New Data Sources for Central Banks

Corinna Ghirelli, Samuel Hurtado, Javier J. Pérez, and Alberto Urtasun

Abstract Central banks use structured data (micro and macro) to monitor and forecast economic activity. Recent technological developments have unveiled the potential of exploiting new sources of data to enhance the economic and statistical analyses of central banks (CBs). These sources are typically more granular and available at a higher frequency than traditional ones and cover structured (e.g., credit card transactions) and unstructured (e.g., newspaper articles, social media posts, or Google Trends) sources. They pose significant challenges from the data management and storage and security and confidentiality points of view. This chapter discusses the advantages and the challenges that CBs face in using new sources of data to carry out their functions. In addition, it describes a few successful case studies in which new data sources have been incorporated by CBs to improve their economic and forecasting analyses.

1 Introduction

Over the past decade, the development of new technologies and social media has given rise to new data sources with specific characteristics in terms of their volume, level of detail, frequency, and structure (or lack of) (see [37]). In recent years, a large number of applications have emerged that exploit these new data sources in the areas of economics and finance, particularly in CBs.

In the specific area of economic analysis, the new data sources have significant potential for central banks (CBs), even taking into account that these institutions 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:

C. Ghirelli · S. Hurtado · J. J. Pérez (✉) · A. Urtasun
Banco de España, Madrid, Spain
e-mail: corinna.ghirelli@bde.es; samuel.hurtado@bde.es; javierperez@bde.es; aurtasun@bde.es

S. Consoli et al. (eds.), *Data Science for Economics and Finance*,
https://doi.org/10.1007/978-3-030-66891-4_8

1. Expanding the base data used to carry out financial stability and banking supervision functions (see, e.g., [14] and [32])
2. The use of new methodologies to improve economic analyses (see, e.g., [33])
3. A better understanding (due to more detailed data) and more agile monitoring (due to shorter time delays—almost real time) of economic activity (see [47] for an overview)
4. Improved measurement of agents' sentiments about the state of the economy and related concepts like uncertainty about key economic and policy variables (e.g., [4])
5. Improved measurement of agents' expectations regarding inflation or economic growth
6. Better assessment of economic policy and more possibilities for simulating alternative measures, owing chiefly to the availability of microdata that could be used to improve the characterization of agents' heterogeneity and, thus, to conduct a more in-depth and accurate analysis of their behavior (e.g., see [22] for application in education and [56] for application on social media)

According to Central Banking's annual survey, in 2019 over 60% of CBs used big data in their operations, and two-thirds of them used big data as a core or auxiliary input into the policy-making process. The most common uses for big data are nowcasting and forecasting, followed, among others, by stress-testing and fraud detection (see [20]). Some examples of projects carried out by CBs with new sources of data are: improving GDP forecasting exploiting newspaper articles [58] or electronic payments data (e.g., [3, 27]); machine learning algorithms to increase accuracy in predicting the future behavior of corporate loans (e.g., [55]); forecasting private consumption with credit card data (e.g., [18, 27]); exploiting Google Trends data to predict unemployment [24], private consumption [34, 19], or GDP [42]; web scraping from accommodation platforms to improve tourism statistics [48]; data from online portals of housing sales to improve housing market statistics [49]; sentiment analysis applied to financial market text-based data to study developments in the financial system [54]; and machine learning for outlier detection [31].

In this chapter, we delve into these ideas. First, in Sect. 2 we give a brief overview of some of the advantages and the challenges that CBs face when using these new data sources, while in Sect. 3 we describe a few successful case studies in which new data sources have been incorporated into a CBs' functioning. In particular, we focus on the use of newspaper data to measure uncertainty (two applications in Sect. 3.1), the link between the qualitative messages about the economic situation in the Bank of Spain's quarterly reports and quantitative forecasts (Sect. 3.2), and forecasting applications by means of machine learning methods and the use of non-standard data sources such as Google Trends (Sect. 3.3). Finally, in Sect. 4, we present some general conclusions.

2 New Data Sources for Central Banks

Central banks make intensive use of structured databases to carry out their functions, whether in the banking supervision, financial stability, or monetary policy domains, to mention the core ones.[1] Some examples of individual data are firms' balance sheets (see, e.g., [51] or [6]), information relating to the volume of credit granted by financial institutions to individuals and firms, or the data relating to agents' financial decisions (see, e.g., [5]). In the area of macroeconomics, the main source of information tends to be the national accounts or the respective central bank sources, although a great deal of other information on the economic and financial situation is also published by other bodies: e.g., social security data, payroll employment data (Bureau of Labor Statistics), stock prices (Bloomberg), and house prices (real estate advertising web platforms).

Thanks to technological developments, sources of information are being expanded significantly, in particular as regards their granularity and frequency. For instance, in many cases one can obtain information in almost real time about single actions taken by individuals or firms, and most of the time at higher frequencies than with traditional sources of data. For example, credit card transaction data, which can be used to approximate household consumption decisions, 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 illustration, Chart 1 shows how credit card transactions performed very similarly to household consumption in Spain (for statistical studies exploiting this feature, see [40] and [13]).

The availability of vast quantities of information poses significant challenges in terms of the management, storage capacity and costs, and security and confidentiality of the infrastructure required. In addition, the optimal management of huge structured and unstructured datasets requires the integration of new professional profiles (data scientists and data engineers) at CBs and conveys the need for fully fledged digital transformations of these institutions. Moreover, the diverse nature of the new information sources requires the assimilation and development of techniques that transform and synthesize data, in formats that can be incorporated into economic analyses. For example, textual analysis techniques enable the information contained in the text to be processed and converted into structured data, as in Google Trends, online media databases, social media (e.g., Facebook and Twitter), web search portals (e.g., portals created for housing or job searches), mobile phone data, or satellite data, among others. From the point of view of the statistical treatment of the data, one concern often quoted (see [28]) is the statistical representativeness of the samples used based on the new data, which are developed without the strict requisites of traditional statistical theory (mainly in the field of surveys).

[1] The discussion in this section relies on our analysis in [37].

