BANCO DE ESPAÑA
Eurosistema

ECONOMIC BULLETIN 2/2019
ANALYTICAL ARTICLES

22 May 2019

Some implications of new data sources for economic analysis and official statistics

Corinna Ghirelli, Juan Peñalosa, Javier J. Pérez and Alberto Urtasun

Abstract

On the back of new technologies, new data sources are emerging. These are of very high frequency, with greater granularity than traditional sources, and can be accessed across the board, in many cases, by the different economic agents. Such developments open up new avenues and new opportunities for official statistics and for economic analysis. From a central bank's standpoint, the use and incorporation of these data into its traditional tasks poses significant challenges, arising from their management, storage, security and confidentiality. Further, there are problems with their statistical representativeness. Given that these data are available to many agents, and not exclusively to official statistics institutions, there is a risk that different measures of the same phenomenon may be generated, with heterogeneous quality standards, giving rise to confusion among the public. Some of these sources, which consist of unstructured data such as text, require new processing techniques so that they can be integrated into economic analysis in an appropriate format (quantitative). In addition, their use entails the incorporation of machine learning techniques, among others, into traditional analysis methodologies. This article reviews, from a central bank's standpoint, some of the possibilities and implications of this new phenomenon for economic analysis and official statistics, with examples of recent studies.

Keywords: new sources of economic information, big data, data science, machine learning, text analysis.

JEL codes: C10, C18, C50, C80 y D80.

SOME IMPLICATIONS OF NEW DATA SOURCES FOR ECONOMIC ANALYSIS AND OFFICIAL STATISTICS

The authors of this article are Corinna Ghirelli, Juan Peñalosa, Javier J. Pérez and Alberto Urtasun, of the Directorate General Economics, Statistics and Research.

Introduction

Over the past decade, the development of new technologies and social media has given rise to new data sources, commonly known as "big data". These new data sources have specific characteristics in terms of volume and level of detail (far greater than those of traditional sources), their high frequency and their often unstructured nature (not necessarily numerical or organised, such as data from text or images). In recent years, a large number of applications have emerged for these new data sources in the area of economics and finance, particularly in central banks, both for expanding the base data they use to carry out their functions and for the economic and prudential analysis and banking supervision work they engage in (see, for example, Broeders and Prenio [2018] and Fernández [2019]). The new data sources have also given rise to key methodological developments (see Fernández-Villaverde *et al.* [2019]).

In the specific area of economic analysis, the new data sources have significant potential, even taking into account that central banks already make very intensive use of statistical data, both individual (microdata) and aggregate (macroeconomic) to perform their functions. In particular, these new sources allow for:

— a better understanding and more agile monitoring of economic reality, with a shorter time lag for the phenomena to be analysed and a greater level of detail. Specifically, relevant information becomes available, in some cases in real time, for the economic forecasting of key variables, such as GDP, private consumption or employment.

— information to be gathered on variables that are difficult to measure (such as sentiment or expectations), but are essential for economic agents' decision-making.

— better assessment of economic policy and more possibilities of simulating alternative measures, owing chiefly to the availability of microdata that could be used to improve the characterisation of agents' heterogeneity and, thus, to conduct a more in-depth and accurate analysis of their behaviour.

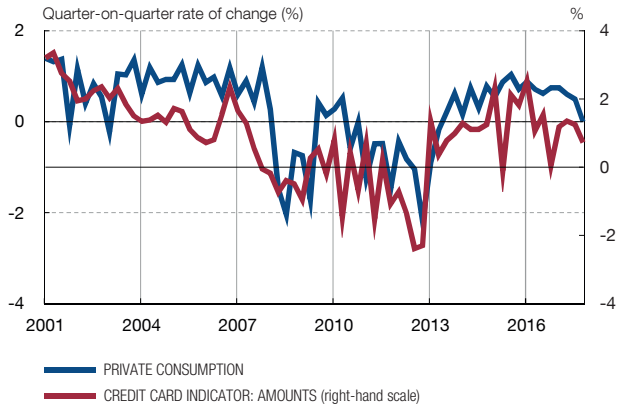
The rest of the article is divided into two parts. The first, which includes boxes 1 and 2, sets out the main contributions of the recent literature, and highlights some innovative applications. The latter part examines the implications of this new phenomenon for official statistics.

