BBVA | Research

N° 18/03

WORKING PAPER

Measuring Retail Trade Using Card Transactional Data

Diego Bodas / Juan R. García / Juan Murillo / Matías Pacce / Tomasa Rodrigo / Pep Ruiz de Aguirre / Camilo Ulloa / Juan de Dios Romero and Heribert Valero

BBVA | Research

Measuring Retail Trade Using Card Transactional Data[1]

Diego Bodas[1] / Juan R. García[2] / Juan Murillo[1] / Matías Pacce[3] / Tomasa Rodrigo[2] / Pep Ruiz de Aguirre[2] / Camilo Ulloa[2] / Juan de Dios Romero[1] and Heribert Valero[1]

1: BBVA Data & Analytics, Madrid, Spain
2: BBVA Research, Madrid, Spain
3: Banco de España, Madrid, Spain*

Abstract

In this paper we present a high-dimensionality Retail Trade Index (RTI) constructed to nowcast the retail trade sector economic performance in Spain, using Big Data sources and techniques. The data are the footprints of BBVA clients from their credit or debit card transactions at Spanish point of sale (PoS) terminals. The resulting indexes have been found to be robust when compared with the Spanish RTI, regional RTI (Spain's autonomous regions), and RTI by retailer type (distribution classes) published by the National Statistics Institute (INE). We also went one step further, computing the monthly indexes for the provinces and sectors of activity and the daily general index, by obtaining timely, detailed information on retail sales. Finally, we analyzed the high-frequency consumption dynamics using BBVA retailer behavior and a structural time series model.

Key words: retail sales, Big Data, electronic payments, consumption, structural time series model

JEL classification: C32; C55; C81; E21

1: We are grateful to Gonzalo de Cadenas-Santiago for his contribution in the first stages of the proyect, Miguel Cardoso, Alvaro Ortiz and for their comments, as well as all of those who give us feedback, which greatly improved the quality of this paper.
The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the Banco de España.

1. Introduction

Recent improvements in data storage, management, and processing have led to an exponential increase in the amount and quality of the information available for economic analysis, both from an individual and a macroeconomic perspective. In particular, the latest developments in Big Data technologies permit a 'quasi-real-time' analysis of the information emerging from citizens, governments, and firms in all interactions that generate a digital footprint. New data sources, such as those originating from social networks and search engines, have been proven to help the forecasting of economic variables, such as employment, consumption, and tourism flows (see (Chamberlain, 2010); McLaren 2011; (Choi & Varian, 2012); (Camacho & Pacce, 2017); among many others). More recently, (Cavallo A. , 2016) and (Cavallo & Rigobon, 2016) have shown that the prices of goods and services offered over the Internet can be used to estimate high-frequency price indices. The authors used web scraping tools for the data compilation and proposed a common methodology for generating a price index for a large sample of countries, to allow for the international comparability of prices. We wish to contribute to this line of research by proposing an alternative method for measuring the business evolution of the retail trade sector based on data from credit and debit card transactions.

The retail trade index (RTI) has traditionally been measured by National Statistics Institutes (NSIs) using surveys conducted with a limited sample of retailers, resulting in the collection of relevant information based on data from the supply side. In this paper, we propose a different focus and show that, as could be expected, data emerging from the demand side can also offer similar measures to the official statistics. In particular, we replicated the evolution of the Spanish RTI released by the Spanish National Institute of Statistics (INE), using information obtained from retail transactions by credit and debit card holders of BBVA (one of the largest banks in Spain). The possibility of studying aggregate economic patterns from individual economic transactions using card transaction data was demonstrated by (Sobolevsky, et al., 2015), who were able to obtain regional socioeconomic signals in Spain using this type of information. Also, electronic payment data has already been shown to be helpful in forecasting the evolution of economic aggregates (see Tkacz, 2013; Galbraith and Tkacz, 2016 or (Duarte, Paulo, & Rua, 2017)). However, to the best of our knowledge, this is the first time that the evolution of an official index published by an NSI has been replicated using information from credit and debit card transactions.

Having accurate estimates of the evolution of retail trade sector activity is of great importance given that this is a key indicator of the current economic situation. In general, its dynamic drives the evolution of aggregate consumption (see Figure 1.1), which in turn represents a high proportion of the gross domestic product (GDP). In this sense, it is not surprising that most short-term macroeconomic forecasting models used by Central Banks or private agencies around the world include the aggregate RTI as an important input. One clear example comes from Stock and Watson ((Stock & Watson, 1989), (Sims, Stock, & Watson, 1990), (Stock & Watson, 1991), who included the RTI as one of the four economic indicators needed to construct a coincident indicator index for the evolution of the US GDP. Besides this important feature, the RTI is also important as an indicator for studying the evolution of the retail sector itself, as the possible disaggregation published by the NSIs (by sectors or regions) is key for a detailed analysis.

[Figure 1.1]

The results of this paper show that developments in Big Data analysis have the potential to replicate the evolution of relevant macroeconomic indicators. In particular, we reproduced the dynamics not only of the aggregate Spanish RTI but also the regional RTI (of the Spanish autonomous regions) and retailer type (distribution classes). A number of benefits emerged from our proposed methodology, which were related to the quasi-real-time availability of the data, the higher frequency with which the index can be computed, and the greater geographical and sectoral disaggregation. In this sense, we were able to construct a RTI for the 50 Spanish provinces, a geographical detail that is not published by the INE, and even a daily aggregate Spanish RTI. In addition, based on this daily index, we were able to analyze consumption dynamics by using a structural time series model like the one proposed by (Harvey, Koopman, & Riani, 1997). We found regular, significant patterns that displayed strong intra-weekly, intra-monthly and intra-yearly seasonalities, which were also affected by holiday effects.

The remainder of this paper is organized as follows: Section 2 describes the methodology followed to replicate the Spanish RTI and the data used, while Section 3 shows how alternative indices are a good way to replicate the dynamics of the official ones. Section 4 describes the daily model used to study regular consumption pattern and Section 5 presents the conclusions.

2. An Alternative Way to Compute Retail Trade Indexes

NSIs around the world use the same pillars to estimate the evolution of business in the retail trade sector, which is in general summarized in the retail trade index (RTI). This index reflects the total gross sales of retailers during a fixed period of time (generally a month) and is constructed by conducting surveys directed to a limited number of companies selected using random sampling techniques2. In other words, the relevant data is collected using information obtained from the supply side. Alternatively, it is possible to consider getting the same information from the demand side by using surveys asking the retailers' customers about their expenditure. Even though the latter was never a real option for statistical offices, a major breakthrough has occurred in view of recent developments in Big Data technologies. In particular, the increase in payments using credit and debit cards makes it possible to use the information recorded whenever a credit or debit card is used for a retail transaction to obtain similar measures to the ones given in the RTI, but using real data on consumption instead of data from surveys. Taking this hypothesis, we propose an alternative measure to the Spanish RTI that is based on the information obtained from retail transactions made by credit and debit card holders of the BBVA's Spanish bank.

[2] In the case of Spain, information is obtained from a sample that covers between 20% and 25% of the 12,500 companies registered in the Central Companies Directory (CCD), which provide data by completing a monthly questionnaire over the telephone, email, fax, or the web.

2.1 The Data Sources

We analyzed a complete set of Point of Sale (PoS) purchases or transactions performed in Spanish retail stores between 2013/01 and 2016/12 by clients of BBVA Spain3. For the purposes of this paper, we focused on information4 relating to the amount of each transaction, the geo-localization and principal activity of each PoS, as well as the company that owns it and the exact time the transaction took place. Following the definition given by INE, retail trade here did not include expenditure on motor vehicles and motorcycles, food service, the hospitality industry, or financial services, while sales at gas stations were taken into account. In other words, retail activities refer to Section G, category 47 of the National Classification of Economic Activities (CNAE-2009). Filtering for that specific category is possible because the dataset on card transactions includes the main activity of each PoS.

Following INE, we also grouped purchases into 5 distribution classes based on the following categories of retail store5:

1. Gas stations.
2. Department stores: premises6 with 2,500 m^2 or more.
3. Large chain stores: chain stores with 25 premises or more.
4. Small chain stores: chain stores with more than one premises and less than 25.
5. Single retail stores: only one premises.

We only considered transactions below €30,000 (the maximum credit limit for BBVA Gold Cards). Values over that threshold were considered to be outliers. The sample contained approximately 12,500 stores and more than 200 million annual transactions by over 4 million cardholders. Columnar databases were used to deal with this huge amount of information.

2.2 Methodology

In order to build the Spanish BBVA-RTI based on card transactions it was necessary to create a data engine capable of regularly delivering the index pursued. With this aim, a number of steps were followed during the process of building the data engine (see Figure 2.1).

