Fourth Statistics Conference: Statistics post Pandemic

Central Bank of Chile

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

September 2021

Introduction

- Recent literature on BigData & Nowcasting .
- Role of Big Data from Financial Transactions including:
  - Consumer-to-Individual Transactions Consumption to mimic Consumption
  - Consumer-to-Individual + Firm-To-Firm Transactions to mimic Investment.
- Test: Out-of-Sample errors of Big Data information in Nowcasting
  - Standard Linear Models (DFM, BVAR)
  - Machine Learning: Linear Non-Linear Models (Linear, Random Forest Gradient Boost) using Bridge Equations
- Results

Recent Literature

- Developing higher frequency models (Weekly/Daily Economic Indexes by Central Banks).
  - FED Weekly Economic Index (Lewis Stock, 2020)
  - BundesBank Weekly Activity Index (Eraslan and Gozt,2020)
  - Central Bank of Portugal Daily GDP (Lourenco and 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).
    - Consumption including Cards Other Transfers (2021).
  - Other
    - Mobility indicators (Woloszko,2020)...

Garanti BBVA Transaction Data: Card & Money Transfers

270 Mill
Card transactions

7.6 Mill.
cardholders

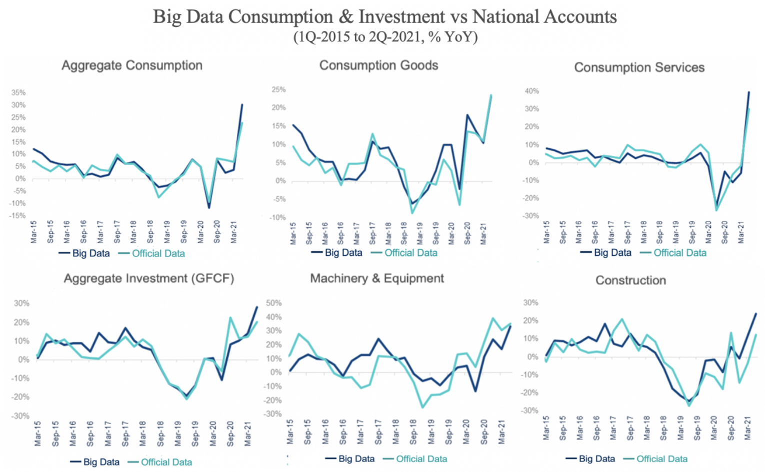
659 Mill
transactions

Table 1 Investment Firms Statistics: Garanti BBVA vs Central Bank of Turkey (CBRT) & Turkstats Survey

	Garanti BBVA			CBRT-Turkstat		
Variable	Tot.	Machinery	Constr.	Tot.	Machinery	Constr.
Transactions(000s)	24.6	22.3	2.3			
Amount(US Bn)	308	280	28	440	257	183
Firms(000s)	179.7	156.5	23.2	730.2	614.4	115.8
Firms(% CBRT)	24.6	25.5	19.8			

Source: Garanti Bank and CBRT- Turkstat Survey.

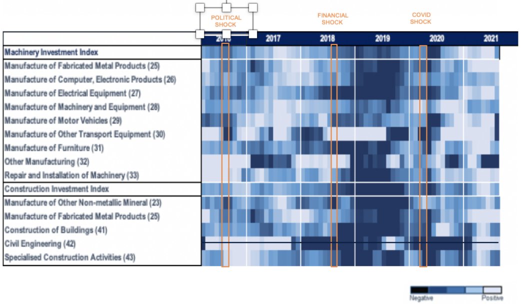
Cross Validation: BigData Consumption & Investment



Source: Own Elaboration & Turkstat

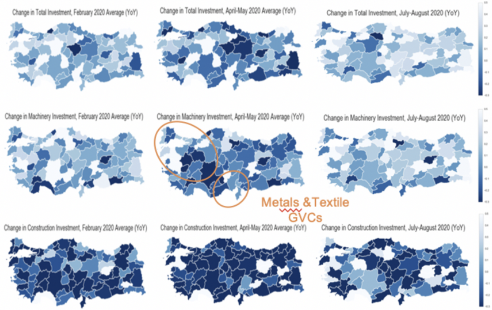
Results by Assets in Real Time & High Definition (Provinces)

