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 / Heribert Valero

Measuring Retail Trade Using Card Transactional Data*

Diego Bodas(1), Juan R. García López(2), Juan Murillo Arias(1), Matías J. Pacce(3), Tomasa Rodrigo López(2), Juan de Dios Romero Palop(1), Pep Ruiz de Aguirre(2), Camilo A. Ulloa(2), and Heribert Valero Lapaz(1)

(1) BBVA Data & Analytics, Madrid, Spain
(2) BBVA Research, Madrid, Spain
(3) Banco de España

This version: November 2018

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.

Keywords: retail sales; Big Data; electronic payments; consumption; structural time series model

JEL Classification: C32; C55; C81; E21

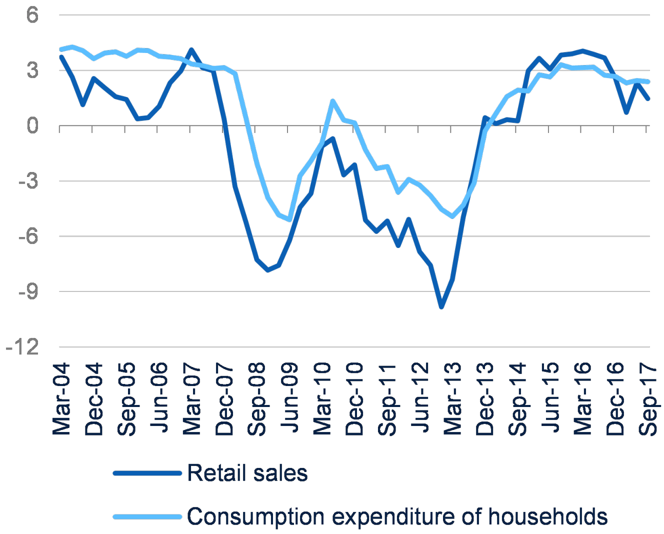
*We are grateful to Gonzalo de Cadenas-Santiago for his contribution in the first stages of the project, Miguel Cardoso and Alvaro Ortiz for their comments, as well as all of those who give us feedback, which greatly improved the quality of this paper. The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the Banco de España or the European System of Central Banks (ESCB).

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 originated from social networks and search engines, have been proven to help the forecasting of economic variables, such as employment, consumption, and tourism flows (see McLaren and Shanbhogue (2011); Choi and Varian (2012); Camacho and Pacce (2018); among many others). More recently, Cavallo and Rigobon (2016) and Cavallo (2018) have shown that the prices of goods and services sold online 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

Figure 1: Spain: Retail Sales vs. Household Consumption Expenditure (%, YoY)



Source: BBVA based on INE.

payment data has already been shown to be helpful in forecasting the evolution of economic aggregates (see Tkacz (2013); Galbraith and Tkacz (2015) or Duarte et al. (2017)). Nonetheless, to the best of our knowledge, this is the first time that the evolution of an official index published by a 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), 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 (1989, 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 economic activity. Besides this important feature, the RTI is also crucial as an indicator for studying the development of the retail sector itself, and the possible disaggregation published by the NSIs (by sectors or regions) is key for a detailed analysis.

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 reproduce the dynamics not only of the aggregate Spanish RTI but also the regional RTI (of the Spanish autonomous regions) and the one by retailer type (distribution classes). A number of benefits emerge from our proposed methodology, which are 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 are 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, base on this daily index, we are able to analyze consumption dynamics by using a daily structural time series model like the one proposed by Harvey et al. (1997). We find regular significant patterns that displayed strong intra-weekly, intra-monthly and intra-yearly seasonalities, which are 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 it is constructed by conducting surveys directed to a limited number of companies selected using random sampling techniques.[1] In other words, the relevant data

[1]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

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 by 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.

2.1 The Data Sources

We analyze a complete set of Point of Sale (PoS) purchases or transactions performed in Spanish retail stores between 2013/01 and 2018/01 by clients of BBVA Spain.[2] For the purposes of this paper, we focus on information[3] 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 does not include expenditure on motor vehicles and motorcycles, food service, the hospitality industry, or financial services, while sales at gas stations are 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 group purchases into 5 distribution classes based on the following categories of retail store:[4]

completing a monthly questionnaire over the telephone, email, fax, or the web.

[2]Both face-to-face and online purchases have been analyzed for this project.

[3]The transactions database has been anonymized and aggregated before analyzing it.

[4]Since information on the number of employees in each company was not available, we did not include

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

We only consider transactions below €$30,000$ (the maximum credit limit for BBVA Gold Cards). Values over that threshold are considered to be outliers. The sample contain approximately 1.2 million of merchants classified in 17 categories and 75 subcategories and more than 900 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).

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 cluster failure or malfunction.

