improve the nowcasting performance.

This work is organized as follows. In section 1.2 we review the literature of nowcasting inflation and the use of data from online retailers to conduct macroeconomic research.

[1] Although we recognize that *inflation* and *CPI* are distinct quantities, in this work we will refer to CPI simply as inflation.

In section 2 we describe the BPP data used in this work. Section 3 data processing, models and model forecasting algorithm. In section 4 we run an out-of-sample exercise to evaluate proposed models, followed by a discussion in section 5. Finally, section 6 concludes.

1.2 Literature review

Nowcasting models are widely used to track GDP and other official statistics along the reference period. It allows analysts monitor and understand macroeconomic conditions before the announcement of official statistics. See (BAŃBURA et al., 2013) for a survey of nowcasting.

(GIANNONE; REICHLIN; SMALL, 2008) proposes using a dynamic factor model to nowcast U.S. GDP using a large set of monthly macroeconomic variables. The authors apply a two-step estimator studied by (DOZ; GIANNONE; REICHLIN, 2011) in order to summarize explanatory variables in few common factors. In the first step a preliminary version of factors are obtained using PCA, in the second they are re-estimated via Kalman Smoother. Expectation of GDP growth is defined as an affine function of the estimated factors.

(BANBURA; GIANNONE; REICHLIN, 2011) generalizes the Factor Model proposed by (GIANNONE; REICHLIN; SMALL, 2008) in order to nowcast GDP using weekly and daily variables.The model is estimated by maximum likelihood applying EM algorithm as proposed by (DOZ; GIANNONE; REICHLIN, 2006). (GIANNONE; REICHLIN; SIMONELLI, 2009) follow this same model and estimation technique to extract factors from disaggregated survey data and nowcast Euro Area GPD.

As an alternative to factor models, (KUZIN; MARCELLINO; SCHUMACHER, 2009) compare the mixed-data sampling (MIDAS) and mixed-frequency VAR (MF-VAR) to nowcast Euro Area GDP using a small sample of 20 monthly indicators. Moreover (CARRIERO; CLARK; MARCELLINO, 2015) nowcast U.S. GDP using a Bayesian linear model with stochastic errors.

The academic literature on nowcasting GDP is extensive, yet scarce on nowcasting inflation. (KNOTEK; ZAMAN, 2017) develop a parsimonious linear model to nowcast U.S. headline and core inflation using eight explanatory variables including subindices of Comsumer Price Index (CPI) and Personal Consumer Expenditure (PCE), and fuel price. Authors mention that BPP data is a potential source of disaggregated information to be used as explanatory variables. However, they argue that U.S. price indexes place a large weight on services while internet purchases constitute a small share of spending on goods, which suggests that e-commerce prices are not able to improve the predictive power of proposed model. This work endorses that simple models with few variables can perform as good as complex models estimated with a large set of time series.

On the other hand, (MODUGNO, 2013) apply the factor model in (GIANNONE; REICHLIN; SMALL, 2008) to nowcast Euro Area inflation using subindices of Harmonised Index of Comsumer Price (HICP), surveys of fuel prices and Raw Material Price Indexes (RMPI). Parameters of state-space form and factors are estimated by maximum likelihood applying EM algorithm as in (BANBURA; GIANNONE; REICHLIN, 2011). (FUNKE; MEHROTRA; YU, 2015) use a factor model with data in different frequencies to track Chinese inflation. Authors estimate the *cumulator* variable suggested by (HARVEY, 1990) in order to construct aggregated variables at the observable lower frequency. The model is written in state-space form and estimated using the Kalman Filter. (AMSTAD; FISCHER, 2009) use a common factor procedure to incorporate up to date information from macroeconomic releases and revisions to construct weekly updates of the monthly Swiss core inflation.

(CAVALLO, 2013) uses BPP data to compute CPI for five Latin American countries. He shows that price indexes computed with online data match official indexes of Brazil, Chile, Colombia, and Venezuela. In Argentina the results are remarkably different, which reinforces the hypothesis that Argentina's government manipulated official statistics. Also exploring BPP data, (APARICIO; BERTOLOTTO, 2020) propose a linear model to forecast CPI using three types of explanatory variables: auto-regressive component, aggregated e-commerce prices and fuel prices. The authors show that the proposed baseline specification outperforms benchmark forecasts for Euro Area, Australia, Canada and United States.

2 Data

Nowcasts of inflation (consumer price index - CPI) are constructed from variables correlated to it that are released intra-monthly or, more generally, at a higher-than-one-month frequency. *The Billion Prices Project* (BPP) is an academic initiative at MIT Sloan and Harvard Business School that collects prices from a large number of e-retailers around the world to conduct research on macro and international economics. Given the frequency the prices are collected it is an ideal candidate for constructing nowcasts of inflation.

The BPP collected brazilian daily prices of approximately 22.000 products, classified into 72 categories, from October 2007 to August 2010. All data come from the website of a specific brazilian retail company that has roughly 15% of market share. Prices were collected daily using web crawlers, and all products and categories are identified only by their codes, hence impossible to know exactly which category corresponds to what. Nevertheless, we assume the products were correctly classified in meaningful categories.

Each product has three distinct series associated to it: `price0` is the original raw price, `nsprice` is the sale price, and `fullprice` is the raw price with missing values treated as described in (CAVALLO, 2018). In this work, we use the raw prices to avoid treatments not related to the nowcast methodology, such as bringing future information.

A large number of products contain some amount of missing data, which usually requires data imputation before processing. To avoid bias incurred from excessive data imputation, products with more than 20% missing values are dropped from the analysis. Hence, only 4,857 series are used in the modeling process. Nevertheless, there are two periods in which all prices are missing. The first one is from February 26, 2009 to May 4, 2009 and the second from September 19, 2009 to November 15, 2009. The missing information in these periods are likely related to a problem in the scraping program, possibly due changes in the website structure of links.

Taking the mean price as reference, the cheapest and the most expansive products cost R$ 0.49 and R$ 240.00, respectively and more than 90% of products cost less than R$ 50.00. Moreover, the mean prices of products within categories vary, which suggests that categories aggregate different value-added products, as we expected.

On the top of e-commerce prices, 25 brazilian financial series are added to the dataset into a new category. These extra variables are prices of commodities, exchange rates, and IBOVESPA stock exchange index, all of which are observed daily. These financial series are downloaded from the official website of B3 (brazilian stock exchange) and CEPEA (bureau of advanced research on applied economics at University of São Paulo).

3 Methodology

Our goal is to nowcast Brazilian CPI using intra-month price data from e-commerce and returns from financial markets. While conventional forecasting models are devoted to compute expectations for months ahead, our approach is able to explore information released daily in order to predict inflation for the current month.

(GIANNONE; REICHLIN; SMALL, 2008) propose a two-steps methodology for extracting information from a large collection of monthly series and use it to nowcast GDP. First, the large collection of series are summarized by dynamic factors, that are subsequently used to construct the nowcasts in the second step.

We use daily data to nowcast monthly inflation. The factors are constructed using the series of monthly relative average price change computed using daily prices and financial returns data. The main role of factors is reduce the set of explanatory variables avoiding *curse of dimensionality*. In a second stage, we explore the daily frequency of factors as features to compute nowcasts.

3.1 Data pre-processing

High-frequency prices series typically contains outliers and missing data that will be handled before constructing the series of relative prices. These steps are performed only on the data used to estimate the model, hence we never use *future* information in this step.

We perform an exploratory analysis to remove outliers. The series are centered around m and outliers are defined, within each product, as observations extrapolating the range $[m - 4IQR, m + 4IQR]$, where m is the median price of a product and IQR is its interquartile range. Outliers are dropped from the series and observations taken as missing.

Missing observations at time t are filled using a centered moving average process of 30 days, i.e., from $t - 30$ to $t + 30$, and, then, decentralize the series adding m back. Precisely, we first set all centered missing observations $p_t^c = 0$. Then, replace the missing observation p_t in each product by $m + (p_{t-30}^c + \cdots + p_{t-1}^c + p_{t+1}^c + \cdots + p_{t+30}^c)/60$, where p_j^c are centered at m. Note that future missing values are treated as 0. Missing observations in financial data correspond to no operations on stock exchange, such as saturdays, sundays and holidays and are replaced by their last observed value.