[Figure 2.1]

We started selecting data sources and useful variables to meet the goals of the project. Big data queries are not a trivial task when using columnar databases and cluster solutions, as they need to be optimized in order to avoid

[3] Both face-to-face and online purchases have been analyzed for this project.
[4] The transactions database has been anonymized and aggregated before analyzing it.
[5] Since information on the number of employees in each company was not available, we did not include the restriction of having more than 50 employees or more in the definition of a "large chain" or "small chain"
[6] According to the INE, a premises is "any structurally separate and independent building that is not dedicated exclusively to family housing, and in which economic activities dependent on a company are carried out, and in which one or more persons work for the company".

cluster failure or malfunction. We obtained daily, weekly, and monthly aggregate data on the total number and total amount of the transactions. At the same time, we queried this information at different levels of granularity, getting data for the entire country, data for each of the 17 autonomous regions of Spain, and the 50 provinces7. Ensuring data quality required that the data be cleaned and formatted. During this process, outliers were deleted before the data were standardized8. In the final stage, the data were tested to check whether these data sources and variables were useful for the project's goals. Finally, the process was automated by implementing a code library.

2.3 Strengths and Weakness of the BBVA-RTI

As we are proposing an alternative way of computing the Spanish RTI (which could potentially be translated to other countries), it is important to point out the advantages and disadvantages of using card transaction data rather than the classical method of estimation used by official statistical offices. The comparison is summarized in Table 19.

Table 2.1 Comparison between RTI Data Sources

	Card Transaction Data (BBVA)	Survey Data (INE)
Cost per observation	Marginally Low	High
Data Frequency	Daily	Monthly
Disaggregation by activity	High	Low
Geographical disaggregation	High	Low
Real-time availability	Yes	No
Retailer sample	12,500	≈ 3,000
Payment methods covered	BBVA's clients credit and debit cards	All
Possible bias of technological trends	Yes	No

The first advantage of using card transaction data is related to the cost of obtaining each extra observation. Even though storing huge amounts of information is not cheap, the economic scale of digital information storage means the cost of marginal observation is close to zero, allowing for an obvious gain as compared to conducting 50 parallel regional monthly surveys (one for each INE provincial delegation) to obtain the relevant information.

A second advantage is related to the frequency of data collection, which allows for a deeper analysis of the behavior of retailers' customers than the one than can be performed when using monthly data. Section 5 shows an example of this applicability.

[7] Spain's autonomous regions and provinces correspond, respectively, to NUTS-2 and NUTS-3 in EUROSTAT nomenclature.
[8] We took natural logs or the first difference of logs, subtracted the mean, and divided by standard deviations.
[9] This table is comparable to Table 1 of (Cavallo & Rigobon, 2016), which describes the advantages and disadvantages of using online data prices to construct an alternative consumer price index.

Thirdly, card transaction data include information on each PoS's main activity, allowing for greater economic activity disaggregation, and not only for each of the 5 groups published by INE. As an example, in Section 4, we show the median expenditure by sector at the end of 2017.

Fourthly, the geographical disaggregation that can be obtained is greater than that published by INE. In particular, with data on the geo-localization of each PoS, it is potentially possible to generate an RTI for a city or even a single postcode. In the present work, we computed the RTI at the provincial level, a disaggregation that is not available for the INE data.

Fifthly, INE publishes its data with a one-month delay, while card transaction data is available almost in real time. This would allow policymakers to access the latest information without any kind of delay.

Lastly, in regard to the sample, the card transaction data include almost the entire sample of companies registered in the CCD (12,500 companies), while INE's data are based on a sample that covers only some 25% of the companies registered.

On the other hand, some disadvantages can be found when using card transaction data. Firstly, the total amount billed only refers to expenditure made using BBVA client credit and debit cards. This means that we excluded all transactions made using cash or non-BBVA cards. Nonetheless, given BBVA's high market share (13.8%) in Spain, we assumed that the sample we were using was sufficiently representative. A second disadvantage was related to the potential bias that technological trends could generate if they affect preferences for using credit or debit cards.

3. The Spanish Retail Trade Index

In this section, we show that information obtained from card transaction data can replicate the dynamics of the official RTI for Spain, not only for the national aggregate, but also for all five sub-divisions of the national index and for all 17 official retail trade indices published for each of the autonomous regions.

3.1 Similarities with the Official Aggregate Indices

By comparing both the official Spanish RTI and the BBVA-RTI, it is easy to see how the two are closely related. In Figure 3.1, we plot the monthly BBVA-RTI next to the nominal non-seasonally adjusted official RTI. As can be seen, even though the dynamic of both indices appears to be similar, the BBVA-RTI shows a steeper trend than the official index. As previously said, this may be associated with the existence of some kind of technological trend affecting consumer behavior and the intensity of use of credit and debit cards. As an example of this pattern, in Figure 3.2 we plot the evolution of the BBVA-RTI's average transaction amount. As shown, the median transaction amount decreased from €45.70 in December 2013 to €40.50 in June 2017. This result, together with the upward trend in the BBVA-RTI, can be interpreted not only as the fact that people are increasingly using credit and debit cards but also that there is a higher number of lower amount transactions. Additionally, the BBVA data may be affected by the addition of new clients or the loss of old ones. This is particularly relevant in the case of mergers and acquisitions, like

BBVA's absorption of UNIM and CatalunyaCaixa in the second quarter of 2013 and the last quarter of 2016[10], respectively.

[Figures 3.1 and 3.2]

Even though Figure 3.1 does not clearly show the official RTI as being a non-stationary series, it does become clear when the index is plotted for a larger sample period[11] (see Figure A1 in the Annex). As both series show a non-stationary pattern (even though with different trends), working with growth rates is a possible solution to continuing with the comparison. In Figure 3.3, we show how the similarities between the indices become stronger when expressed in monthly standardized[12] growth rates, giving strong support to the BBVA-RTI as a very close approximation to the official index. This became even more evident when we analyzed the 5 distribution channels for which the INE gives retail trade indices, besides the aggregate one. In Figure 3.4, it is possible to see that those similarities are high enough for all 5 disaggregations, even though with some heterogeneity between them. The case of "department stores" is the one that shows most proximity between the indices, even though the agreement is also very high for "large chain stores". When analyzing the cases of "small chain stores" and "single retail stores", it can be seen that the dynamics of the series (in monthly growth rates) are similar, but the indices built on card transaction data appear to be more volatile. In contrast, "gas stations" is where greater differences appear. One possible explanation for the greater similarities found for the indices relating to larger retailers could be explained by a more intense use of credit and debit cards in these kinds of stores. The differences that emerge for "gas stations" could be due to inflows of cash payments. In the left pane of table 3.1, we show the R-squared from the linear regression between the BBVA-RTI and the official indices. In the right pane of the table, we show the Hansen stability test p values for these regressions. The results indicate that, although the correlation between the series levels is high, the relationship is statistically stable for all the distribution channels only when growth rates are taken into account. Altogether, the results clearly show that co-movement between the series is robust enough to reinforce the idea that card transaction data provide suitable information for correctly replicating the official RTI.

[Figure 3.3]

[Figure 3.4]

[10] BBVA Research estimates indicate that the relationship between BBVA-RTI's growth rates and the official RTI deviate by 0.47 and 0.59 sd in September and October 2016 as a result of the absorption of CatalunyaCaixa. No statistically significant effects were found for national transactions after the UNIMUnnim takeover.
[11] The official data start in 1995/01.
[12] We rescale each of the month to month growth rate series to have a mean of zero and a standard deviation of one.

Table 3.1 INE-RTI and BBVA-RTI

	R-squared		Hansen stability test P-Value (H0: parameter stability)	
	Levels	Monthly growth rate	Levels	Monthly growth rate
Total	**0.89**	**0.94**	**0.22**	**0.90**
Department Stores	0.89	0.95	0.03	0.99
Large chain stores	0.73	0.91	0.32	0.93
Small chain stores	0.48	0.91	0.30	0.67
Single Retail stores	0.57	0.92	0.21	0.85
Service stations	0.05	0.79	0.01	0.18

Note: sample: 2013.01-2017.12 period

3.2 Similarities with Regional Indices

As mentioned above, the INE publishes an RTI for each of Spain's 17 autonomous regions. Taking advantage of the high geographical disaggregation permitted by the card transaction data, we constructed each of these 17 indices based on that information. Figure 3.5 shows the dynamic of all the regional INE-RTI and BBVA-RTI, once again expressing the indices as monthly growth rates for better comparability. The figures show the great similarities in the dynamics of the indices, which are also reflected in the high R-squared and Hansen stability test p values (see Table 3.2).

[Figure 3.5]

One of the bi-products of using card transaction data for computing regional RTIs is the possibility of obtaining 5 distribution groups for each of the autonomous regions, which are not publicly available from INE[13].

[13] All figures regarding the 5 distribution groups by regions are available upon request

Table 3.2 INE-RTI and BBVA-RTI by Autonomous Region (R-squared)

	R-squared		Hansen stability test P-value (H0: parameter stability)	
	Levels	Monthly growth rate	Levels	Monthly growth rate
Andalusia	0.72	0.96	0.14	0.32
Aragon	0.52	0.91	0.05	0.24
Asturias	0.76	0.94	0.20	0.70
Balearic Islands	0.72	0.90	0.19	0.96
Canary Islands	0.87	0.94	0.16	0.21
Cantabria	0.87	0.96	0.10	0.87
Castile and León	0.70	0.94	0.05	0.96
Castilla-La Mancha	0.65	0.93	0.09	0.86
Catalonia	0.38	0.83	0.37	0.23
Valencian Community	0.81	0.94	0.06	0.69
Extremadura	0.54	0.94	0.06	0.48
Galicia	0.82	0.95	0.07	0.51
Community of Madrid	0.79	0.96	0.09	0.88
Region of Murcia	0.65	0.91	0.27	0.15
Navarre	0.78	0.91	0.08	0.27
La Rioja	0.78	0.95	0.05	0.82
Basque Country	0.82	0.95	0.47	0.38

Note: sample period: 2013/01-2017/12

4. Higher Dimensionality: Granular Data by Time Span, Geography and Further Dimensions

After checking that the constructed BBVA-RTI replicated the official figures published by the INE at all levels in which is available, we went one step further in taking advantage of the BBVA quasi-real-time transaction data, getting insights into retail sales at higher frequencies (e.g., daily) with greater geographical detail (i.e., at the provincial level), as well as exploiting 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., socioeconomic characteristics of consumers, such as sex, age, and income level).

