BBVA Research

The use of Big Data for the statistical production. The experience of BBVA Research

European Statistics Day

Spanish National Statistics Institute

October 2019

Creating Opportunities

BBVA Research

Index

01 Opportunities and challenges in the digital era. Main takeaways working with Big Data at BBVA Research

02 Some applications:

- Economic indicators in Real Time
- Economic Analysis in High Definition
- Social & Economic Networks

Main takeaways working with Big Data at BBVA Research

It helps us to …

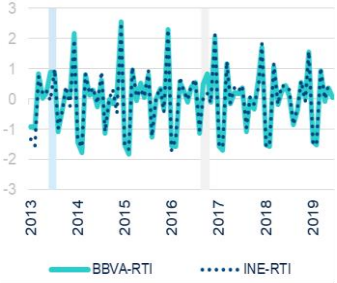
Complement and enrich our traditional databases with high dimensional data:

- Quantifying **new trends** and exploiting new **dimensions.**
- Having **timely answers** on the impact of different events, **providing early warning signals** indicators.
- Improving our models performance at **nowcasting.**

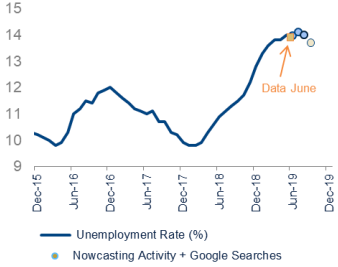
… but still some challenges

- **Data challenges**: missing data, data sparsity, data quality,…
- In most cases, there's not enough **time horizon** to improve our models performance at **forecasting.**
- **Legal and regulatory** issues for data sharing.

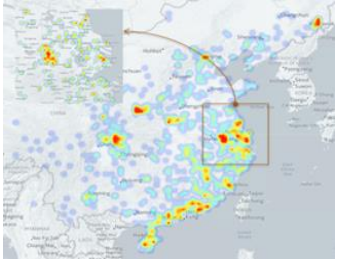
RETAIL SALES IN SPAIN: INE & BBVA



UNEMPLOYMENT NOWCASTING IN TURKEY



HOUSING PRICE MEDIA SENTIMENT IN CHINA 2018



CENTRAL BANK NARRATIVE WORDCLOUD

Data treatment and robustness check became the most time consuming parts of the working process

To face with new and high dimensional data

1

Data treatment and analysis:

Data cleaning, missing values, outlier detection, high heterogeneity, sparsity,…

New methodologies to face data challenges: dimensionality reduction, clustering, regularization,…

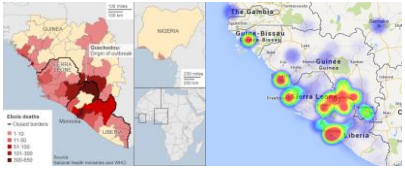
Massive and unstructured datasets:
Importance of making the right questions

2

Robustness check:

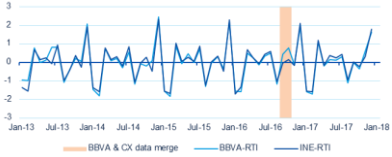
Cross-check of Big Data outcome with traditional data and methodologies.

Ebola Outbreak:
WHO and GDELT



Protectionism:
GTA and GDELT

Retail sales:
INE and BBVA



How to exploit the potential of Big Data?

New framework in the digital era…

- New availability of data
- Better and faster infrastructure
- New answers to old questions
- Combination of historical data with real time data
- Advanced data science techniques and algorithms
- Higher computational abilities to face more data granularity

…which needs the development of new competences to take advantage of it

- Economic and business knowledge to guide the question.
- Developing the data management and programming capabilities to work with large-scale datasets.
- Deepening the statistical and econometric skills to analyze and deal with high-dimensional data.
- Interpreting the results: summarize, describe and analyze the information.

New data may end up changing the way in which economists approach empirical questions and the tools they use to answer them.

We use Big Data at BBVA Research to provide a better, ”Real Time” and “High Definition” economic analysis

Some examples of our products

Economic indicators in **Real Time**

Nowcasting:

1. Activity using bank’s data (Retail Sales Index).
2. Unemployment using Google data.

Economic Analysis in **High Definition**

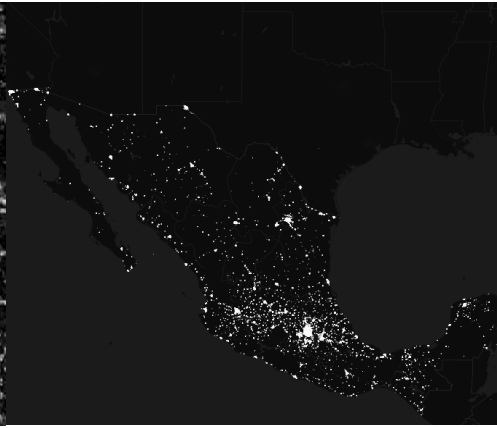
Real time analysis with high granular data to analyse sentiment towards corporates using the media information.

Social & Economic **Networks**

Using NLP to understand monetary policy narrative for European Central Bank, Federal Reserve and Central Bank of Turkey.

Economic Indicators in **Real Time**

Internal databases: working with aggregated and anonymized BBVA Data



900M card transactions from 1.2M PoS, made by 60M people, representing €37.000M

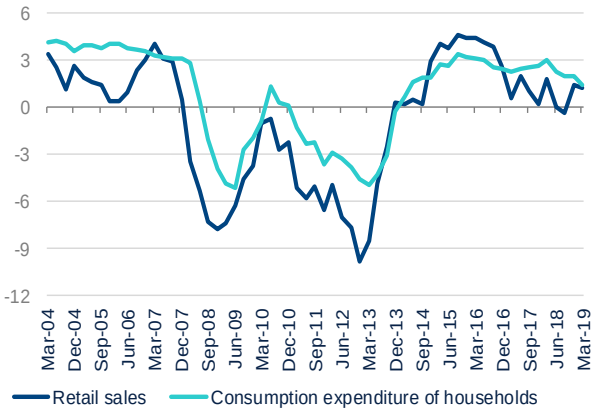
300M card transactions made by 14M people

1.500M card transactions from 1,1M PoS, made by 88M people, representing €41.000M

Economic Indicators in **Real Time**

Retail trade sector dynamic leads the evolution of consumption, which represents a high share of the GDP. The case of Spain

SPAIN: RETAIL SALES VS. CONSUMPTION EXPENDITURE OF HOUSEHOLDS
(%, YOY)



RTI has traditionally been measured by National Statistics Institutes using surveys conducted with a limited sample of retailers

We propose an alternative method for measuring the business evolution of the retail trade sector based on data from credit and debit card transactions

Source: BBVA Research and BBVA Data & Analytics
Bodas et al. (2018). Measuring retail trade through card's transaction data.
Further information here

Having accurate estimates of the retail trade evolution is of great importance given that this is a key indicator of the economic situation and its dynamic drives the evolution of aggregate consumption

Economic Indicators in **Real Time**

We replicate INE data treatment and methodology using transactional data

Methodology

Internal taxonomy - Spain (BBVA)

Category	Subcategory	Ramo / Giro	CIF / RFC	FUC / Afiliación	POS ID
(Fashion)	(Fashion-big)	(Textiles and clothing)	(Cadena Zara)	(Zara, Gran Vía, Madrid)	(TPV)