The data imputation process is applied to missing values excluding those allocated in end of the series, which are interpreted as not released information.

The series of prices are highly nonstationary and are not used directly in nowcasting inflation. Instead, we use their monthly relative average price change, denote $\mathbf{x}_t$, which is stationary. This transformation is an adaptation of (GIANNONE; REICHLIN; SMALL, 2008) to our setting. Moreover, it put prices and inflation in the same scale. The series of prices within each product j are transformed as follows:

$$x_{tj} = \frac{\text{average}(p_{t-30:t,j})}{\text{average}(p_{t-60:t-30,j})} - 1, \tag{3.1}$$

where $p_{s:t,j} = (p_{sj}, p_{s+1,j}, ..., p_{tj})$ for $s < t$, and average$(\cdot)$ is the sample average of the series between brackets.

The new series $\{\mathbf{x}_t = (x_{t1}, ..., x_{tN})\}$ are defined as the variation of the mean prices for two consecutive periods of 30 days. This transformation ensures that at the end of the month each series corresponds to mean prices variation between two consecutive periods, as in the official price index. The first 60 observations of each series are lost after this transformation.

3.2 Dynamic Factor Model

The factors are modeled using the state space representation. In state space models, we assume the observations are conditionally independent, given a set of latent variables which are the factors in our case. The latent variables are Markov in a sense that depend only on its immediate past.

More precisely, the measurement equation explains prices relatives at time t, $\mathbf{x}_t$ $(N \times 1)$, as a linear combination of r factors F_t and respective loadings, $\mathbf{\Lambda}$ $(N \times rp)$. Taking F_t as coefficients we have a model with dynamic coefficients evolving according to a transition equation. The transition equation for the factors is a p^{th} order Vector Auto-Regressive process, VAR(p), where $\mathbf{A}$ $(rp \times rp)$ and $\mathbf{B}$ $(rp \times q)$ are coefficient matrices and u_t $(r \times 1)$ is a vector of idiosyncratic shocks. A state space model with two VAR(2) factors is given in (3.2)-(3.3):

$$\underbrace{\begin{pmatrix} x_{t1} \\ x_{t2} \\ \vdots \\ x_{tN} \end{pmatrix}}_{\mathbf{x}_t} = \underbrace{\begin{pmatrix} | & | & | & | \\ \lambda_1 & \lambda_2 & \mathbf{0} & \mathbf{0} \\ | & | & | & | \end{pmatrix}}_{\mathbf{\Lambda}} \underbrace{\begin{pmatrix} f_t^1 \\ f_t^2 \\ f_{t-1}^1 \\ f_{t-1}^2 \end{pmatrix}}_{F_t} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \epsilon_{t2} \\ \vdots \\ \epsilon_{tN} \end{pmatrix}}_{\epsilon_t} \tag{3.2}$$

$$\underbrace{\begin{pmatrix} f_t^1 \\ f_t^2 \\ f_{t-1}^1 \\ f_{t-1}^2 \end{pmatrix}}_{F_t} = \underbrace{\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} f_{t-1}^1 \\ f_{t-1}^2 \\ f_{t-2}^1 \\ f_{t-2}^2 \end{pmatrix}}_{F_{t-1}} + \underbrace{\begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ 0 & 0 \\ 0 & 0 \end{pmatrix}}_{\mathbf{B}} \underbrace{\begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}}_{u_t}, \tag{3.3}$$

where $E[\epsilon_t \epsilon_t'] = \mathbf{\Psi}_t = \text{diag}(\tilde{\psi}_{1,t}, \tilde{\psi}_{2,t}, ..., \tilde{\psi}_{N,t})$, $E[\epsilon_t \epsilon_{t-s}'] = 0$, $s > 0$ and $E[\epsilon_t u_{t-s}'] = 0$ for all s.

Typically, $\mathbf{x}_t$'s are high dimensional. (GIANNONE; REICHLIN; SMALL, 2008) and (DOZ; GIANNONE; REICHLIN, 2011) suggest a two-steps estimator for the factors. In the first step, loadings $\mathbf{\Lambda}$, individual variances ψ_i, $i = 1, ..., N$, and auto-regressive parameters $\mathbf{A}$ and $\mathbf{B}$ are estimated using a fixed set of r factors calculated using principal components from all available information at time t. Effectively we truncate our data in the last row without missing. The diagonal matrix $\mathbf{\Psi}_t$ is calculated as

$$\tilde{\psi}_{i,t} = \begin{cases} \psi_i, & \text{if } p_{i,t} \text{ is available} \\ \infty, & \text{if } p_{i,t} \text{ is not available} \end{cases}. \tag{3.4}$$

As discussed in (GIANNONE; REICHLIN; SMALL, 2008), this procedure enables the computation of factors even if some observations are not available in the end of the series. In the second step, the means and variances of factors are recovered using the Kalman Smoother.[1] Denote $F_{t|T} = (f_{t|T}^1, f_{t|T}^2, ..., f_{t|T}^r)$, $t = 1, ..., T$, the sequence of factor means calculated using information up to time T.

We are required to specify three hyperparameters of the model: the number of factors r, the order of VAR model p and the number of idiosyncratic shocks q. Above, we set $r = p = q = 2$. In our empirical exercise we select these values using information criteria. First, r is determined as in (BAI; NG, 2002), then p is selected minimizing BIC of the model. The dimension q is estimated conditional on $r = r*$ and $p = p*$, following steps in (BAI; NG, 2007). Let $c_1 > c_2 \geq \cdots \geq c_{r^*} \geq 0$ be the first ordered eigenvalues of the residual covariance matrix of the VAR(p^*) model followed by 0, and at least one is positive. The eigenvalues are estimated from the model using PCA factors. Let $\hat{D}_k = \left(\frac{c_{(k+1)}^2}{\sum_{j=1}^{r^*} c_j^2} \right)^{1/2}$ denote

[1] For description of Kalman Smoother algorithm see (HARVEY, 1990).

the sequence of normalized eigenvalues and $K = \{k : \hat{D}_k < 1/\min[N^{(1/2)-0.1}, T^{(1/2)-0.1}]\}$. The number of primitive shocks is $q^* = \min\{k \in K\}$.

3.3 Computing Nowcasts

3.3.1 Model 1: GIANNONE; REICHLIN; SMALL

Given the series of factors $F_{t|T}$ up to time T, we move to estimating the nowcasting equation for inflation. Let y_m denote the inflation at time m observed monthly, and $y_{m|T} = E(y_m|\mathcal{F}_T)$, where $\mathcal{F}_T$ is all information up to time T. If $m = M \equiv M(T)$ is the current month at time T, $y_{m|T}$ is called the nowcast of inflation in month m given daily information available up to T. Naturally, $y_{m|T}$ is unobserved and model is estimated using monthly observations $(y_m, F'_{m|T})$ for each month $m = 0, ..., M-1$. The inflation $y_{m|T}$ is assumed to follow a linear process

$$y_{m|T} = \alpha_T + \beta'_T F_{m|T} + \epsilon_{m|T}, \tag{3.5}$$

where the parameters α_T and β_T depend on the information up to T. We use ordinary least squares to recover $\widehat{\alpha}_T$ and $\widehat{\beta}_T$.

Given estimated parameters and factors we nowcast the inflation $\widehat{y}_{M|T} = \widehat{\alpha}_T + \widehat{\beta}'_T F_{T|T}$. Note that the model has to be estimated each time T based on its past. We use an expanding window of observations so the number of months M in our dataset increase.

The final nowcast algorithm requires estimating all parameters and factors every day:

1. Estimate factor means $F_{t|T}$, $t = 1, ..., T$;
2. Estimate nowcasting equation parameters $\widehat{\alpha}_T$ and $\widehat{\beta}_T$ using low frequency data $(y_m, F_{m|T})$, $m = 0, \ldots, M-1$, where M is the current month at time T.
3. Calculate inflation nowcast $\widehat{y}_{M|T} = \alpha_T + \beta'_T F_{T|T}$.