The high frequency of the BBVA-RTI (Figure 4.1) covered the one-month lag in publication by INE, providing timely answers on retail sales for particular events. It also permitted a deeper analysis of the retailers' customers' behavior,

uncovering the aggregate consumption dynamic using a structural time series model like the one proposed by (Harvey, Koopman, & Riani, 1997) (an example is shown in Section 5).

[Figure 4.1]

The geo-located information from the PoSs gave a higher geographical disaggregation to the BBVA-RTI, providing information on the evolution of retail sales that is not published by the INE. We had the RTI for the 50 provinces (NUTS 3 geographical division in the EUROSTAT nomenclature), but it would potentially be possible to generate an RTI at the city or postcode level. Figure 4.2 shows the evolution of the RTI in December 2017 in each province as compared to December 2016 (yoy levels). Although it was not feasible to compare the dynamics of these series with the INE statistics (the INE-RTI information is published by autonomous region), the consistency of the index at the national level and by autonomous region, as well as the correspondence between the INE-RTI and the BBVA-RTI for autonomous regions with only one province, brings a high likelihood to the rest of the provincial indices.

[Figure 4.2]

The transaction data included information on the main activity of each PoS, allowing for more economic activity disaggregation than the information published by INE. The analysis of the BBVA-RTI by sector of activity showed that the lowest median ticket in December 2017 was for healthcare, other services, and the bar and restaurant sectors, respectively. In contrast, technology, sports and toys, and automotive were among the sectors in which we found the highest median expenditure. (Figure 4.3).

[Figure 4.3]

5. Working with Daily Data

One of the most important features of working with card transaction data is the possibility of studying aggregate consumption patterns. In other words, given that high frequency data is available, it is feasible to study the actors' decisions regarding daily, or even hourly, expenditure. The daily BBVA-RTI displays weekly, monthly and annual seasonalities, plus some calendar effects. Even though the figure shows a very volatile pattern, it is clear that Saturdays are in general the day on which people consume the most, while on Sundays they consume the least. In addition, it seems that, within a year, December is the month with the highest consumption, followed by July, while calendar effects relating to public holidays or Easter Week can be found where the blue line becomes thicker. Modeling all those patterns into one single model that operates at a daily frequency was not an easy task and several issues had to be taken into account. Not only the numbers of days within a month or within a year[14] change, but also the position of the date on a specific day of the week[15] or for holidays like Easter is not the same from year to year. As mentioned in Cabrero et al. (2007), two major approaches exist for dealing with those and other problems in the context of daily time series: the ARIMA model suggested by (Bell & Hillmer, 1984) and the structural time series (STS)

14: The number of days in a year depends on its being a leap year or not.
15: e.g., January 1st is not always on a Monday.

approach of (Harvey, Koopman, & Riani, 1997). In this paper, we used the second approach, which includes *periodic cubic splines* to model some of the seasonal components exhibited by the daily BBVA-RTI data[16].

[Figure 5.1]

Using the Harvey et al. (1997) notation, the basic STS model can be described for a univariate time series (y_t) where

$$y_t = \mu_t + \gamma_t + \varepsilon_t, \qquad \varepsilon_t \sim NID(0, \sigma_\varepsilon^2)$$

where μ_t, γ_t and ε_t are, respectively, the stochastic trend, the stochastic or deterministic seasonal components, and the irregular component, while $t = 1, \dots, T$. The dynamic of the stochastic trend is defined by,

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t, \qquad \eta_t \sim NID\left(0, \sigma_\eta^2\right)$$

$$\beta_t = \beta_{t-1} + \zeta_t, \qquad \zeta_t \sim NID(0, \sigma_\zeta^2)$$

where the level and the slope of the trend are given by μ_t and β_t, while η_t and ζ_t are mutually independent processes. The seasonal component is characterized as the sum of the day of the week effect (γ_t^d), the intra-monthly effects (γ_t^m), the intra-yearly effects (γ_t^y), and moving and fixed holidays (γ_t^h)[17]:

$$\gamma_t = \gamma_t^d + \gamma_t^m + \gamma_t^y + \gamma_t^h \qquad (3.1)$$

Each of the seasonal components is described by its own dynamics. In particular, to model the day of the week effect, we rely on stochastic dummies of the form:

$$\gamma_t^d = -\sum_{j=1}^{s-1} \gamma_{t-j}^d + \omega_t, \qquad \omega_t \sim NID(0, \sigma_\omega^2) \qquad (3.2)$$

where $s = 7$ is the number of days in a week[18]. By imposing $\sigma_\omega^2 = 0$, the seasonality becomes deterministic (the main results remain unchanged when this is done).

Intra-monthly and intra-yearly effects are both modeled by using time-varying cubic splines. For setting a spline it is necessary to choose h^i knots in the range of $[0, N^i]$, where N^i is the number of the days in a month or in a year ($i = \{m, y\}$). Once again, following Harvey et al. notation, we define

$$\gamma_d^i = \mathbf{z}_d^{i\prime} \gamma_t^i \qquad d = 1, \dots, N^i \quad ; \quad i = \{m, y\} \qquad (3.3)$$

where $\mathbf{z}_d^i$ is vector of dimension $(h^i - 1) \times 1$, which depends on the number and positioning of the knots and should be defined in a way that guarantees continuity from one period to the next. By letting the vector γ_t^i follow a random

16: Harvey et al. (1997) indicate that, by giving more scope for a parsimonious parameterization, this approach better captures periods with sharp peaks, like the one that can be observed surrounding Christmas.
17: For notational simplicity, we have included the fixed and moving festivals as seasonal components, even though both are calendar effects.
18: Dummies (z_{jt}) in equation 3.1 are introduced as $\gamma_t^d = -\sum_{j=1}^{s-1} \gamma_{t-j}^d z_{jt} + \omega_t$ for $t = 1, \dots, T$ where for $t = i, i+s, i+2s, \dots$ and $i = 1, \dots, s-1$ the variable z_{jt} is one for $j = i$ and zero for $j \neq i$, while for $t = s, 2s, 3s, \dots$ and $j = 1, \dots, s-1$ the value of z_{jt} is equal to minus one.

path, γ_d^i becomes stochastic. For a detailed explanation on the modeling of periodic cubic splines, see Harvey et al. (1997).

In order to specify the length N^i , it was necessary to take into account the fact that not all months or all years have the same number of days. To deal with this problem, we followed the strategy of Cabrero et al. (2009) and set $N^m = 31$ for all months and $N^y = 366$ for all years. The days that do not exist (February 29th when the year is not a leap year or days like April 31st) were considered to be missing values and were easily handled, given that our estimation strategy relied on Kalman filter iterations. When using periodic cubic splines, a second issue to be taken into account was related to the number and position of the knots. Moreover, to obtain periodicity, the value of the first and the last knot within a period should be equal[19]. Therefore, two consecutive days with similar characteristics should be chosen for placing the starting and final knots. Cabrero et al. (2009) correctly highlight that setting the first and last knot as January 1st and December 31st for intra-yearly seasonality gives the particularities of those days, while for intra-monthly patterns the first and last days of the month are less likely to be similar than two days in the middle of the month[20]. The dates finally chosen for the placement of the first and last knot for the annual periodicity were February 18th and 19th, while the 22nd and 23rd of each month were selected for monthly splines. The decision on the final number and position of the knots was based on the analysis of residual correlograms, goodness-of-fit performance, and visual observation. In particular, after trying many different specifications, we decided to include 7 knots for intra-month patterns and 18 knots for intra-year seasonality. As in (Harvey, Koopman, & Riani, 1997), when dealing with annual seasonality, we needed to impose a relatively larger number of knots in the short period of time surrounding Christmas, while fewer knots were needed when seasonal patterns changed slowly. The knots for the intra-yearly spline were placed at 1, 5, 9, 13, 17, 21, 25 and 28[21] and for the intra-monthly spline at 1, 25, 50, 75, 125, 150, 200, 225, 251, 286, 296, 301, 306, 313, 320, 330, 345, and 365[22].