New data sources for economic analysis

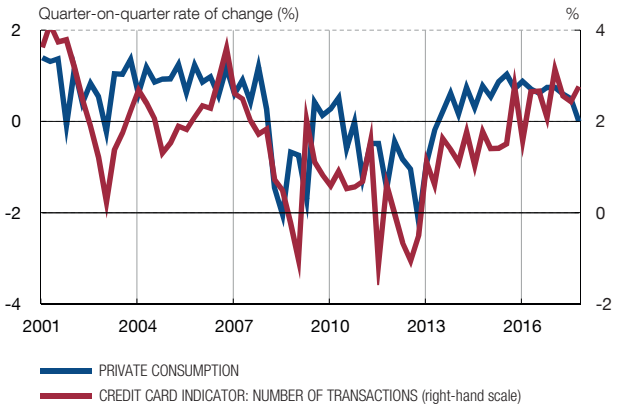
Central banks make intensive use of structured databases to carry out their functions. Structured data are data that have already been classified and organised, and can be readily used by central bank economists. Some examples of individual (microeconomic) data are firms' balance sheets (for example, from the Banco de España's Central Balance Sheet Data Office; see Menéndez and Mulino [2018] or Banco de España [2018]), information relating to the volume of credit granted by financial institutions to individuals and firms (for example, the studies carried out using the Banco de España's Central

PRIVATE CONSUMPTION AND CREDIT CARD TRANSACTIONS CHART 1

1 CREDIT/DEBIT CARD INDICATOR: AMOUNTS (a)



2 CREDIT/DEBIT CARD INDICATOR: NUMBER OF TRANSACTIONS (a)



SOURCES: Quarterly National Accounts (INE) and Gil, Pérez, Sánchez and Urtasun (2018).

a Aggregate for expenditure at points of sale and cash withdrawals at ATMs.

Download

Credit Register[1]) or the data relating to agents' financial decisions in the Spanish Survey of Household Finances (see Banco de España, 2017). These databases are normally published annually or quarterly. In the area of macroeconomics, the main source of information is the National Accounts which, in the case of Spain, are prepared and published by the National Statistics Institute (INE, by its Spanish abbreviation) and the Banco de España, although a great deal of other information on the economic and financial situation is also published, primarily on a monthly basis, but also with a higher frequency.

On the back of new technological developments, the sources of information can be expanded, with greater granularity and higher frequency. Greater granularity is associated with the volume of available information. Thanks to new technologies, information can be obtained about every single action taken by an individual or firm (that is, at the most disaggregated level), for example, transactions conducted using bank cards. Higher frequency refers to the speed at which such data are updated, daily or sometimes even in real time. To continue with the above example, credit card transaction data, which can be used to approximate household consumption patterns, are potentially available in real time at a very reduced cost in terms of use, particularly when compared with the cost of conducting country-wide household surveys. By way of example, Chart 1 shows how credit card transactions performed very similarly to household consumption in Spain (see Gil *et al.* [2018] and Bodas *et al.* [2018]).

The availability of vast quantities of new information poses significant challenges in terms of the management, storage, security and confidentiality infrastructure required:

— First, the orderly management of different types of data requires a flexible infrastructure that stores both structured (table format) and unstructured (no

1 The Central Credit Register of the Banco de España (CIRBE, by its Spanish abbreviation) provides confidential data and highly detailed information on almost all loans extended by credit institutions in Spain (all loans exceeding €6,000).

specific format) data, and provides analytical tools to view any relevant information contained in such data.

— It is also necessary to develop adequate security systems to protect the privacy of the different databases.

— Owing to their high frequency, the volume of unstructured data grows exponentially and thus requires a system that minimises storage costs, without compromising security needs.

