BBVA Research

Measuring Retail Trade Using Card Transactional Data

May 2018

Creando Oportunidades

Main takeaways

- We present a high-dimensionality Retail Trade Index (RTI) constructed to nowcast the retail trade sector economic performance in Spain
- Results are robust when compared with the Spanish RTI, regional RTI (Spain’s autonomous regions), and RTI by retailer type (distribution classes) published by the INE
- We got monthly indexes for the provinces and sectors of activity and the daily general index, by obtaining timely, detailed information on retail sales
- We analyzed the high-frequency consumption dynamics using BBVA retailer behavior and a structural time series model

Research motivation

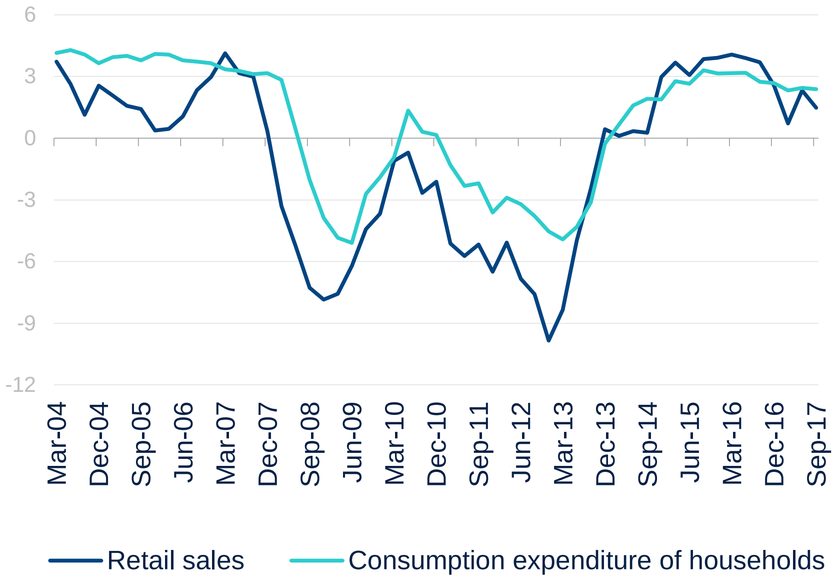
?

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

Spain: Retail Sales vs. Household Consumption Expenditure

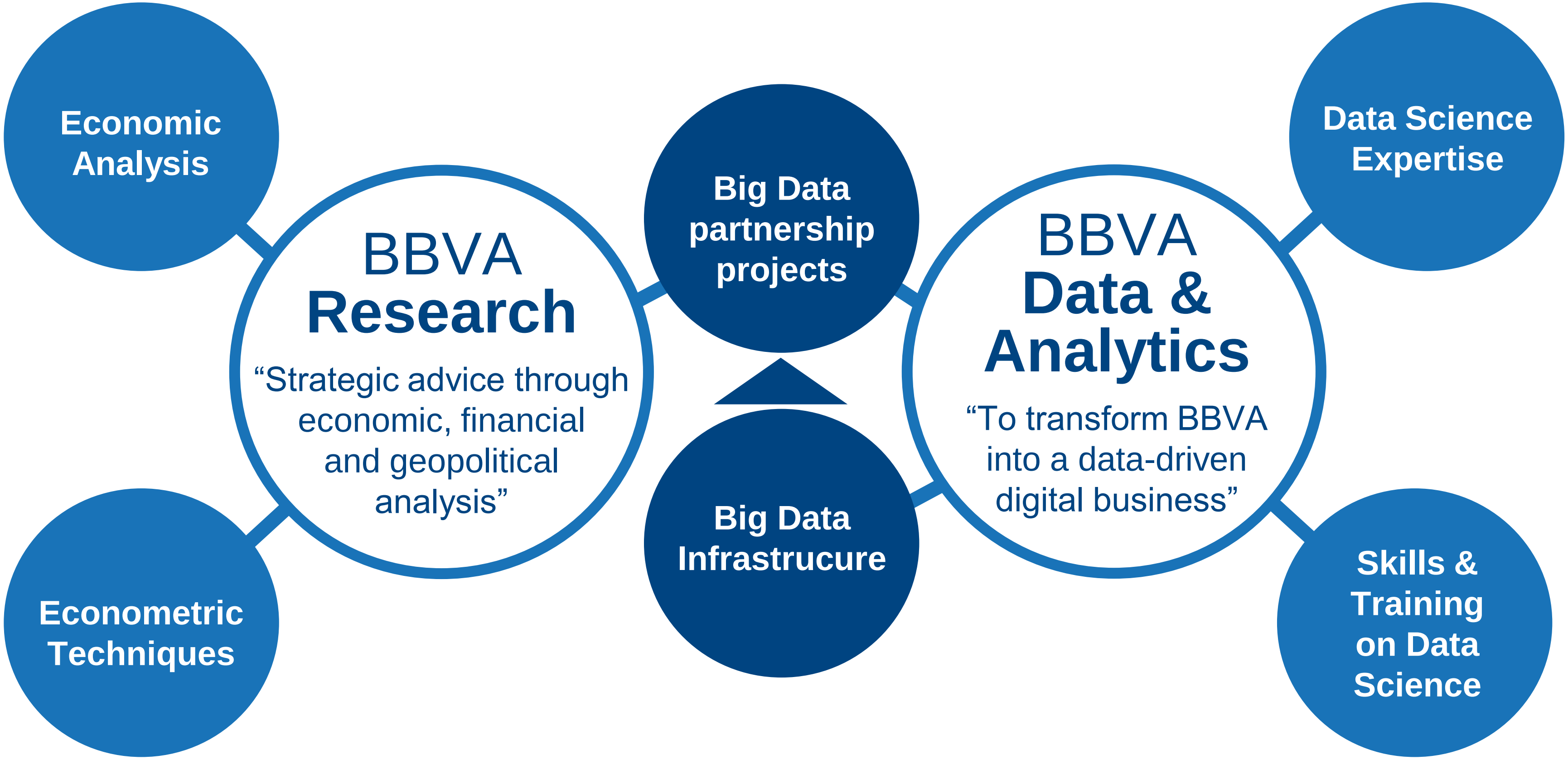
(%, YoY)



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

Collaboration BBVA Research and D&A

Different but complementary abilities, strengths and aims...



Economic Analysis

Econometric Techniques

BBVA **Research**

"Strategic advice through economic, financial and geopolitical analysis"

Big Data partnership projects

Big Data Infrastrucure

BBVA **Data & Analytics**

"To transform BBVA into a data-driven digital business"

Data Science Expertise

Skills & Training on Data Science

Collaboration BBVA Research and D&A

... shared spirit to achieve innovative results

Develop a **high resolution economic activity indicator** (analyzing high frequency and high granularity internal data sources) based on BBVA transactional information of electronic payments

BBVA Research

Analysis: explore the macroeconomic consistency of internal high resolution data

Techniques: enrich classical modeling techniques acquiring new methodological capabilities in the field of Data Science

BBVA Data & Analytics

Data Science: foster the bank's digital transformation upon the value enclosed in massive and dynamic data sources and new analytic methodologies. Promote the acquisition of analytic capabilities and model-oriented programming among BBVA BUs teams working in integrated teams.