To model holidays effects (γ_t^h), we used a deterministic approach and included dummy variables for each of those specific days[23]. It should be noted that $\gamma_t^h = \sum_{i=1}^{I} \gamma_t^{h,i}$ where $i = 1, \dots, I$ is an indicator for each holiday. Under this notation, the holidays effect was modeled as

$$\gamma_t^{h,i} = \omega_i(B) h(\tau_i, t) \qquad (3.4)$$

where $\omega_i(B)$ is a polynomial lag operator and $h(\tau_i, t)$ is an indicator function that takes the value 1 when $\tau_i = t$ and zero otherwise. The presence of a polynomial lag operator is related to the fact that days surrounding a holiday could

19: As mentioned in Cabrero et al. (2009), this is strictly true only for the case of deterministic periodic cubic splines.
20: It should also be remembered that in 5 out of the 12 months of the year, the last day of the month does not really exist (e.g., April 31st) and it is considered as a missing value for estimation purposes. Also, the end of the month displays a sharp trend given the monthly seasonal pattern.
21: Notice that knots 1 and 365 correspond, respectively, to February 18th and 19th
22: Knots 1 and 28 correspond, respectively, to the 22nd and 23rd day of the month.
23: In Spain, there are three classes of public holidays: national holidays, holidays specific to each autonomous region and municipal holidays. As the daily model will be applied to the national aggregate RTI, only national holidays were taken into account.

also show some peculiarities (e.g., people going to the supermarket the day before a holiday)[24] (see table 5.1 to see the polynomial lag operator set for each holiday). When a National holiday falls on a Sunday, we opted not to include a dummy for that day.

Table 5.1 Fixed Holiday Lag Polynomials

	Card Transaction Data (BBVA)
Good Friday	$(w_0 + w_1 B + \cdots + w_{12} B^{12}) B^{-6}$
New Year, Epiphany (Jan 6th), St. Joseph (Mar 19th), Labor Day (May 1st), Assumption (Aug 15th), Spain's National Holiday (Oct 12th) and All Saints' Day	$(w_0 + w_1 B + \cdots + w_4 B^4) B^{-1}$
Immaculate Conception (Dec 8th)	$(w_0 + w_1 B + \cdots + w_8 B^8) B^{-3}$

Since the whole model described in 3.1-3.4 can be written in state space form, maximum likelihood estimation in combination with a Kalman filter and Kalman smoothed could be used. The main results are summarized in Figures 3.2 to 3.7.

Figure 5.1 shows the intra-weekly effects for a week of the year[25]. As highlighted by the raw data (Figure 4.1), Sundays are the days of the week with the lowest consumption while Saturdays have the highest. This behavior is not surprising in a country like Spain, where Sunday is a day for the family and rest, meaning that most retail shops are closed. On the other hand, it seems that Saturdays are used for doing the shopping that is harder to do on weekdays, maybe because of restrictions caused by working hours. For weekdays, the consumption pattern looks to be very similar between Monday and Thursday, while it rises on Fridays.

The intra-monthly effects are displayed in Figure 5.2. The results show that consumption is higher during the first two weeks and the last three days of a month, suggesting a consumption pattern linked to salary payment[26]. Working with statistics on the daily banknotes in circulation in Europe, Cabrero et al. (2009) found a similar intra-monthly behavior. This result is in line with Stephens (2003, 2006) (Shapiro, 2005), (Mastrobuoni & Weinberg, 2009) and (Aguila, Kapteyn, & Pérez-Arce, 2017), who found monthly increase in consumption during the week of and the week after payroll. Alternative explanations for this kind of consumption behavior rely on the existence of credit restrictions, liquidity constraints, myopia, or the existence of hyperbolic discounting in the actor's preferences.

Figure 5.3 shows the intra-yearly seasonality. As can be observed, there is a sharp peak starting in the first few days of December and ending around January 10th. This period is related to the Christmas holidays, when the increase in

24: To model holidays, we also took into account the fact that, to be treated as a seasonal effect, the holidays effects plus the non-holiday factor should be null (the dummy variables were altered to get this kind of effect).
25: As we are working with stochastic dummies, the consumption pattern is not exactly similar for every week of the year. Figure 3 shows the intra-weekly behavior for the first week of June, but results are very similar for any week of the year.
26: In Spain, wages are paid monthly, normally on the first or last day of the month.

retail sales may be associated with purchasing Christmas gifts. Another period of high retail consumption is during July, which may be related to the summer sales period. The rest of the year displays a very similar pattern, although February and March appear to be the months with the lowest sales. The same kind of intra-yearly seasonality is the one that the INE found in the monthly data for the RTI.

The holiday effects are shown in Figure 5.4. As expected, all national holidays have a negative effect on retail sales, which is obviously related to the fact that most retail stores are closed on those days[27]. Also, the days before and after a holiday show a positive pattern. This could be explained by a distribution of consumption around the holiday date if it coincides with a working day. December 25th and January 1st and 6th are the holidays with the highest negative effects on retail sales followed by May 1st.

Given the importance of Easter, Figure 5.5 shows consumption during a period of two weeks around those days. As can be seen, consumption increases during the week prior to the Easter weekend and falls on Good Friday. The fall observed on Easter Monday is not surprising as it is a holiday in some of the biggest autonomous regions (e.g., Catalonia). Finally, Figure 5.6 shows the estimated stochastic daily trend. As expected from dynamic observed in Figure 3.1, a positive trend was found during the period in which the model was estimated.

The results obtained using the daily STS model should be considered carefully. Since we only had four years of data, the intra-yearly and fixed holiday effects were mostly indicative. As time passes and we amass more data, a better estimate will be possible.

6. Conclusions

The new digital era, together with the development of data infrastructure, technologies, and data science techniques, presents a chance for economic research to take advantage of unprecedented amounts of data. In this paper, 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 point of sale (PoS) terminal.

The results of this paper show that card transaction data replicate with great precision the evolution of the Spanish RTI, an important macroeconomic indicator showing the evolution of aggregate consumption and, therefore, of economic activity. The RTI indicator developed replicated the dynamics 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).

We also investigated the behavior of retailers' customers to analyze the high frequency consumption dynamics using a structural time series model. We found regular, significant patterns that displayed strong intra-weekly (Sundays are

27: Note the absence of some holidays during the period s plotted (October 12th, for example). As previously explained, we did not include dummies for a holiday when it fell on a Sunday.

the days of the week with the lowest consumption while Saturdays are the ones with the highest), intra-monthly (consumption is higher during the first two weeks and last three days of a month) and intra-yearly seasonalities (we found a sharp peak in retail sales starting in the first few days of December and ending around January 10th, and also in July), which are also affected by holiday effects.

This line of research could be extended to exploit further dimensions offered by the data, such as the credit consumption behavior of BBVA clients or the socioeconomic features of online versus offline payments. Deseasonalizing the index to work with real values instead of nominal ones, and testing its predictive power at nowcasting, is left for further research.

References

Aguila, E., Kapteyn, A., & Pérez-Arce, F. (2017). Consumption Smoothing and Frequency of Benefit Payments of Cash Transfer Programs. *American Economic Review, 107*(5), 430-35.

Bell, W., & Hillmer, S. (1984). Issues Involved with the Seasonal Adjustment of Economic Time Series. *Journal of Business & Economic Statistics, 2*(4).

Camacho, M., & Pacce, M. (2017). Forecasting Travellers in Spain with Google's search volumen indices. *Tourism Economics*.

Cavallo, A. (2016). Scraped Data and Sticky Prices. *Review of Economics and Statistics*.

Cavallo, A., & Rigobon, R. (2016). The Billion Prices Project: Using Online Prives for Measurement and Research. *Journal of Economic Perspectives, 30*(2), 151-78.

Chamberlain, G. (2010). Binary Response Models for Panel Data: Identification and Information. *Econometrica, 78*(2), 159-168.

Choi, H., & Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record, 88*, 2-9.

Duarte, C., Paulo, M., & Rua, A. (2017). A Mixed Frequency Approach to the Forecasting of Private Consumption with ATM/POS Data. *International Journal of Forecasting*, 61-75.

Harvey, A., Koopman, S., & Riani, M. (1997). The Modelling and Seasonal Adjustment of Weekly Observations. *Journal of Business & Economic Statistics, 15*(3), 354-68.

Mastrobuoni, G., & Weinberg, M. (2009). Heterogeneity in Intra-monthly Consumptions Patterns, Self-Control, and Savings at Retirement. *American Economic Journal: Economic Policy, 1*(2), 163-89.

Shapiro, J. (2005). Is there a daily discount rate? Evidence from the food stamp nutrition cycle. *Journal of Public Economics, 89*(2-3), 303-325.

Sims, C. A., Stock, J., & Watson, M. (1990). Inference in Linear Time Series Models with some Unit Roots. *Econometrica, 58*(1), 113-144.

Sobolevsky, S., Bojic, I., Belyi, A., Sitko, I., Hawelka, B., Murillo Arias, J., & Ratti, C. (2015). Scaling of City Attractiveness for Foreign Visitors through Big Data of Human Economical and Social Media Activity. *IEEE International Congress on Big Data*, *Proceedings*, pp. 600-07.

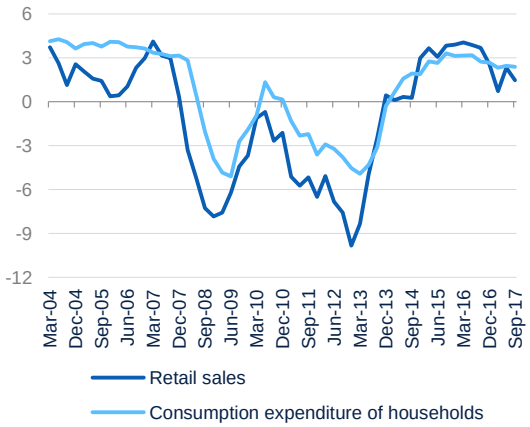
Stock, J., & Watson, M. (1989). New Indexes of Coincident and Leading Economic Indicators. *NBER Macroeconomics Annual, 4*, 351-409.

