Conference on Non-traditional Data, Machine Learning and Natural Language Processing in Macroeconomics
Bank of Canada-Federal Reserve Board-Bank of Italy

Big Data Information & Nowcasting

Consumption & Investment from Bank Transactions in Turkey

A. B. Barlas [1], S. Guler [1], B. Orkun [1],
A. Ortiz [1], T. Rodrigo [1],
B. Soybilgen [2] & E. Yazgan [2]

[1]BBVA Research & [2]Bilgi University

November 2021

Introduction

- Some recent literature on BigData and Nowcasting after COVID.
- The role of Big Data from Financial Transactions including:
  - Consumer-to-Individual Transactions to mimic Consumption
  - Consumer-to-Individual + Firm-To-Firm Transactions to mimic Investment.
- Horse Racing: Out-of-Sample Test for Big Data information in Nowcasting
  - Standard Linear Models (DFM, BVAR)
  - Machine Learning: Linear and Non-Linear Models using Bridge Equations (Linear, Random Forest & Gradient Boost)
- Results

Recent Literature Spurred by Covid Crisis

- Developing of higher frequency (Weekly/Daily) Economic Activity Now-casting Models by Central Banks).
  - FED Weekly Economic Index (Lewis & Stock, 2020)
  - BundesBank Weekly Activity Index (Eraslan & Gozt,2020)
  - Central Bank of Portugal Daily GDP (Lourenco & Rua, 2020)
- Developing New Big Data Indicators: (Banking Transactions, Mobility...)
  - Financial Transactions
    - Alternative Sources for US. Cards PoS: Chetty et Al (2020).
    - Developed and EM countries. Cards PoS: Carvalho et al (2020).
    - National Accounts Consumption (from cards, direct debits, transfers, imputed rents) (Carvalho et al, 2021 Forthcoming)
    - National Accounts Investment (including Individual-to-firm & firm-to-firm transactions) (Carvalho et al, 2022 Forthcoming)
  - Other
    - Mobility indicators and others (Woloszko,2020)...

Financial Transactions Big Data: Garanti-BBVA Database

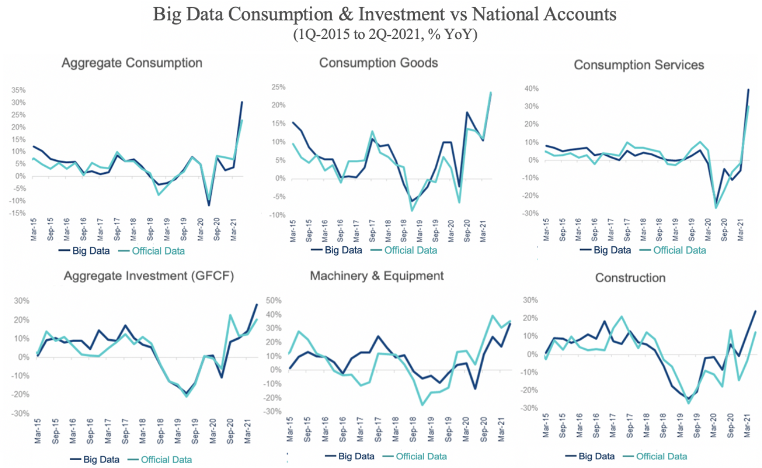
Garanti-BBVA: Consumption Transactions (2020)

395.7 Million	7.6 Million	1.67 Million
Card Transactions	**Card Holders**	**Merchants**

Garanti-BBVA: Investment Transactions (2020)

31.1 Million	367 Thousand.
Firm to Firm Transactions	**Firms**

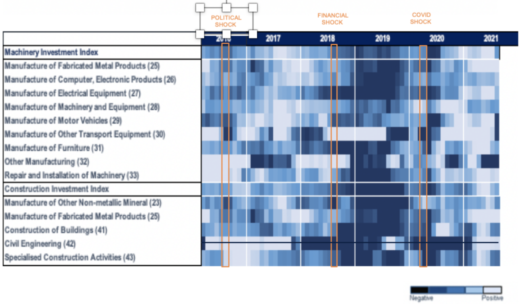
Big Data Indicators: Consumption & Investment in Real Time...