New data sources are expanding the frontier of statistics, in particular (but not exclusively) in the field of non-financial statistics. Examples are the initiatives to acquire better price measures in the economy using web-scraping techniques or certain external trade items, such as the estimation of tourist movements by tracking mobile networks (see [44]). Developing countries, which face greater difficulties in setting up solid statistics infrastructures, are starting to use the new data sources, even to conduct estimates of some national accounts aggregates (see [43]). The boom in new data sources has also spurred the development of technical tools able to deal with a vast amount of information. For instance, Apache Spark and Apache Hive are two very popular and successful products for processing large-scale datasets.[2] These new tools are routinely applied along with appropriate techniques (which include artificial intelligence, machine learning, and data analytics algorithms),[3] not only to process new data sources but also when dealing with traditional problems in a more efficient way. For example, in the field of official statistics, they can be applied to process structured microdata, especially to enhance their quality (e.g., to detect and remove outliers) or to reconcile information received from different sources with different frequency (e.g., see [60] and the references therein).

Finally, it should be pointed out that, somehow, the public monopoly over information that official statistical agencies enjoy is being challenged, for two main reasons. First, vast amounts of information are held by large, private companies that operate worldwide and are in a position to efficiently process them and generate, for example, indicators of economic and financial developments that "compete" with the "official" ones. Second, and related to the previous point, new techniques and abundant public-domain data can also be used by individuals to generate their own measures of economic and social phenomena and to publish this information. This is not a problem, per se, but one has to take into account that official statistics are based on internationally consolidated and comparable methodologies that serve as the basis for objectively assessing the economic, social, and financial situation and the response of economic policy. In this context, thus, the quality and transparency framework of official statistics needs to be strengthened, including by statistical authorities disclosing the methods used to compile official statistics so that other actors can more easily approach sound standards and methodologies. In addition, the availability of new data generated by private companies could be used to enrich official statistics. This may be particularly useful in nowcasting, where official

[2] Hive is a data warehouse system built for querying and analyzing big data. It allows applying structure to large amounts of unstructured data and integrates with traditional data center technologies. Spark is a big-data framework that helps extract and process large volumes of data.

[3] Data analytics refers to automated algorithms that analyze raw big data in order to reveal trends and metrics that would otherwise be lost in the mass of information. These techniques are typically used by large companies to optimize processes.

statistics are lagging: e.g., data on credit card transactions are an extremely useful indicator of private consumption.[4]

3 Successful Case Studies

3.1 Newspaper Data: Measuring Uncertainty

Applications involving text analysis (from text mining to natural language processing)[5] have gained special significance in the area of economic analysis. With these techniques, relevant information can be obtained from texts and then synthesized and codified in the form of quantitative indicators. First, the text is prepared (preprocessing), specifically by 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.[6] Second, the information contained in the words is synthesized 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.

The rest of this section presents two examples of studies that use text-based indicators to assess the impact of economic policy uncertainty on the economy in Spain and the main Latin American countries: Argentina, Brazil, Chile, Colombia, Mexico, Perú, and Venezuela. These indicators have been constructed by the authors of this chapter based on the Spanish press and are currently used in the regular economic monitoring and forecasting tasks of the Bank of Spain.

3.1.1 Economic Policy Uncertainty in Spain

A recent branch of the literature relies on newspaper articles to compute indicators of economic uncertainty. Text data are indeed a valuable new source of information

[4] Data on credit card transactions are owned by credit card companies and, in principle, are available daily and with no lag. An application on this topic is described in Sect. 3.3.3.

[5] Text mining refers to processes to extract valuable information from the text, e.g., text clustering, concept extraction, production of granular taxonomies, and sentiment analysis. Natural language processing (NLP) is a branch of artificial intelligence that focuses on how to program computers to process and analyze large amounts of text data by means of machine learning techniques. Examples of applications of NLP include automated translation, named entity recognition, and question answering.

[6] The newest NLP models (e.g., transformer machine learning models) do not necessarily require preprocessing. For instance, in the case of BERT, developed by Google [25], the model already carries out a basic cleaning of the text by means of the tokenization process, so that the direct input for the pre-training of the model should be the actual sentences of the text.

since they reflect major current events that affect the decisions of economic agents and are available with no time lag.

In their leading paper, Baker et al. (see [4]) constructed an index of economic policy uncertainty (Economic Policy Uncertainty (EPU) index) for the United States, based on the volume of newspaper articles that contain words relating to the concepts of uncertainty, economy, and policy. Since this seminal paper, many researchers and economic analysts have used text-based uncertainty indicators in their analyses, providing empirical evidence of the negative effects on activity in many countries (e.g., see [50] for Germany, France, Italy, and Spain, [35] for China, or [23] for the Euro area). The authors of this chapter constructed an EPU index for Spain based on two leading Spanish newspapers: (*El País* and *El Mundo*). [38] recently developed a new Economic Policy Uncertainty index for Spain, which is based on the methodology of [4] but expands the press coverage from 2 to 7 newspapers, widens the time coverage starting from 1997 rather than from 2001, and fine-tunes the richness of the keywords used in the search expressions.[7]

The indicator shows significant increases or decreases relating to events associated, ex ante, with an increase or decrease in economic uncertainty, such as the terrorist attacks of September 11, 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 the Spanish region of Catalonia in October 2017.

[38] found a significant dynamic relationship between this indicator and the main macroeconomic variables, such that unexpected increases in the economic policy uncertainty indicator have adverse macroeconomic effects. Specifically, an unexpected rise in uncertainty leads to a significant reduction of GDP, consumption, and investment. This result is in line with the findings in the empirical literature on economic uncertainty.

In addition, the authors of this chapter provide evidence on the relative role of enriching the keywords used in search expressions and widening both press and time coverage when constructing the index. Results are shown in Fig. 1, which compares macroeconomic responses to unexpected shocks in alternative EPU versions in which they vary in one of the aforementioned dimensions at a time, moving from the EPU index constructed by [4] to the new index. All of these dimensions are important since they all contribute to obtaining the expected negative sign in the responses. Expanding the time coverage is key to improving the precision of the estimates and to yielding significant results. The press coverage is also relevant.

[7] The new index is based on the four most widely read general newspapers in Spain and its three leading business newspapers: *El País*, *El Mundo*, *El Economista*, *Cinco Días*, *Expansión*, *ABC*, and *La Vanguardia*.