An important aspect to the model is that current inflation may be released after last day of month. We synchronize $F_{m|T}$ and y_m selecting the factor estimated at the last day of the month m instead of the day of release.

3.3.2 Model 2: ARMA-augmented GIANNONE; REICHLIN; SMALL

(GIANNONE; REICHLIN; SMALL, 2008) nowcasting equation is limited in a sense that information about inflation and error in previous month is ignored. We propose augment the nowcasting equation by incorporating an ARMA(1, 1) term to capture recent dynamics.

Following the notation in the previous section, the inflation $y_{m|T}$ is assumed to follow the linear process

$$y_{m|T} = \alpha_T + \beta_T' F_{m|T} + \varphi_T y_{m-1} + \epsilon_{m|T} + \theta_T \epsilon_{m-1|T}, \tag{3.6}$$

where the parameters α_T, β_T, φ_T and θ_T depend on the information up to T. We use quasi maximum likelihood using the Kalman Filter[2] to estimate the parameters $\widehat{\alpha}_T$, $\widehat{\beta}_T$, $\widehat{\varphi}_T$ and $\widehat{\theta}_T$.

After estimating the regression parameters, the nowcast for inflation at month M given information up to T is $\widehat{y}_{M|T} = \widehat{\alpha}_T + \widehat{\beta}_T' F_{T|T} + \widehat{\varphi}_T y_{M-1} + \widehat{\theta}_T (y_{M-1} - \widehat{y}_{M-1|T})$. The model has to be estimated each time T based on its past. We use an expanding window of observations so the number of months M in our dataset increase.

3.4 Base Model

The *Base model* differs from Model 1 in section 3.3.1 in the way factors are computed. We follow (CAVALLO, 2013) to compute every day the 30-days, category-wise, supermarket inflation index. Then, indexes are converted into factors that are used in our nowcasting algorithm.

Let $n_{t,j}$, $j = 1, ..., 72$ denote the number of non-missing product-price in each category j at time t. In each category, let p_{ti} denote the price of product i at time t. The monthly inflation rate is calculated as follows. First calculate the geometric average of relative prices

$$R_{t,j} = \prod_i \left(\frac{p_{t,i}}{p_{t-1,i}} \right)^{1/n_{t,j}}$$

for each category j and compute the cumulative product for each t:

$$I_{t,j} = \prod_{s=0}^{t} R_{s,j},$$

for $j = 1, ..., 72$ corresponding to all categories. The respective category inflation index is

$$S_{t,j} = \frac{\text{average}(I_{t:t-30,j})}{\text{average}(I_{t-30:t-60,j})} - 1, \tag{3.7}$$

for each t. Naturally we will have to drop the first two-months of observations.

Given the series of category inflation, we construct factors extracting the r principal components that explain at least 80% of variance. The estimated factors, at time t given information up to T are denoted $F_{t|T} = (f^1_{t|T}, \cdots, f^r_{t|T})'$, where r is the number of selected factors.

[2] See (HARVEY, 1990) for details. This function is implemented in the statsmodel package in Python or `arima` function in Stata.

We calculate the nowcasts as in Model 1 in section 3.3.1. The final nowcasting algorithm also requires estimating all parameters and factors every day.

3.5 Nowcasting Combination

Estimating factors using all individual products is computationally demanding as the number of individual products ranges thousands. The Kalman Smoother requires inverting very large matrices for all instants of time. We overcome this challenge using a forecasting combination strategy, which have the added advantage of often reducing the prediction error. In short, we construct B stratified samples with N products, and follow the methodology to calculate nowcasts that are combined using simple average.

The idea of combining forecasts dates to (BATES; GRANGER, 1969) who show that combining forecast improve the predictive error. The are has developed since and we refer to (TIMMERMANN, 2006) for a literature review. There has been some development on how to weigh forecasts. In practice, uniform weights tend perform well. Estimating optimal weights, although appealing, usually has poor performance in applications even when compared to uniform weights (CLAESKENS et al., 2016).

Let $k = 1, ..., B$ denote the index of sampled product series constructed using stratified sample by category without replacement, hence ensuring adequate representation. The corresponding nowcasting at time T is $\widehat{y}^{(j)}_{M|T}$ and is calculated for each sampled series and each day T following the algorithms in sections 3.3.1, 3.3.2. The combined nowcast is

$$\widehat{y}^{B}_{M|T} = \frac{1}{B}\sum_{j=1}^{B} \widehat{y}^{(j)}_{M|T}. \tag{3.8}$$

In practical terms, we have to select a number of series to be sampled. For practical reasons we use 200 series to nowcast inflation, which is the same number of series in (GIANNONE; REICHLIN; SMALL, 2008). The algorithm is not very sensitive to the number of series. The number of subsamples B is 200.

4 Empirical Results

We use the dataset in Section 2 to compute daily nowcasts of inflation. Prices from October, 2007 until July. 2009 are used only for training purposes. We use an expanding window scheme to nowcast inflation from August 1, 2009 until July 31, 2010, using all three models presented in Section 3. The dynamic factor models (Model 1 and Model 2) are estimated using the forecasting combination approach presented in Section 3.5.

The performance of the model is measured both by the Root Mean Square Error (RMSE), in equation (4.1), and Mean Absolute Error (MAE), in equation (4.2), between daily nowcasts and observed inflation both at the end of the respective month and also in all days in our out-of-sample horizon. Recall that $M \equiv M(T)$ is the current month at time T, y_m is the observed inflation for month m and $\widehat{y}_{M|T}$ is the nowcast of inflation on month M given information up to T. Then, the year-average RMSE and MAE are

$$RMSE = \sqrt{\frac{1}{\tau}\sum_{T=1}^{\tau}(y_M - \widehat{y}_{M|T})^2}, \tag{4.1}$$

and

$$MAE = \frac{1}{\tau}\sum_{T=1}^{\tau}\left|y_M - \widehat{y}_{M|T}\right|, \tag{4.2}$$

where τ is either 12 for monthly comparison and 365 for the entire nowcasting period.

We use the forecasts provided by Focus survey as the benchmark to assess our results. Focus is a survey of professional forecasters organized by the Central Bank of Brazil (BCB) that compiles monthly inflation forecasts, updated at any day of the month. Participants of the survey can update their expectations of inflation for the current month and months ahead any day before the official inflation release. Central Bank of Brazil ranks these outputs and publishes the top 5 in recognition to their analytical effort and to encourage participants improve their predictions.

Our results are compared with the average of all Focus' nowcasts. We use both the daily releases and the one updated at the last business day before the announcement of inflation preview, which are the forecasts used by BCB to compute the rank. This data are available in the official website of BCB and is a form of forecasting combination, however, using possibly distinct information sets each candidate.

The hyperparameters in the model and forecasting combination algorithm are set as follows. The nowcasting algorithm is robust to the number of product series $N = 200$. We set the number of samples $B = 200$. For Model 1, hyperparameters are selected in the range $p \in \{1, .., 5\}$, $r \in \{1, ..., 10\}$ using information criteria; for Model 2, $p \in \{1, .., 5\}$ but

we fix $r = 1$. The number of primitive shocks is selected conditional on r and p following (BAI; NG, 2007). Details are found in Section 3.2.

The RMSE and MAE for all models and Focus research are shown in Table 1. The Focus nowcasts are naturally closer to the target CPI than competing models. Model 1 ranks worst both in terms of RMSE and MAE, whereas Model 2 ranks closer to Focus. In the monthly performance measurement, base model and Model 1 are almost indistinguishable in both metrics. In the yearly performance measurement, Base model distinguishes slightly from Model 1. Model 2 has a better performance in all measures when comparing the three alternatives. Nevertheless, it is not able to capture CPI dynamics as well as Focus.

Figures 1, 2 and 3 show pointwise comparison between the realized inflation, Focus' nowcasts, and out-of-sample predictions from Model 1, Model 2 and Base Model, respectively. Dynamics of Model 1 and Model 2 are very similar, but Model 2 has a clear change of level due auto-regressive term. Base model displays a very different dynamics and oscilates more freely. It also captures the same decline in inflation as observed by Focus.