— Lastly, optimal management of unstructured data requires the integration of new professional profiles (data scientists) at central banks and closer collaboration between the different areas, such as information systems, statistics and economic analysis and research.

Moreover, the diverse nature of the new information sources requires the assimilation and development of techniques that transform and synthesise data, in formats that can be incorporated into economic analysis. In the case of unstructured data, the information gathered is of no particular value until it has been processed. For example, textual analysis techniques enable the information contained in text to be processed and converted into structured data (for a brief description of these methodologies, see Box 2). These new information types notably include:

— *Google Trends.* This Google tool provides access to the searches carried out by web users on keywords of interest.

— *Media online databases.* Many web servers store text from different information sources, such as newspaper and magazine articles. Information can be extracted on topics published by newspapers and on the treatment they are given.

— *Social media* (for example, Facebook and Twitter). From the messages posted by social media users, the prevailing opinions and the general tone of online debates can be extracted.

— *Web search portals.* These include, for example, portals created for housing or job searches and allow information to be extracted about the real estate and labour markets, respectively.

— *Mobile phone data.* Through mobile applications and the combination of geolocation data, the use of the mobile provides insight into the habits, activities and movements of users.

— *Satellite data.* Satellite images allow, for example, to measure agricultural areas in less developed countries or night-time electricity consumption[2].

However, the new data sources entail certain problems and should be used in an informed manner. In particular (see Einav and Levin, 2014), there may be problems with the

2 Night-time electricity consumption is considered a good proxy for a country's level of development and has proved to be particularly useful in the case of developing countries with scant availability of traditional national accounts data.

representativeness of samples, for example, when they provide information about users of new technologies but do not necessarily represent the target population. Further, the abundance of unstructured information available requires the use of statistical and machine learning techniques to summarise it, and thus, automated predictive models gain prominence over the approximation approach of traditional empirical analysis in which economic theory usually guides analysts in their choice of variables. Thus, in predictive models using big data, it is usually the model that selects the most relevant variables[3]. Lastly, these sources are new and it is difficult to determine to what extent their representativeness will be maintained over time. This poses risks for investment, in terms of the research that can be carried out to exploit specific sources.

Box 1 reviews a broad range of recent papers in which new techniques and new data sources are applied to economic analysis.

Implications of new data sources for official statistics

The new data sources also open up untapped possibilities for the compilation of statistics[4]. Admittedly, the ability of these sources to affect the area that is most specific to central banks, that of financial statistics, is limited, since statistics are largely based on administrative registers (such as bank balances or firms' reporting on their cross-border transactions) that cannot be easily replaced by alternative sources and whose availability is assured since agents are obliged to regularly report such information to regulatory and supervisory authorities. Compared with these structured sources, which have long been used for statistical purposes, the new data (mostly unstructured) would have a largely complementary role, possibly geared towards very specific or more qualitative purposes (such as detecting information gaps in certain segments).

It is important to note that the field of statistics has been working with granular information for a long time. Some statistics have a very high level of detail, as in the case of issuance and holdings of securities (which provide a breakdown by security), individual loans extended by the banking system, transactions and positions of domestic agents vis-à-vis the rest of the world, individual balance sheets of non-financial corporations, etc. Also, in central bank statistics, there is a growing trend to obtain and use such microdata, which may contribute to improving the quality of official statistics, since they enable a more accurate comparison of source data and analysis of consistency with the aggregate variables. Moreover, the availability of this information means that different features can be analysed (instruments, maturities, counterparties or denomination currencies, among many others) without having to request the information again, saving costs for those reporting it, since microdata normally have multiple dimensions that can be used for analysis. At the same time, the increased availability of microdata requires the development of new technical resources to ensure that their use leads to higher quality statistics.

The boom in new data sources has had a positive effect for official statistics in that technical tools are currently being developed to deal with the vast amount of information. These new tools (which include artificial intelligence techniques, machine learning and data analytics) can be used by official statistics to process structured microdata, especially to enhance their quality (for example, to detect and remove outliers) or to reconcile information received from different sources with different frequency.