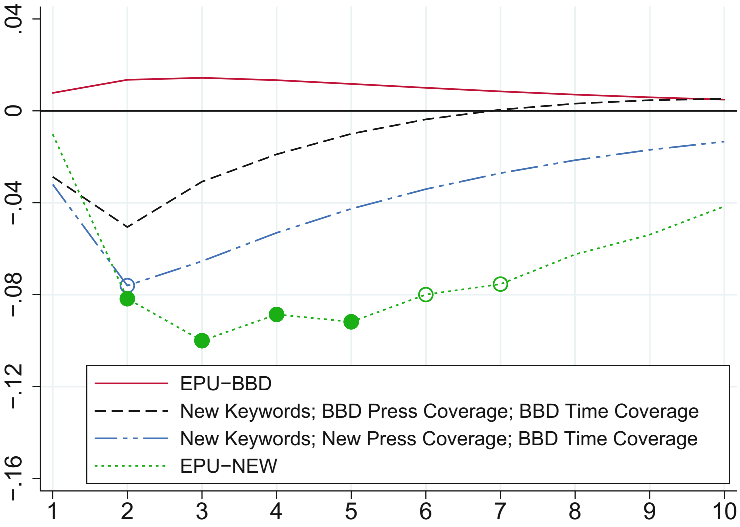

Fig. 1 The graph shows the impulse response function of the Spanish GDP growth rate up to 10 quarters after a positive shock of one standard deviation in the EPU for Spain. The x-axis represents quarters since the shock. The y-axis measures the Spanish GDP growth rate (in percentage points). Full (empty) circles indicate statistical significance at the 5 (10)% level; the solid line indicates no statistical significance. EPU-BBD: EPU index for Spain provided by [4]. EPU-NEW: EPU index for Spain constructed by [38]. Vector autoregression (VAR) models include the EPU index, spread, GDP growth rate, and consumer price index (CPI) growth rate; global EPU is included as an exogenous variable

3.1.2 Economic Policy Uncertainty in Latin America

By documenting the spillover effects of rising uncertainty across countries, the literature also demonstrates that rising economic uncertainty in one country can have global ramifications (e.g., [8, 9, 23, 59]). In this respect, [39] develop Economic Policy Uncertainty indexes for the main Latin American (LA) countries: Argentina, Brazil, Chile, Colombia, Mexico, Peru, and Venezuela. The objective of constructing these indexes is twofold: first, to measure economic policy uncertainty in LA countries in order to get a narrative of "uncertainty shocks" and their potential effects on economic activity in LA countries, and second, to explore the extent to which those LA shocks have the potential to spillover to Spain. This latter country provides an interesting case study for this type of "international spillover" given its significant economic links with the Latin American region.

The uncertainty indicators are constructed following the same methodology used for the EPU index for Spain [38], i.e., counting articles in the seven most important national Spanish newspapers that contain words related to the concepts of *economy*, *policy*, and *uncertainty*. In addition, however, we customize the text searches for the

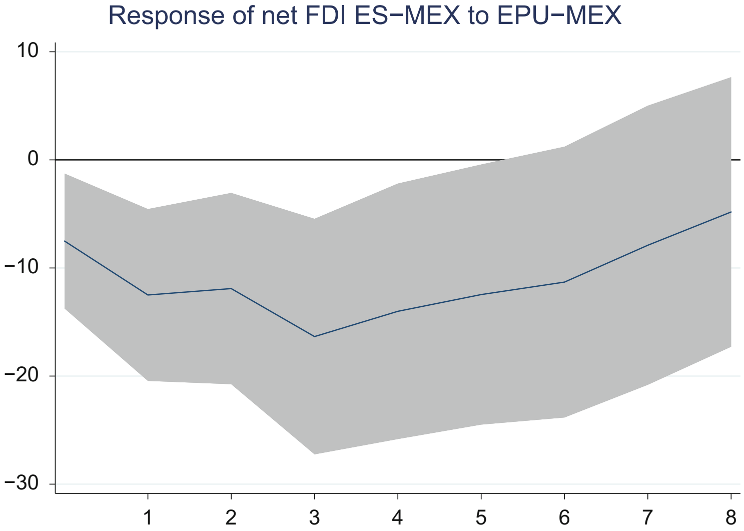

Fig. 2 The graph shows the impulse response function of Spanish net foreign direct investment (FDI) up to 10 quarters after a positive shock of one standard deviation in the Mexican EPU. The x-axis represents quarters since the shock. The y-axis measures the Spanish net FDI growth rate (in percentage points). Confidence intervals at the 5% level are reported

Latin American countries case by case.[8] Note that these indicators are also based on the Spanish press and thereby purely reflect variation in uncertainty in LA countries that is relevant to the Spanish economy, given the importance of the region to the latter. The premise is that the Spanish press accurately reflects the political, social, and economic situation in the LA region, given the existing close economic and cultural ties—including a common language for a majority of these countries. In this respect, one may claim that the indexes provide sensible and relevant measures of policy uncertainty for those countries. This is also in line with a branch of the literature that uses the international press to compute text-based indicators for broad sets of countries (see, e.g., [2] or [53]).

To explore the extent to which LA EPU shocks have the potential to spillover to Spain, the empirical analysis relies on two exercises. A first exercise studies the impact of LA EPU shocks on the performance of Spanish companies operating in the LA region. The underlying assumption is that higher uncertainty in one LA country would affect the investment decisions of Spanish companies that have subsidiaries in this Latin American country: i.e., investment in the LA country may be postponed due to the "wait-and-see effect" and/or the local uncertainty

[8] In particular, (1) we require that each article also contains the name of the LA country of interest; (2) among the set of keywords related to *policy*, we include the name of the central bank and the name of the government's place of work in the country of interest. For more details, see [39].

may foster investment decisions toward other foreign countries or within Spain. To carry out this exercise, the authors consider the stock market quotations of the most important Spanish companies that are also highly exposed to LA countries, controlling for the Spanish macroeconomic cycle. Results show that an unexpected positive shock in the EPU index of an LA country generates a significant drop in the companies' quotation growth rate in the first 2 months. This holds for all LA countries considered in the study and is confirmed by placebo tests, which consider Spanish companies that are listed in the Spanish stock market but do not have economic interests in the Latin American region. This suggests that, as expected, economic policy uncertainty in LA countries affects the quotations of Spanish companies that have economic interests in that region.