Stock, J., & Watson, M. (1991). A probability Model of the Coincident Economic Indicators. In G. Moore, & K. Lahiri, *The Leading Economic Indicators: New Approaches and Forecsting Records* (pp. 63-90). Cambridge University Press.

Tkacz, G., (2013). Predicting Recessions in Real-time: Mining Google Trends and Electronic Payments Data for Clues. Working Paper

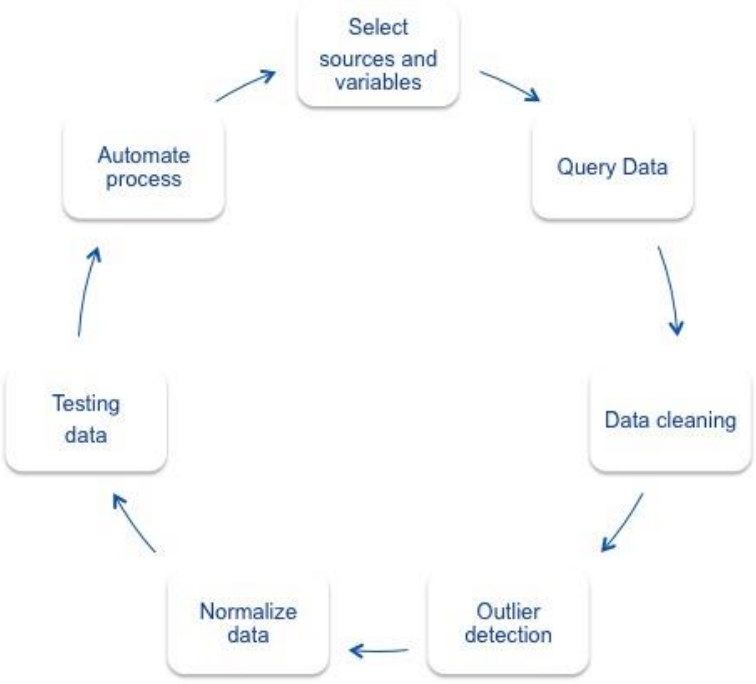
Figures

Figure 1.1 Spain: Retail Sales vs. Household Consumption Expenditure (%, YoY)



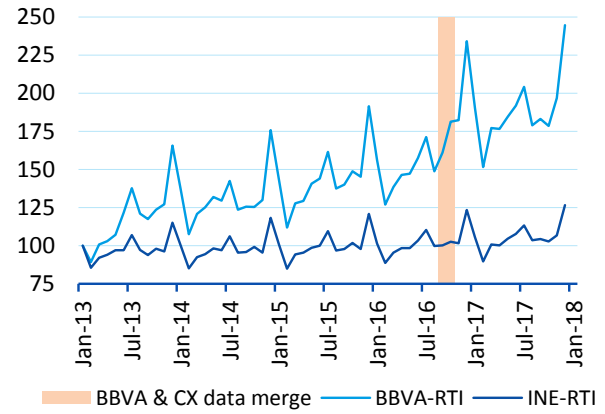
Source: BBVA based on INE

Figure 2.1 Data Engine Building Process. Data extraction, cleansing, and transformation



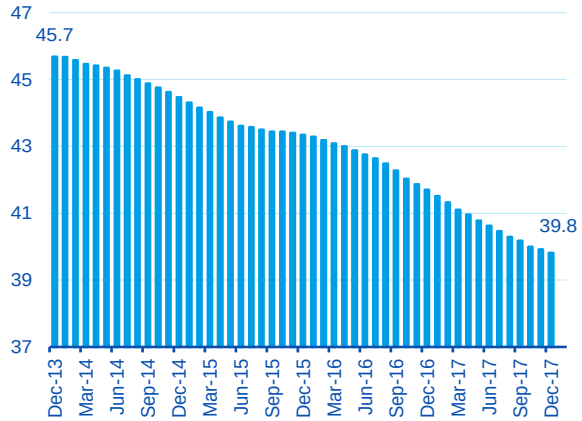
Source: BBVA

Figure 3.1 Aggregate Retail Trade Indices (Jan-13 = 100)



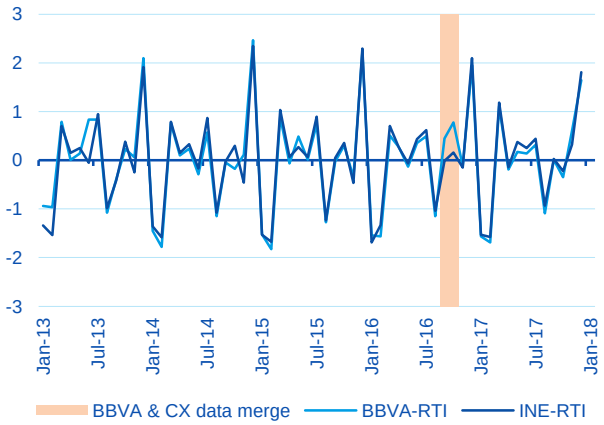
Source: BBVA based on INE

Figure 3.2 BBVA-RTI, average transaction amounts (euros, 12 months moving average)



Source: BBVA

Figure 3.3 Aggregate Retail Trade Indices (standardized monthly growth rate)



Source: BBVA based on INE

Figure 3.4 RTI by Distribution Classe

Figure 3.4a Department Store RTI
(standardized monthly growth rate)

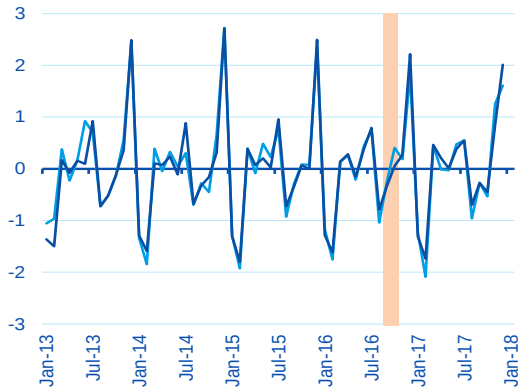


Figure 3.4b Large Chain Store RTI
(standardized monthly growth rate)

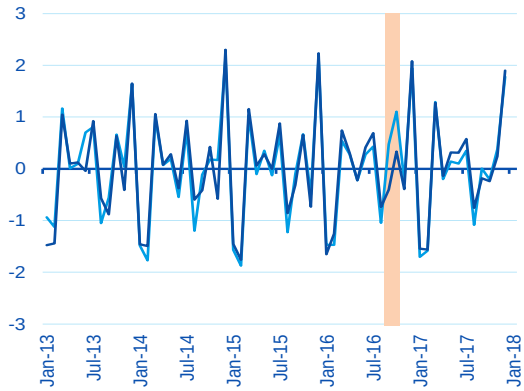


Figure 3.4c Small Chain Store RTI
(standardized monthly growth rate)

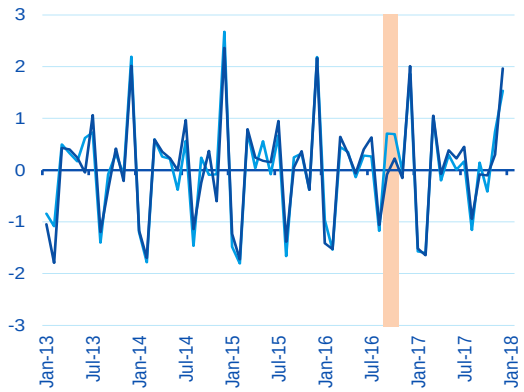


Figure 3.4d Single Retail Store RTI
(standardized monthly growth rate)

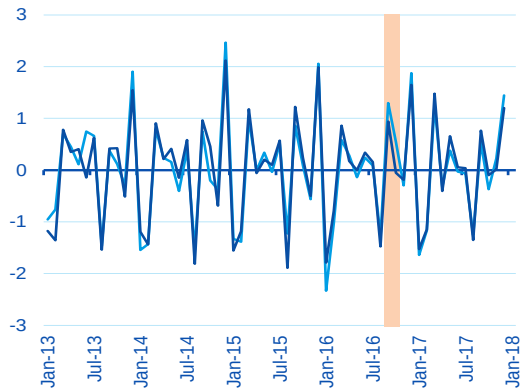
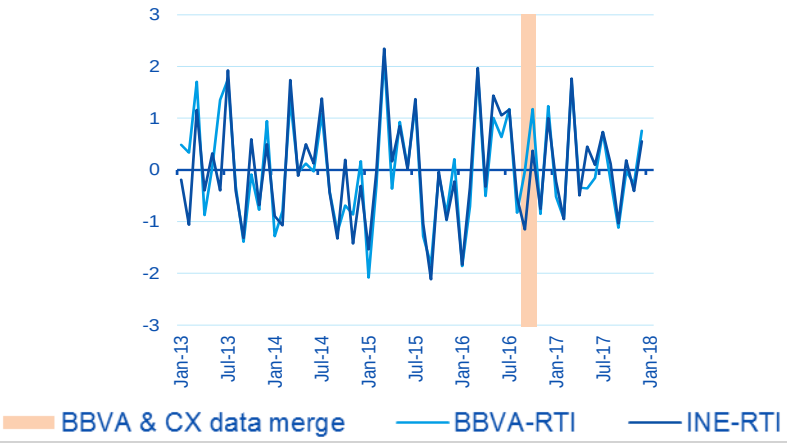