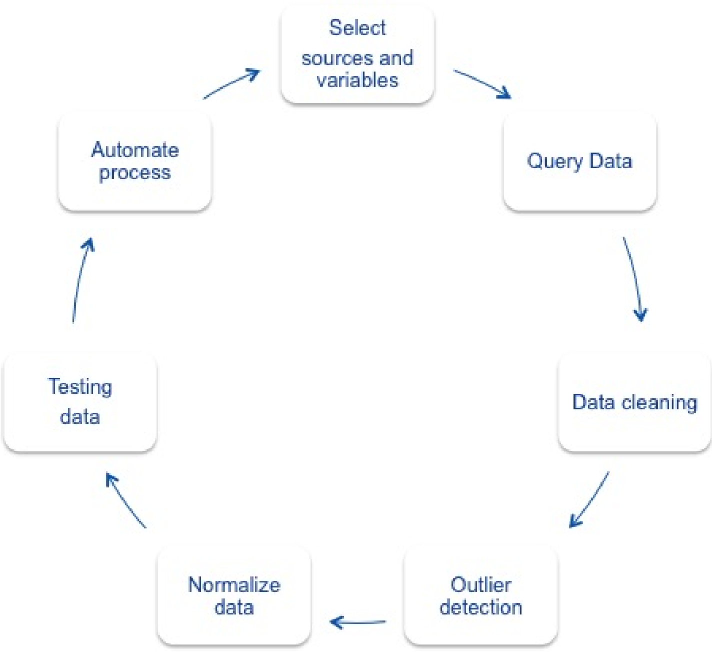
We obtained daily, weekly, and monthly aggregate data on the total number and the total amount of transactions. At the same time, we queried this information at different levels of granularity, getting data for the entire country, data for each of the 17 autonomous regions of Spain, and the 50 provinces.[6] Ensuring data quality required the data to be

the 50 employees' restriction in the definition of a "large chain" or "small chain".

[5]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".

[6]Spain's autonomous regions and provinces correspond, respectively, to NUTS-2 and NUTS-3 in EUROSTAT nomenclature.

Figure 2: Data Engine Building Process.
Data extraction, cleansing, and transformation



Source: BBVA.

cleaned and formatted (during this process, outliers were deleted). 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 1.

The first advantage of using card transaction data is related to the cost of obtaining additional observations. Even though storing huge amounts of information is not cheap, the economic scale of digital information storage means the cost of a 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

Table 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	$\approx 1,2$ million	$\approx 12,000$
Payment methods covered	BBVA's clients credit and debit cards	All
Possible bias of technological trends	Yes	No

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.

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.

It should noticed that, potentially, the INE could also publish a higher disaggregation at the activity or geographical level. However, for the disaggregated series to be accurate, it would be needed a larger sample size –or possible a new design of the sample– which, in turn, would increase the cost of the survey.

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, while INE's data are based on a sample that covers only $12,500$ companies.[7]

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 exclude all transactions made using cash or non-BBVA cards. Nonetheless, given BBVA's high market share (13.8%) in Spain, we assume that the sample we are using is sufficiently representative. A second disadvantage is 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

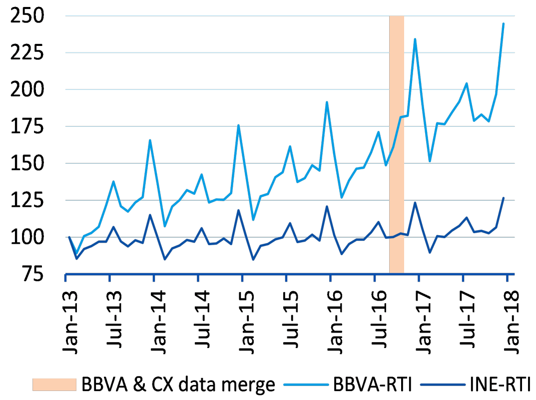
In this section, we show that information obtained from card transaction data can replicate the dynamics of the official RTI for Spain, not only for the national aggregate, but also for all five sub-divisions of the national index and for all 17 official retail trade indices published for each of the 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, we plot the monthly BBVA-RTI next to the nominal non-seasonally adjusted official RTI. As it 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

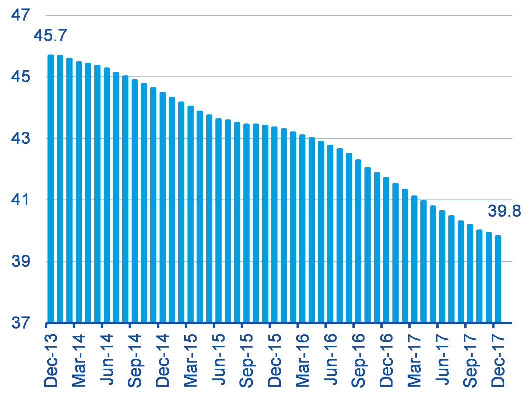
[7] According to the last release of the "Structural Business Statistics" published by the INE (which corresponds to 2016) the are $486,684$ companies that operate at the retail sector. We account for $\approx 300,000$ different tax identification codes (CIF) that amounts for 1.2 million of retailers.

Figure 3: Aggregate Retail Trade Indices (Jan-13 = 100)



Source: BBVA based on INE.

Figure 4: BBVA-RTI, average transaction amounts (euros, 12 months moving average)



Source: BBVA.

debit cards. As an example of this pattern, in Figure 4 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 €39.8 in December 2018. 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,[8] respectively.

Even though Figure 3 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[9] (see Figure A.1 in the Appendix). As both series show a non-stationary pattern, and that our main interested is regarding the signals we can extract from the data, comparing the series in terms of monthly growth rates is even more important than the previous exercise in terms of levels. In Figure 5a, we show how the similarities between the indices become stronger when expressed in standardized[10] monthly growth rates, giving strong support

[8]Our estimates indicate that the relationship between BBVA-RTI's growth rates and the official RTI deviate by 0.47 and 0.59 standard deviations 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.

[9]The official data start in January 1995.

[10]We took natural logs or the first difference of logs, subtracted the mean, and divided by standard

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 Figures 5b-5f, it is possible to see that those similarities are high enough for all 5 disaggregation, 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 coincidence 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 panel of Table 2, we show the R-squared from the linear regression between the BBVA-RTI and the official indices. In the right panel 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.