Infrastructure, data sources and Big Data Partnership Projects: BBVA D&A triggers the process of BBVA's data democratization throughout BBVA

Outline

01 Data Sources & Research Methodology

02 The Spanish Retail Trade Index

03 Daily Model Development & Results

04 Conclusions

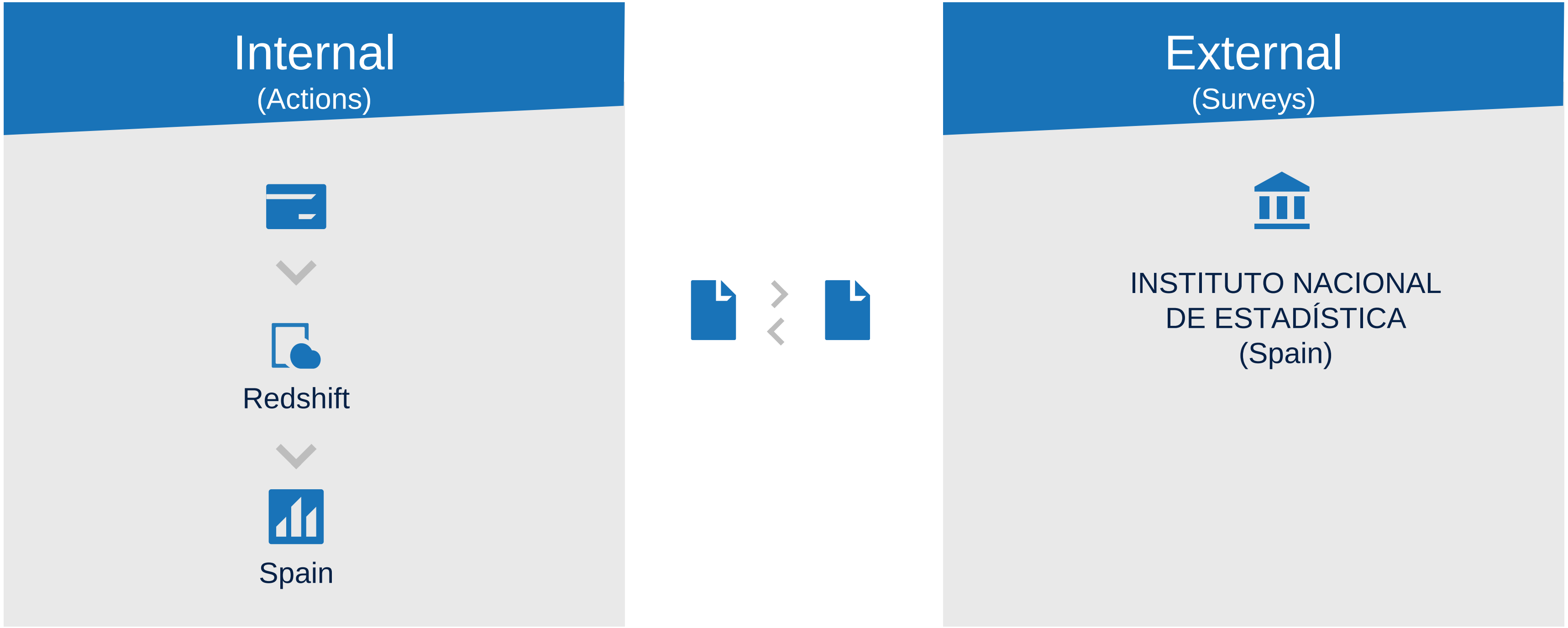
05 Annex

01

Data Sources & Research Methodology

Data Sources

Replicating INE data treatment and methodology using transactional data



Internal
(Actions)

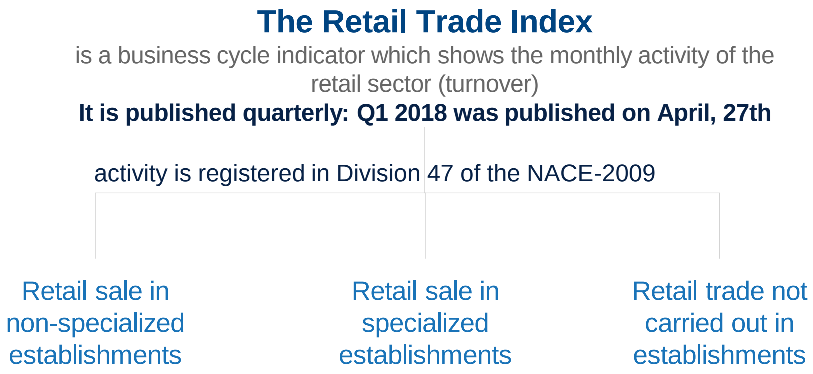
Redshift

Spain

External
(Surveys)

INSTITUTO NACIONAL
DE ESTADÍSTICA
(Spain)

External sources: Spanish National Statistics Institute



The Retail Trade Index

is a business cycle indicator which shows the monthly activity of the retail sector (turnover)

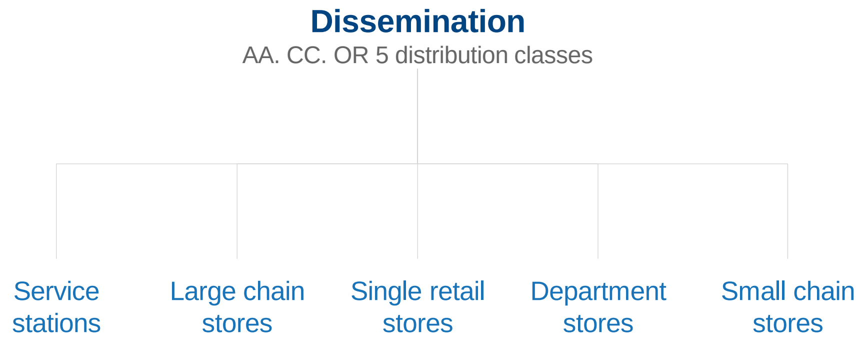
It is published quarterly: Q1 2018 was published on April, 27th

activity is registered in Division 47 of the NACE-2009

Retail sale in non-specialized establishments

Retail sale in specialized establishments

Retail trade not carried out in establishments



Dissemination

AA. CC. OR 5 distribution classes

Service stations

Large chain stores

Single retail stores

Department stores

Small chain stores

It does not include:

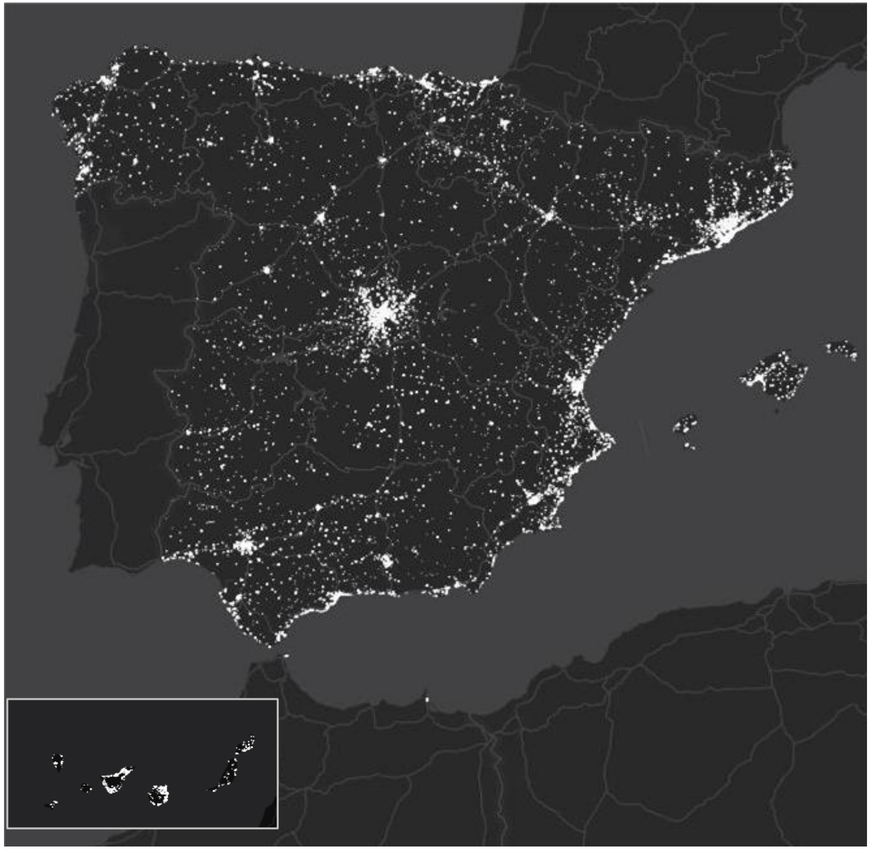
Sale of motor vehicles, Foodservice, hospitality industry, financial services, etc.,

Sample:

12,500

stores

Internal sources: BBVA transactional data



15.3% Transactions made by BBVA cards at any PoS

21.1% BBVA PoS

$1.2 \cdot 10^6$ merchants, classified in 17 categories and 75 subcategories (from~400 “ramos” recsys)

$900 \cdot 10^6$ transactions/ year

$4.3 \cdot 10^6$ cardholders

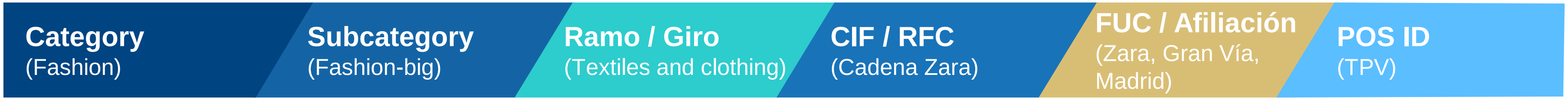
300.000 CIFs

900M card transactions at 1.2M PoS, made by 60M people, representing €37.000M.
We focus on purchases made by BBVA cardholders at any PoS between 2016/01 and 2017/12

Matching internal and external sources

Methodology

Internal taxonomy - Spain (BBVA)



Category (Fashion)

Subcategory (Fashion-big)

Ramo / Giro (Textiles and clothing)

CIF / RFC (Cadena Zara)

FUC / Afiliación (Zara, Gran Vía, Madrid)

POS ID (TPV)

External taxonomy - Spain (INE)

5 distribution classes:

01 service stations

02 single retail stores (one premise)

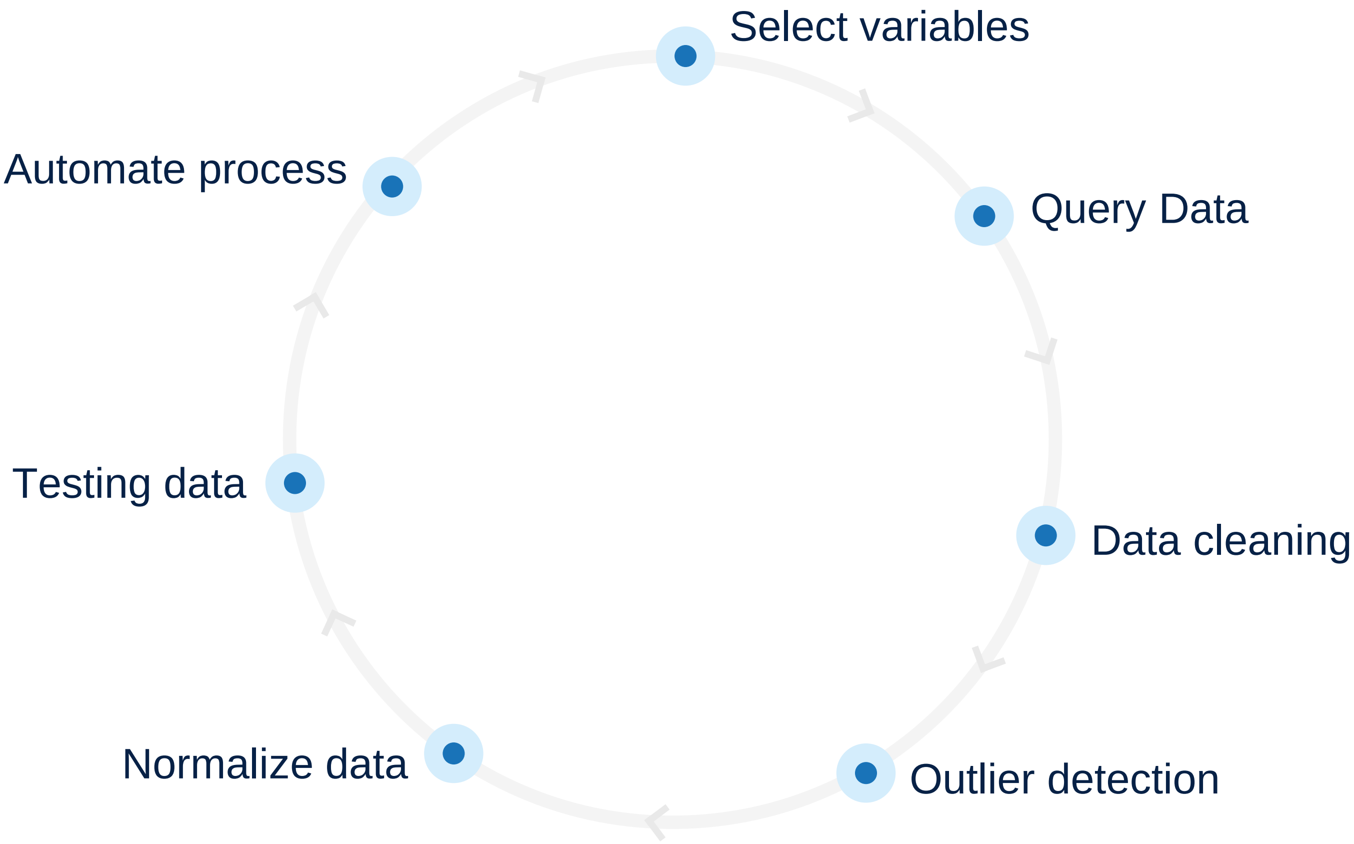
03 small chain stores (2-24 premises & <50 employees)