3 See Belloni *et al.* (2012) and Belloni, Chernozhukov and Hansen (2014), as examples of applications in which automatic learning models improve causal inference studies, identifying the ideal number of control variables or instruments from a vast number of potential variables

4 See Bean (2016), inter alia.

Perhaps more efforts have been made to exploit the new data sources in the area of non-financial statistics. Initiatives such as measuring prices (using web scraping techniques) or certain external trade items (for example, estimating tourist movements by tracking mobile networks). Developing countries, which face greater difficulties in setting up solid statistics infrastructures, are starting to use the new data sources, even to make estimates of some National Accounts aggregates (see Hammer *et al* [2017]).

Finally, it should be pointed out that the huge quantities of available data also pose a challenge for official statistics, since they can be used by individuals to generate their own measures of economic phenomena and to publish this information. To counteract the potential competition from the private sector to generate "statistical" information, which could be based on sources as unreliable as the social media, the quality and transparency framework of official statistics needs to be strengthened. Official statistics are based on an internationally consolidated and comparable methodology that serves as the basis for objectively assessing the economic situation and the response of economic policy. In this regard, statistical authorities should be transparent in disclosing the methods used to compile official statistics, so as to counter the potential effects of "fake news" on measures of economic variables, and draw up a communication policy that will debunk any falsehoods before they have a chance to spread.

22.5.2019.

REFERENCES

M. ACCORNERO and M. MOSCATELLI (2018). *Listening to the Buzz: Social Media Sentiment and Retail Depositors,* Working Paper 1165, Banca d'Italia, February.

V. APRIGLIANO, G. ARDIZZI and L. MONTEFORTE (2017). *Using the Payment System Data to Forecast the Italian GDP,* Working Paper 1098, Banca d'Italia.

C. ARTOLA and E. GALÁN (2012). *Tracking the future on the web: construction of leading indicators using Internet searches,* Occasional Paper 1203, Banco de España.

S.R. BAKER, N. BLOOM and S. J. DAVIS (2016). "Measuring Economic Policy Uncertainty", *The Quarterly Journal of Economics,* 131, pp. 1593-1636.

BANCO DE ESPAÑA (2018). *Central Balance Sheet Data Office. Annual results of non-financial corporations, 2017.*

— (2017). *Survey of Household Finances, 2014:* methods, results and changes since 2011, Analytical Articles, Banco de España, January.

BEAN, C. (2016). *Independent Review of UK Economic Statistics,* ONS, United Kingdom.

BELLONI, A., D. CHEN, V. CHERNOZHUKOV and C. HANSEN, (2012). "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain", *Econometrica,* 80 (6), pp. 2369-2429.

BELLONI, A., V. CHERNOZHUKOV and C. HANSEN (2014). "Inference on Treatment Effects After Selection Among High-Dimensional Controls", *The Review of Economic Studies,* 81 (2), pp. 608-650.

BHOLAT, D., S. HANSEN, P. SANTOS and C. SCHONHARDT-BAILEY (2015). *Text mining for central banks: handbook,* Centre for Central Banking Studies (33), pp. 1-19.

BODAS, D., J. GARCÍA, J. MURILLO, M. PACCE, T. RODRIGO, P. RUIZ, C. ULLOA, J. ROMERO and H. VALERO (2018). *Measuring Retail Trade Using Card Transactional Data,* Working Paper 18/03, BBVA Research.

BROEDERS, D., and J. PRENIO (2018). "Innovative Technology in Financial Supervision (Suptech) - The Experience of Early Users", *Financial Stability Institute Insights on policy implementation,* no. 9, Bank for International Settlements, July.

CENTRAL BANKING (2019). *UK Statistics Body to Launch High-Speed Indicators,* available at https://www.centralbanking.com/central-banks/economics/data/4094151/uk-statistics-body-to-launch-high-speed-indicators.

CHOI, H., and H. VARIAN (2012). "Predicting the Present with Google Trends", *Economic Record,* 88(1), pp. 2-9.