	Monthly		Entire period	
	RMSE	MAE	RMSE	MAE
Focus	0.0701	0.0567	0.1251	0.0951
Model 2	0.1817	0.1362	0.1815	0.1370
Base model	0.2642	0.2030	0.2326	0.1813
Model 1	0.2550	0.2085	0.25109	0.2024

Table 1 – Comparing models in terms of RMSE and MAE.

Figure 4 compares the number of observations and explanatory variables used to fit models in-sample. Model 1 always selected the maximum number of factors whereas Base model started with 13 factors and increased with the number of observations. Model 2 has a fixed number of factors.

In Appendix A and B we show the distribution of individual nowcasts before combination and the Adjusted R^2. In Model 1, individual nowcast before combination vary symmetrically around their mean, and the range is mostly constant within months, and not always covering the observed inflation. The R^2 tells a similar story presenting a wide range with means between 5% and 30%, indicating factors explain only a small amount of variation. Model 2 individual nowcasts do not present a symmetric distribution around their mean and seems to have more extreme nowcasts, presenting a fat tail. Within each month, the series are more dynamic but due strong mean concentration, it often fails to cover observed inflation value. Similarly, adjusted R^2 have a wider variation, but factor + ARMA components improve upon only factors, presenting mean values ranging from 20% to 50% of variance explained.

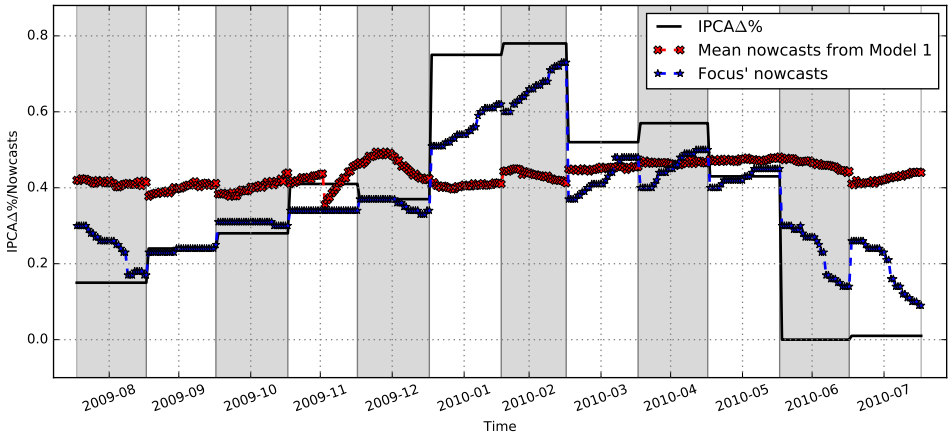

Figure 1 – Realized inflation, Focus' nowcasts and out-of-sample nowcasts from Model 1.

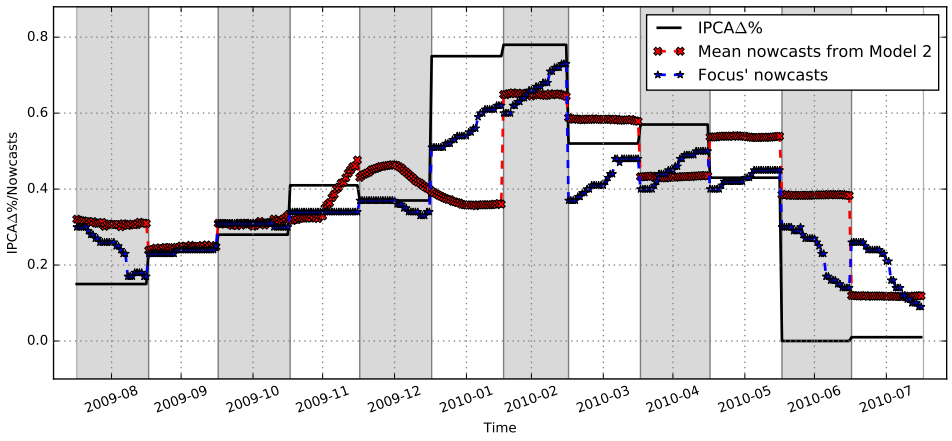

Figure 2 – Realized inflation, Focus' nowcasts and out-of-sample nowcasts from Model 2.

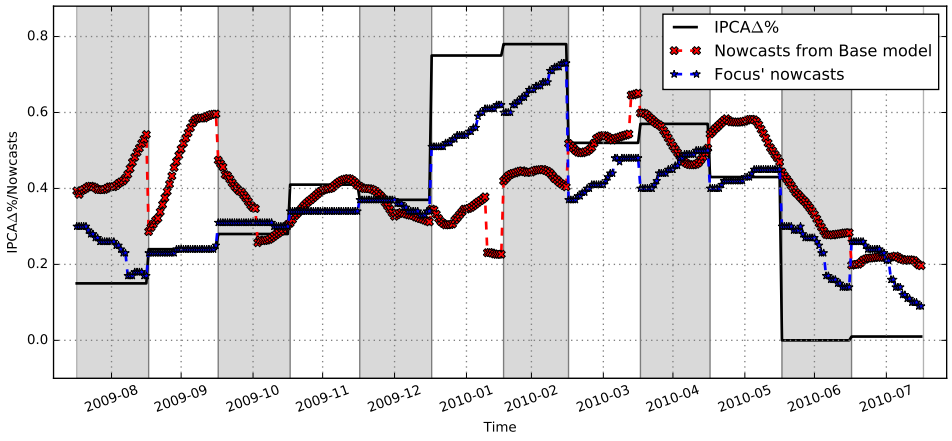

Figure 3 – Realized inflation, Focus' nowcasts and out-of-sample nowcasts from Base model.

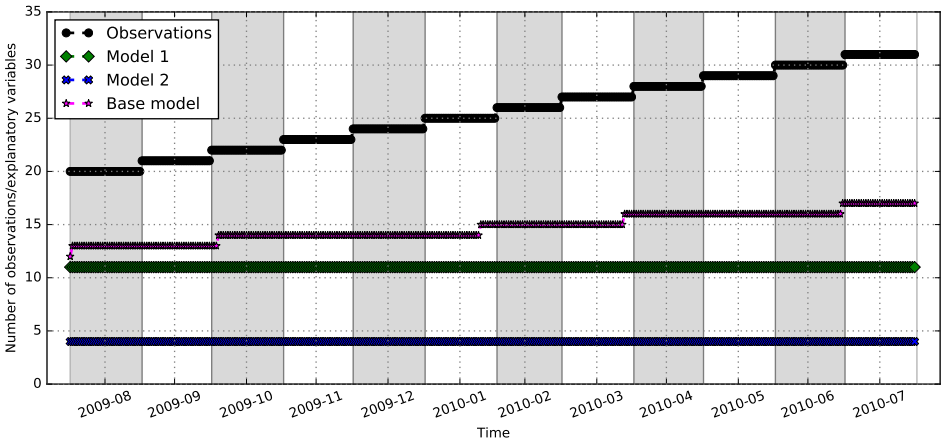

Figure 4 – Comparing the number of observations and explanatory variables used in each model.

5 Discussion

The discussion is based on results from the dataset from BPP that contains its biases and, therefore, should be read conditional on our experiment.

The out-of-sample exercise indicated that factors extracted from online prices were not able to explain most part of brazilian inflation dynamics. It is observed through the nowcasts provided by the three models and adjusted R^2 computed for all regressions. The poor predictive power of factors can be related to the weight of e-commerce prices on the computation of brazilian CPI.

Before our work, only (KNOTEK; ZAMAN, 2017) recognized BPP dataset as a source of information to nowcast CPI. However, these authors avoided the use of such data arguing that, in U.S., online purchases represent a small share of spending on goods, and by this fact the dataset would not be able to model inflation.

The results also show that out-of-sample performance change significantly depending on the way how factors are extracted from online prices. Figure 1 shows that Model 1 failed to capture changes in inflation dynamics, performing similar to historical average. Base model, however, was able to adjust the level of nowcasts in some months, as in figure 3. It conveys that factors extracted from category-level indexes, computed based on BPP data, track inflation better than those extracted directly from price series.