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

Data extraction, cleaning and transformation



Select variables

Query Data

Data cleaning

Outlier detection

Normalize data

Testing data

Automate process

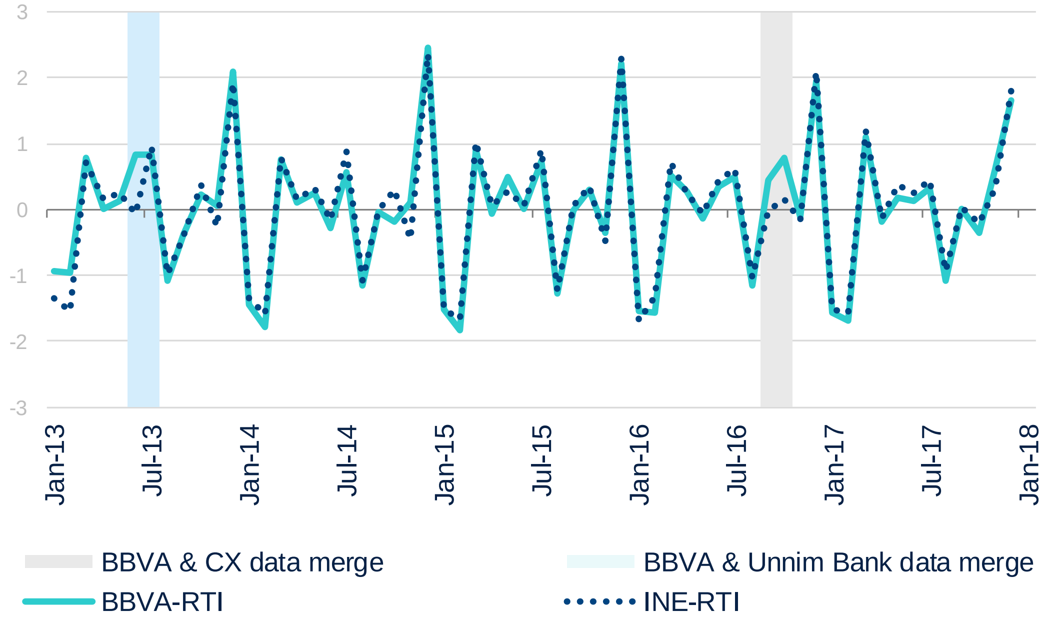
02

The Spanish Retail Trade Index

Macroeconomic consistency of BBVA data

Retail Trade Indices: BBVA vs INE

(standardized monthly growth rate)



→

Retail sales by distribution class

Retail sales by AA. CC.

High correlation between retail sales index and BBVA data (~95%)

“Nowcasting” monthly retail sales using BBVA data

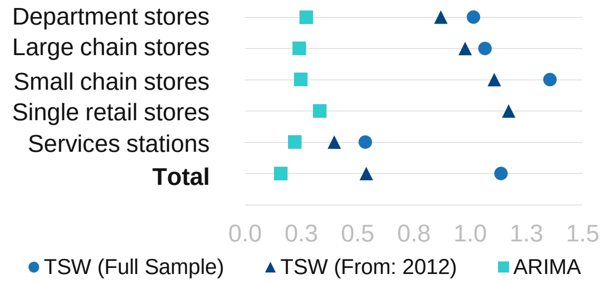
It also holds for AA.CC

$$\underbrace{N(\Delta^1 \log(y_{it}))} = \alpha + \beta \cdot \underbrace{N(\Delta^1 \log(x_{it}))} + \varepsilon_{it}$$

RS and BBVA’s transaction data MoM standardized growth rates for $t \geq 2013m1$

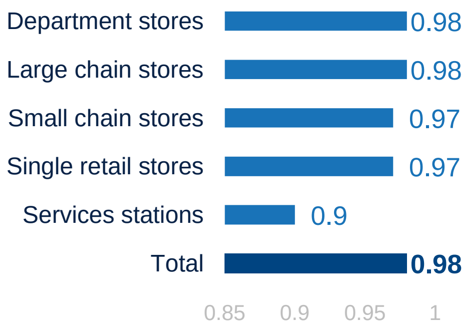
Quasi real time forecast accuracy

(relative RMSE last 12 periods)

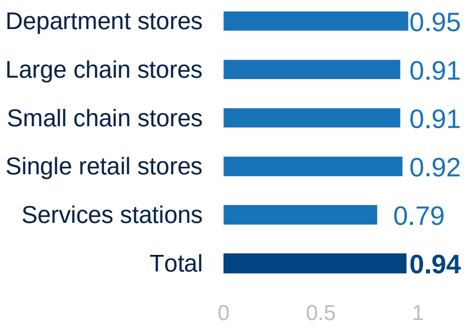


Monthly Model Regression Results

ß

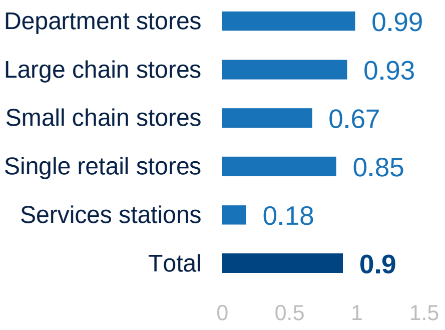


R^2



Hansen test P-Value

(H0: parameter stability)