External taxonomy - Spain (INE)

5 distribution classes:

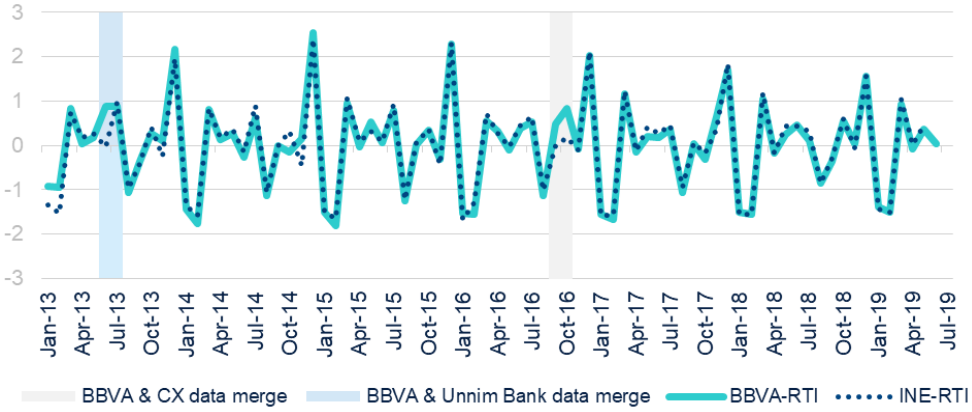
- 01 service stations
- 02 single retail stores (one premise)
- 03 small chain stores (2-24 premises & <50 employees)
- 04 large chain stores (25 or more premises, and 50 or more employees)
- 05 department stores (sales area greater than or equal to 2.500m^2)

Comparison between RTI Data Sources	Card Transaction Data (BBVA)	Survey Data (INE)
Cost per observation	Marginally Low	High
Data Frequency	Timestamp HH:MM/DD/MM/AAAA	Monthly
Disaggregation by activity	High: 17 categories and 73 subcategories	Low
Geographical disaggregation	High (lat, long)	Low
Real-time availability	3 days delay on ETL	No
Retailer sample	1,2 million	≈ 12,500
Payment methods covered	BBVA's clients credit and debit cards	All
Possible bias of technological trends	Yes	No

Economic Indicators in **Real Time**

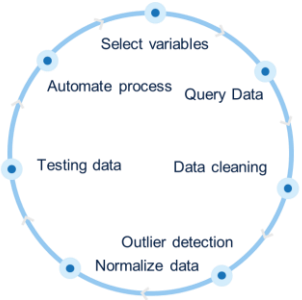
High correlation between retail sales index and BBVA data at national, regional and distribution classes levels

RETAIL TRADE INDICES: BBVA VS INE
(STANDARDIZED MONTHLY GROWTH RATE)



Retail sales by distribution class

Retail sales by AA. CC.



Source: BBVA Research and BBVA Data & Analytics
Bodas et al. (2018). Measuring retail trade through card's transaction data.
Further information here

High granularity: Dynamics down to subnational level

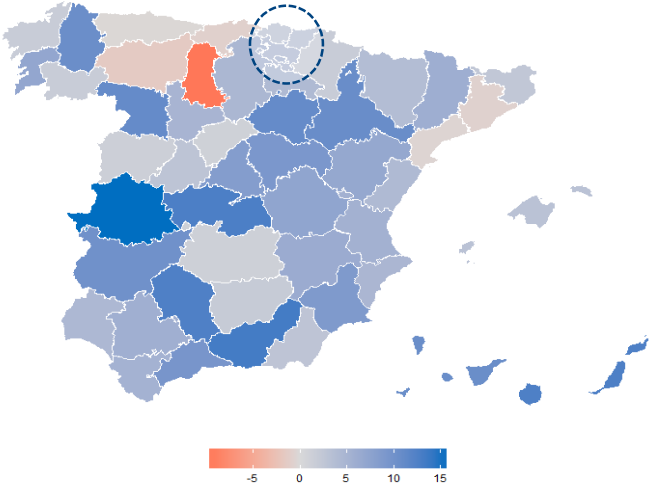
Multi Dimensional: More detailed socioeconomic features

Ultra High Frequency: Dynamics up to sub-monthly frequency

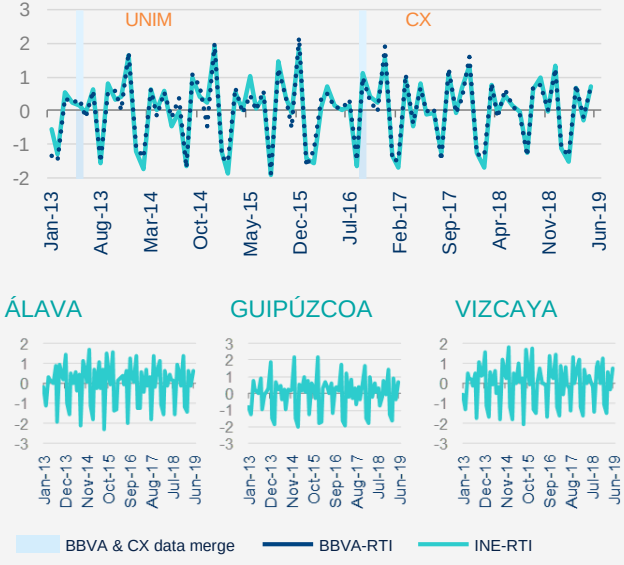
Economic Indicators in **Real Time**

The granularity of the information can be really valuable for the analysis: regional level

BBVA RETAIL SALES INDEX GROWTH in 1H19
(% YOY)



BASQUE COUNTRY
(STANDARDIZED % MOM)

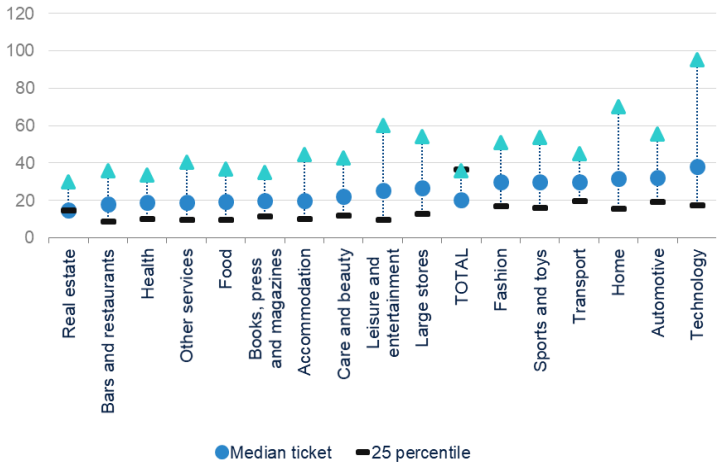


Source: BBVA Research and BBVA Data & Analytics
Bodas et al. (2018). Measuring retail trade through card's transaction data.
Further information here