3.2 Similarities with Regional Indices

As mentioned above, the INE publishes a RTI for each of Spain's 17 autonomous regions. Taking advantage of the high geographical disaggregation allowed by the card transaction data, we constructed each of these 17 indices based on that information. Figure 6 shows the dynamic of all the regional INE-RTI and BBVA-RTI, once again expressed in monthly growth rates. The figures show great similarities in the dynamics of the indices, which

deviations.

Figure 5: Retail Trade Index (RTI): aggregate and by distribution Classes (standardized monthly growth rate)

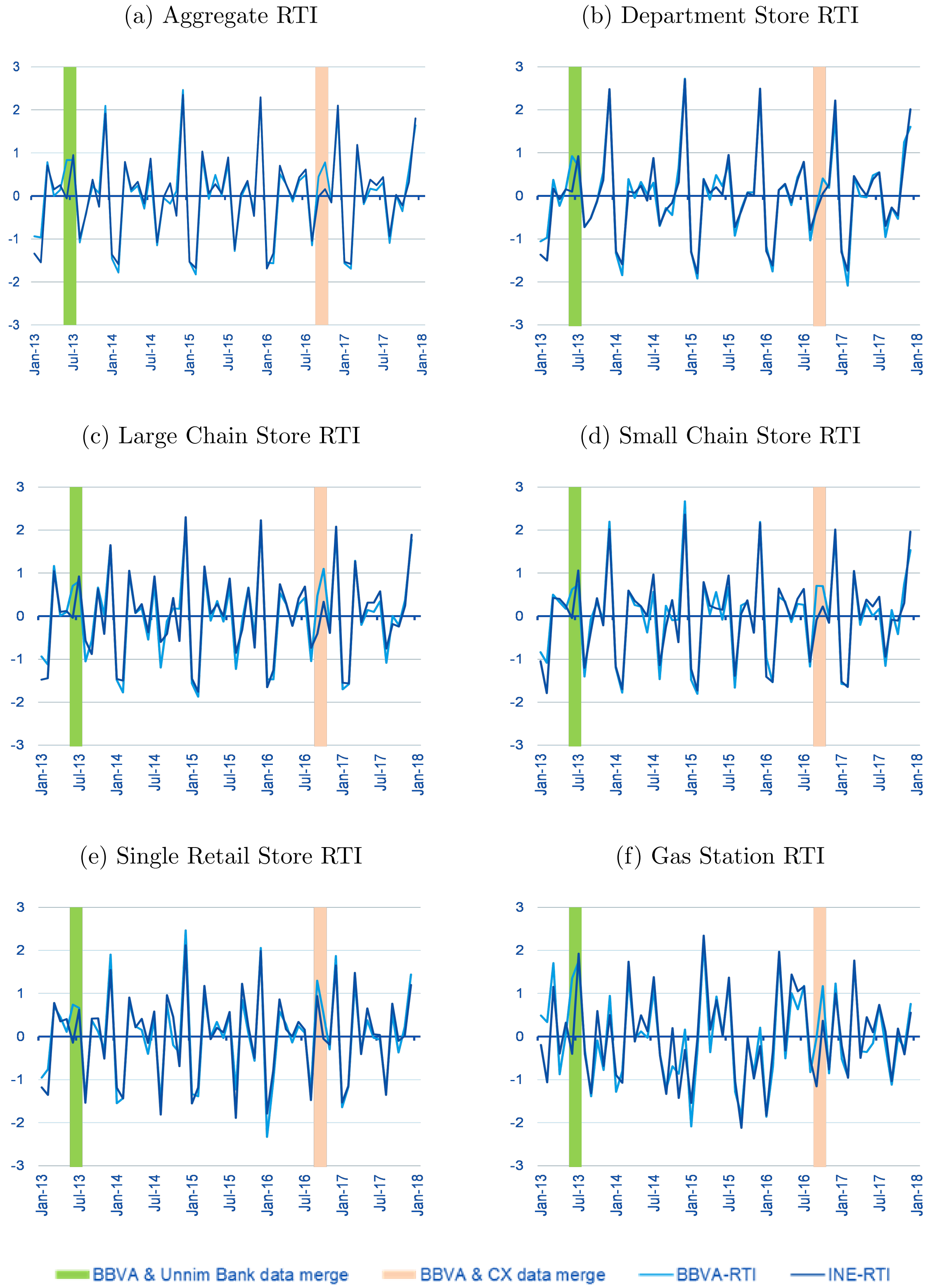


Table 2: INE-RTI and BBVA-RTI

	R-squared		Hansen stability test P-value (H0: parameter stability)	
	Levels	Monthly growth rate	Levels	Monthly growth rate
Total	.89	.94	.22	.90
Department stores	.89	.95	.03	.99
Large chain stores	.73	.91	.32	.93
Small chain stores	.48	.91	.30	.67
Single Retail stores	.57	.92	.21	.85
Service stations	.05	.79	.01	.18

Note: the sample period for the test is 2013.01-2017.12.

are also reflected in the high R-squared and Hansen stability test p values (see Table 3).

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.[11]

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

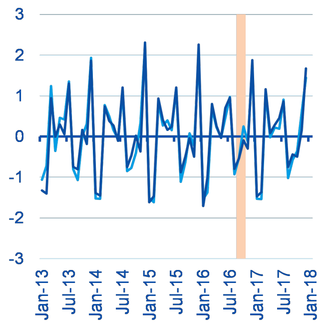
After checking the constructed BBVA-RTI replicates the official figures published by the INE at all levels in which is available, we went one step further in taking advantage of the BBVA quasi-real-time transaction data, getting insights of the retail trade at higher frequencies (e.g., daily) with greater 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).