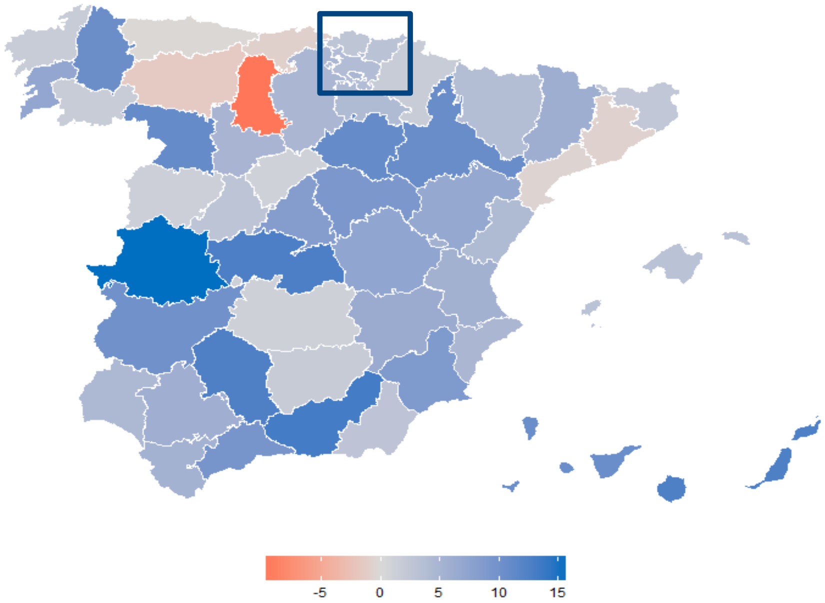


Source: BBVA

Data by provinces

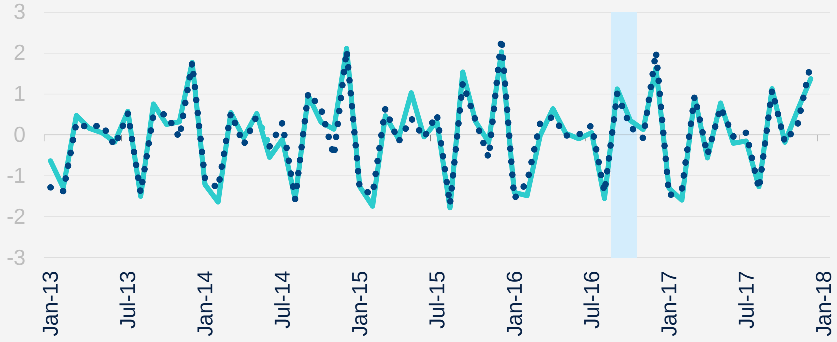
BBVA RTI growth in Dec-17

(% yoy)

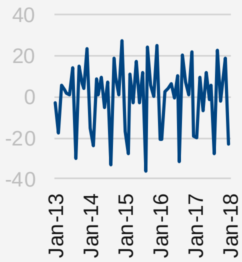


Source: BBVA

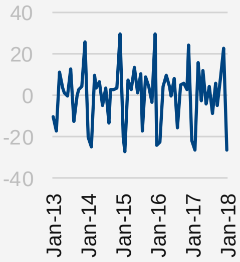
Basque Country (% mom)



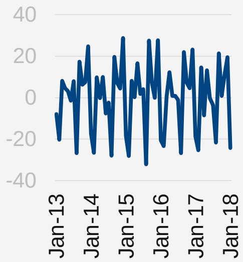
Álava



Guipúzcoa



Vizcaya

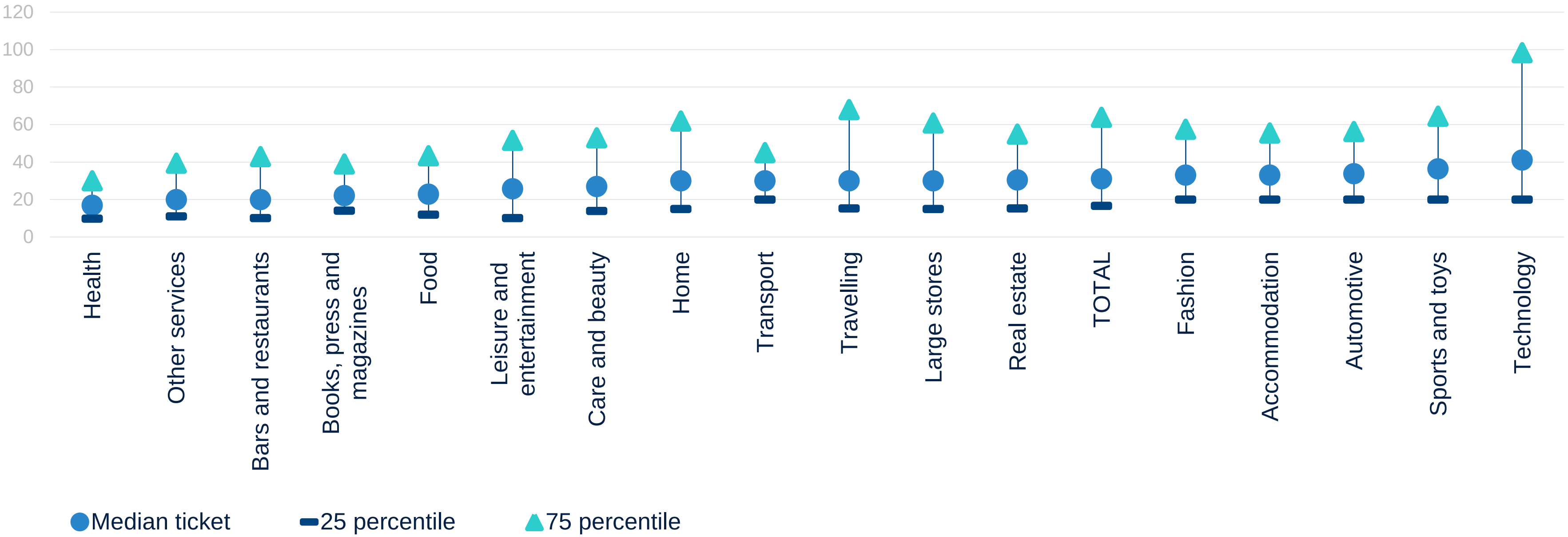


▭ BBVA & CX data merge — BBVA-RTI — INE-RTI

Data by merchant

BBVA RTI by merchant

(median ticket in Dec-17, €)



Data granularity allows us to exploit new dimensions that the INE-RTI does not provide, both on the supply side (e.g., sector of activity) and the demand side (e.g., clients' socioeconomic features)