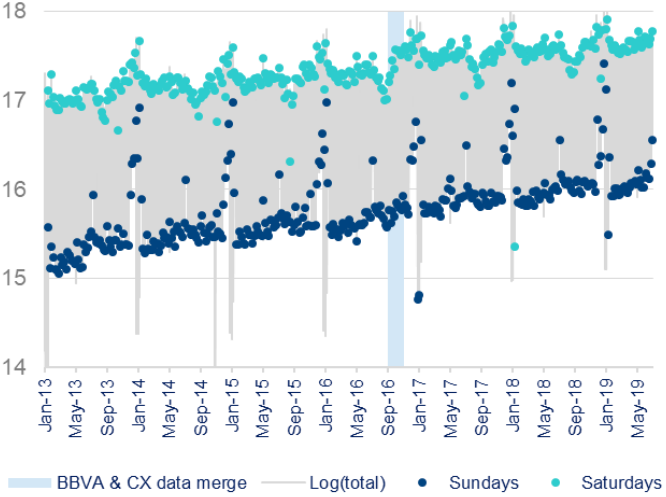
Economic Indicators in **Real Time**

The granularity of the information can be really valuable for the analysis: sector of activity and daily data

BBVA RETAIL SALES INDEX BY MERCHANT
(MEDIAN TICKET IN JUN-19, €)



BBVA RETAIL SALES INDEX – DAILY FRECUENCY (LOGARITHMS)



Source: BBVA Research and BBVA Data & Analytics
Bodas et al. (2018). Measuring retail trade through card's transaction data.
Further information here

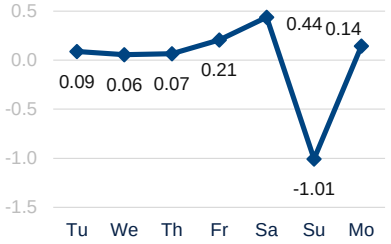
Economic Indicators in **Real Time**

The need of Analysts: Dealing with seasonalities to analyze daily data

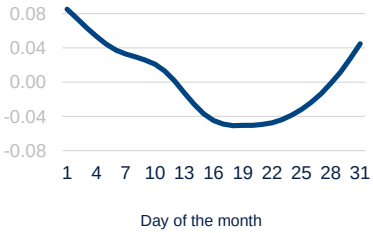
PERIODIC EFFECTS (SEASONALITIES)

$$\log(y_t) = \mu_t + \gamma_t^w + \gamma_t^m + \gamma_t^y + \gamma_t^h + \varepsilon_t$$

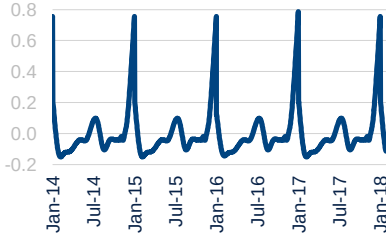
INTRA-WEEKLY SEASONALITY (γ_t^w)



INTRA-MONTHLY SEASONALITY (γ_t^m)



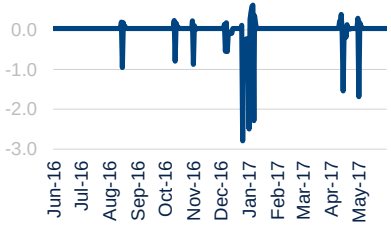
INTRA-ANNUAL SEASONALITY (γ_t^y)



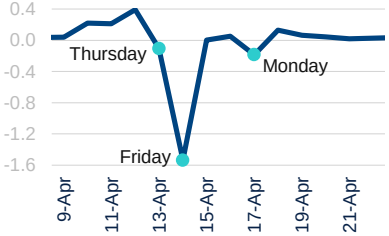
FIXED AND MOVING HOLIDAYS

$$\log(y_t) = \mu_t + \gamma_t^w + \gamma_t^m + \gamma_t^y + \gamma_t^h + \varepsilon_t$$

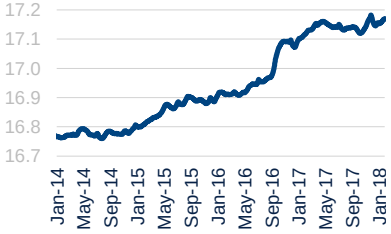
BBVA RTI: HOLIDAY'S EFFECTS (γ_t^h)



BBVA RTI: EASTER 2016



BBVA RTI: TREND (μ_t)

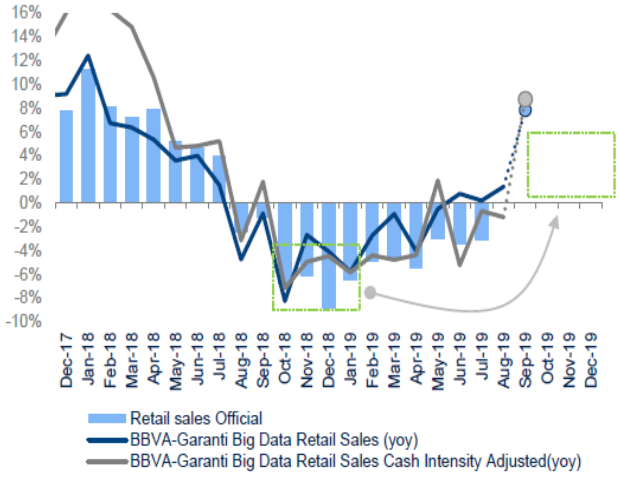


Bodas et al. (2018). Measuring retail trade through card's transaction data.
Source: BBVA Research and BBVA Data & Analytics

Economic Indicators in **Real Time**

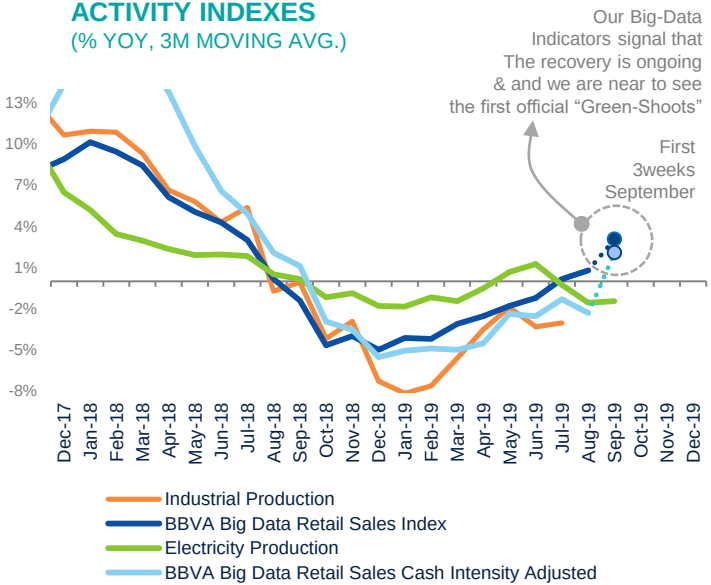
Our real time indicators give us some advantages to track the business cycle. The case of Turkey

BBVA-GB BIG DATATRANSACTIONS VS OFFICIAL DATA
(REAL TERMS, YOY)



Source: BBVA Research

TURKEY: HARD & BIG DATA ACTIVITY INDEXES
(% YOY, 3M MOVING AVG.)