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: Testing BigData 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
GDP	Hard	Quarterly	2003	YoY Growth
Industrial Production	Hard	Monthly	2006	YoY Growth
Auto Imports	Hard	Monthly	2006	YoY Growth
Auto Sales	Hard	Monthly	2003	YoY Growth
Auto Exports	Hard	Monthly	2006	YoY Growth
Non Metalllic Minerals	Hard	Monthly	2006	YoY Growth
Electricity Production	Hard	Daily	2003	YoY Growth
Number of Employed	Hard	Monthly	2006	YoY Growth
NUmber of Unemployed	Hard	Monthly	2006	YoY Growth
PMI	Soft	Monthly	2006	Level
Real Sector Confidence	Soft	Monthly	2003	Level
Loans (Credit)	Hard	Weekly	2006	Ann 13-week Growth
Big Data Consumption	Hard	Daily	2015	YoY Growth
Big Data Investment	Hard	Daily	2015	YoY Growth

Source: Own Elaboration

Description of Models: Linear and Non-Linear Bridge Equations

- Monthly Vector: $x_{t_m} = (x_{1,t_m}, x_{2,t_m}, \ldots, x_{n,t_m})', t_m = 1, 2, \ldots, T_m$ as n monthly standardized explanatory variables.
- Quarterly Vector: $x_{t_q} = (x_{1,t_q}, x_{2,t_q}, \ldots, x_{n,t_q})', t_q = 1, 2, \ldots, T_q$, 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 Linear & Non-linear function between the Output y_{t_q} and the Input x_{t_q} will be given by $g()$:

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

- Where $g()$ defines a linear (OLS) or or a nonlinear functional form random forests (RF) and gradient boosted decision trees (GBM).

Description of Models: Dynamic Factor Models (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:

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

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

Description of Models: BVAR

- $y_{t_m}^Q$ denotes a partially observed monthly counterpart of GDP growth rates that can only be observed in the third month of the respective quarter and linked its unobserved monthly counterpart as follows:

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

- 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:

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

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

Nowcasting Performance: Mean Absolute Errors (MAE)

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

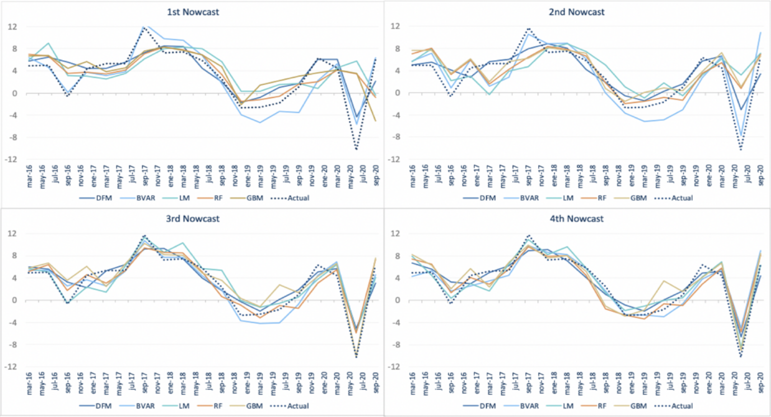
Table 3 MAEs of the models for successive nowcasting horizons between 2006Q1 and 2020Q3

	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.

Nowcasting Performance: Alternative Models vs Official

Figure: Alternative Models vs Official(2006Q1 to 2020Q3)



Source: Own Elaboration

Nowcasting Performance: Combination vs Individual Models

Figure: MAE for Alternative Models (2008Q1 to 2020Q3)

Table 4 MAEs of nowcasting combinations for successive nowcasting horizons between 2008Q2 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 Averaging (Simple), Median (Median), Relative Performance Weight (RPW) , Rank based Weight (Rank)
**Individual Models: Dynamic Factor Model (DFM), Bayesian VAR (BVAR) , Bridge Linear (LM), Bridge Random Forest (RF), Bridge Gradient Boost Model (GBM)

Source: Own Elaboration

Nowcasting Performance: Pre Selection of Variables (Lasso)

Figure: MAE for Models with Pre-Selection of Variables (2006Q1 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

Individual Models: Dynamic Factor Model (DFM), Bayesian VAR (BVAR) , Bridge Linear (LM), Bridge Random Forest (RF), Bridge Gradient Boost Model (GBM)
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))

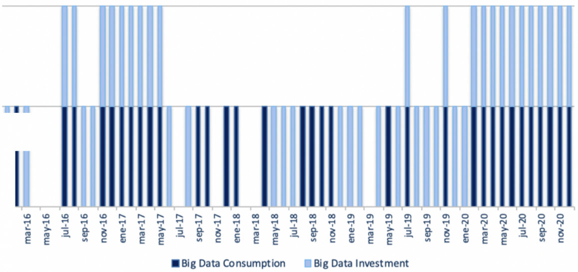
Name	Selection Ratio
IP	100.0%
Car Imports	0.0%
Ind. Production Non-Metallic Minerals	98.3%
Car Total Sales	1.7%
Electricity Demand	48.3%
Number of Employed	8.3%
Number of Unemployed	15.0%
Car Exports	0.0%
PMI	98.3%
Total Loans 13week	83.3%
Real Sector Confidence Index	100.0%
Big Data Consumption	55.0%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



Contribution BigData 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 2006Q1 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.