EINAV, L., and J. LEVIN (2014). "The Data Revolution and Economic Analysis", *Innovation Policy and the Economy,* 14, pp. 1-24.

FERNÁNDEZ, A. (2019). "Artificial intelligence in financial services", Analytical Articles, *Economic Bulletin,* Banco de España .

FERNÁNDEZ VILLAVERDE, J., S. HURTADO and G. NUÑO (2019). *Financial Frictions and the Wealth Distribution,* mimeo, Banco de España.

GARCÍA-URIBE, S. (2018). *The Effects of Tax Changes on Economic Activity: a Narrative Approach to Frequent Anticipations,* Working Paper 1828, Banco de España.

GHIRELLI, C., J. J. PÉREZ and A. URTASUN (2019). *A New Economic Policy Uncertainty Index for Spain,* Working Paper 1906, Banco de España. Revised version forthcoming in *Economic Letters*.

GIL, M., J. J. PÉREZ, A. J. SÁNCHEZ and A. URTASUN (2018). *Nowcasting Private Consumption: Traditional Indicators, Uncertainty Measures, Credit Cards and Some Internet Data,* Working Paper 1842, Banco de España.

HAMMER, C. L., D. C. KOSTROCH and G. QUIRÓS (2017). *Big Data: Potential, Challenges and Statistical Implications,* IMF Staff Discussion Note, 17/06.

LOBERTO, M., A. LUCIANI and M. PANGALLO (2018). *The Potential of Big Housing Data: an Application to the Italian Real-Estate Market,* Working Paper 1171, Banca d'Italia.

MENÉNDEZ, Á. and M. MULINO (2018). "Results of non-financial corporations in the first half of 2018" in *Economic Bulletin*, 3/2018, Banco de España.

NYMAN, R., S. KAPADIA, D. TUCKETT, D. GREGORY, P. ORMEROD and R. SMITH (2018). *News and Narratives in Financial Systems: Exploiting Big Data for Systemic Risk Assessment,* Staff Working Paper 704, Bank of England.

THORSRUD, L. (2018). "Words are the New Numbers: A Newsy Coincident Index of the Business Cycle», *Journal of Business & Economic Statistics.*

TURRELL, A., B. SPEIGNER, J. DJUMALIEVA, D. COPPLE and J. THURGOOD (2018). *Using Job Vacancies to Understand the Effects of Labour Market Mismatch on UK Output and Productivity,* Staff Working Paper 737, Bank of England, July.

Newspaper articles and other financial texts. Baker, Bloom and Davis (2016) have constructed an index of economic policy uncertainty *(Economic Policy Index,* EPU) for the United States, based on the volume of newspaper articles that contain words relating to the concepts of uncertainty, economy and policy (for Spain, see Ghirelli *et al.,* 2019). Thorsrud (2018) uses the articles published in Norway's leading business newspaper to construct an indicator of significant economic and financial topics and shows that this information, together with traditional variables, improves GDP forecasting. García-Uribe (2018), on the basis of data relating to United States news, develops an indicator of announcements relating to tax change approvals, showing that news about anticipated tax cuts stimulate economic activity. Lastly, Nyman *et al.* (2018) construct an index relating to financial market sentiment ("anxiety" and "excitement") based on communications issued by the Bank of England, reports by financial institutions and Reuters news items. This indicator has explanatory power to anticipate movements in economic and financial variables.

Twitter. Accornero and Moscatelli (2018) develop an economic sentiment indicator based on tweets that mention the main Italian financial institutions, and show how analysis of the tone of tweets can serve to enhance forecasts of changes in retail deposits.

Google Trends. Choi and Varian (2012) show how Google Trends indicators can improve short-term predictions of private consumption in the case of purchases planned in advance (for example, durable goods). Gil *et al.* (2018) use this tool to develop indices of durable and non-durable goods consumption which can anticipate agents' consumption decisions. Artola and Galán (2012) employ the same tool to construct an indicator of British tourist inflows to Spain (the Spanish tourist industry's main customers), which enhances the prediction of demand and activity variables in Spain.