Source: Own Elaboration & Turkstat

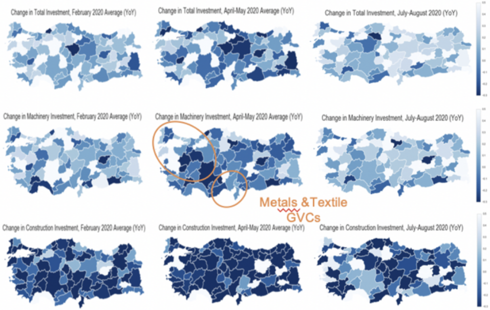
...and High Definition: By Assets or Goods & Provinces

Figure 3 Big Data Investment Sectoral HeatMap
(% YoY Light Colours stand for positive growth rates and Dark Colours for negative rates)



Source: Own Elaboration.

Figure 4 Big Data Regional Investment Maps
(% YoY Light Colours stand for positive growth rates and Dark Colours for negative rates)



Source: Own Elaboration.

Methodology: Big Data information in a Horse Race of Models

- A Horse Race including Bridge Linear (OLS) and Non-Linear Bridge equation models (Random Forest (RF) & Gradient Boost (GB)), Dynamic Factor Models (DFM), and Bayesian Vector Autoregressive models (BVAR) to nowcast GDP YoY growth rates.
- While DFM can deal with the missing data at the start of the dataset, we need to have a balanced dataset to estimate Bridge Equation models and BVAR.
- As our dataset is highly unbalanced, we follow Stekhoven and Bühlmann (2012) to fill out the missing data at the beginning of the dataset.

Data included in the Model

Table 2: Detail of Variables Included in the Nowcasting Models

Variable	Type	Frequency	StartDate	Transformation	Release Lags (M)
GDP	Hard	Quarterly	2003	YoY Growth	2-3
Industrial Production	Hard	Monthly	2006	YoY Growth	2
Auto Imports	Hard	Monthly	2006	YoY Growth	2
Auto Sales	Hard	Monthly	2003	YoY Growth	2
Auto Exports	Hard	Monthly	2006	YoY Growth	2
Non Metalllic Minerals	Hard	Monthly	2006	YoY Growth	2
Electricity Production	Hard	Daily	2003	YoY Growth	0
Number of Employed	Hard	Monthly	2006	YoY Growth	3
NUmber of Unemployed	Hard	Monthly	2006	YoY Growth	3
PMI	Soft	Monthly	2006	Level	1
Real Sector Confidence	Soft	Monthly	2003	Level	0
Loans (Credit)	Hard	Weekly	2006	Ann 13-week Growth	1
Big Data Consumption	Hard	Daily	2015	YoY Growth	0
Big Data Investment	Hard	Daily	2015	YoY Growth	0

Source: Own Elaboration

Data included in the Model Prevalence could penalize Big Data

Table 2: Detail of Variables Included in the Nowcasting Models

Variable	Type	Frequency	StartDate	Transformation	Release Lags (M)
GDP	Hard	Quarterly	2003	YoY Growth	2-3
Industrial Production	Hard	Monthly	2006	YoY Growth	2
Auto Imports	Hard	Monthly	2006	YoY Growth	2
Auto Sales	Hard	Monthly	2003	YoY Growth	2
Auto Exports	Hard	Monthly	2006	YoY Growth	2
Non Metalllic Minerals	Hard	Monthly	2006	YoY Growth	2
Electricity Production	Hard	Daily	2003	YoY Growth	0
Number of Employed	Hard	Monthly	2006	YoY Growth	3
NUmber of Unemployed	Hard	Monthly	2006	YoY Growth	3
PMI	Soft	Monthly	2006	Level	1
Real Sector Confidence	Soft	Monthly	2003	Level	0
Loans (Credit)	Hard	Weekly	2006	Ann 13-week Growth	1
Big Data Consumption	Hard	Daily	2015	YoY Growth	0
Big Data Investment	Hard	Daily	2015	YoY Growth	0

Source: Own Elaboration

Big Data Information only present from 2015 (30% of the Sample) therefore results could be subject to prevalence (i. e Penalizing the Big Data Info)

The Models: Linear & Non-Linear Bridge Equations

- We use Bridge Equations to convert:
  - A Monthly Vector:$x_{t_m} = (x_{1,t_m}, x_{2,t_m}, \dots, x_{n,t_m})'$, $t_m = 1, 2, \dots, T_m$ of n (std) variables
  - In Quarterly ones: $x_{t_q} = (x_{1,t_q}, x_{2,t_q}, \dots, x_{n,t_q})'$, $t_q = 1, 2, \dots, T_q$, by taking simple averages of x_{t_m}. Missing data for the reference quarter(s) will be filled by an AR(p) model (p chosen according to AIC)
- The functional form between the Output y_{t_q} and Input x_{t_q} is given by $g()$:

$$y_{t_q} = g(x_{t_q}) + \varepsilon_{t_q} \tag{1}$$

- In our case $g()$ can take linear (OLS) or nonlinear functional forms as Random Forests (RF) & Gradient Boost decision trees (GBM).