These same category-level indexes are used in (CAVALLO, 2013) to compute official CPI for five Latin American countries including Brazil. Moreover, (APARICIO; BERTOLOTTO, 2020) explore agregations of BPP data as explanatory variable to predict CPI, outperforming benchmarks for Euro Area, Australia, Canada and United States. Such works suggest that aggregations of BPP dataset are more informative than product-level series to model and forecast inflation, which justifies the improvements obtained from Base model compared to Model 1.

The best out-of-sample performance is provided by Model 2, which combines the recent inflation dynamics with factors to predict CPI, as also explored in (APARICIO; BERTOLOTTO, 2020). In figure 2, it is clear that the ARMA(1, 1) component helped regulating the level of the nowcasts, which is not observed for the remaining models. This result corroborates that inflation and error in previous month cannot be ignored from the information set used to estimate current CPI.

Concerning the estimation process, Figure 4 shows that the number of observations used to fit models is small compared with the number of explanatory varibales, which contributes to increase the uncertainty of predictions. For the first estimation window, for

instance, Model 1 and Base model use 11 and 12 explanatory variables, respectively, to fit 20 observations of inflation. Thus, the limited out-of-sample performance of proposed models can also be related to the short in-sample horizon used in estimation process.

6 Conclusion

This work contributes to nowcasting literature by exploiting data from online retailers in order to nowcast inflation. While the existing literature on nowcasting inflation use parsimonious sets of explanatory variables composed basically by off-line series, we exploit prices scraped from websites to compute daily estimates of current brazilian inflation. Moreover, our work adapts the popular dynamic factor model in (GIANNONE; REICHLIN; SMALL, 2008), to handle a extremely large set of price series provided by *The Billion Price Project*, not yet used to nowcast macroeconomic variables.

Taking into account the dataset explored, its biases associated to the way how prices are collected and treated, as well as the period covered by the data, our main findings are (1) prices from e-commerce are not able to explain most part of brazilian inflation dynamics, (2) the information set used to compute nowcasts of inflation should include auto-regressive components.

Concerning the use of online prices, our results show that factors extracted from category-level indexes are more efficient to nowcast inflation than factors extracted from stationarized price series. It suggests that future analysis should be developed based on category-level aggregations, as done in Base model, instead of product-level, as the case of Model 1 and 2.

Future analysis should also include: rerun the proposed models using a dataset which cover a longer time horizon. Investigate alternative ways to extract factors from BPP data, as Machine Learning methods. Explore diffenrent techniques to combine nowcasts, as the approach proposed by (AASTVEIT et al., 2014). And incorporate off-line series in the set of explanatory variables, as suggested by the literature.

Bibliography

AASTVEIT, K. A. et al. Nowcasting gdp in real time: A density combination approach. *Journal of Business & Economic Statistics*, Taylor & Francis, v. 32, n. 1, p. 48–68, 2014. Citado na página 35.

AMSTAD, M.; FISCHER, A. M. Are weekly inflation forecasts informative? *Oxford Bulletin of Economics and Statistics*, Wiley Online Library, v. 71, n. 2, p. 237–252, 2009. Citado na página 19.

APARICIO, D.; BERTOLOTTO, M. I. Forecasting inflation with online prices. *International Journal of Forecasting*, Elsevier, v. 36, n. 2, p. 232–247, 2020. Citado 2 vezes nas páginas 19 and 33.

BAI, J.; NG, S. Determining the number of factors in approximate factor models. *Econometrica*, Wiley Online Library, v. 70, n. 1, p. 191–221, 2002. Citado na página 25.

BAI, J.; NG, S. Determining the number of primitive shocks in factor models. *Journal of Business & Economic Statistics*, Taylor & Francis, v. 25, n. 1, p. 52–60, 2007. Citado 2 vezes nas páginas 25 and 30.

BAŃBURA, M. et al. Now-casting and the real-time data flow. In: *Handbook of economic forecasting*. [S.l.]: Elsevier, 2013. v. 2, p. 195–237. Citado na página 18.

BANBURA, M.; GIANNONE, D.; REICHLIN, L. Nowcasting with daily data. *European Central Bank, Working Paper*, Citeseer, 2011. Citado 2 vezes nas páginas 18 and 19.

BATES, J. M.; GRANGER, C. W. The combination of forecasts. *Journal of the Operational Research Society*, Taylor & Francis, v. 20, n. 4, p. 451–468, 1969. Citado na página 28.

CARRIERO, A.; CLARK, T. E.; MARCELLINO, M. Realtime nowcasting with a bayesian mixed frequency model with stochastic volatility. *Journal of the Royal Statistical Society. Series A,(Statistics in Society)*, Wiley-Blackwell, v. 178, n. 4, p. 837, 2015. Citado na página 18.

CAVALLO, A. Online and official price indexes: measuring argentina's inflation. *Journal of Monetary Economics*, Elsevier, v. 60, n. 2, p. 152–165, 2013. Citado 5 vezes nas páginas 7, 9, 19, 27, and 33.

CAVALLO, A. Scraped data and sticky prices. *Review of Economics and Statistics*, MIT Press, v. 100, n. 1, p. 105–119, 2018. Citado na página 21.

CLAESKENS, G. et al. The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, Elsevier, v. 32, n. 3, p. 754–762, 2016. Citado na página 28.

DOZ, C.; GIANNONE, D.; REICHLIN, L. *A Quasi Maximum Likelihood Approach for Large Dynamic Factor Models*. [S.l.], 2006. Citado na página 18.

DOZ, C.; GIANNONE, D.; REICHLIN, L. A two-step estimator for large approximate dynamic factor models based on kalman filtering. *Journal of Econometrics*, Elsevier, v. 164, n. 1, p. 188–205, 2011. Citado 2 vezes nas páginas 18 and 25.

FUNKE, M.; MEHROTRA, A.; YU, H. Tracking chinese cpi inflation in real time. *Empirical Economics*, Springer, v. 48, n. 4, p. 1619–1641, 2015. Citado 2 vezes nas páginas 17 and 19.

GIANNONE, D.; REICHLIN, L.; SIMONELLI, S. Nowcasting euro area economic activity in real time: the role of confidence indicators. *National Institute Economic Review*, SAGE Publications Sage UK: London, England, v. 210, n. 1, p. 90–97, 2009. Citado na página 18.

GIANNONE, D.; REICHLIN, L.; SMALL, D. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, Elsevier, v. 55, n. 4, p. 665–676, 2008. Citado 12 vezes nas páginas 7, 9, 15, 17, 18, 19, 23, 24, 25, 26, 28, and 35.

HARVEY, A. C. *Forecasting, structural time series models and the Kalman filter*. [S.l.]: Cambridge university press, 1990. Citado 3 vezes nas páginas 19, 25, and 27.

KNOTEK, E. S.; ZAMAN, S. Nowcasting us headline and core inflation. *Journal of Money, Credit and Banking*, Wiley Online Library, v. 49, n. 5, p. 931–968, 2017. Citado 3 vezes nas páginas 17, 18, and 33.

KUZIN, V.; MARCELLINO, M. G.; SCHUMACHER, C. Midas versus mixed-frequency var: nowcasting gdp in the euro area. Bundesbank Series 1 Discussion Paper, 2009. Citado na página 18.

MODUGNO, M. Now-casting inflation using high frequency data. *International Journal of Forecasting*, Elsevier, v. 29, n. 4, p. 664–675, 2013. Citado 2 vezes nas páginas 17 and 19.

TIMMERMANN, A. Forecast combinations. *Handbook of economic forecasting*, Elsevier, v. 1, p. 135–196, 2006. Citado na página 28.