Credit cards. Aprigliano *et al.* (2017) show how, based on electronic payment system data, information about electronic payment flows improves GDP forecasting compared with models using traditional economic cycle indicators. Gil *et al.* (2018) assess the contribution of a broad range of new data sources to the short-term prediction of household consumption. The results of this study underline the high predictive capacity of credit card transaction indicators (see Chart 1 of the main text). Other relevant papers include Bodas *et al.* (2018). These authors use data relating to all BBVA debit and credit card transactions in Spain to replicate the INE's standard retail sales index, which also allows them to construct retail consumption indices for areas for which no official statistics are available, such as the regions or sectors.

Online search portals. On the basis of housing sales and rental advertisements in Italy's main real estate services online portal, Loberto *et al.* (2018) analyse house prices in specific real estate market segments and identify the housing features that are most relevant to explaining their price. Regarding the labour market, Turrell *et al.* (2018) use job adverts posted on a recruitment portal to analyse labour market mismatch, the increase in occupational or regional mismatch, and recent developments in the UK labour market.

Applications involving text analysis (text mining) have gained special significance in the area of economic analysis. With these techniques, relevant information can be obtained from texts, and then synthesised and codified in the form of quantitative indicators. First, the text is prepared (pre-processing), specifically removing the part of the text that does not inform analysis (articles, non-relevant words, numbers, odd characters) and word endings, leaving only the root. Second, the information contained in the words is synthesised using quantitative indicators obtained mainly by calculating the frequency of words or word groups. Intuitively, the relative frequency of word groups relating to a particular topic allows for the relative significance of this topic in the text to be assessed.

Text mining techniques can be summarised as follows:

— *Searches using logical operators* (Boolean)[1]. Searches for keywords in texts using logical operators (notably "and" and "or"), which establish straightforward relationships between the search terms. The result consists of the total number of texts that meet the search requirements. Such searches can be carried out using search engines or database directories and do not require the researcher to have access to full texts.

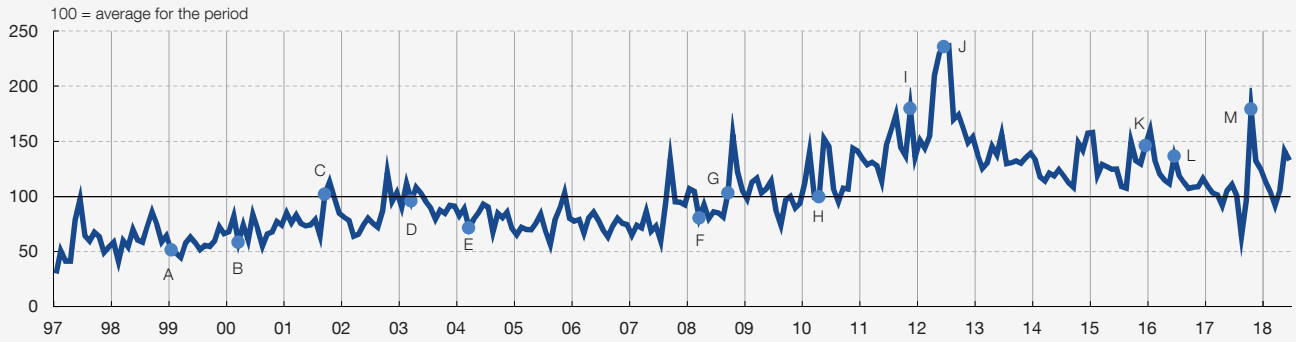
— *Dictionary analysis.* Dictionary analysis is also based on text database searches, but requires the researcher to have access to texts. In this case, searches provide the relative frequency of keywords in each text, as a proxy for the relevance of the topics they represent. It is used in economic sentiment analysis. Specifically, a list of positive and negative words is required to calculate the frequency of positive and negative terms in a text or part of a text. The sentiment index is defined as the difference between the two frequencies, that is, a text has a positive (negative) sentiment when the frequency of positive terms is higher (or lower) than that of the negative terms.