The Models: Dynamic Factor Model (DFM)

- We model the DFM with idiosyncratic components $\epsilon_{i,t}$ as:

$$x_{t_m} = \Lambda f_{t_m} + \epsilon_{t_m}; \tag{2}$$

$$\epsilon_{t_m} = \alpha \epsilon_{t_m - 1} + v_{t_m}; \quad v_{t_m} \sim i.i.d.\ \mathcal{N}(0, \sigma^2), \tag{3}$$

- The unobserved common factors vector f_t evolves as:

$$f_{t_m} = \varphi(L) f_{t_m - 1} + \eta_{t_m}; \quad \eta_{t_m} \sim i.i.d.\ \mathcal{N}(0, R), \tag{4}$$

- We transform to quarterly GDP growth rates by:

$$y^Q_{t_m} = \bar{\Lambda}_Q [f'_t \, f'_{t-1} \, f'_{t-2}] + \bar{\epsilon}^Q_{t_m} \tag{5}$$

$$\bar{\epsilon}^Q_{t_m} = \alpha^Q \bar{\epsilon}^Q_{t_m - 1} + \bar{v}^Q_{t_m}; \quad \bar{v}^Q_{t_m} \sim i.i.d.\ \mathcal{N}(0, \bar{\sigma}^2), \tag{6}$$

The Models: BVAR

- Define $y_{t_m}^{Q}$ as the monthly GDP growth rates (partially observed in the third month of the quarter & linked to its unobserved monthly counterpart as:

$$y_{t_m}^{Q} = \frac{1}{3}(x_{t_m}^{Q} + x_{t_m-1}^{Q} + x_{t_m-2}^{Q}). \tag{7}$$

- We assume $x_{t_m}^{QM}$ follow a VAR(p) process as:

$$x_{t_m} = \varphi(L)x_{t_m-1} + u_{t_m}; \quad u_{t_m} \sim i.i.d.\ \mathcal{N}(0, \Sigma), \tag{8}$$

- The BVAR's state-space transition and measurement equation evolves as:

$$z_{t_m} = \pi + \Pi z_{t_m-1} + \zeta_{t_m}; \quad \zeta_{t_m} \sim i.i.d.\ \mathcal{N}(0, \Omega); \tag{9}$$

$$X_{t_m} = M_t \alpha z_{t_m} \tag{10}$$

Results: Nowcasting Models Mean Absolute Errors (MAE)

$$\text{MAE}^{(i)} = (1/n) \sum_{t_q=2016Q1}^{2020Q3} |y_{t_q} - \hat{y}_{t_q}^{(i)}|; \quad i = 1, 2, ..., 5.$$

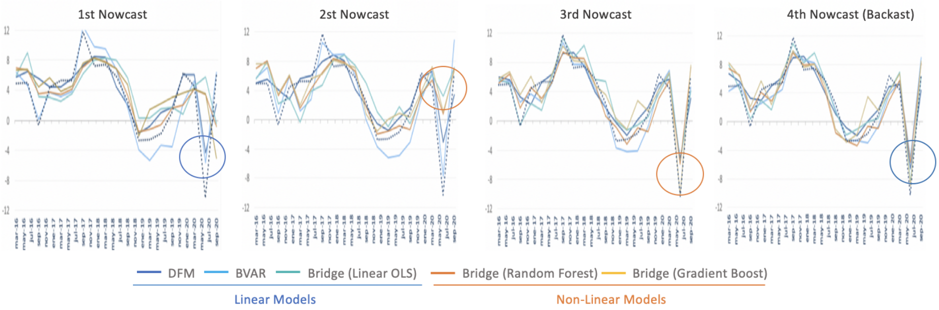
Table 3 MAEs of the models for successive nowcasting horizons between 2016Q1 and 2020Q3