APPENDIX A – Distribution of nowcasts

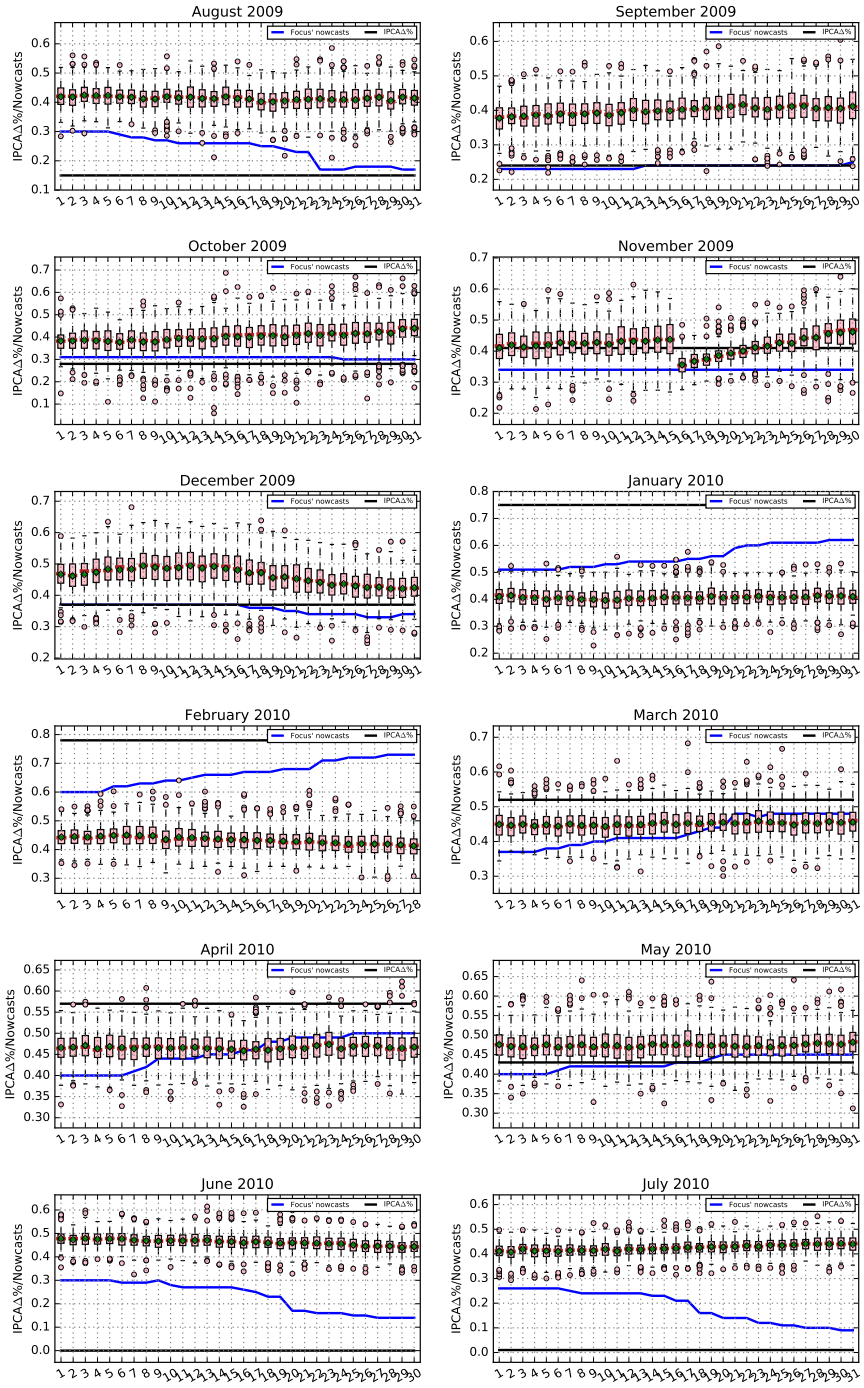

Figure 5 – Distribution of nowcasts generated by Model 1.

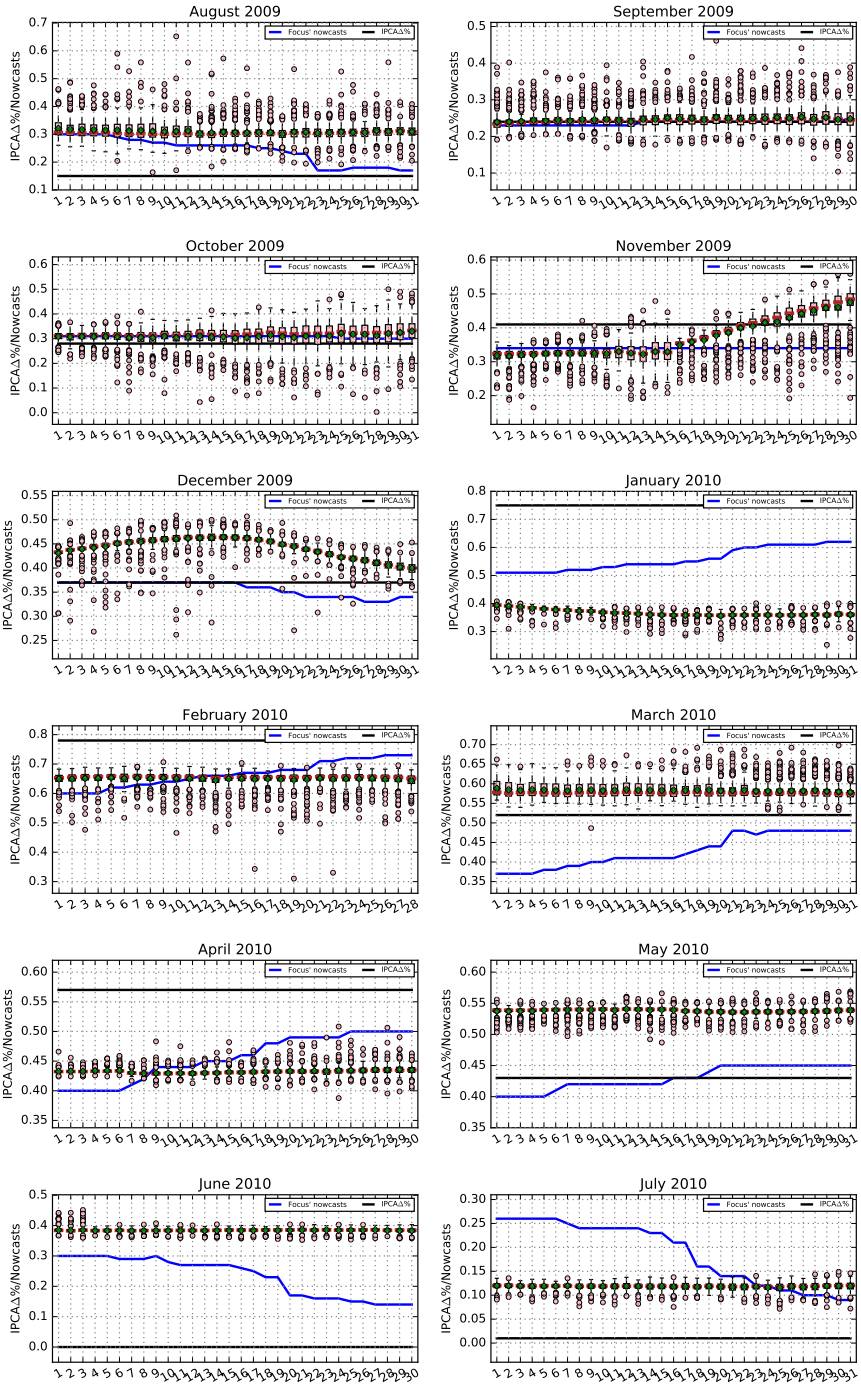

Figure 6 – Distribution of nowcasts generated by Model 2.