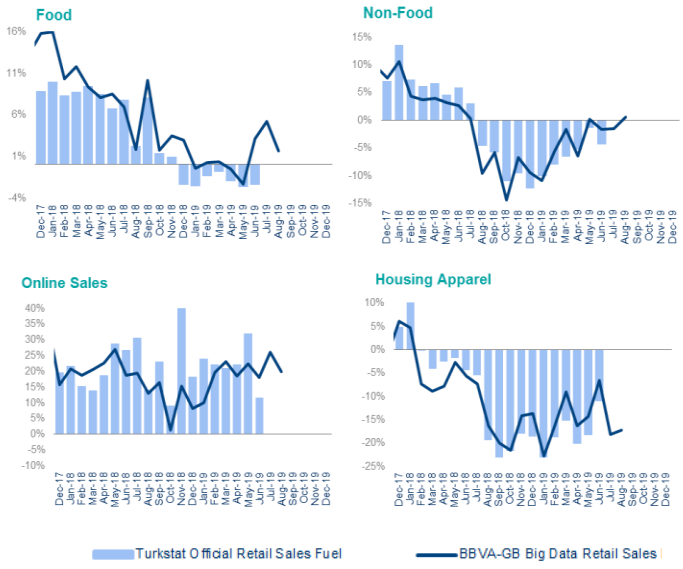
Source: CBRT, TURKSTAT, BBVA Research Turkey

Our Big Data indicators using bank's transactions data signal that the recovery is gaining momentum (consistent with soft data).

Economic Indicators in **Real Time**

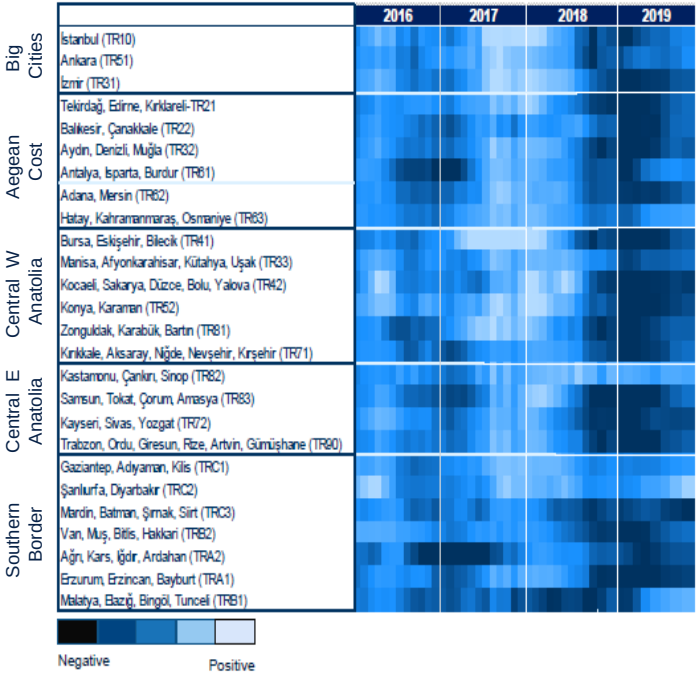
BigData allows us to define high definition indexes and build new statistics as provincial data

RETAIL TRADE INDICES: BBVA-GB BIG DATA



Source: BBVA Research

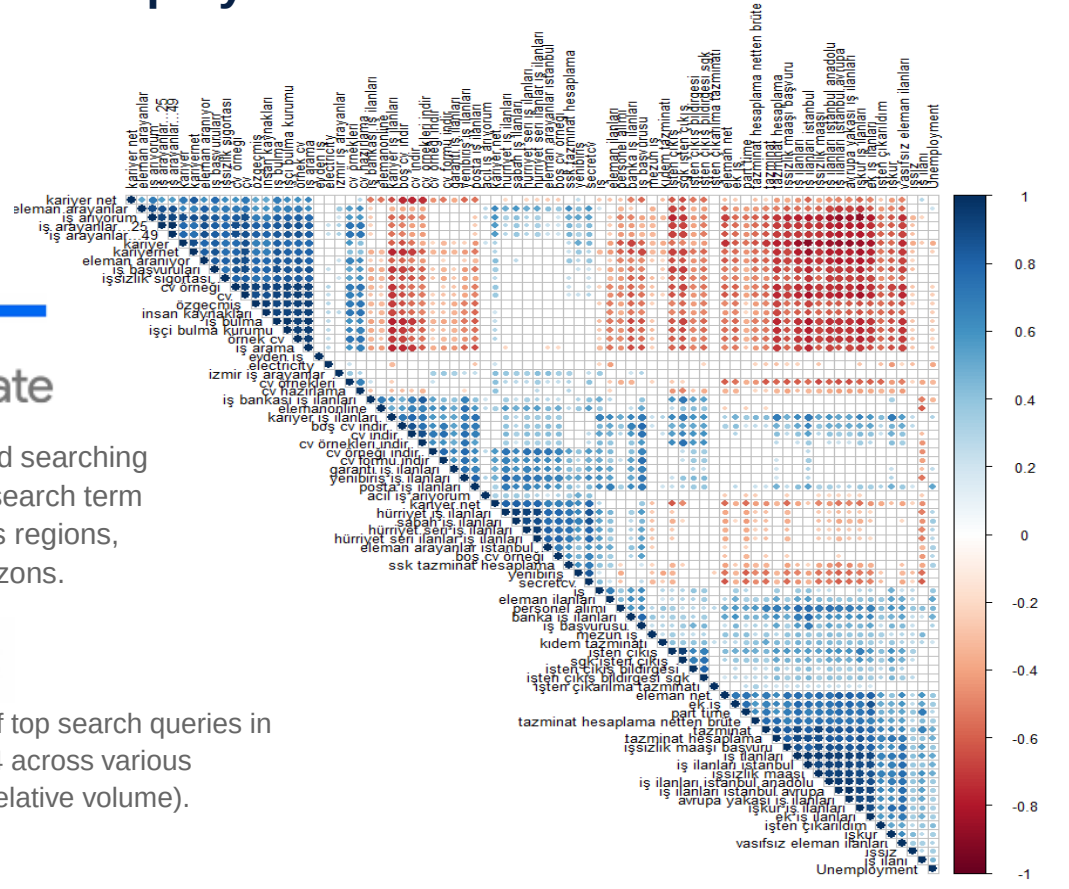
TURKEY: RETAIL SALES REGIONAL HEAT MAP



Source: CBRT, TURKSTAT, BBVA Research Turkey

Economic Indicators in **Real Time**

Google Correlate &Trends can help us to find terms related to employment searches



Google Correlate

Finds the most correlated searching topics (100) for a given search term in interest across various regions, languages and time horizons.

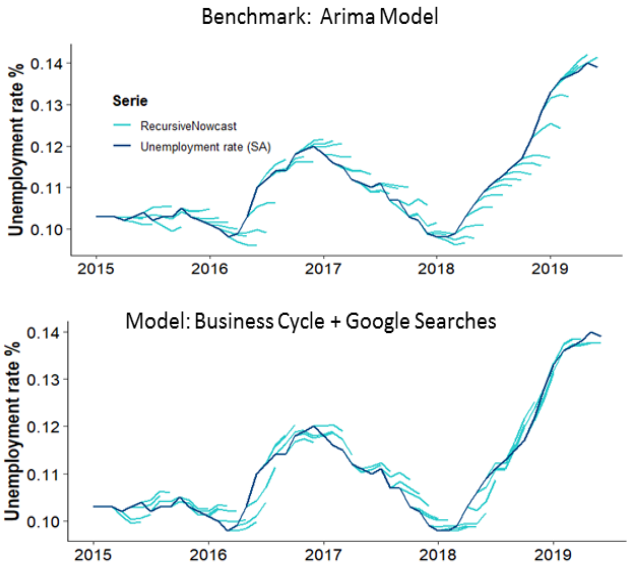
Google Trends

Analyzes the popularity of top search queries in Google Search from 2004 across various regions and languages (relative volume).