03

Daily Model Development & Results

BBVA transactions at daily frequencies

See the annex for further model details

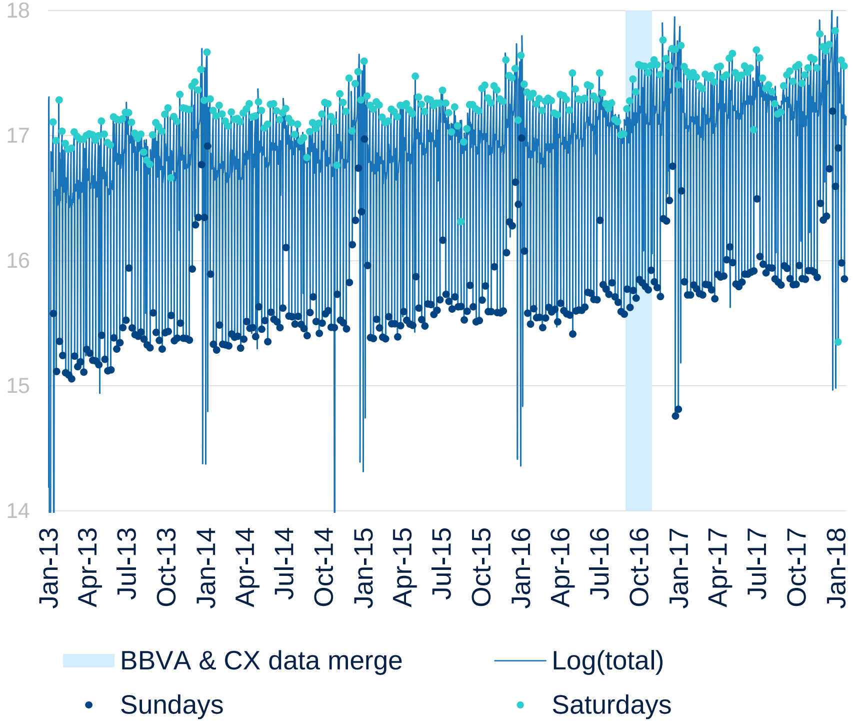
Daily data dynamic modeling is not common in the economic literature. Many sources of variability need to be accounted for:

- 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

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

Aggregate Retail Trade - Daily Frequency
(logarithms)



Source: BBVA

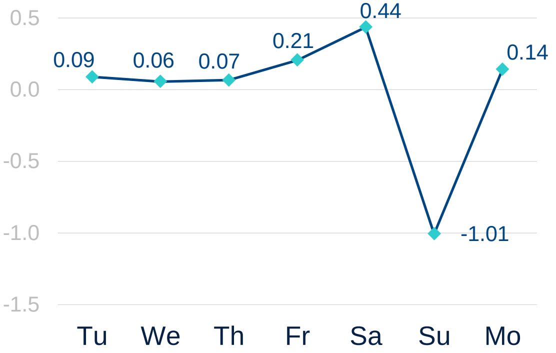
BBVA transactions at daily frequencies: Periodic effects (seasonalities)

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

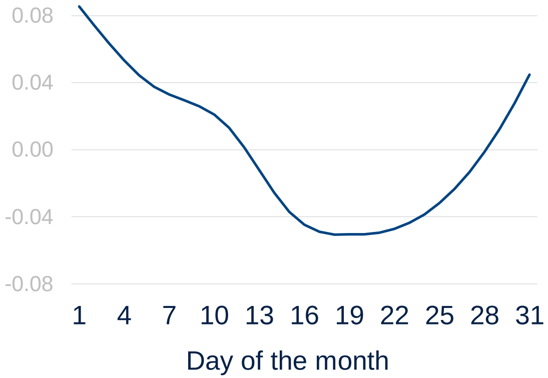
- The day of the week effect is modeled using stochastic dummies $\gamma_t^w = \sum_{j=1}^{s-1} \gamma_{t-j}^w + \omega_t$.
- The intra-monthly and intra-year seasonality is captured using “splines”

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

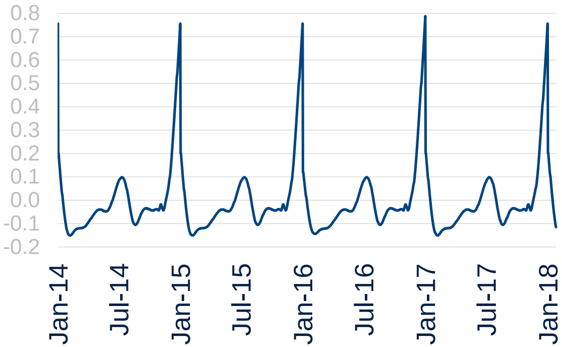
Intra-weekly seasonality (γ_t^w)
(logarithms)



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



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



Source: BBVA

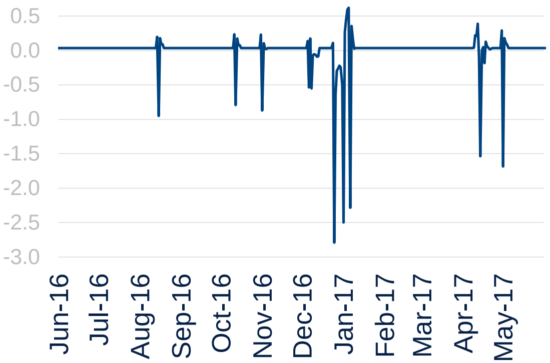
BBVA transactions at daily frequencies: Fixed and moving holidays

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

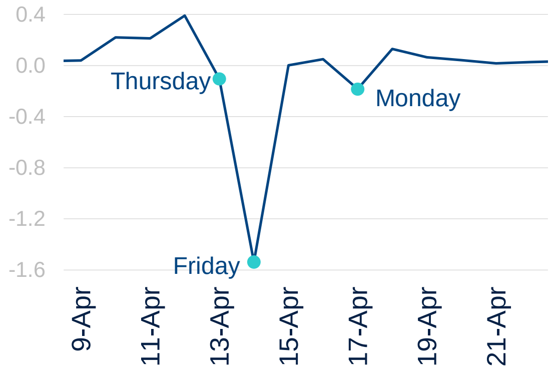
- Holiday’s are modeled using deterministic seasonal dummies (sum zero over the year)
- The trend is stochastic: $\mu\_(t+1)=\nu\_(t+1)+\mu\_t+\xi\_t$ where $\nu\_(t+1)=\nu\_t+\zeta\_t$

Encouraging results: We could analyze the period surrounding each holiday

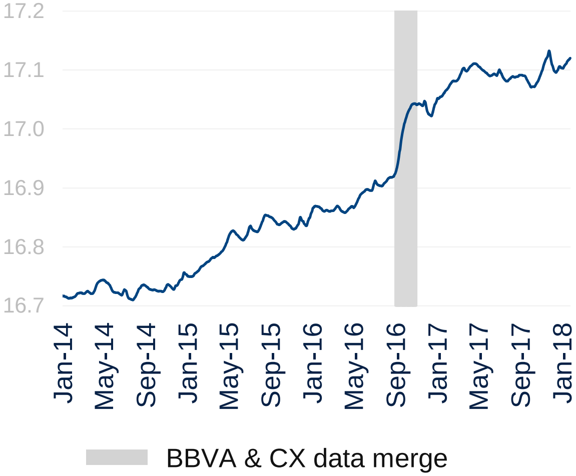
BBVA RTI: Holiday’s effects (γ_t^h)
(logarithms)



BBVA RTI: Easter 2016
(logarithms)



BBVA RTI: Trend (μ_t)
(logarithms)



Source: BBVA

04

Conclusions

Conclusions

- We developed an alternative way of measuring the retail trade in Spain using high dimensional data collected from the digital footprint of BBVA clients using their credit or debit card transactions at a Spanish PoS