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

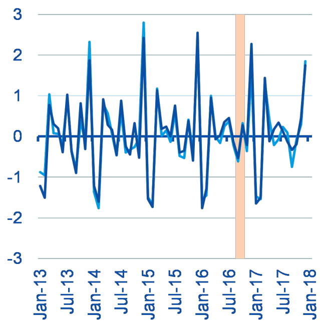
Figure 6: RTI by Autonomous Region (standardized monthly growth rate)

Figure 6: RTI by Autonomous Region (standardized monthly growth rate)

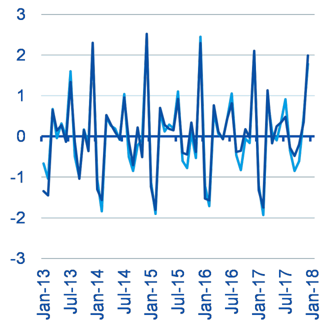
(j) Valencian Community



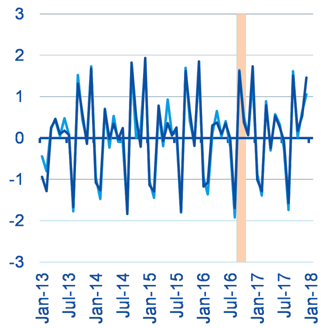
(k) Extremadura



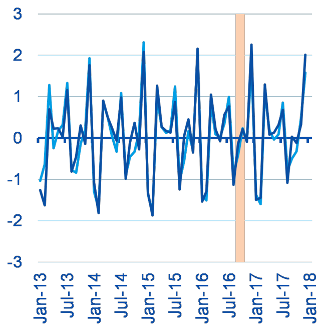
(l) Galicia



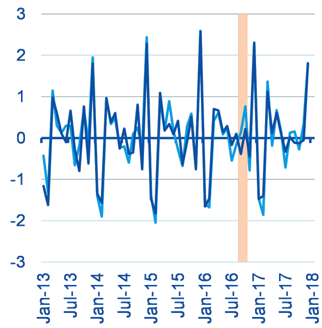
(m) Community of Madrid



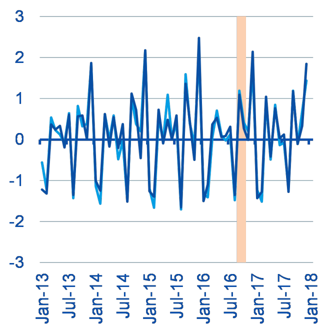
(n) Region of Murcia



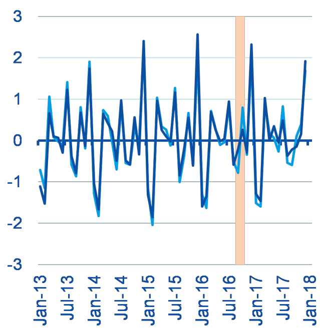
(o) Navarre



(p) Basque Country



(q) Rioja



(r) Spain

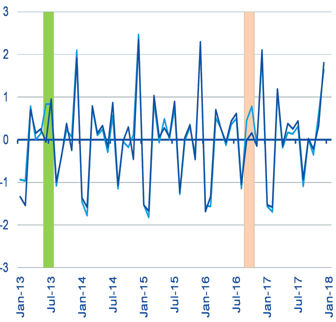


Table 3: INE-RTI and BBVA-RTI

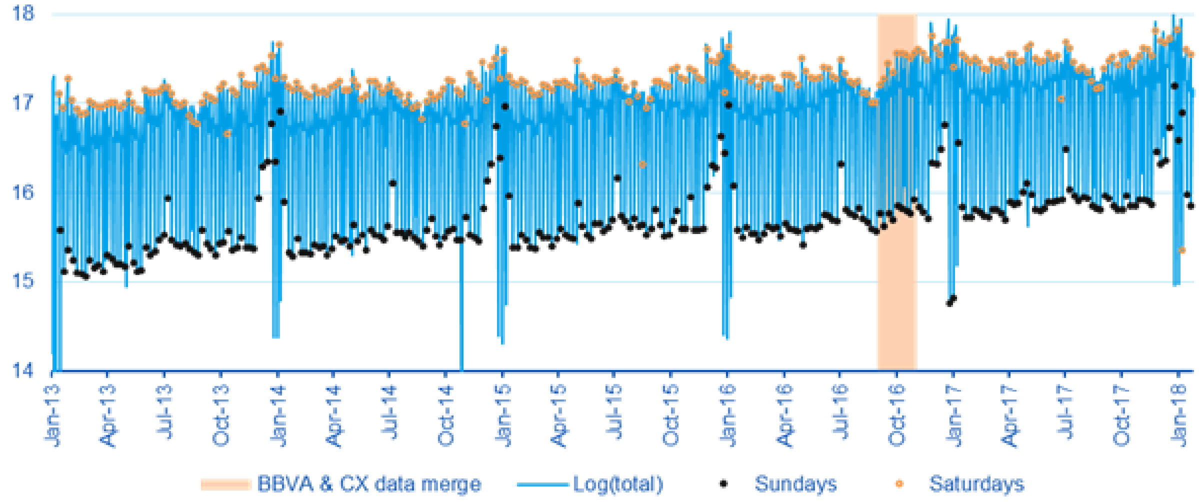
	R-squared		Hansen stability test P-value (H0: parameter stability)	
	Levels	Monthly growth rate	Levels	Monthly growth rate
Andalusia	.72	.96	.14	.32
Aragon	.52	.91	.05	.24
Asturias	.76	.94	.20	.70
Balearic Islands	.72	.90	.19	.96
Canary Islands	.87	.94	.16	.21
Cantabria	.87	.96	.10	.87
Castile and León	.70	.94	.05	.96
Castilla-La Mancha	.65	.93	.09	.86
Catalonia	.38	.83	.37	.23
Valencian Community	.81	.94	.06	.69
Extremadura	.54	.94	.06	.48
Galicia	.82	.95	.07	.51
Community of Madrid	.79	.96	.09	.88
Region of Murcia	.65	.91	.27	.15
Navarre	.78	.91	.08	.27
La Rioja	.78	.95	.05	.82
Basque Country	.82	.95	.47	.38