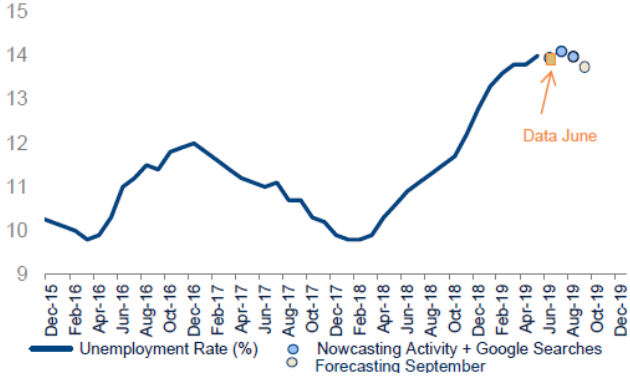
Economic Indicators in **Real Time**

Google searches provide extra information to nowcast unemployment with an advantage of 3 months

UNEMPLOYMENT OUT-OF-SAMPLE FORECASTS
(3M RECURSIVE OUT OF SAMPLE FORECASTS)



TURKEY: UNEMPLOYMENT RATE (SA) NOWCAST



TURKEY: UNEMPLOYMENT CHANGES

Source: BBVA Research

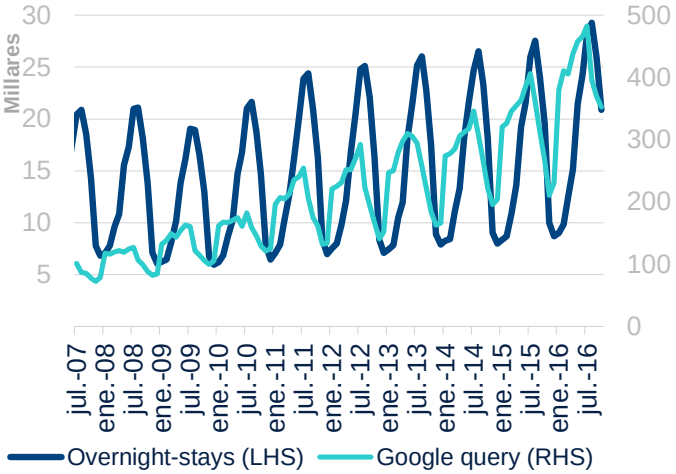
Here you can find a similar project to nowcast Spanish tourism

Economic Indicators in **Real Time**

We also use Google queries three years ago to nowcast Spanish tourism

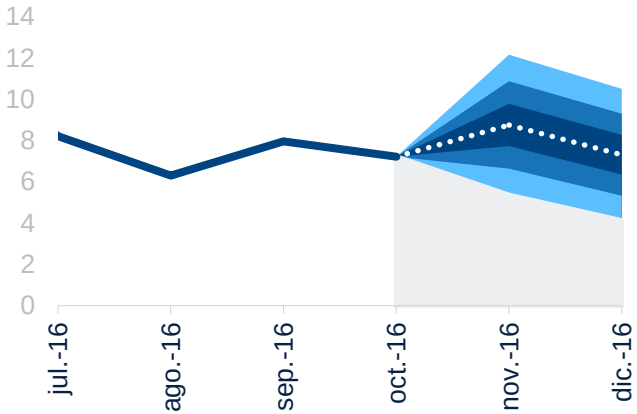
OVERNIGHTS OF NON-RESIDENT TOURISTS IN HOTELS AND SEARCH TRENDS IN GOOGLE

(OVERNIGHT STAYS IN THOUSANDS, SEARCHES INDEX = 100, JULY 2007)



OVERNIGHTS OF NON-RESIDENT IN HOTELS

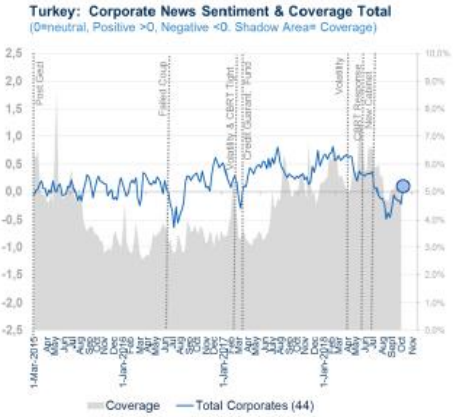
(% YOY, LATEST FORECAST AS OF NOVEMBER 30, 2016)

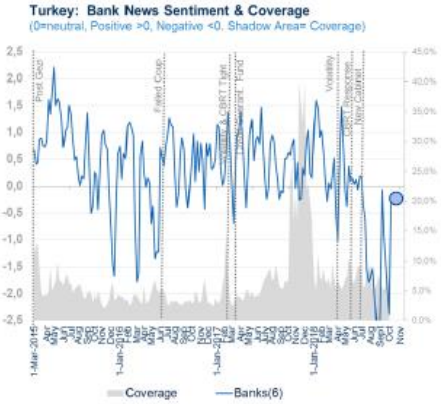


(More information can be found in the following link).
Source: BBVA Research, INE and Google

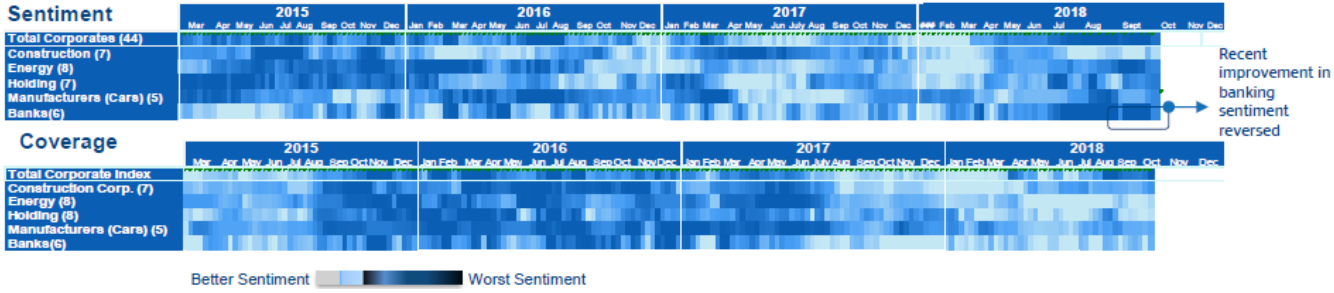
Economic Analysis In **High Definition**

Corporate news Sentiment give us "Early Warning Signals" of corporate balance sheet health…