- Card transaction data replicates with great precision the evolution of the aggregate Spanish RTI, the RTI by region (Spain’s autonomous regions) and the RTI by retailer type (distribution classes). In addition, the high granularity of the data allowed us to reproduce the evolution of daily retail sales, with timely answers on the impact of any retail sales event, great geographical detail (by province or even by postcode) and information on further dimensions (such as the sector of activity)

- Analyzing the behavior of retailers’ customers to study the high frequency consumption dynamics we found regular, significant patterns that displayed strong intra-weekly, intra-monthly and intra-yearly seasonalities, which are also affected by holiday effects

05

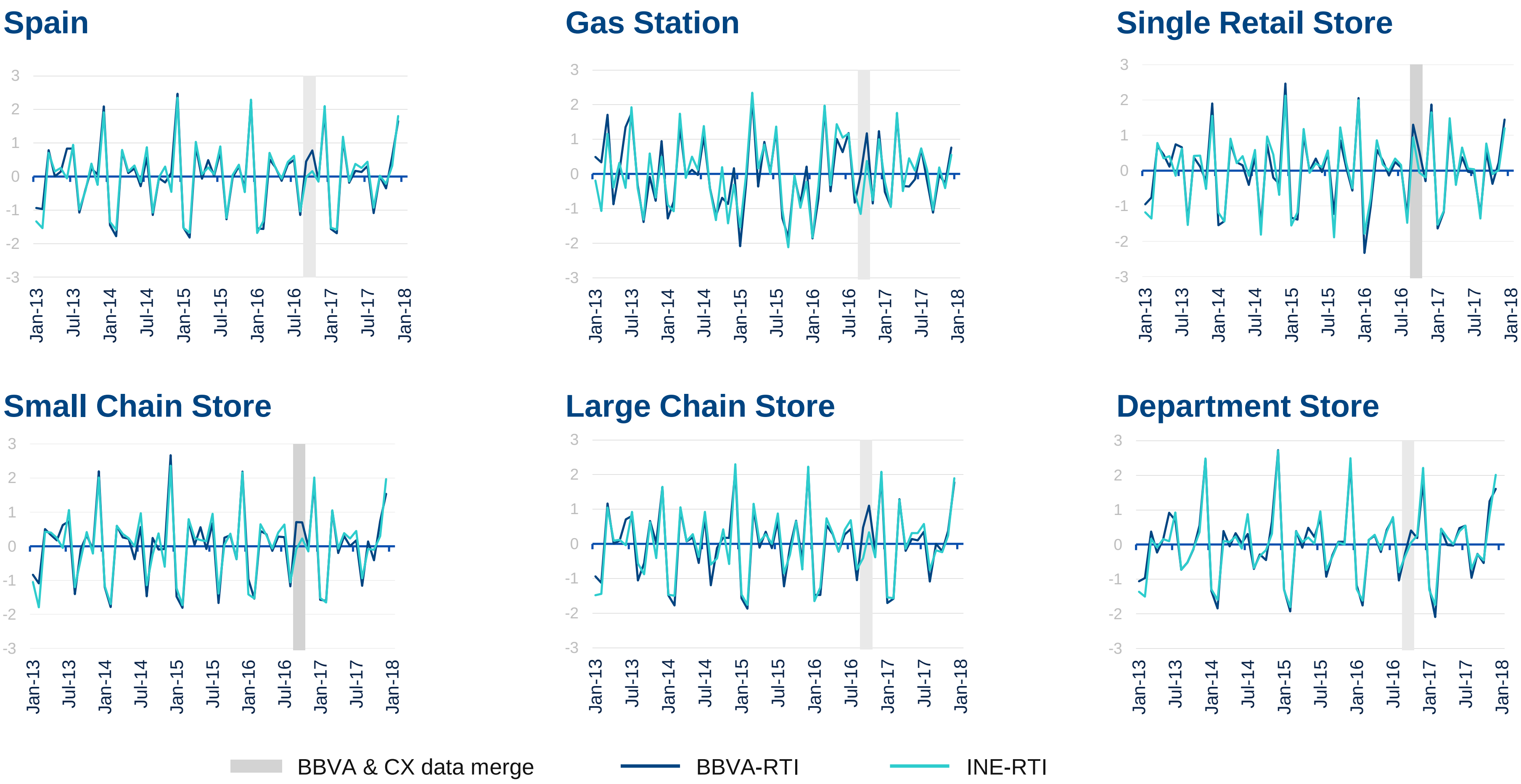
Annex

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)

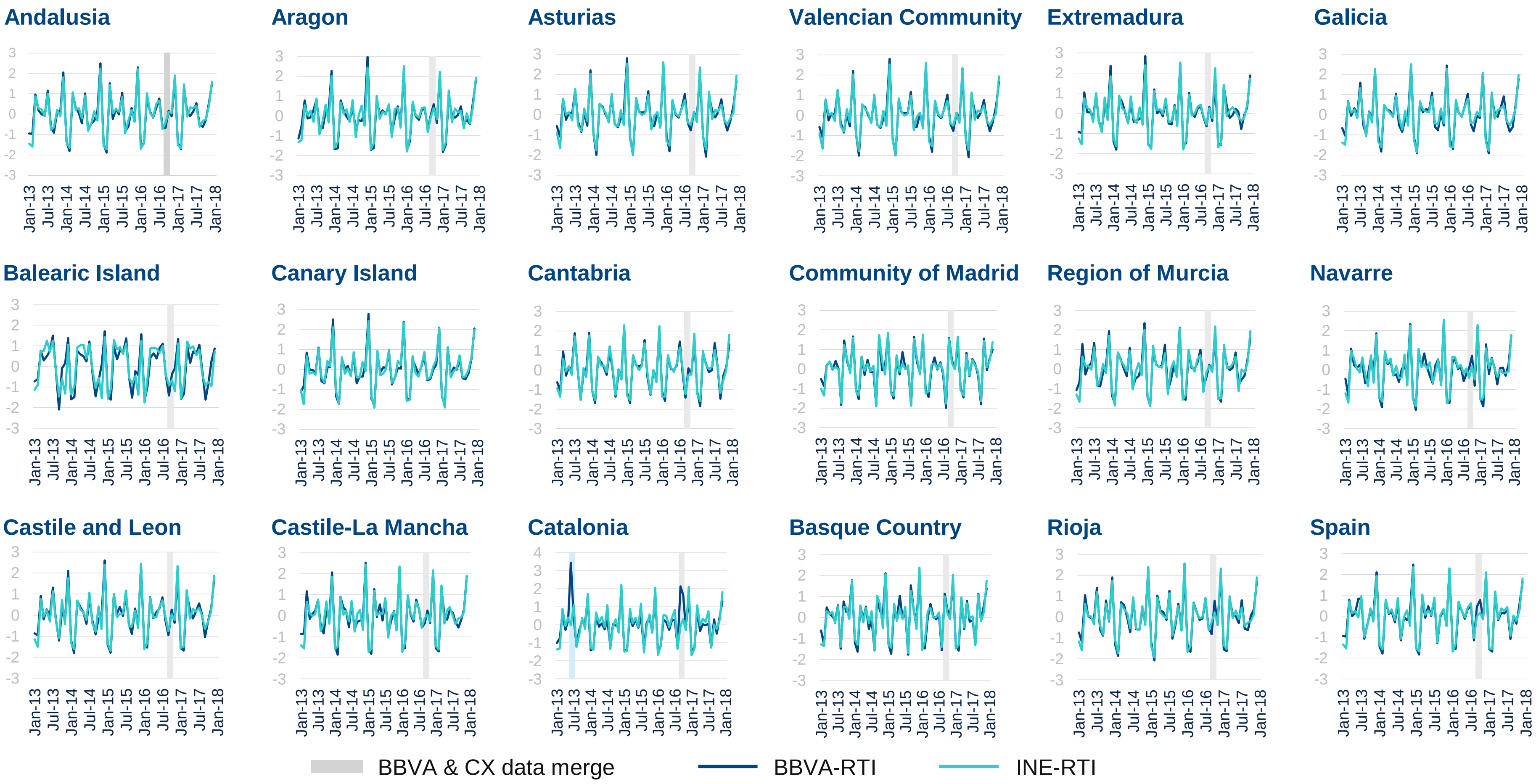
Spain: Macroeconomic consistency of BBVA data by distribution class