1 Basic, logical or Boolean search operators are "and", "or" and "not".

Chart 1
ECONOMIC AND POLICY UNCERTAINTY: SIGNIFICANT EVENTS

The indicator shows more pronounced movements on the dates of the events associated with significant changes in uncertainty in Spain.

INDICATOR OF ECONOMIC AND POLICY UNCERTAINTY



Selection of significant events	
A January 1999	Creation of the euro area
B March 2000	General elections in Spain
C September 2001	11/9 terrorist attacks
D March 2003	Invasion of Iraq
E March 2004	11 March terrorist attacks and general elections in Spain
F March 2008	Bailout of Bear Stearns and general elections in Spain
G September 2008	Collapse of Lehman Brothers
H April 2010	Greece requests financial assistance
I November 2011	General elections in Spain
J June 2012	Financial assistance to Spain
K December 2015	General elections in Spain
L June 2016	United Kingdom referendum (Brexit) and general elections in Spain
M October 2016	Catalan crisis

SOURCE: Ghirelli, Pérez and Urtasun (2019).

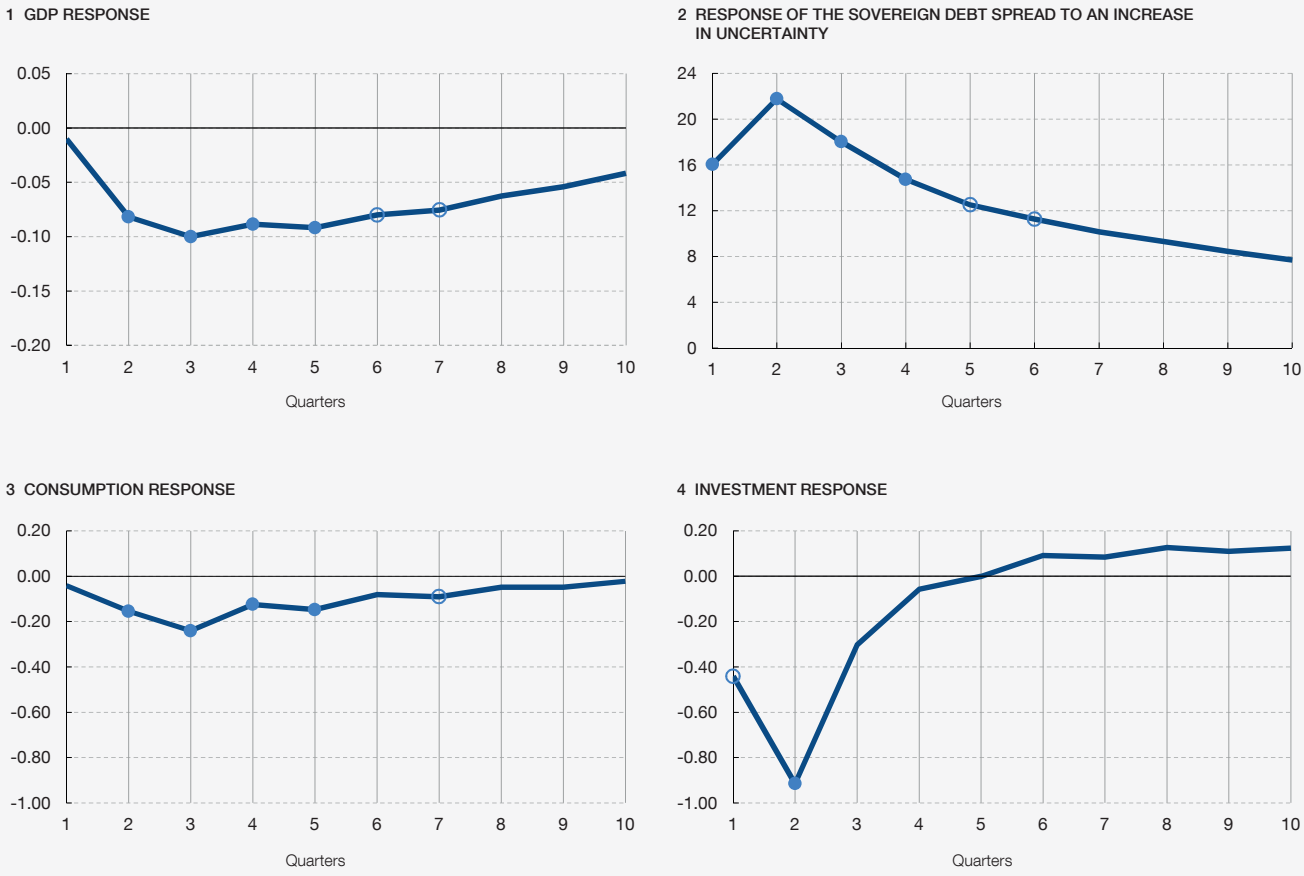
— "Latent topic" techniques (topic modelling). These techniques identify the main topics in a body of text without the researcher having to specify keywords of interest. This is what limits dictionary analysis, since keywords are restricted to the researcher's pre-existing ideas before analysing the topics. Instead, latent topic modelling enables topics not previously identified by the researcher to be discovered. The basic assumption is that each text is represented by a mixture of unobserved topics, which the researcher must identify. To this end, the frequency of words in the body of the text is analysed. Intuitively, topics are identified by repeated patterns in which the same words coincide in the same texts. For example, if words such as "hospital", "sick" and "health" appear in the same texts, the model recognises this as a topic in itself, represented by those words[2].