Figure 3.4e Gas Station RTI
(standardized monthly growth rate)



Source: BBVA based on INE

Figure 3.5 RTI by Autonomous Region (standardized monthly growth, %) rate) (1/2)

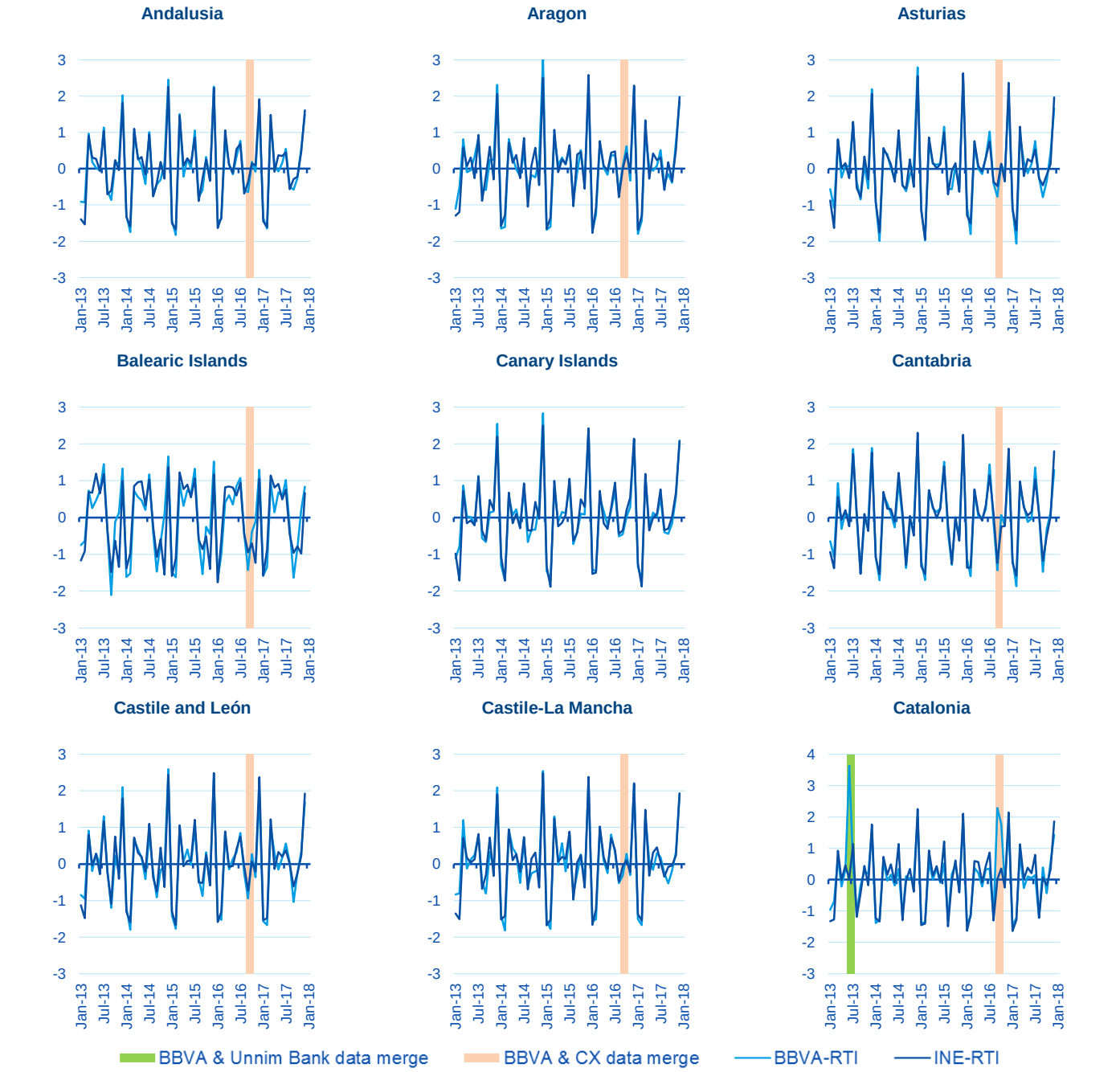
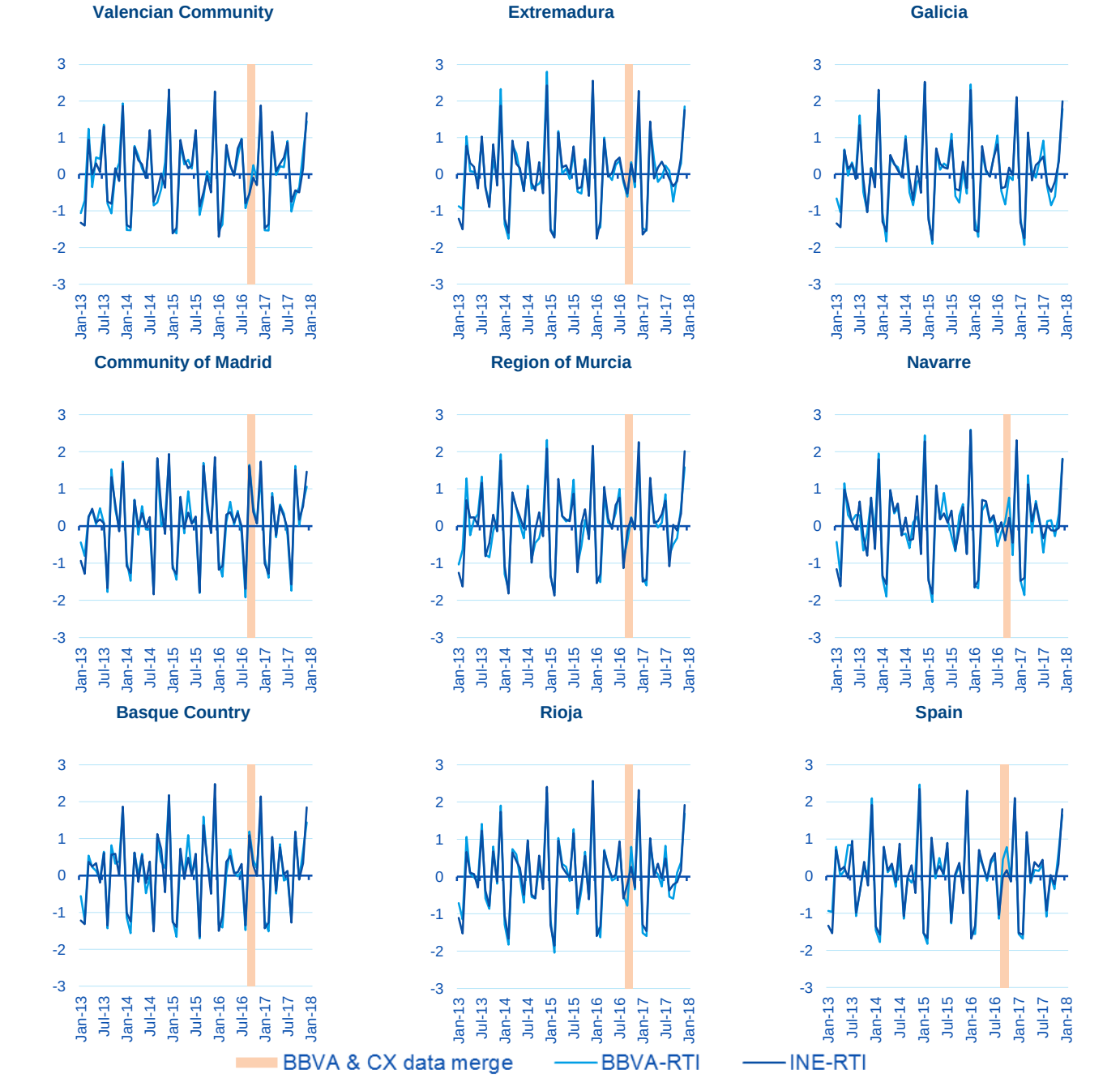
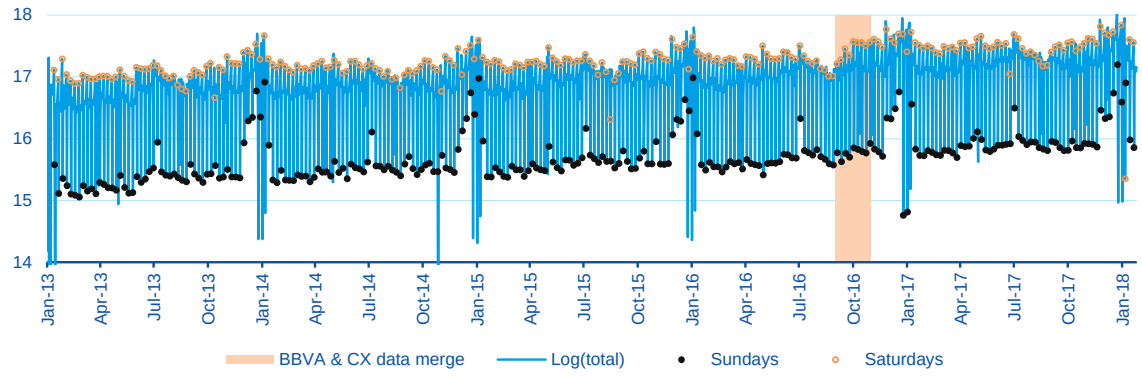