The second exercise studies the impact of Latin American EPU shocks on the following Spanish macroeconomic variables: the EPU index for Spain, exports and foreign direct investment (FDI) from Spain to Latin America, and the Spanish GDP. In this case as well, one would expect the spillover from one LA country's EPU to the Spanish EPU to be related to commercial relationships between both countries. The higher the exposure of Spanish businesses to a given country, the higher the spillover. To the extent that the EPU reflects uncertainty about the expected future economic policy situation in the country, unexpected shocks in the EPU of one LA country may affect the export and FDI decisions of Spanish companies. Finally, the relation between Latin American EPUs and the Spanish GDP is expected to be driven by the reduction in exports (indirect effect) and by the business decisions of multinational companies that have economic interests in the region. In particular, multinational companies take into account the economic performance of their subsidiaries when deciding upon investment and hiring in Spain. This, in turn, may affect the Spanish GDP. This second exercise is carried out at the quarterly level by means of VAR models, which document the spillover effects from Latin American EPU indexes to the Spanish EPU. Unexpected shocks in Latin American EPUs significantly dampen the commercial relationship between Spain and the Latin American countries in question. In particular, Spanish firms decrease their exports and FDI toward the countries that experience negative shocks in their EPU index. As an example, Fig. 2 shows the impulse response functions of Spanish net FDI to unexpected shocks in the Mexican EPU index.

3.2 *The Narrative About the Economy as a Shadow Forecast: An Analysis Using the Bank of Spain Quarterly Reports*

One text mining technique consists in the use of dictionary methods for sentiment analysis. To put it simply, a dictionary is a list of words associated with positive and negative sentiments. These lists can be constructed in several ways, ranging from purely manual to machine learning techniques.[9] Sentiment analysis is based on text database searches and requires the researcher to have access to the texts. In its simplest version, the searches allow calculating the frequency of positive and negative terms in a text. The sentiment index is defined as the difference (with some weights) between the two frequencies, that is, a text has a positive (negative) sentiment when the frequency of positive terms is higher (lower) than that of the negative terms. The newest applications of sentiment analysis are more sophisticated than this and rely on neural network architectures and transformer models, which are trained on huge datasets scraped from the web (e.g., all texts in Wikipedia), with the objective of predicting words based on their context. Many of these models take into account negations and intensifiers when computing the sentiment of the text, i.e., improving the results of dictionary-based sentiment exercises. As an example, the paper by [57] sets up a tool to extract opinions from a text by also taking into account the structure of sentences and the semantic relations between words.

In this section, we provide an example of sentiment analysis to show the usefulness of text data (following [26]). We rely on the most basic sentiment analysis technique, i.e., the simple counting of words contained in our own dictionary. Our application is based on the *Quarterly Economic Bulletin* on the Spanish economy by the Bank of Spain, published online since the first quarter of 1999. We consider the Overview section of the reports. The aim of the exercise is to construct an indicator (from Q1 1999) that reflects the sentiment of the Bank of Spain economic outlook reports, and the analysis shows that it mimics very closely the series of Bank of Spain GDP forecasts. This means that the (qualitative) narrative embedded in the text contains similar information to that conveyed by quantitative forecasts.[10]

To carry out the analysis, we create a dictionary of positive and negative terms in Spanish (90, among which some are roots, i.e., we have removed word endings) that are typically used in the economic language to describe the economy, e.g., words like *crecimiento* (growth) or *aumento* (increase) among positive terms, and *disminución* (decrease) or *reducción* (reduction) among negative ones. In order to control for wrong signs, we ignore these terms when they appear around (within nine words before or after) the words “unemployment” or “deficit.” We assign a weight of $+1$ (-1) to the resulting counts of positive (negative) terms. Then, for each bulletin, we sum up all of the weighted counts of terms in the dictionary and divide the resulting number by the length of the bulletin. Then, we compare the resulting text-based index with the GDP growth projections conducted each quarter by the Bank of Spain, which in most of the samples under consideration were recorded internally but not published.

[9] Examples for English include the Bing Liu Opinion Lexicon [46] or SentiWordNet [30]. [52] created a Spanish dictionary based on the Bing Liu Opinion Lexicon: this list was automatically translated using the Reverso translator and subsequently corrected manually.

[10] Researchers at the Bank of Canada carried out a similar exercise: they applied sentiment analysis by means of machine learning methods on the monetary policy reports of the Bank of Canada. See [10].

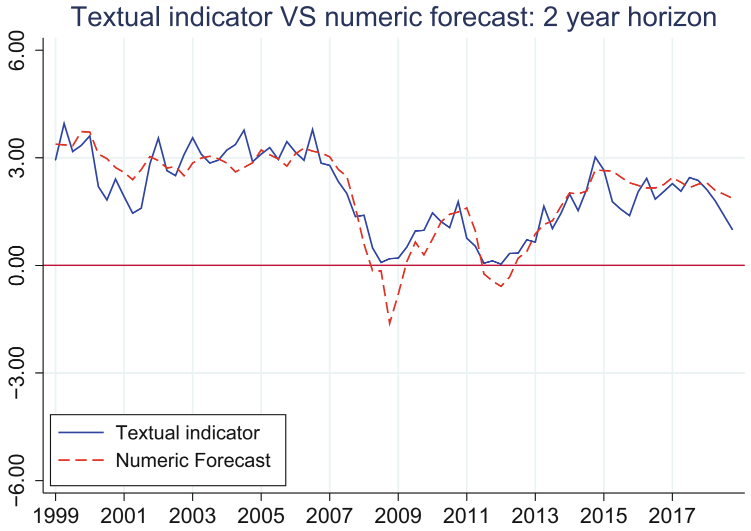

Fig. 3 The graph shows the textual indicator (solid blue line) against the numerical forecasts of the Bank of Spain (dashed red line). The y-axis measures the GDP growth rate (in percentage points). The black dotted line represents the observed GDP growth rate (the target variable of the forecast exercise)

We find a significant dynamic relationship between both series: the narrative text-based indicator follows the Spanish cycle and increases or decreases with the quantitative projections. In addition, the comparison shows that the economic bulletins are informative not only at the short-term forecast horizon but even more so at the 1-to-2-year forecast horizon. The textual indicator shows the highest correlation with the projections for a 2-year horizon. Figure 3 reports the textual indicator (solid blue line) against the GDP growth projection carried out by the Bank of Spain for the 2-year horizon (dashed red line). This evidence suggests that the narrative reflected in the text of the economic bulletins by the Bank of Spain follows very closely the underlying story told by the institution's GDP growth projections. This means that a "sophisticated" reader could infer GDP growth projections based on the text of the reports.