Turkey: BBVA Research Big Data Sentiment and Coverage on Corporates

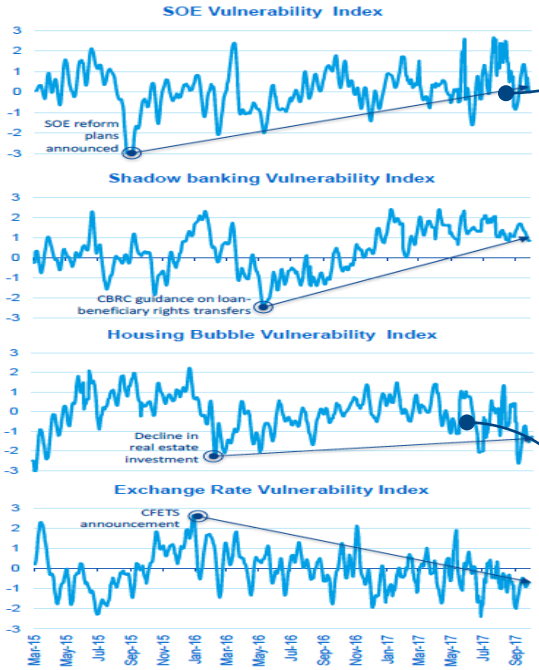


Source: BBVA Research

Economic Analysis In **High Definition**

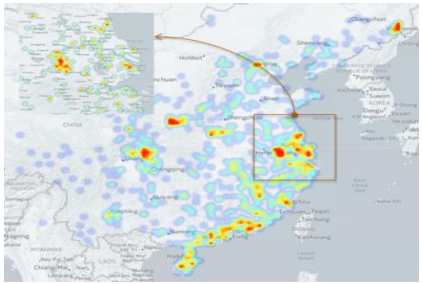
We also developed hybrid Indicators (Hard Data & Sentiment Data) to disentangle risks in China

CHINESE VULNERABILITY SENTIMENT INDEX
(CVSI)



CHINA SOE MAP
(SENTIMENT ON SOE)

GEOGRAPHICAL ANALYSIS HOUSING PRICES
(SENTIMENT ON HOUSING PRICES IN 2018)



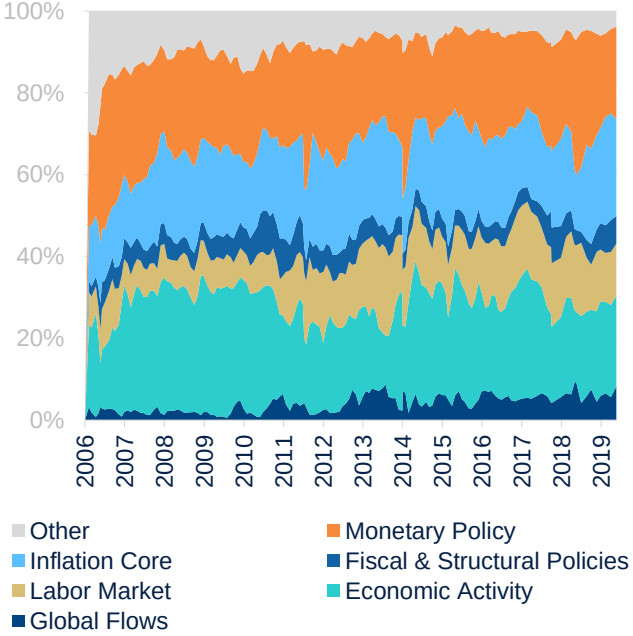
Source: BBVA Research. Further information could be found here.

Social & Economic Networks

NLP to analyze monterary policy narrative of Central Banks, identifying what they are talking about…

IDENTIFIED TOPICS IN CENTRAL BANKS REPORTS

CENTRAL BANK OF TURKEY: EVOLUTION OF TOPICS



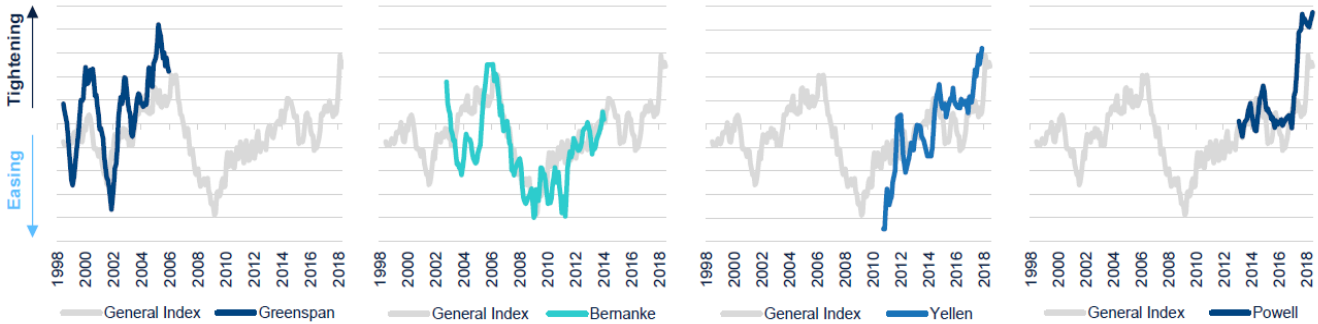
Source: BBVA Research. Further information can be found here.

Social & Economic **Networks**

… and how they are talking, even focusing in personal tone according to particular speeches …

GENERAL AND FED GOVERNOR HAWKISH/DOVISH INDEX BY SPEAKER OVER TIME

(TONE. 12 MONTHS MOVING AVERAGE TONE)



From a EM Crisis Reactive and tigtening… (Mr Greenspan) 1987-2003

To a Governor Managing the crisis… (Mr Bernanke) 2016-2014

To a Lady managing the Exit Strategy… (Mrs Yellen) 2014-2018

To a Normalization Policy (Mr Powell) 2018-

Source: BBVA Research. Further information can be found here.

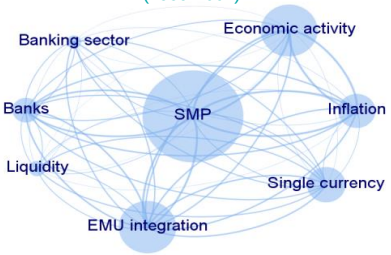
Social & Economic **Networks**

…Or understand the inter-connexions between topics and Central Banks

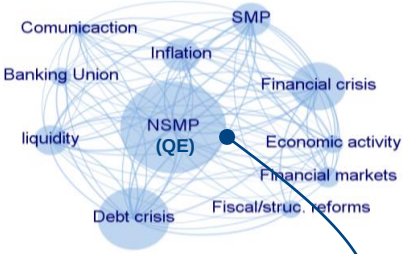
MONETARY POLICY IN DEVELOPED ECONOMIES AND RESPONSE IN THE EMERGING MARKETS
(NETWORKS)

ECB

PRE LEHMAN
(1999-2007)



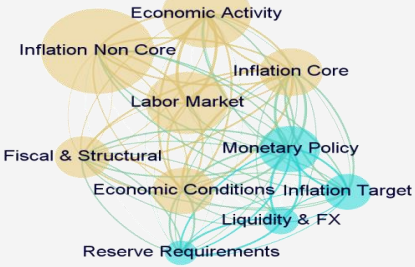
CRISIS
(2007 -2013)



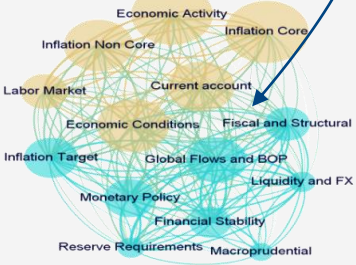
QE & RECALIBRATION
(2013-2018)

CBRT

INFLATION TARGET
(2006-2008)



FINANCIAL CRISIS
(2009-2012)



THE GLOBAL POST-CRISIS
(2012-2017)