Figure 3.5 RTI by Autonomous Region (standardized monthly growth, %) rate) (2/2)



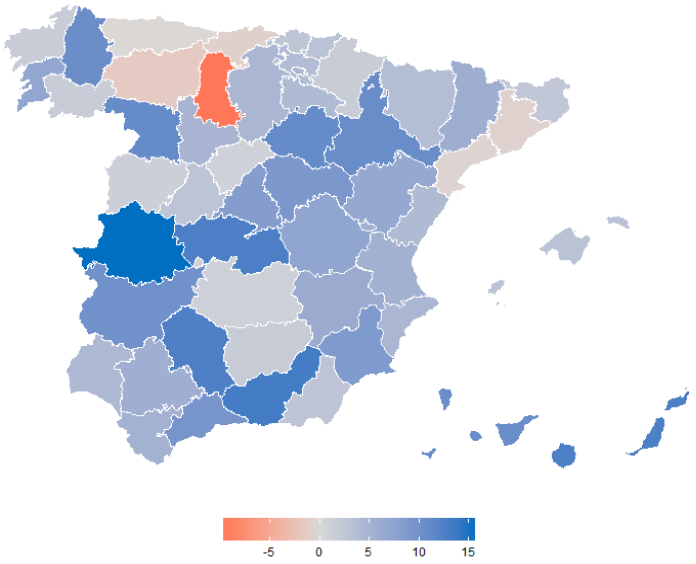
Source: BBVA based on INE

Figure 4.1 Aggregate Retail Trade - Daily Frequency (logarithms)



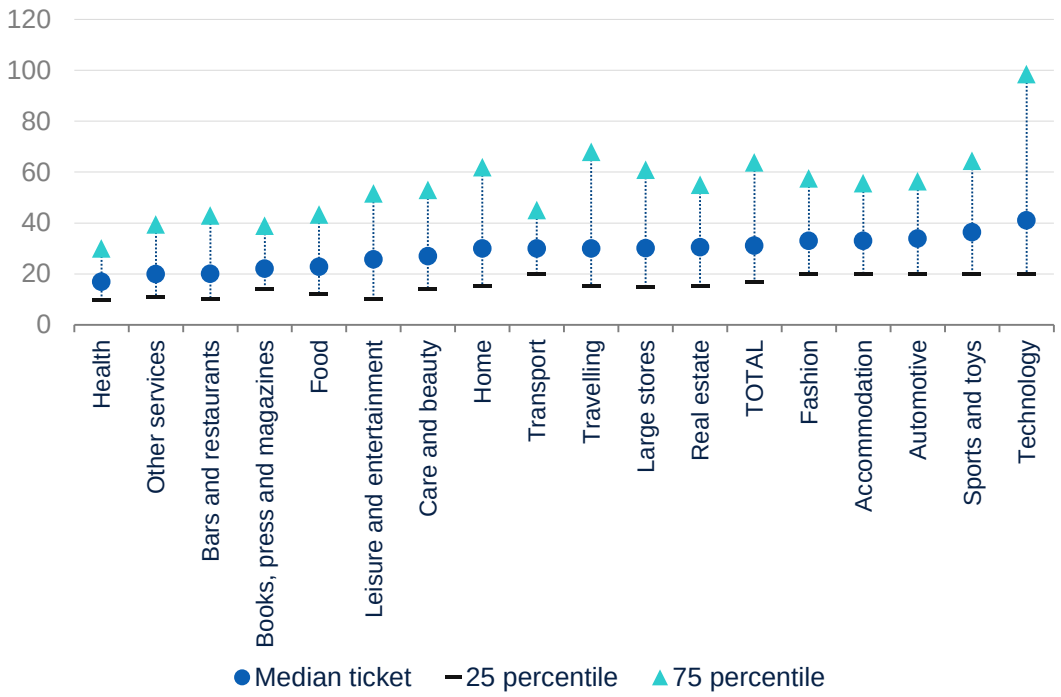
Source: BBVA

Figure 4.2 BBVA-RTI by Province in Dec-17 (% YoY)



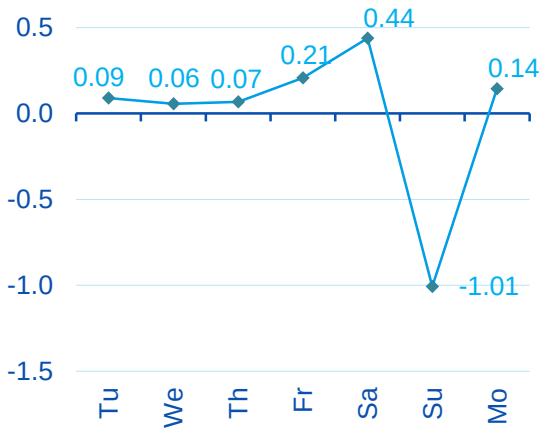
Source: BBVA

Figure 4.3 BBVA-RTI: Median Transaction by Sector of Activity in Dec-17 (Euros)



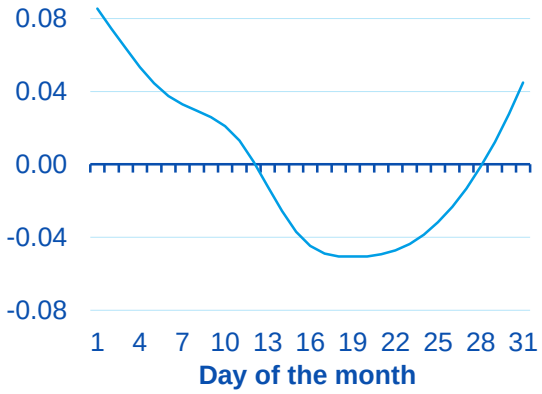
Source: BBVA

Figure 5.1 Intra-Weekly Effects (γ_t^d)



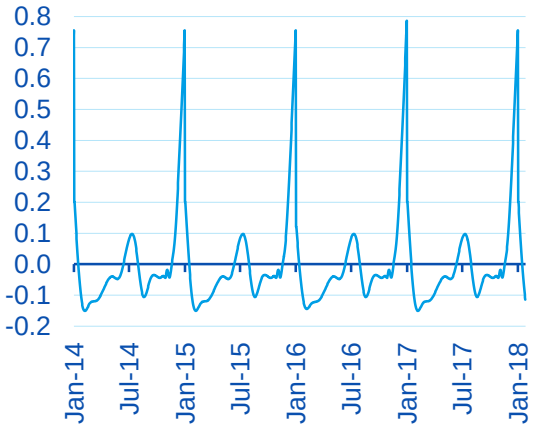
Note: Parameters as estimated for the second week of January 2018.
Source: BBVA

Figure 5.2 Intra-Monthly Effect (γ_t^m)



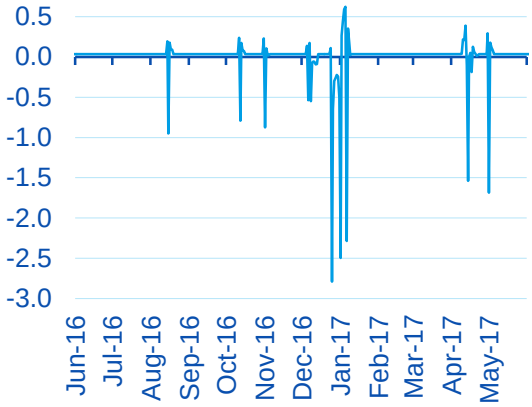
Note: Seasonal pattern for October 2017.
Source: BBVA

Figure 5.3 Intra-Yearly Effect (γ_t^y)



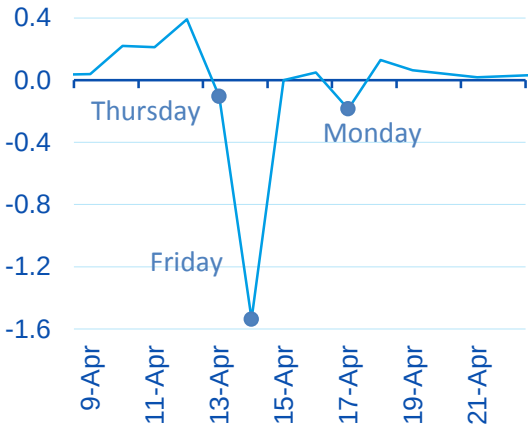
Source: BBVA

Figure 5.4 Holiday Effect (γ_t^h)