3.3 Forecasting with New Data Sources

Typically, central banks' forecasting exercises are carried out by combining soft indicators with the set of information provided by hard indicators (e.g., data from government statistical agencies such as the main macroeconomic variables: GDP, private consumption, and private investment, for instance).[11] The main limitation posed by hard data is that they are typically published with some lag and at a low frequency (e.g., quarterly). Soft indicators include, for instance, business and consumer confidence surveys. As such, these data provide qualitative information (hence, of a lower quality than hard data) typically available at a higher frequency than hard data. Thus, they provide additional and new information especially at the beginning of the quarter, when macroeconomic information is lacking, and their usefulness decreases as soon as hard data are released [34]. Text indicators are another type of soft indicator. Compared to the traditional survey-based soft indicators, text-based indicators show the following features:

1. They are cheaper from an economic point of view, in that they do not rely on monthly surveys but rather on subscriptions to press repository services.
2. They provide more flexibility since one can select the keywords depending on specific needs and get the entire time series (spanning backward), whereas in a survey, the inclusion of a new question would be reflected in the time series from that moment onward.

The rest of this section presents three applications aimed at improving forecasting. The first is based on sentiment analysis. The second application shows how machine learning can improve the accuracy of available forecasting techniques. Finally, the second application assesses the relative performance of alternative indicators based on new sources of data (Google Trends and credit card transactions/expenses).

3.3.1 A Supervised Method

As an empirical exercise, we construct a text-based indicator that helps track economic activity, as described in [1]. It is based on a similar procedure that is used to elaborate the economic policy uncertainty indicator, i.e., it relies on counting the number of articles in the Spanish press that contain specific keywords. In this case, we carry out a dictionary analysis as in the previous section, i.e., we set up a dictionary of positive and negative words that are typically used in portions of texts related to the GDP growth rate, the target variable of interest, so as to also capture the tone of the articles and, in particular, to what extent they describe upturns or downturns. For instance, words like "increase," "grow," or "raise are listed among the positive terms, while "decrease" and "fall" appear in the negative list. As with the EPU indicators, this one is also based on the Factiva Dow Jones repository of Spanish press and relies on seven relevant Spanish national newspapers: *El País*, *El Mundo*, *Cinco Días*, *Expansión*, *El Economista*, *La Vanguardia*, and *ABC*.

[11] Recently, [16] set up a model to efficiently exploit—jointly and in an efficient manner—a rich set of economic and financial hard and soft indicators available at different frequencies to forecast economic downturns in real time.

We place the following restrictions on all queries: (1) the articles are in Spanish; (2) the content of the article is related to Spain, based on Factiva's indexation; and (3) the article is about corporate or industrial news, economic news, or news about commodities or financial markets, according to Factiva's indexation. We then perform three types of queries for each newspaper:[12]

1. We count the number of articles that satisfy the aforementioned requirements. This will serve as the denominator for our indicator.
2. We count the number of articles that, in addition to satisfying the aforementioned conditions, contain upswing-related keywords. That is, the articles must contain the word *recuperacion** (recovery) or one of the following words, provided that they are preceded or followed by either *economic** (economic) or *economia* (economy) within a distance of five words: *aceler** (acceleration), *crec** (increase), *increment** (rise), *aument** (boost), *expansi** (growth), and *mejora** (improvement). In addition, in order to ensure that the news items are about the Spanish business cycle, we also require articles to contain the word *Españ** (Spain).
3. Similarly, we count the number of articles that, in addition to satisfying the aforementioned conditions, are about downswings. In particular, the articles have to contain the word *recession** (recession) or *crisis* (crisis), or one of the following words, provided that they are preceded or followed by either *economic** or *economia* within a distance of five words: *descen** (decrease), *ralentiz** (slowdown), *redu** (reduction), *disminu** (fall), *contraccion** (contraction), *decrec** (downturn), and *desaceler**(deceleration). The articles should also contain *Españ**.

Then, for each newspaper, we take the difference between the upturn- and downturn-related counts and scale the difference by the total number of economic articles in the same newspaper/month. Finally, we standardize the monthly series of scaled counts, average them across newspapers, rescale the resulting index to mean 0, and average it at the quarterly level.

The right panel in Fig. 4 shows the resulting textual indicator (solid blue line) against the GDP growth rate (red and dashed line).

Next, we test whether our textual indicator has some predictive power to nowcast the Spanish GDP growth rate. We perform a pseudo-real-time nowcasting exercise at the quarterly level as follows.[13] First, we estimate a baseline nowcasting model in which the GDP growth rate is nowcasted by means of an AR(1) process. Second, we estimate an alternative nowcasting model that adds our textual indicator and its lag to the GDP AR(1) process. Finally, we compare the forecast accuracy of both models. The alternative model provides smaller mean squared errors of predictions than the baseline one, which suggests that adding textual indicators to the AR(1)

[12] The search is carried out in Spanish. English translations are in parentheses.

[13] We use unrevised GDP data, so that our data should be a fair representation of the data available in real time.

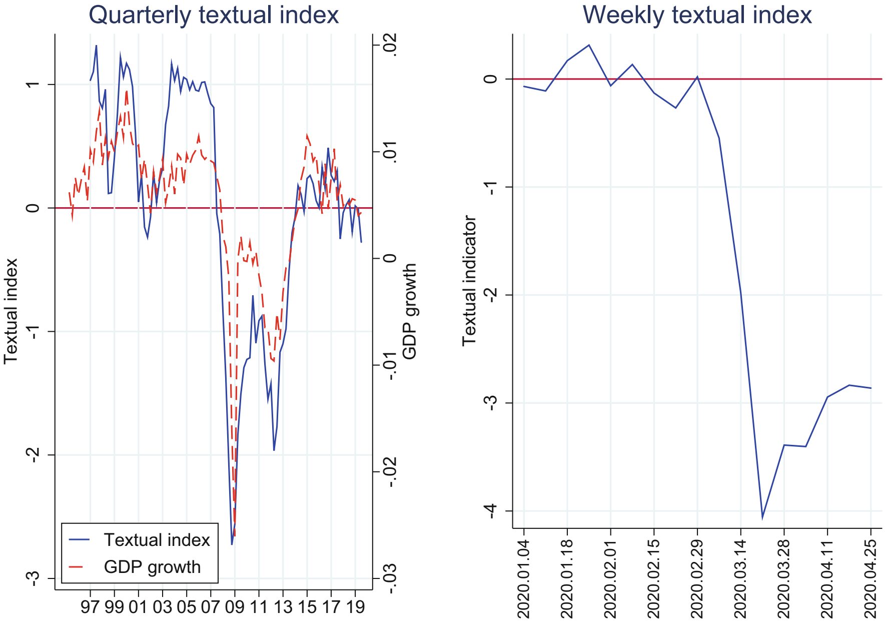

Fig. 4 The figure on the right shows the quarterly textual indicator of *economy* (blue and solid line) against the Spanish GDP growth rate (red and dashed line) until June 2019. The figure on the left shows the weekly textual indicator from January to March 2020

process improves the predictions of the baseline model. In addition, according to the Diebold–Mariano test, the forecast accuracy of the model improves significantly in the alternative model. The null hypothesis of this test is that both competing models provide the same forecast accuracy. By comparing the baseline with the alternative model, this hypothesis is rejected at the 10% level with a p-value of 0.063.[14]