BBVA Research

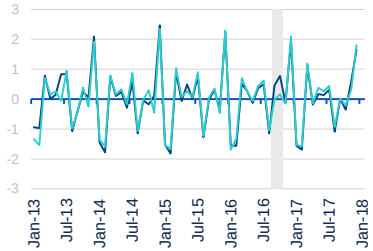
ANNEX

Creating Opportunities

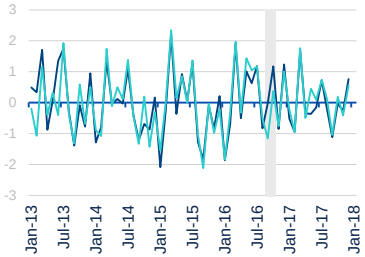
Spain: Macroeconomic consistency of BBVA data by distribution class

Return

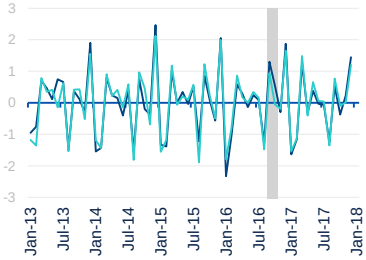
Spain



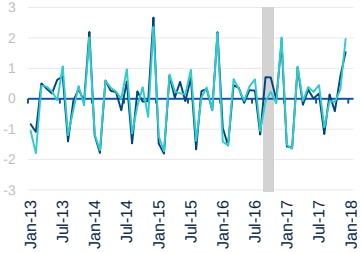
Gas Station



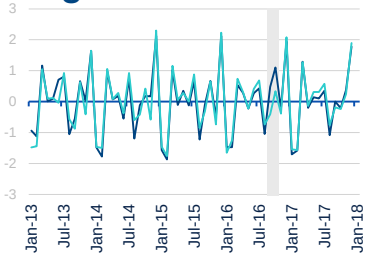
Single Retail Store



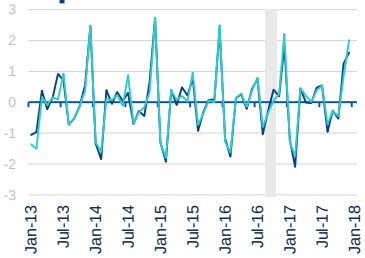
Small Chain Store



Large Chain Store



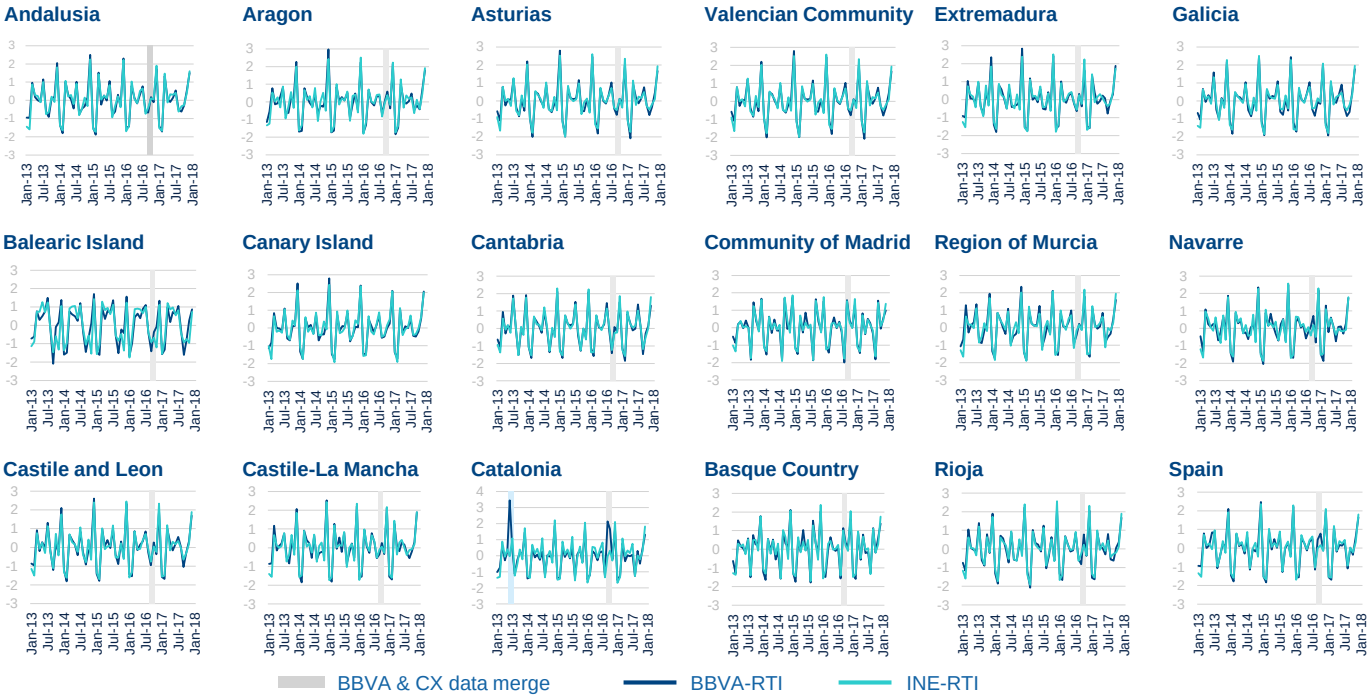
Department Store



BBVA & CX data merge BBVA-RTI INE-RTI

Spain: Macroeconomic consistency of BBVA data by AA.CC

Return



External sources: the case of Spain

- The Retail Trade Index is a business cycle indicator which shows the monthly activity of the retail sector (turnover)
- Population scope: stores whose main activity is registered in Division 47 of the NACE-2009, which includes the following groups:
  - Retail sale in non-specialized establishments (supermarkets, deparment stores, etc.)
  - Retail sale in specialized establishments (food, beverages and tobacco; fuel; IT equipment and communications; personal goods, such as fabric, clothing and footwear; household items, such as textiles, hardware, electrical appliances and furniture; cultural and recreational items, such as books, newspapers and software; pharmaceutical products; etc.)
  - Retail trade not carried out in establishments (eCommerce, home delivery, vending machines, etc.)
- Sale of motor vehicles, Foodservice, hospitality industry, financial services, etc., are not included in RTS!
- Sample: 12,500 stores (Random stratified sampling <50 employees + exhaustive>=50)
- Dissemination: AA. CC. OR 5 distribution classes:
  - service stations,
  - single retail stores (one premises),
  - small chain stores (2-24 premises & <50 employees),
  - large chain stores (25 or more premises, and 50 or more employees)
  - department stores (sales area greater than or equal to 2500 m2)

BBVA transactions at daily frequencies

Daily data dynamic modeling is not common in the economic literature. Many sources of variability need to be accounted for:

- Day-of-week effect
- Day-of-month effect
- Day-of-year effect
- Fixed and moving holidays' effect
- Long-lasting effects (Christmas)

We base on Harvey et al (1997) structural time series modeling

$$\log(y_t) = \underbrace{\mu_t}_{\text{Stochastic Trend}} + \underbrace{\gamma_t^w + \gamma_t^m + \gamma_t^y}_{\text{Seasonalities}} + \underbrace{\gamma_t^h}_{\text{Holidays}} + \varepsilon_t$$