Note: the sample period for the test is 2013.01-2017.12.

The high frequency of the BBVA-RTI (Figure 7) covers the one-month lag in publication by INE, providing timely answers on retail sales for particular events. It also allows a deeper analysis of the retailers' customers behavior, uncovering the aggregate daily consumption dynamic using a structural time series model like the one proposed by Harvey et al. (1997) (an example is shown in Section 5).

The geo-located information from the PoSs gives a higher geographical disaggregation to the BBVA-RTI, providing information on the evolution of retail sales that is not published by the INE. We have the RTI for the 50 provinces (NUTS 3 geographical division in the EUROSTAT nomenclature), but it would potentially be possible to generate an RTI

Figure 7: Aggregate Retail Trade - Daily Frequency (logarithms)



Source: BBVA.

at the city or postcode level. Figure 8 shows the evolution of the RTI in December 2017 in each province as compared to December 2016 (% YoY). Although it is 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 those autonomous regions with only one province, brings a high likelihood to the rest of the provincial indices.

The transaction data include information on the main activity of each PoS, allowing for a higher economic activity disaggregation than the information published by INE. The analysis of the BBVA-RTI by sector of activity show 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 9).[12]

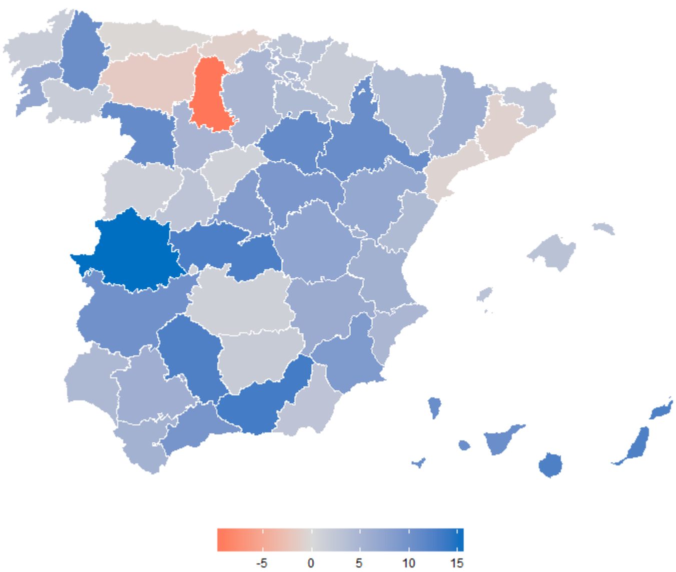
[12]For the sector's analysis we decide to include some sectors that are not taking into account in the official RTI, as "Bars and restaurants" or the "Accommodation" sectors

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 days 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 the Easter Week can be found where the blue line becomes thicker. Modeling all those patterns into one single model that operates at daily frequency is not an easy task and several issues need to be taken into account. Not only the numbers of days within a month or within a year change,[13] but also the position of the date on a specific day of the week[14] or for holidays like Easter is not the same from year to year. As mentioned in Cabrero et al. (2009), two major approaches exist for dealing with those and

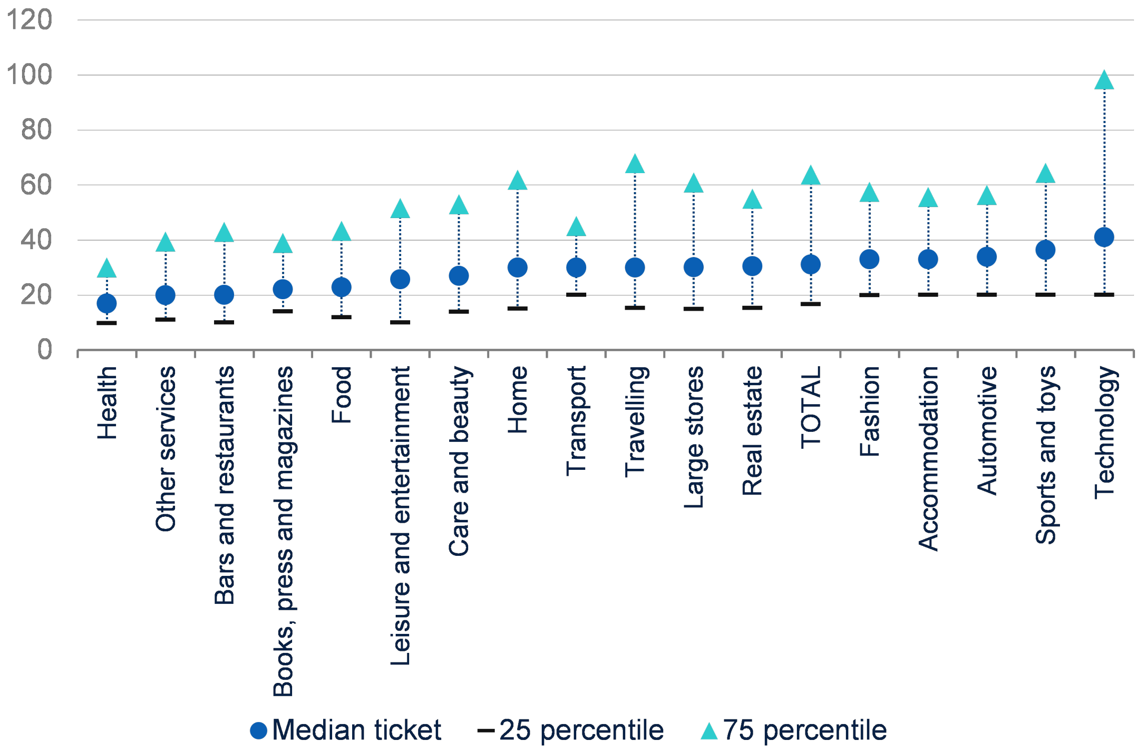
[13]The number of days in a year depends on its being a leap year or not.
[14]E.g., January 1st is not always on a Monday.