APPENDIX B – Distribution of R^2 and adjusted R^2

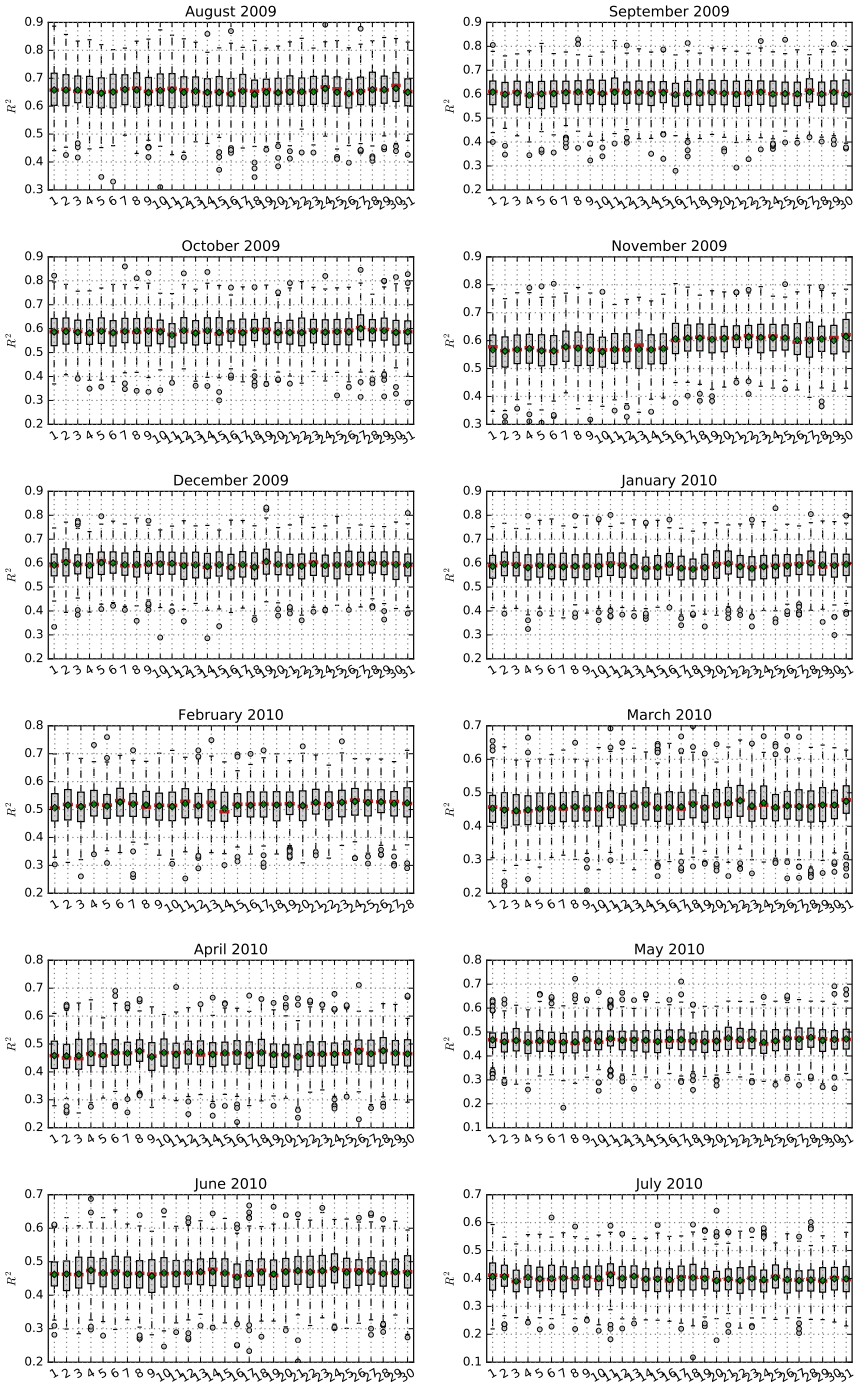

Figure 7 – Distribution of R^2 for Model 1.

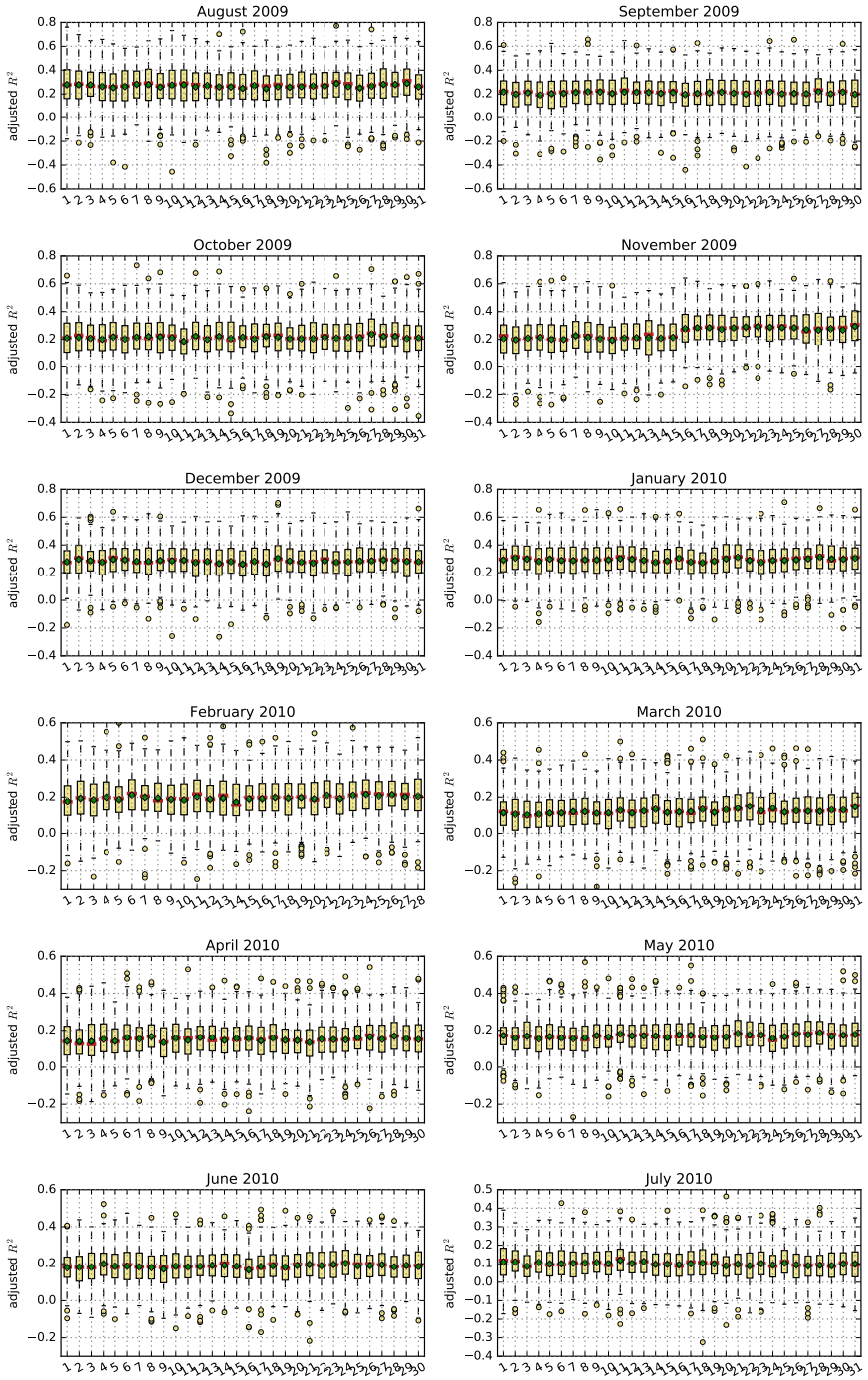

Figure 8 – Distribution of adjusted R^2 for Model 1.