A major advantage of newspaper-based indicators is that they can be updated in real time and are of high frequency. This has been extremely valuable since the Covid-19 outbreak, when traditional survey-based confidence indicators failed to provide timely signals about economic activity.[15] As an example, the right panel in Fig. 4 depicts the textual indicator at a weekly frequency around the Spanish lockdown (14 March 2020) and correctly captures the drastic reduction in Spanish economic activity around that time.

[14] A natural step forward would be to incorporate this text-based indicator into more structured nowcasting models that combine hard and soft indicators to nowcast GDP (e.g., [16]). The aim of the current exercise was to show the properties of our text-based indicator in the simplest framework possible.

[15] In [1], we compare this text-based indicator with the economic sentiment indicator (ESI) of the European Commission and show that, for Spain, the former significantly improves the GDP nowcast when compared with the ESI.

3.3.2 An Unsupervised Method

The latent Dirichlet allocation or LDA (see [11]) method can be used to estimate topics in text data. This is an unsupervised learning method, meaning that the data do not need to include a topic label and that the definition of the topics is not decided by the modeler but is a result of running the model over the data. It is appealing because, unlike other methods, it is grounded in a statistical framework: it assumes that the documents are generated according to a generative statistical process (the Dirichlet distribution) so that each document can be described by a distribution of topics and each topic can be described by a distribution of words. The topics are latent (unobserved), as opposed to the documents at hand and the words contained in each document.

The first step of the process is to construct a corpus with text data. In this instance, this is a large database of more than 780,000 observations containing all news pieces published by *El Mundo* (a leading Spanish newspaper) between 1997 and 2018, taken from the Dow Jones repository of Spanish press. Next, these text data have to be parsed and cleaned to end up with a version of the corpus that includes no punctuation, numbers, or special characters and is all lowercase and excludes the most common words (such as articles and conjunctions). This can then be fed to a language-specific stemmer, which eliminates variations of words (e.g., verb tenses) and reduces them to their basic stem (the simpler or, commonly, partial version of the word that captures its core meaning), and the result from this is used to create a bag-of-words representation of the corpus: a big table with one row for each piece of news and one column for each possible stemmed word, filled with numbers that represent how many times each word appears in each piece of news (note that this will be a very sparse matrix because most words from an extensive dictionary will not appear in most pieces of news).

This bag-of-words representation of the corpus is then fed to the LDA algorithm, which is used to identify 128 different topics that these texts discuss[16] and to assign to each piece of news an estimate of the probability that it belongs to each one of those topics. The algorithm analyzes the texts and determines which words tend to appear together and which do not, optimally assigning them to different topics so as to minimize the distance between texts assigned to any given topic and to maximize the distance between texts assigned to different topics.

The result is a database that contains, for each quarter from 1997 to 2018, the percentage of news pieces that fall within each of the 128 topics identified by the unsupervised learning model. A dictionary of positive and negative terms is also applied to each piece of news, and the results are aggregated into quarterly series that indicate how positive or negative are the news pieces relating to each topic.

[16] In LDA models, the number of topics to be extracted has to be chosen by the researcher. We run the model by varying the number of topics (we set this parameter equal to numbers that can be expressed as powers of two: 16, 32, 64, 128) and choose the model with 128 topics since it provides better results. Typically, the number of topics is chosen by minimizing the perplexity, which is a measure of the goodness-of-fit of the LDA.

We can now turn to a machine learning model using the data resulting from the analysis of Spanish newspapers to forecast Spanish GDP.[17] The term "machine learning" encompasses a very wide range of methods and algorithms used in different fields such as machine vision, recommender systems, or software that plays chess or go. In the context of economics, support vector machines, random forests, and neural networks can be used to analyze microdata about millions of consumers or firms and find correlations, patterns of behavior, and even causal relationships. CBs have incorporated machine learning techniques to enhance their operations, for instance, in the context of financial supervision, by training models to read banks' balance sheets and raise an alert when more scrutiny is required (e.g., see [21]). For time-series forecasting, ensemble techniques, including boosting and bagging, can be used to build strong forecasting models by optimally combining a large number of weaker models. In particular, ensemble modeling is a procedure that exploits different models to predict an outcome, either by using different modeling algorithms or using different training datasets. This allows reducing the generalization error of the prediction, as long as the models are independent. [7] provides an extensive evaluation of some of these techniques. In this subsection, we present one such ensemble model: a doubly adaptive aggregation model that uses the results from the LDA exercise in the previous subsection, coined DAAM-LDA. This model has the advantage that it can adapt to changes in the relationships in the data.

The ingredients for this ensemble forecasting model are a set of 128 very simple and weakly performing time-series models that are the result of regressing quarterly Spanish GDP growth on its first lag and the weight, positiveness, and negativeness of each topic in the current quarter. In the real-time exercise, the models are estimated every quarter and their first out-of-sample forecast is recorded. Since the share of each topic in the news and its positiveness or negativeness will tend to be indicators with a relatively low signal-to-noise ratio, and since most topics identified in the LDA exercise are not actually related to economics, most of these models will display a weak out-of-sample performance: only 4 out of the 128 outperform a simple random walk. Ensemble methods are designed specifically to build strong models out of such a set of weak models. One advantage is that one does not have to decide which topics are useful and which are not: the model automatically discards any topic that did not provide good forecasts in the recent periods.

One possible way to combine these forecasts would be to construct a nonlinear weight function that translates an indicator of the recent performance of each model at time t into its optimal weight for time $t + 1$. We constructed such a model, using as a weight function a neural network with just three neurons in its hidden layer, in order to keep the number of parameters and hyperparameters relatively low. We

[17] Basically, we rely on novel data to forecast an official statistic. An example of another application in which novel data replace official statistics is The Billion Prices Project, an academic initiative that computes worldwide real-time daily inflation indicators based on prices collected from online retailers (see http://www.thebillionpricesproject.com/). An alternative approach would be to enhance official statistics with novel data. This is not the target of this application.