Figure 8: BBVA-RTI by Province in Dec-17 (% YoY)



Source: BBVA.

Figure 9: BBVA-RTI: Median Transaction by Sector of Activity in Dec-17(Euros)



Source: BBVA.

other problems in the context of daily time series: the ARIMA model suggested by Bell and Hillmer (1984) and the structural time series (STS) approach of Harvey et al. (1997). In this paper, we rely on the second approach, which includes periodic cubic splines to model some of the seasonal components exhibited by the daily BBVA-RTI data.[15]

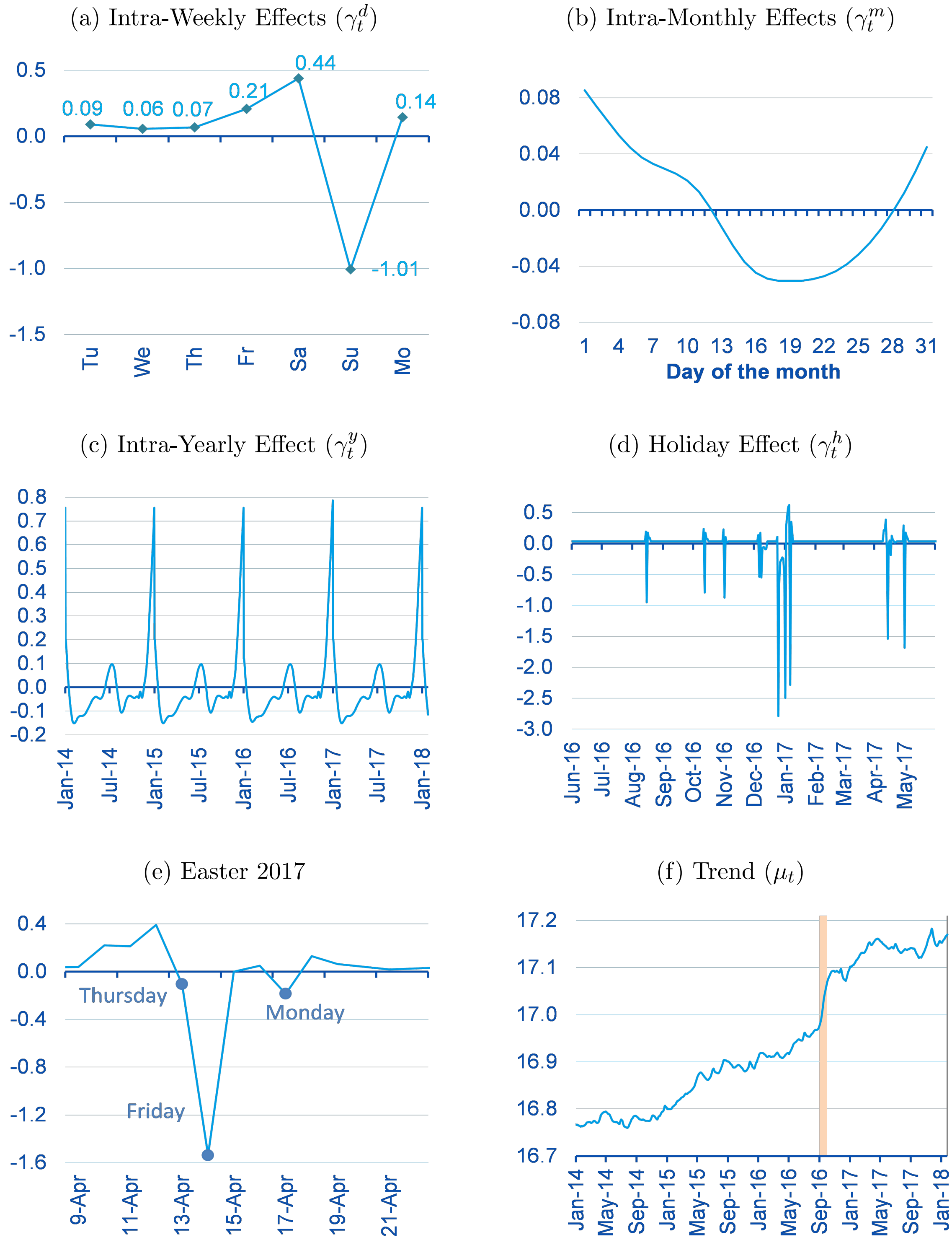
Using Harvey et al. (1997) notation, the basic STS model can be described for a univariate time series (y_t) as

$$y_t = \mu_t + \gamma_t + \varepsilon_t \qquad \varepsilon_t \sim NID\left(0, \sigma_\varepsilon^2\right) \tag{1}$$

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

[15] 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.