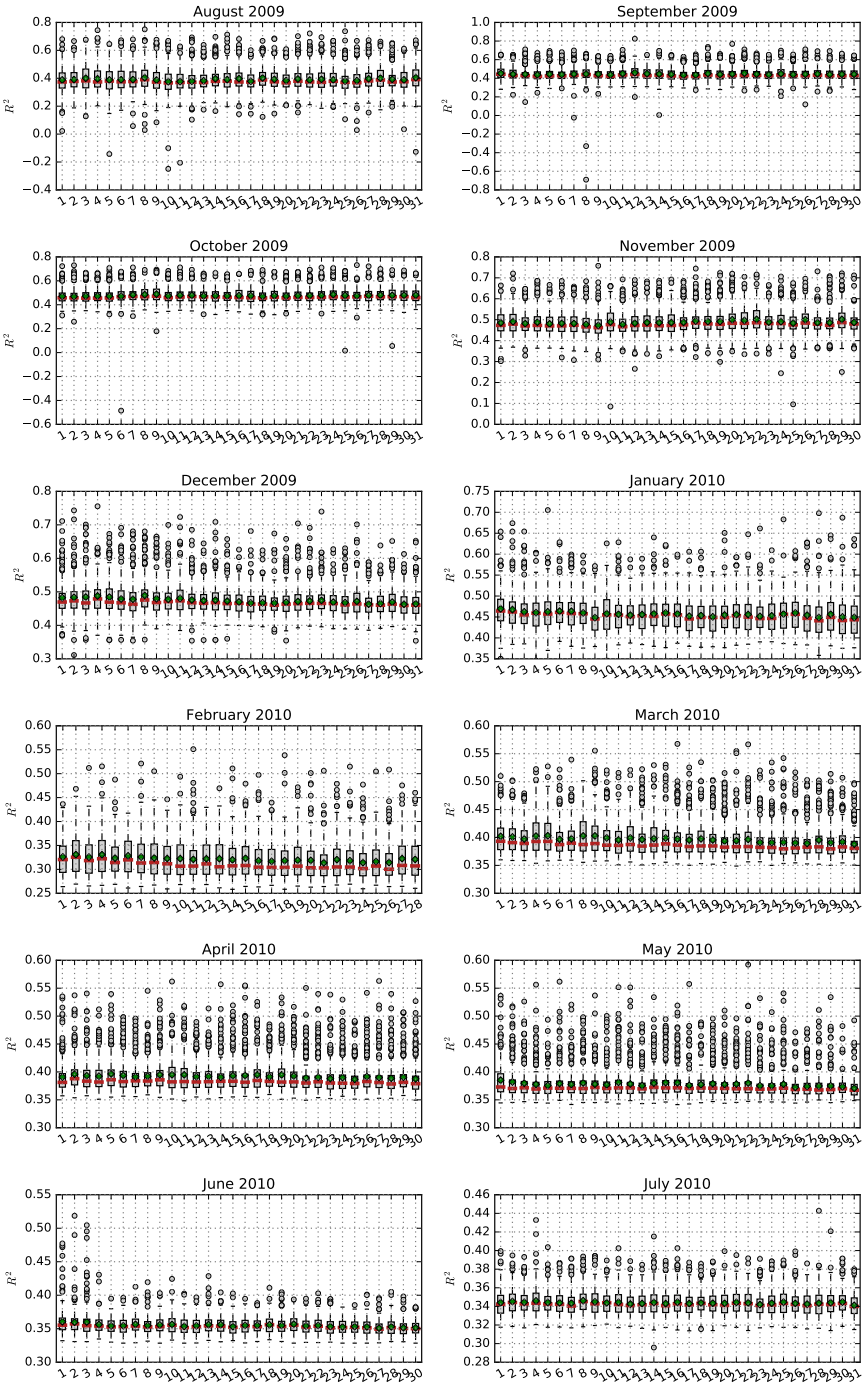

Figure 9 – Distribution of R^2 for Model 2.

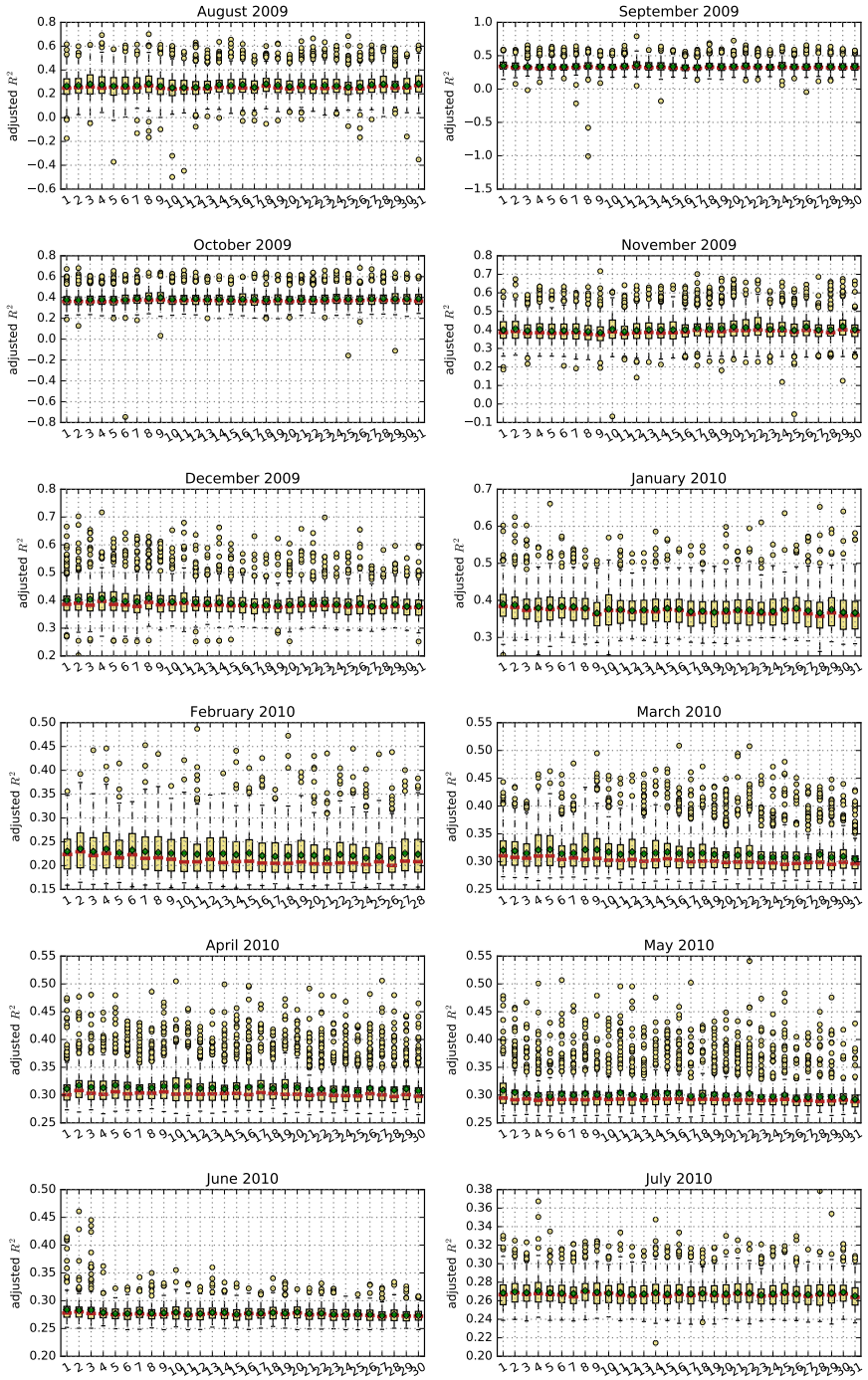

Figure 10 – Distribution of adjusted R^2 for Model 2.

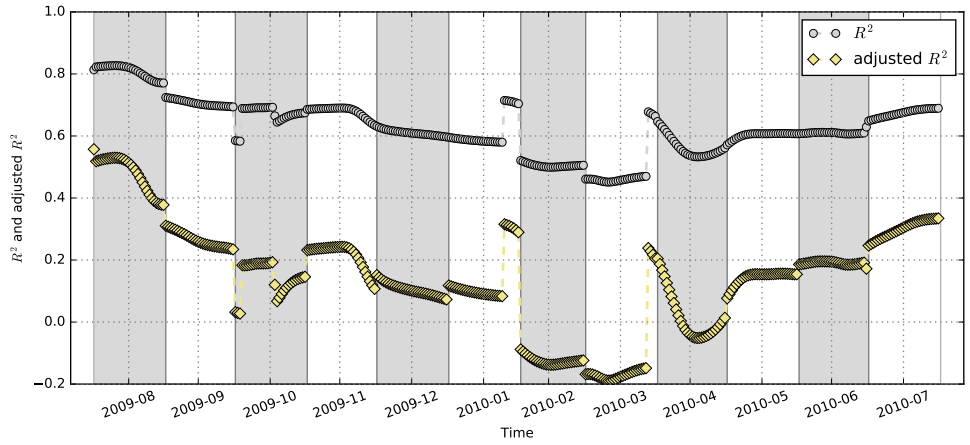

Figure 11 – R^2 and adjusted R^2 for Base model.