	AR	DFM	BVAR	LM	RF	GBM
1st Nowcast	3.71	1.92	1.77	3.46	2.60	3.13
2nd Nowcast	3.71	1.85	2.29	3.07	2.32	2.55
3rd Nowcast	3.80	1.72	1.52	1.70	1.53	1.71
4th Nowcast	3.80	1.58	1.45	1.42	1.74	1.83
5th Nowcast	3.80	1.38	1.64	1.46	1.65	1.49

Abbreviations: AR, the benchmark autoregressive model; DFM, the dynamic factor model; BVAR, the Bayesian vector autoregressive model; LM, the linear bridge equation model; RF, the random forest based bridge equation model; GBM, the gradient tree boosted bridge equation model.

Results: Alternative Nowcasting Models vs Official Data

Figure: Alternative Nowcasting Models vs Official: Linear vs Non-Linear (2016Q1 to 2020Q3)



The Standard Linear DFM & BVAR fitted the sharp swing (COVID) thanks to Big Data info..

The NonLinear Model (RF & GB) needed some time to fit the Official Data..

...But finally catched up with the Official Data as more information were included...

...Once all the relevant HardData is also present Most model fit relatively well

Results: Combination vs Individual Nowcasting Models

Table 4 MAEs of nowcasting combinations for successive nowcasting horizons between 2018Q1 and 2020Q3

	Averaging Models*				**Individual Nowcasting Models				
	Simple	Median	RPW	Rank	DFM	BVAR	LM	RF	GBM
1st Nowcast	2.67	3.29	2.53	2.30	2.01	2.16	4.37	3.18	4.20
2nd Nowcast	2.03	2.40	1.95	1.89	2.09	2.65	3.40	2.20	2.69
3rd Nowcast	1.39	1.65	1.32	1.34	1.92	1.95	1.99	1.80	1.68
4th Nowcast	1.44	1.43	1.44	1.45	1.59	1.57	1.22	2.06	1.88
5th Nowcast	1.36	1.43	1.38	1.43	1.48	1.82	1.44	1.77	1.75

*Averaging Models: Simple (average), Median (median), Relative Performance or RPW (weights calculated as the inverse of the error), Rank (weigths according the rank of the model).
** DFM(Dynamic Factor Model), BVAR (Bayesian VAR), LM(Machine Learning Linear Model), RF(Random Forest) and GBM (Gradient Boost)
Source: Own Elaboration

BigData & Nowcasting: Pre-selection of Variables (Lasso)

Figure: MAE for Models with Pre-Selection of Variables (2016Q1 to 2020Q3)

	AR	DFM	BVAR	LM	RF	GBM
1st Nowcast	3.71	2.52	2.17	3.24	2.81	3.47
2nd Nowcast	3.71	2.15	1.45	2.63	2.07	2.57
3rd Nowcast	3.80	1.72	1.64	1.36	1.48	1.62
4th Nowcast	3.80	1.73	1.38	1.28	1.73	1.76
5th Nowcast	3.80	1.64	1.36	1.08	1.56	1.56

First Step: We first select variables at any moment of time by using a linear regression with L1 regularization as known as Lasso regression and restricti variables at to use only the variables selected b by Lasso (significant non zero coefficient). Second Step: We use the variables wselected by Lasso in the main nowcasting models.

Source: Own Elaboration

BigData & Nowcasting:Variable Selection (Linear&Non-Linear)

Table 7: Selection Ration by Linear Model (Lasso)
(% Periods variables chosen by Linear Model (Lasso))

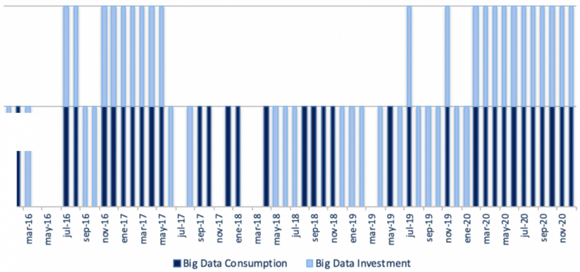
Name	Selection Ratio
IP	100.0%
Car Imports	0.0%
Ind. Production Non-Metallic Minerals	98.3%
Car Total Sales	1.7%
Electricity Demand	48.3%
Number of Employed	8.3%
Number of Unemployed	15.0%
Car Exports	0.0%
PMI	98.3%
Total Loans 13week	83.3%
Real Sector Confidence Index	100.0%
Big Data Consumption	55.0%
Big Data Investment	68.3%