Figure 10: Seasonal effects for the daily RTI



(a) Intra-Weekly Effects (γ_t^d)

(b) Intra-Monthly Effects (γ_t^m)

(c) Intra-Yearly Effect (γ_t^y)

(d) Holiday Effect (γ_t^h)

(e) Easter 2017

(f) Trend (μ_t)

Note: parameters for the intra-weekly effects are the ones that correspond to the second week of January 2018. The monthly seasonal pattern correspond to October 2017.

the stochastic trend is defined by,

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

$$\beta_t = \beta_{t-1} + \varsigma_t \qquad \varsigma_t \sim NID\left(0, \sigma_\varsigma^2\right) \tag{3}$$

where the level and the slope of the trend are given by μ_t and β_t, while ν_t and ς_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):[16]

$$\gamma_t = \gamma_t^d + \gamma_t^m + \gamma_t^y + \gamma_t^h \tag{4}$$

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

$$\gamma_t^d = -\sum_{j=1}^{s-1} \gamma_{t-j}^d + \omega_t \qquad \omega_t \sim NID\left(0, \sigma_\omega^2\right) \tag{5}$$

where $s = 7$ is the number of days in a week.[17] 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. (1997) notation, we define

$$\gamma_d^i = \mathbf{z}_d^{i\prime} \gamma_t^i \qquad d = 1, \ldots, N^i \; ; \; i = m, y \tag{6}$$

where $\mathbf{z}_d^i$ is vector of dimension $(h^i - 1) \times 1$, which depends on the number and positioning

[16] For notational simplicity, we have included the fixed and moving festivals as seasonal components, even though both are calendar effects.

[17] Dummies (z_{jt}) in equation 5 are introduced as $\gamma_t^d = -\sum_{j=1}^{s-1} \gamma_{t-j}^d z_{jt} + \omega_t$ for $t = 1, \ldots, T$ where for $t = i, i+s, i+2s, \ldots$ and $i = 1, \ldots, s-1$ the variable z_{jt} is one for $j = i$ and zero for $j \neq i$, while for $t = s, 2s, 3s, \ldots$ and $j = 1, \ldots, s-1$ the value of z_{jt} is equal to minus one.

of the knots and it 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 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 29^{th} when the year is not a leap year or days like April 31^{st}) were considered to be missing values and were easily handled, given that our estimation strategy relied on Kalman filter. 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.[18] 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 1^{st} and December 31^{st} 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.[19] The dates finally chosen for the placement of the first and last knot for the annual periodicity were February 18^{th} and 19^{th}, while the 22^{nd} and 23^{rd} 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 et al. (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

[18] As mentioned in Cabrero et al. (2009), this is strictly true only for the case of deterministic periodic cubic splines.

[19] 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 31^{st}) 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.

Table 4: Fixed Holiday Lag Polynomials

	Ploynomial
Good Friday	$\left(w_0 + w_1 B + \cdots + w_1 2 B^{12}\right) B^{(-6)}$
New Year (Jan 1st), Epiphany (Jan 6th), Labor Day (May 1st), Assumption (Aug 15th), Spain's National Holiday (Oct 12th) and All Saints' Day (Nov 1st)	$\left(w_0 + w_1 B + \cdots + w_4 B^{4}\right) B^{(-1)}$
Immaculate Conception (Dec 8th)	$\left(w_0 + w_1 B + \cdots + w_8 B^{8}\right) B^{(-3)}$

intra-monthly spline were placed at $1, 5, 9, 13, 17, 21, 25$ and 28[20] and for the intra-yearly spline at $1, 25, 50, 75, 125, 150, 200, 225, 251, 286, 296, 301, 306, 313, 320, 330, 345$ and 365.[21] To model holidays effects (γ_t^h), we use a deterministic approach and we include dummy variables for each of those specific days.[22] It should be noted that $\gamma_t^h = \sum_{i=1}^{I} \gamma_t^{h,i}$ where $i = 1, \ldots, I$ is an indicator for each holiday. Under this notation, the holiday's effect was modeled as

$$\gamma_t^{h,i} = w_i(B) h(\tau_i, t) \tag{7}$$

where $w_i(B)$ 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 also show some peculiarities (e.g., people going to the supermarket the day before a holiday)[23] (see table 4 for a description of 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.

Since the whole model described in (1)-(7) can be written in state space form, maximum likelihood estimation in combination with Kalman filter could be used. The main results are summarized in Figures 10a to 10f.

[20] Knots 1 and 28 correspond, respectively, to the 22nd and 23rd day of the month.

[21] Notice that knots 1 and 365 correspond, respectively, to February 18th and 19th

[22] 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.

[23] To model holidays, we also took into account the fact that the holidays effect plus the non-holiday factor should be null (the dummy variables were altered to get this kind of effect).

Figure 10a shows the intra-weekly effects for a week of the year.[24] As observed by the raw data (Figure 7), Sundays are the days of the week with the lowest consumption while Saturdays the highest one. This behavior is not surprising in a country like Spain, where a Sunday is a rest day, 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 10b. 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.[25] 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 and Weinberg (2009) and Aguila et al. (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 actors preferences.