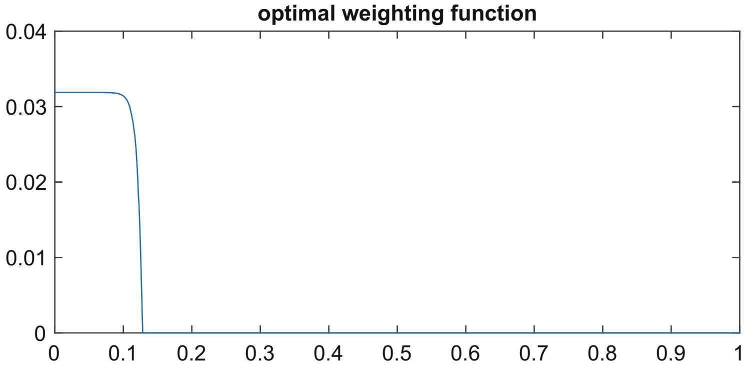

Fig. 5 This is the optimal function for transforming previous performance (horizontal axis) into the current weight of each weak model (vertical axis). It is generated by a neural network with three neurons in its hidden layer, so it could potentially have been highly nonlinear, but in practice (at least for this particular application), the optimal seems to be a simple step function

used a k-fold cross-validation procedure[18] to find the optimal memory parameter for the indicator of recent performance and the optimal regularization, which restricts the possibility that the neural network would overfit the data. The problem is that after all of this, even if the small neural network was able to generate all sorts of potentially very nonlinear shapes, the optimal weighting function would end up looking like a simple step function, as seen in Fig. 5.

To some extent, this was to be expected as it is already known in the forecasting literature that sophisticated weighting algorithms often have a hard time beating something less complex, like a simple average (see, e.g., [29]). In our case, though, since our weak models are really not well-performing, this would not be enough. So instead of spending the degrees of freedom on allowing for potentially highly nonlinear weights, the decision taken was to use a simple threshold function with just one parameter and then add complexity in other areas of the ensemble model, allowing the said threshold to vary over time.

This doubly adaptive aggregation model looks at the recent performance of each weak model in order to decide if it is used for $t + 1$ or not (i.e., weak models either enter into the average or they do not, and all models that enter have equal weight). The threshold is slowly adapted over time by looking at what would have been optimal in recent quarters, and both the memory coefficient (used for the indicator

[18] The k-fold cross-validation process works as follows: we randomly divide the data into k bins, train the model using $k - 1$ bins and different configurations of the metaparameters of the model, and evaluate the forecasting performance in the remaining bin (which was not used to train the model). This is done k times, leaving out one bin at a time for evaluation. The metaparameters that provide the best forecasting performance are selected for the final training, which uses all of the bins.

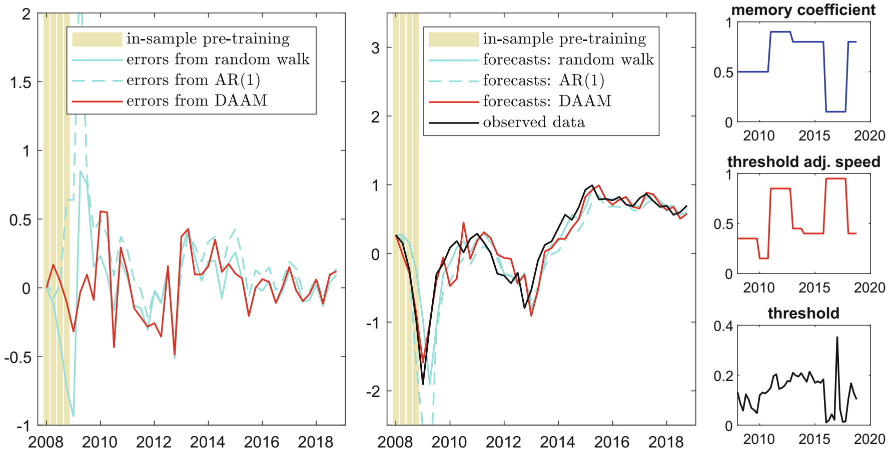

Fig. 6 Results from the real-time forecast exercise for Spanish quarterly GDP growth. DAAM-LDA is the doubly adaptive aggregation model with LDA data presented in this subsection

of recent performance of each weak model) and the allowed speed of adjustment of the threshold are re-optimized at the end of each year.

Importantly, the whole exercise is carried out in real time, using only past information in order to set up the parameters that are to be used for each quarter. Figure 6 summarizes the results from this experiment and also displays the threshold that is used at each moment in time, as well as the memory parameter and speed of adjustment of the threshold that are found to be optimal each year.

As seen in Table 1, the forecasts from DAAM-LDA can outperform a random walk, even if only 4 out of the 128 weak models that it uses as ingredients actually do so. If we restrict the comparison to just the last 4 years in the sample (2015–2018), we can include other state-of-the-art GDP nowcasting models currently in use at the Bank of Spain. In this restricted sample period, the DAAM-LDA model performs better than the random walk, the simple AR(1) model, and the Spain-STING model (see [17]). Still, the Bank of Spain official forecasts display unbeatable performance compared to the statistical methods considered in this section.

Table 1 Spanish GDP forecasting: root mean squared error in real-time out-of-sample exercise

	RW	AR(1)	BdE	DAAM-LDA	Spain-STING
2009–2018	0.290	0.476	0.082	0.240	–
2015–2018	0.110	0.155	0.076	0.097	0.121

Notes: Out-of-sample root mean squared error (RMSE) for different forecasts of Spanish quarterly GDP growth: random walk, simple AR(1) model, official Bank of Spain forecast, doubly adaptive aggregation model with LDA data, and Spain-STING

3.3.3 Google Forecast Trends of Private Consumption

The exercise presented in this section follows closely our paper [40]. In that paper, the question is whether new sources of information can help predict private household consumption. Typically, benchmark data to approximate private household spending decisions are provided by the national accounts and are available at a quarterly frequency ("hard data"). More timely data are usually available in the form of "soft" indicators, as discussed in the previous subsection of this chapter. In this case, the predictive power of new sources of data is ascertained in conjunction with the traditional, more proven, aforementioned "hard" and "soft" data.[19] In particular, the following sources of monthly data are considered: (1) data collected from automated teller machines (ATMs), encompassing cash withdrawals at ATM terminals, and point-of-sale (POS) payments with debit and credit cards; (2) Google Trends indicators, which provide proxies of consumption behavior based on Internet search patterns provided by Google; and (3) economic and policy uncertainty measures,[20] in line with another recent strand of the literature that has highlighted the relevance of the level of uncertainty prevailing in the economy for private agents' decision-making (e.g., see [12] and the references therein).