Table C1: Selection Ration by Non-Linear Model (RF)
(% mean decrease in MSE calculated from out-of-bag sample in Random Forest Model)

Name	Selection Ratio
IP	17.4%
Car Imports	-0.2%
Ind. Production Non-Metallic Minerals	11.2%
Car Total Sales	-0.5%
Electricity Demand	2.6%
Number of Employed	3.2%
Number of Unemployed	6.3%
Car Exports	5.8%
PMI	5.1%
Total Loans 13week	4.3%
Real Sector Confidence Index	4.3%
Big Data Consumption	5.1%
Big Data Investment	9.8%

Source: Turkstat, Markitt, OSD and Own Elaboration

Figure 6 Big Data Investment and Consumption variables selection by Lasso Regression



BigData Contribution to Nowcasting: Models & Periods

Mean Absolute Error Difference (MAED): Traditional Information vs Big Data

$$\mathrm{MAED}^{(i)} = \mathrm{MAE}^{(i)} - \mathrm{MAE}^{(i)}_{RD}; \quad i = 1, 2, ..., 5.$$

RD: Models without Big data

Table 8 MAEDs of the models for successive nowcasting horizons between 2016Q1 and 2020Q3

	Linear Models			Non-Linear Models	
	DFM	BVAR	LM	RF	GBM
1st Nowcast	0.09	0.57	0.39	0.51	0.28
2nd Nowcast	0.09	-0.60	0.26	0.22	0.03
3rd Nowcast	0.07	-0.13	0.01	0.12	-0.04
4th Nowcast	0.06	0.11	0.00	0.02	-0.29
5th Nowcast	0.05	-0.01	0.06	-0.20	0.01

Abbreviations: DFM, the dynamic factor model; BVAR, the Bayesian vector autoregressive model; LM, the linear bridge equation model; RF, the random forest based bridge equation model; GBM, the gradient tree boosted bridge equation model.

Source: Own Elaboration

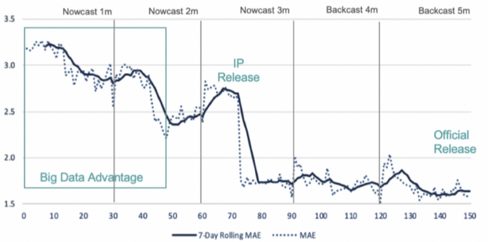
BigData Contribution to Nowcasting: Time Advantage

Table A.1 Announcement days and delays of the monthly variables

Name	Announcement Lag in Months	Announcement Day
Industrial Production (IP)	2	13
Car Imports	2	15
IP Non Metallic Minerals	2	13
Car Sales	2	15
Electricity Demand	0	30
Number of Employed	3	12
Number of Unemployed	3	12
Car Exports	2	15
Manufacturing PMI	1	1
Total Loans 13week	1	10
Real Sector Confidence Index	0	26
Big Data Consumption	0	Daily
Big Data Investment	0	Daily

Source: Own Elaboration through Turkstat, OSD, Markit, CBRT and own Big Data

Figure 7 Daily MAEs of equally weighted nowcast combinations betwee 2016Q1 and 2020Q3



* We run the models on daily basis assuming that big data variables are released daily but the rest of variables are announced at a specific date as shown in Table A1. For the sake of simplicity, we assume that each month consists of 30 days and calculate nowcasts for the reference quarter for 150 days until GDP is announced.Instead of showing each model individually, we take simple averages of all models' nowcasts.

Conclusions

- Financial Transactions´ BigData improve accuracy of Nowcasting models in Turkey. It is useful more than 50% of the time (even with prevalence).
- The contribution is more relevant during the first 45 days (when Hard relevant Data is scarce) and uncertain crisis times.
- The Standard Nowcasting Models as Dynamic Factor Model (DFM) & Bayesian VARs (BVAR) appears to be a good alternative model even in a volatile environment (Turkey has been exposed to relevant shocks during last 4 years).
- Nowcast combination outperform most of the single models in many cases but not in short term.
- Non-Linear Models could be more useful during Turning Points but no evidence they supposed an advantage during COVID in this exercise.(Ragged edge problem? Short Sample?)

Thanks

Thanks!!

Follow this and other indicators in real time here

https://www.bbvaresearch.com/en/special-section/charts/