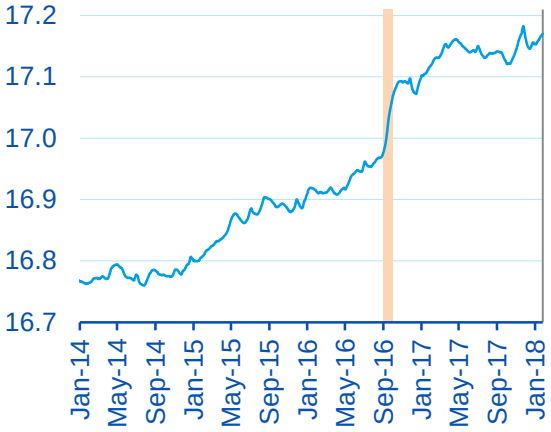
Source: BBVA

Figure 5.5 Easter 2017



Source: BBVA

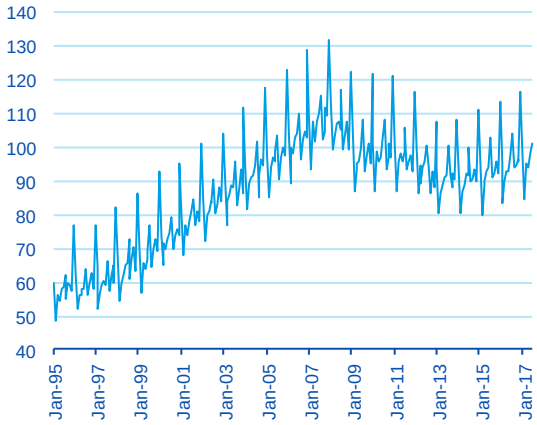
Figure 5.6 Trend (μ_t)



Source: BBVA

Annex

Figure A.1 INE-RTI (nominal and non-seasonal adjusted, base 2010=100)



BBVA | Research

Working Papers

2018

18/03 **Diego Bodas, Juan R. García López, Juan Murillo Arias, Matías Pacce, Tomasa Rodrigo López, Pep Ruiz de Aguirre, Camilo Ulloa, Juan de Dios Romero Palop and Heribert Valero Lapaz:** Measuring Retail Trade Using Card Transactional Data

18/02 **Máximo Camacho and Fernando Soto:** Consumer confidence's boom and bust in Latin America.

18/01 **Ana I. Segovia Domingo and Álvaro Martín Enríquez:** Digital Identity: the current state of affairs.

2017

17/24 **Joaquín Iglesias, Álvaro Ortiz and Tomasa Rodrigo:** How Do the Emerging Markets Central Bank Talk? A Big Data Approach to the Central Bank of Turkey.

17/23 **Ángel de la Fuente:** Series largas de algunos agregados económicos y demográficos regionales: Actualización de RegData hasta 2016.

17/22 **Ángel de la Fuente:** Series enlazadas de algunos agregados económicos regionales, 1955-2014. Parte II: Otras variables de empleo, rentas del trabajo y paro.

17/21 **Ángel de la Fuente:** La evolución de la financiación de las comunidades autónomas de régimen común, 2002-2015.

17/20 **Maximo Camacho, Matias Pacce and Camilo Ulloa:** Business cycle phases in Spain.

17/19 **Ángel de la Fuente:** La liquidación de 2015 del sistema de financiación de las comunidades autónomas de régimen común.

17/18 **Víctor Adame y David Tuesta:** The labyrinth of the informal economy: measurement strategies and impacts.

17/17 **Víctor Adame y David Tuesta:** El laberinto de la economía informal: estrategias de medición e impactos.

17/16 **Liliana Rojas-Suárez y Lucía Pacheco:** Índice de prácticas regulatorias para la inclusión financiera en Latinoamérica: Facilitadores, Promotores y Obstaculizadores.

17/15 **Liliana Rojas-Suárez y Lucía Pacheco:** An Index of Regulatory Practices for Financial Inclusion in Latin America: Enablers, Promoters and Preventers.

17/14 **Ángel de la Fuente:** Las finanzas autonómicas en 2016 y entre 2003 y 2016.

17/13 **Carlos Casanova, Joaquín Iglesias, Álvaro Ortiz, Tomasa Rodrigo y Le Xia:** Tracking Chinese Vulnerability in Real Time Using Big Data.

17/12 **José E. Boscá, Rafael Doménech, Javier Ferri y José R. García:** Los Desplazamientos de la Curva de Beveridge en España y sus Efectos Macroeconómicos.

17/11 **Rafael Doménech y José Manuel González-Páramo:** Budgetary stability and structural reforms in Spain: lessons from the recession and options for the future.

17/10 **Ángel de la Fuente:** Series enlazadas de algunos agregados económicos regionales, 1955-2014. Parte I: Metodología, VAB, PIB y puestos de trabajo.

17/09 **José Félix Izquierdo:** Modelos para los flujos de nuevo crédito en España.

17/08 **José María Álvarez, Cristina Deblas, José Félix Izquierdo, Ana Rubio y Jaime Zurita:** The impact of European banking consolidation on credit prices.

17/07 **Víctor Adame García, Javier Alonso Meseguer, Luisa Pérez Ortiz, David Tuesta:** Infrastructure and economic growth from a meta-analysis approach: do all roads lead to Rome?

17/06 **Víctor Adame García, Javier Alonso Meseguer, Luisa Pérez Ortiz, David Tuesta:** Infraestructuras y crecimiento: un ejercicio de meta-análisis.

17/05 **Olga Cerqueira Gouveia, Enestor Dos Santos, Santiago Fernández de Lis, Alejandro Neut y Javier Sebastián:** Monedas digitales emitidas por los bancos centrales: adopción y repercusiones.

17/04 **Olga Cerqueira Gouveia, Enestor Dos Santos, Santiago Fernández de Lis, Alejandro Neut and Javier Sebastián:** Central Bank Digital Currencies: assessing implementation possibilities and impacts.

17/03 **Juan Antolín Díaz and Juan F. Rubio-Ramírez:** Narrative Sign Restrictions for SVARs.

17/02 **Luis Fernández Lafuerza and Gonzalo de Cadenas:** The Network View: applications to international trade and bank exposures.

17/01 **José Félix Izquierdo, Santiago Muñoz, Ana Rubio and Camilo Ulloa:** Impact of capital regulation on SMEs credit.

2016

16/21 **Javier Sebastián Cermeño:** Blockchain in financial services: Regulatory landscape and future challenges for its commercial application

16/20 **Máximo Camacho and Matías Pacce:** Forecasting travelers in Spain with Google queries.

16/19 **Javier Alonso, Alfonso Arellano, David Tuesta:** Factors that impact on pension fund investments in infrastructure under the current global financial regulation.

16/18 **Ángel de la Fuente:** La financiación regional en Alemania y en España: una perspectiva comparada.

16/17 **R. Doménech, J.R. García and C. Ulloa:** The Effects of Wage Flexibility on Activity and Employment in the Spanish Economy.

16/16 **Ángel de la Fuente:** La evolución de la financiación de las comunidades autónomas de régimen común, 2002-2014.

16/15 **Ángel de la Fuente:** La liquidación de 2014 del sistema de financiación de las comunidades autónomas de régimen común: Adenda.

16/14 **Alicia García-Herrero, Eric Girardin and Hermann González:** Analyzing the impact of monetary policy on financial markets in Chile.

16/13 **Ángel de la Fuente:** La liquidación de 2014 del sistema de financiación de las comunidades autónomas de régimen común.

16/12 **Kan Chen, Mario Crucini:** Trends and Cycles in Small Open Economies: Making The Case For A General Equilibrium Approach.

16/11 **José Félix Izquierdo de la Cruz:** Determinantes de los tipos de interés de las carteras de crédito en la Eurozona.

16/10 **Alfonso Ugarte Ruiz:** Long run and short run components in explanatory variables and differences in Panel Data estimators.

16/09 **Carlos Casanova, Alicia García-Herrero:** Africa's rising commodity export dependency on China.

16/08 **Ángel de la Fuente:** Las finanzas autonómicas en 2015 y entre 2003 y 2015.

16/07 **Ángel de la Fuente:** Series largas de algunos agregados demográficos regionales, 1950-2015.

16/06 **Ángel de la Fuente:** Series enlazadas de Contabilidad Regional para España, 1980-2014.

16/05 **Rafael Doménech, Juan Ramón García, Camilo Ulloa:** Los efectos de la flexibilidad salarial sobre el crecimiento y el empleo.

16/04 **Angel de la Fuente, Michael Thöne, Christian Kastrop:** Regional Financing in Germany and Spain: Comparative Reform Perspectives.

16/03 **Antonio Cortina, Santiago Fernández de Lis:** El modelo de negocio de los bancos españoles en América Latina.

16/02 **Javier Andrés, Ángel de la Fuente, Rafael Doménech:** Notas para una política fiscal en la salida de la crisis.

16/01 **Ángel de la Fuente:** Series enlazadas de PIB y otros agregados de Contabilidad Nacional para España, 1955-2014.

Click here to Access the Working Paper published

Spanish

and English

Contact details:
BBVA Research
Azul Street, 4
La Vela Building - 4th and 5th floors
28050 Madrid (Spain)

Tel.: +34 91 374 60 00 and +34 91 537 70 00
Fax: +34 91 374 30 25
bbvaresearch@bbva.com
www.bbvaresearch.com