Figure 10c 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 10^{th}. This period is related to the Christmas holidays, when the increase in 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 10d. As expected, all national holidays have a negative effect on retail sales, which is obviously related to the fact that most retail

[24] As we are working with stochastic dummies, the consumption pattern is not exactly similar for every week of the year. Figure 10a shows the intra-weekly behavior for the second week of January, but results are very similar for any week of the year.

[25] In Spain, wages are paid monthly, normally during the first or last week of the month.

stores are closed on those days.[26] 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 25^{th} and January 1^{st} and 6th are the holidays with the highest negative effects on retail sales followed by May 1^{st}.

Given the importance of Easter, Figure 10e 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 10f shows the estimated stochastic daily trend. As expected from dynamic observed in Figure 3, a positive trend was found during the period for which the model is estimated.

The results obtained using the daily STS model should be considered carefully. Since we only have five 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 we develop replicates the dynamics of the aggregate Spanish RTI, the RTI by region (Spains

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

autonomous regions) and the RTI by retailer type (distribution classes). In addition, the higher data granularity allows 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 investigate 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 the days of the week with the lowest consumption while Saturdays are the ones with the highest one), 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 10^{th}, 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. and Perez-Arce, F. (2017), 'Consumption smoothing and frequency of benefit payments of cash transfer programs', *American Economic Review* **107**(5), 430–35.

Bell, W. R. and Hillmer, S. C. (1984), 'Issues involved with the seasonal adjustment of economic time series', *Journal of Business & Economic Statistics* **2**(4), 291–320.

Cabrero, A., Camba-Mendez, G., Hirsch, A. and Nieto, F. (2009), 'Modelling the daily banknotes in circulation in the context of the liquidity management of the european central bank', *Journal of Forecasting* **28**(3), 194–217.

Camacho, M. and Pacce, M. J. (2018), 'Forecasting travellers in spain with googles search volume indices', *Tourism Economics* **24**(4), 434–448.

Cavallo, A. (2018), 'Scraped data and sticky prices', *Review of Economics and Statistics* **100**(1), 105–119.

Cavallo, A. and Rigobon, R. (2016), 'The billion prices project: Using online prices for measurement and research', *Journal of Economic Perspectives* **30**(2), 151–78.

Choi, H. and Varian, H. (2012), 'Predicting the present with google trends', *Economic Record* **88**, 2–9.

Duarte, C., Rodrigues, P. M. and Rua, A. (2017), 'A mixed frequency approach to the forecasting of private consumption with atm/pos data', *International Journal of Forecasting* **33**(1), 61–75.

Galbraith, J. W. and Tkacz, G. (2015), Nowcasting gdp with electronic payments data, Technical report, ECB Statistics Paper.

Harvey, A., Koopman, S. J. and Riani, M. (1997), 'The modeling and seasonal adjustment of weekly observations', *Journal of Business & Economic Statistics* **15**(3), 354–68.

Mastrobuoni, G. and Weinberg, M. (2009), 'Heterogeneity in intra-monthly consumption patterns, self-control, and savings at retirement', *American Economic Journal: Economic Policy* **1**(2), 163–89.

McLaren, N. and Shanbhogue, R. (2011), 'Using internet search data as economic indicators', *Bank of England Quarterly Bulletin* **51**(2), 134–140.

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

Sobolevsky, S., Bojic, I., Belyi, A., Sitko, I., Hawelka, B., Arias, J. M. and Ratti, C. (2015), Scaling of city attractiveness for foreign visitors through big data of human economical and social media activity, *in* 'big data (BigData congress), 2015 IEEE international congress on', IEEE, pp. 600–607.

Stephens, M. (2003), '"3rd of tha month": Do social security recipients smooth consumption between checks?', *American Economic Review* **93**(1), 406–422.

Stephens, M. (2006), 'Paycheque receipt and the timing of consumption*', *The Economic Journal* **116**(513), 680–701.

Stock, J. and Watson, M. (1989), New indexes of coincident and leading economic indicators, *in* 'NBER Macroeconomics Annual 1989, Volume 4', NBER Chapters, National Bureau of Economic Research, Inc, pp. 351–409.

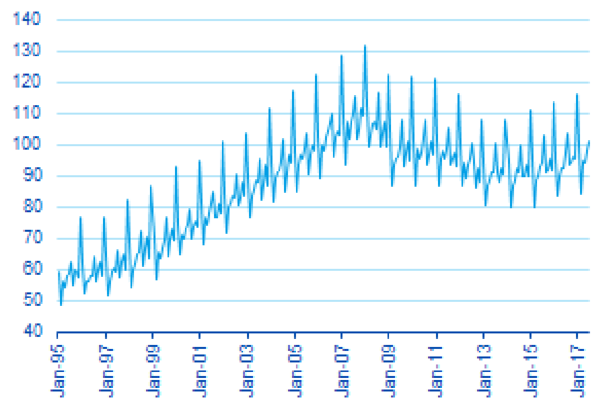
Stock, J. and Watson, M. (1991), A probability model of the coincident economic indicators, *in* G. Moore and K. Lahiri, eds, 'The Leading Economic Indicators: New Approaches and Forecasting Records', Cambridge University Press, pp. 63–90.

Tkacz, G. (2013), 'Predicting recessions in real-time: Mining google trends and electronic payments data for clues', *C.D. Howe Institute Commentary* (387).

Appendices

Appendix A Figures

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



Source: INE.

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