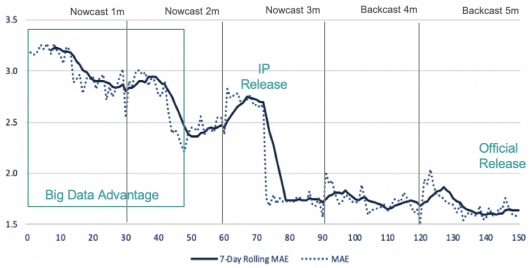
Contribution BigData to Nowcasting: Time Advantage

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 between 2006Q1 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 will be more useful during shocks and Turning Points.

References

Akcigit, Ufuk et al. (2019). *Facts on Business Dynamism in Turkey*. CBRT Working Paper No. 19/30.

Alexander D; Karger, E (2020). *Do stay-at-home orders cause people to stay at home? Effects of stay-at-home orders on consumer behavior*. Federal Reserve Bank of Chicago Working Paper No. 2020-12.

Andersen, Asger Lau et al. (2020). *Consumer Responses to the COVID-19 Crisis: Evidence from Bank Account Transaction Data*. CEBI Working Paper Series No. 1820.

Ankargren, Sebastian and Yukai Yang (2019). "Mixed-Frequency Bayesian VAR Models in R: the mfbvar package". In:

Aprigliano, Valentina, Guerino Ardizzi, and Libero Monteforte (2017). *Using the payment system data to forecast the Italian GDP*. Bank of Italy Temi di Discussione Working Paper No. 1098.

Babii, Andrii, Eric Ghysels, and Jonas Striaukas (2021). "Machine learning time series regressions with an application to nowcasting". In: *Journal of Business & Economic Statistics*, pp. 1–23.

Baker, Scott R et al. (2020a). *How Does Household Spending Respond to an Epidemic? Consumption During the 2020 COVID-19 Pandemic*. NBER Working Paper No. 26949.

Baker, Scott R et al. (2020b). *Income, Liquidity, and the Consumption Response to the 2020 Economic Stimulus Payments*. NBER Working Paper No. 27097.

Bańbura, Marta and Michele Modugno (2014). "Maximum Likelihood Estimation of Factor Models on Datasets with Arbitrary Pattern of Missing Data". In: *Journal of Applied Econometrics* 29(1), pp. 133–160.

Barlas, Ali B et al. (2020). *Investment in real time and High definition: A Big data Approach*. BBVA Research Working Papers No. 2013.

Barnett, William et al. (2016). *Nowcasting Nominal GDP with the Credit-Card Augmented Divisia Monetary Aggregates*. University of Kansas Working Papers Series in Theoretical and Applied Economics No. 201605.

Bodas, Diego et al. (2019). *Measuring Retail Trade using Card Transactional Data*. Tech. rep.

Bounie D. Camara, Y. and J. Galbraith (2020). *Consumers' Mobility, Expenditure and Online- Offline Substitution Response to COVID-19: Evidence from French Transaction Data*. HAL Working Papers No. 02566443.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45(1), pp. 5–32.

Carvalho, Vasco M, Stephen Hansen, Alvaro Ortiz, Juan Ramon Garcia, et al. (2020). *Tracking the Covid-19 crisis with high resolution transaction data*. CEPR DP No 14642.

Carvalho, Vasco M, Stephen Hansen, Alvaro Ortiz, Tomasa Rodrigo, et al. (2021). *National accounts in a world of naturally occurring data*. MiMEO.

Chetty, Raj et al. (2020). *How Did COVID-19 and Stabilization Policies Affect Spending and Employment? A New Real-Time Economic Tracker Based on Private Sector Data*. NBER WP No 27431.

Chronopoulos, Dimitris K, Marcel Lukas, and John OS Wilson (2020). *Consumer Spending Responses to the COVID-19 Pandemic: An Assessment of Great Britain*. Working Papers in Responsible Banking and Finance No.20-012.