2 The family of topic models includes many different models, the most popular of which is the LDA *(Latent Dirichlet Allocation)* model. For an overview of models, see Bholat *et al.* (2015).

Measuring economic uncertainty caused by economic policy is an issue that is of relevance to macroeconomic analysis addressed in recent literature. In the case of Spain, the most recent paper is by Ghirelli *et al.* (2019). Following the methodology set out in Baker *et al.* (2016), the authors of this article have constructed an uncertainty indicator for Spain's economic policies and have found a significant dynamic relationship between this indicator and the main macroeconomic variables. Specifically, based on a search of keywords in Spanish newspapers, every month the total number of newspaper articles containing terms relating to the concepts of uncertainty, economy and policy are counted. The index is based on seven of the most widely-read national newspapers (the country's four most widely-read general newspapers and its three leading business newspapers), thus providing very broad coverage.

The indicator shows significant increases or decreases relating to events associated, ex ante, with an increase or decrease in

Chart 2
EFFECTS OF AN INCREASE IN ECONOMIC AND POLICY UNCERTAINTY (a)



SOURCE: Ghirelli, Pérez and Urtasun (2019).
NOTE: ● indicates statistical significance at the 5%,level and ○ indicates statistical significance at the 10% level.

a The VAR model includes: as endogenous variables, uncertainty as measured by the synthetic indicators of financial markets, disagreement and economic policy uncertainty, GDP/consumption/investment, the Spanish sovereign debt spread over the German Bund and a price index; and as exogenous variables, EURO STOXX 50 volatility, the EPU for the EU as a whole and a synthetic indicator of European uncertainty (calculated in a similar manner to that used for Spain's synthetic indicators).

economic uncertainty. It should be noted that the indicator takes major events in recent decades which could be associated with significant changes in uncertainty (see Chart 1). For example, the terrorist attacks of 11 September 2001 in the United States, the collapse of Lehman Brothers in September 2008, the request for financial assistance by Greece in April 2010, the request for financial assistance to restructure the banking sector and savings banks in Spain in June 2012, the Brexit referendum in June 2016, or the episodes of political tension in Catalonia in October 2017.

In addition, unexpected increases in the economic policy uncertainty indicator are estimated to have adverse macroeconomic effects. Charts 1 and 2 illustrate the main results of the exercises carried out and provide the response of the main macroeconomic variables (GDP, the Spanish risk premium, private consumption and investment in capital goods) to an unexpected uncertainty shock (see Chart 2). Specifically, an unexpected rise in uncertainty would lead to a significant reduction of GDP, consumption and investment, and to a higher risk premium.