Return



Spain: Macroeconomic consistency of BBVA data by AA.CC

Return

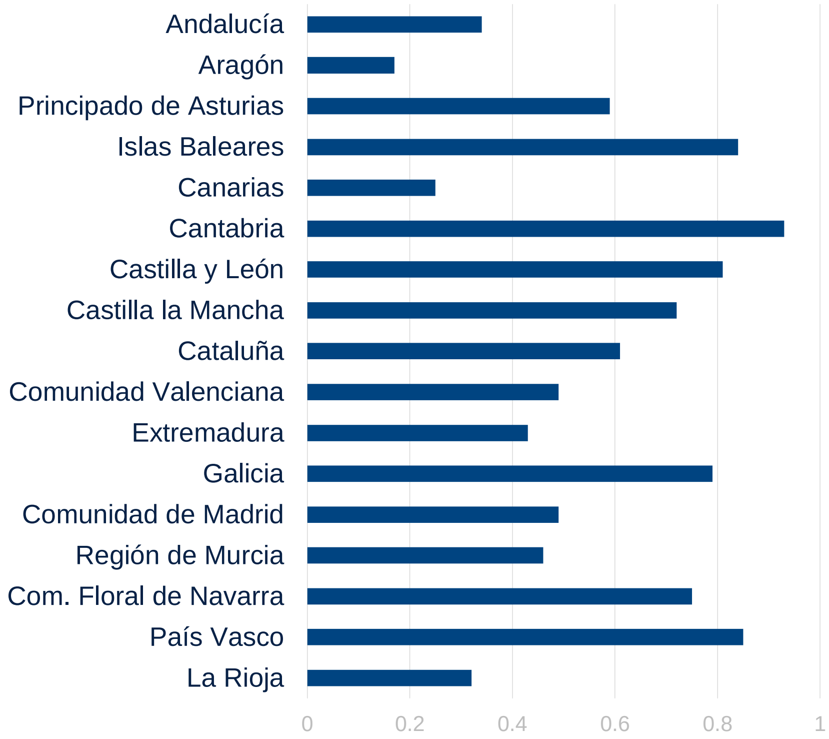


Spain: “Nowcasting” monthly retail sales using BBVA data

Return

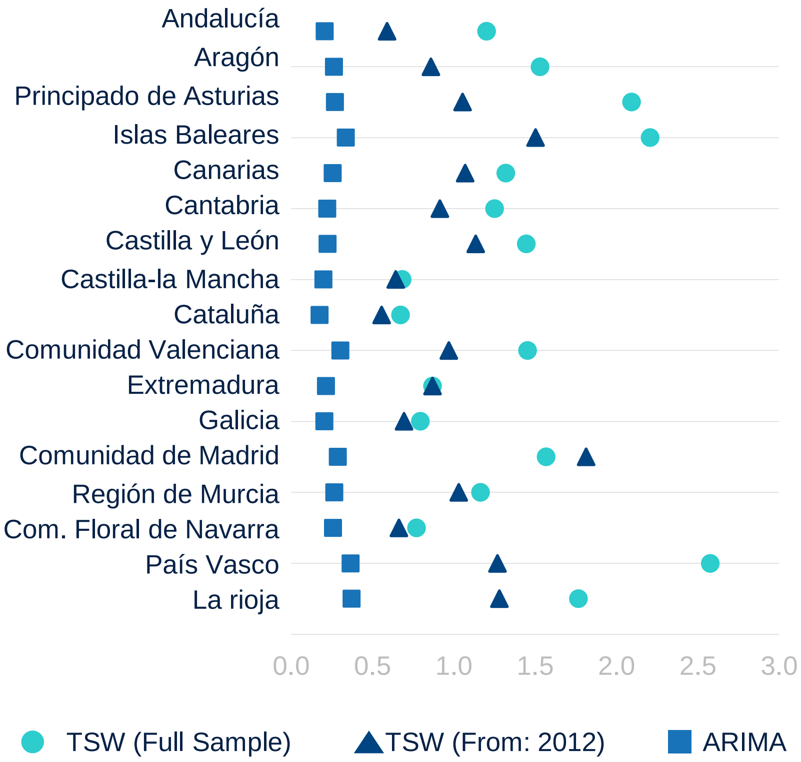
Hansen test P-Value

(H0: parameter stability)



Quasi real time forecast accuracy

(relative RMSE last 12 periods)



Source: BBVA

Daily model

$$\log(y_t) = \overbrace{\mu_t}^{\text{Stochastic Trend}} + \overbrace{\gamma_t^w + \gamma_t^m + \gamma_t^y}^{\text{Seasonalities}} + \overbrace{\gamma_t^h}^{\text{Holidays}} + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$$

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

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

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).

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

where $\varpi_i(B)$ 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

$$\log(y_t) = \overbrace{\mu_t}^{\text{Stochastic Trend}} + \overbrace{\gamma_t^w + \gamma_t^m + \gamma_t^y}^{\text{Seasonalities}} + \overbrace{\gamma_t^h}^{\text{Holidays}} + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$$

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

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

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:

$\gamma_d = \boldsymbol{w}'_{\boldsymbol{d}} \gamma^\dagger \qquad d = 1, \dots, N$ where $\boldsymbol{w}'_{\boldsymbol{d}}$ 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 $\boldsymbol{z}'_{\boldsymbol{d}}$ (replacing $\boldsymbol{w}'_{\boldsymbol{d}}$) where each element "$i$" of $\boldsymbol{z}'_{\boldsymbol{d}}$ is equal to:

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

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

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