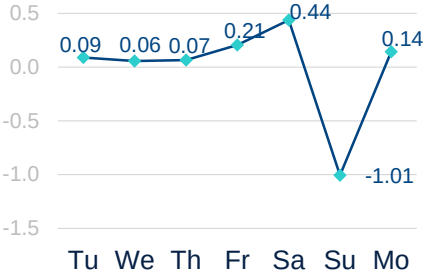
BBVA transactions at daily frequencies: Periodic effects (seasonalities)

$$\log(y_t) = \mu_t + \gamma_t^w + \gamma_t^m + \gamma_t^y + \gamma_t^h + \varepsilon_t$$

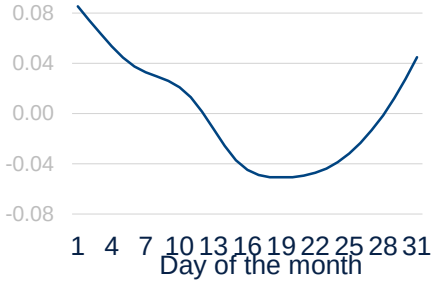
- The day of the week effect is modeled using stochastic dummies $\gamma_t^w = \sum_{j=1}^{s-1} \gamma_{t-j}^w + \omega_t$.
- The intra-monthly and intra-year seasonality is captured using "splines"

Encouraging results: Seasonalities are as expected, but the data is proving it

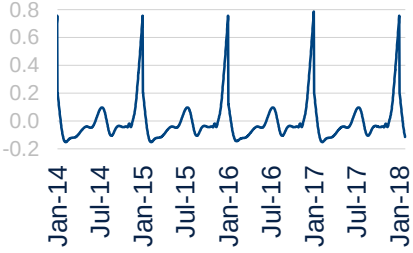
Intra-weekly seasonality (γ_t^w)
(logarithms)



Intra-monthly seasonality (γ_t^m)
(logarithms)



Intra-annual seasonality (γ_t^y)
(logarithms)



Source: BBVA

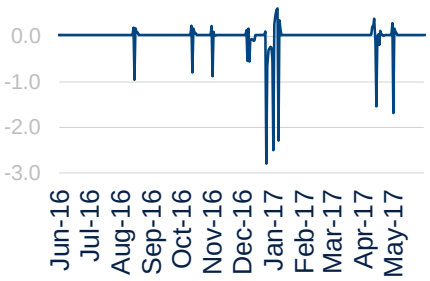
BBVA transactions at daily frequencies: Fixed and moving holidays

$$\log(y_t) = \mu_t + \gamma_t^w + \gamma_t^m + \gamma_t^y + \gamma_t^h + \varepsilon_t$$

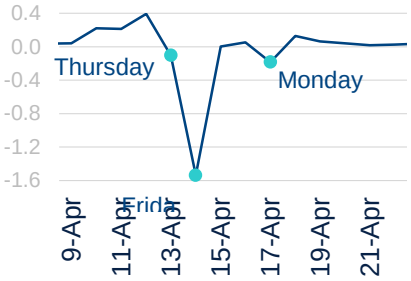
- Holiday's are modeled using deterministic seasonal dummies (sum zero over the year)
- The trend is stochastic: $\mu\_(t+1)=\nu\_(t+1)+\mu\_t+\xi\_t$ where $\nu\_(t+1)=\nu\_t+\zeta\_t$

Encouraging results: We could analyze the period surrounding each holiday

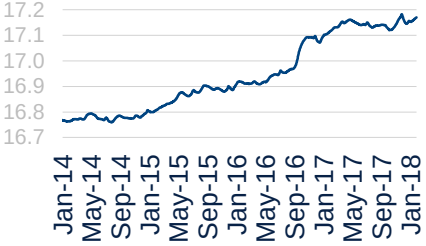
BBVA RTI: Holiday's effects (γ_t^h)
(logarithms)



BBVA RTI: Easter 2016
(logarithms)



BBVA RTI: Trend (μ_t)
(logarithms)



Source: BBVA

Daily model

Stochastic Trend Seasonalities Holidays

$$\log(y_t) = \mu_t + \gamma_t^w + \gamma_t^m + \gamma_t^y + \gamma_t^h + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$$

$$\mu_t = \mu_{t+1} + \nu_t + \xi_t \qquad \xi_t \sim N(0, \sigma_\xi^2)$$

$$\nu_t = \nu_{t+1} + \zeta_t \qquad \zeta_t \sim N(0, \sigma_\zeta^2)$$

Intra-weekly effect (γ_t^w):

There are various alternatives to model the day of the week effect (we try three alternatives). We finally use the following one:

$$\gamma_t^w = \sum_{j=1}^{s-1} \gamma_{t-j}^w + \omega_t \qquad \omega_t \sim N(0, \sigma_\omega^2)$$

Holidays effect (γ_t^h):

We base on a deterministic approach. We include dummy variables for the holiday specific day and some days previous and after the holiday (pending to check which is the best number of days surrounding each holiday).

$$\gamma_t^{h,i} = w_i(B) h(\tau_i, t)$$

where $w_i(B)$ is a polynomial lag operator and $h(\tau_i, t)$ is an indicator function that takes the value 1 when $t = \tau_i$ and zero otherwise. In our model, seasonality is also takes into account regarding holidays by making the sum of the days of the year to be equal zero (the dummy variables are altered to get this kind of effect).

Daily model

Return

Stochastic Trend Seasonalities Holidays

$$\log(y_t) = \mu_t + \gamma_t^w + \gamma_t^m + \gamma_t^y + \gamma_t^h + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$$

$$\mu_t = \mu_{t+1} + \nu_t + \xi_t \qquad \xi_t \sim N(0, \sigma_\xi^2)$$

$$\nu_t = \nu_{t+1} + \zeta_t \qquad \zeta_t \sim N(0, \sigma_\zeta^2)$$

Intra-month and intra-year effect (γ_t^m and γ_t^y):
Two possible alternatives, trigonometric or "spline" approaches. We try both of them with the same qualitative results. The one showed here is the "spline" type of modeling.

Splines: choose h knots in the range $[0, N]$, where N is the number of the days in a month or in a year. Then:

$$\gamma_d = \boldsymbol{w}_d' \gamma^\dagger \qquad d = 1, \dots, N$$

where $\boldsymbol{w}_d'$ is a $h \times 1$ vector that depends on the knots and it is also define to guarantee continuity from period to period

To guarantee seasonality define $\boldsymbol{z}_d'$ (replacing $\boldsymbol{w}_d'$) where each element "i" of $\boldsymbol{z}_d'$ is equal to:

$$z_{di} = w_{di} - w_{dh} w_{*i} / w_{*h} \qquad d = 1, \dots, N \quad ; \quad i = 1, \dots, g \quad ; \quad \boldsymbol{w}_* = \sum_{d=1}^N \boldsymbol{w}_d$$

To allow the splines to evolve over time:

$$\gamma_t^\dagger = \gamma_{t-1}^\dagger + \chi_t \qquad t = 1, \dots, T_d$$

where T_d is the total number of observations

$$\mathrm{var}(\chi_t) = \sigma_\chi^2 I$$

BBVA Research

The use of Big Data for the statistical production. The experience of BBVA Research

European Statistics Day

Spanish National Statistics Institute

October 2019

Creating Opportunities