Clemen, Robert T. (1989). "Combining Forecasts: A Review and Anotated Bibliography". In: *International journal of forecasting* 5(4), pp. 559–583.

Cox, Natalie et al. (2020). *Initial Impacts of the Pandemic on Consumer Behavior: Evidence from Linked Income, Spending, and Savings Data*. University of Chicago, Becker Friedman Institute for Economics Working Paper.

References II

Duarte, Cláudia, Paulo MM Rodrigues, and António Rua (2017). "A mixed frequency approach to the forecasting of private consumption with ATM/POS data". In: *International Journal of Forecasting* 33(1), pp. 61–75.

Eraslan, S. and T. Götz (2020). *An unconventional weekly economic activity index for Germany.* Deutsche Bundesbank Technical Paper No. 02/202.

Friedman, Jerome (2002). "Stochastic gradient boosting". In: *Computational statistics & data analysis* 38(4), pp. 367–378.

Galbraith, John and Greg Tkacz (2015). *Nowcasting GDP with electronic payments data.* European Central Bank Statistics Paper Series No. 10.

Genre, Véronique et al. (2013). "Combining Expert Forecasts: Can Anything Beat the Simple Average?" In: *International Journal of Forecasting* 29(1), pp. 108–121.

Gezici, Ferhan, Burçin Yazgı Walsh, and Sinem Metin Kacar (2017). "Regional and structural analysis of the manufacturing industry in Turkey". In: *The Annals of Regional Science* 59(1), pp. 209–230.

Giannone, Domenico, Silvia Miranda Agrippino, and Michele Modugno (2013). *Nowcasting China Real GDP.* Tech. rep.

Hacioglu, Sinem, Diego R Känzig, and Paolo Surico (2020). *Consumption in the time of Covid-19: Evidence from UK transaction data.* CEPR Discussion Papers No. 14733.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

Hendry, David F and Michael P Clements (2004). "Pooling of Forecasts". In: *The Econometrics Journal* 7(1), pp. 1–31.

Huang, Huiyu and Tae-Hwy Lee (2010). "To Combine Forecasts or to Combine Information?" In: *Econometric Reviews* 29(5-6), pp. 534–570.

Lewis, Daniel J et al. (2020). *Measuring Real Activity Using a Weekly Economic Index.* New York FED WP No 920.

Lourenco, N. and A. Rua (2020). *The DEI: tracking economic activity daily during the lockdown.* Banco de Portugal Working Paper No 13.

Modugno, Michele, Barış Soybilgen, and Ege Yazgan (2016). "Nowcasting Turkish GDP and news decomposition". In: *International Journal of Forecasting* 32(4), pp. 1369–1384.

Olafsson, Arna and Michaela Pagel (2018). "The liquid hand-to-mouth: Evidence from personal finance management software". In: *The Review of Financial Studies* 31(11), pp. 4398–4446.

Schorfheide, Frank and Dongho Song (2015). "Real-time forecasting with a mixed-frequency VAR". In: *Journal of Business & Economic Statistics* 33(3), pp. 366–380.

Sebastian, Ankargren, Unosson Måns, and Yang Yukai (2020). "A Flexible Mixed-Frequency Vector Autoregression with a Steady-State Prior". In: *Journal of Time Series Econometrics* 12(2).

Soybilgen, Barış and Ege Yazgan (2018). "Evaluating nowcasts of bridge equations with advanced combination schemes for the Turkish unemployment rate". In: *Economic Modelling* 72, pp. 99–108.

Soybilgen, Barış and Ege Yazgan (2021). "Nowcasting US GDP Using Tree-Based Ensemble Models and Dynamic Factors". In: *Computational Economics* 57(1), pp. 387–417.

Stekhoven, Daniel J and Peter Bühlmann (2012). "MissForest—non-parametric missing value imputation for mixed-type data". In: *Bioinformatics* 28(1), pp. 112–118.

Stock, James and Mark W. Watson (2004). "Combination forecasts of output growth in a seven-country data set". In: *Journal of Forecasting* 23(6), pp. 405–430.

Timmermann, Allan (2006). "Forecast combinations". In: *Handbook of economic forecasting*, pp. 135–196.

Verbaan, Roy, Wilko Bolt, and Carin van der Cruijsen (2017). *Using debit card payments data for nowcasting Dutch household consumption.* De Nederlandsche Bank Working Paper No. 571.