To exploit the data in an efficient and effective manner, [40] build models that relate data at quarterly and monthly frequencies. They follow the modeling approach of [45]. The forecasting exercise is based on pseudo-real-time data, and the target variable is private consumption measured by the national accounts. The sample for the empirical exercises starts by about 2000 and ends in 2017Q4.[21] As ATM/POS data are not seasonally adjusted, the seasonal component is removed by means of the TRAMO-SEATS software [41].

In order to test the relevant merits of each group of indicators, we consider several models that differ in the set of indicators included in each group. The estimated models include indicators from each group at a time, several groups at a time, and different combinations of individual models. As a mechanical benchmark, [40] use a random walk model whereby they repeat in future quarters the latest quarterly growth rate observed for private consumption. They focus on the forecast performance at the nowcasting horizon (current quarter) but also explore forecasts

[19] A growing literature uses new sources of data to improve forecasting. For instance, a number of papers use checks and credit and debit card transactions to nowcast private consumption (e.g., [36] for Canada, [27] for Portugal, [3] for Italy) or use Google Trends data (e.g., see [61], [19], and [34] for nowcasting private consumption in the United States, Chile, and France, respectively, or [15] for exchange rate forecasting).

[20] Measured alternatively by the IBEX stock market volatility index and the text-based EPU index provided by [4] for Spain

[21] The sample is restricted by the availability of some monthly indicators, i.e., Google Trends, the EPU index, and the Services Sector Activity Indicator are available from January 2004, January 2001, and January 2002, respectively.

at 1 to 4 quarters ahead of each of the current quarter forecast origins (first month of the quarter, second, and third).

The analysis yields the following findings. First, as regards models that use only indicators from each group, the ones that use quantitative indicators and payment cards (amounts) tend to perform better than the others in the nowcasting and, somewhat less so, in forecasting (1-quarter- and 4-quarters-ahead) horizons (see Panel A in Table 2). Relative root mean squared errors (RMSEs) are in almost all cases below one, even though from a statistical point of view, they are only different from quarterly random walk nowcasts and forecasts in a few instances. In general, the other models do not systematically best the quarterly random walk alternative. The two main exceptions are the model with qualitative indicators for the nowcasting horizons and the Google-Trends-based ones for the longer-horizon forecasts. The latter results might be consistent with the prior that Google-Trends-based indicators deliver information for today on steps to prepare purchases in the future.

Second, Panel B in Table 2 shows the results of the estimation of models that include quantitative indicators while adding, in turn, variables from the other groups (qualitative, payment cards (amounts), uncertainty, Google Trends). The improvement in nowcast accuracy is not generalized when adding more indicators, with the exception of the "soft" ones. Nonetheless, there is a significant improvement for longer forecast horizons when expanding the baseline model. In particular, for the 4-quarters-ahead one, uncertainty and Google-Trends-based indicators add significant value to the core "hard"-only-based model.

Finally, it seems clear that the combination (average) of models with individual groups of indicators improves the forecasting performance in all cases and at all horizons (see Panel C in Table 2). Most notably, the combination of the forecasts of models including quantitative indicators with those with payment cards (amounts) delivers, in general, the best nowcasting/forecasting performance for all horizons. At the same time, adding the "soft" forecasts seems to add value in the nowcasting phase. In turn, the combination of a broad set of models produces the lowest RMSE relative to the quarterly random walk in the 4-quarters-ahead forecast horizon.

So, to conclude, this study shows that even though traditional indicators do a good job nowcasting and forecasting private consumption in real time, novel data sources add value—most notably those based on payment cards but also, to a lesser extent, Google-Trends-based and uncertainty indicators—when combined with other sources.

Table 2 Relative RMSE statistics: ratio of each model to the quarterly random walk[a]

Panel A: models including indicators of only one group									
	Nowcast			1-q-ahead			4-q-ahead		
	m1	m2	m3	m1	m2	m3	m1	m2	m3
Quantitative (“hard”) indicators[b]	0.84	0.75 *	0.79	0.75 **	0.81	0.80	0.98	0.97	1.00
Qualitative (“soft”) indicators[c]	1.01	0.85	0.85	1.11	1.05	1.05	1.09	1.10	1.29 *
Payment cards (amounts, am)[d]	0.79	0.82	0.88	0.65 ***	0.84	0.69**	0.74 **	0.84	0.83
Payment cards (numbers)[d]	1.05	1.15	1.13	0.90	1.10	0.98	0.75 **	0.81	0.79
Uncertainty indicators[e]	1.06	0.97	0.99	1.00	1.05	1.06	0.94	1.00	1.02
Google: aggregate of all indicators	1.04	1.06	1.06	0.85	1.03	1.03	0.71 **	0.79	0.79
Google: durable goods (lagged)	1.04	0.97	0.98	0.96	1.04	1.04	0.85 *	0.93	0.93
Panel B: models including indicators from different groups									
	Nowcast			1-q-ahead			4-q-ahead		
	m1	m2	m3	m1	m2	m3	m1	m2	m3
Quantitative and qualitative	0.69 **	0.78	0.77	0.67 ***	0.76 *	0.72 *	0.79 *	0.82 *	0.80 *
Quantitative and payment cards (am)[d]	0.90	0.82	0.91	0.67 ***	0.79	0.78	0.86	0.89	0.91
Quantitative and uncertainty	0.88	0.86	0.75	0.74 **	0.91	0.93	0.69 **	0.76	0.76
Quantitative and Google (aggregate)	0.85	0.76	0.77	0.81 *	0.94	0.89	0.77 **	0.81 *	0.82
Quantitative and Google (durables)	0.91	0.95	0.87	0.69 **	0.83	0.88	0.72 **	0.76 *	0.77 *

(continued)

Table 2 (continued)

Panel C: combination of models

	Nowcast			1-q-ahead			4-q-ahead		
	m1	m2	m3	m1	m2	m3	m1	m2	m3
All models[f]	0.66 **	0.71 **	0.69 **	0.68 ***	0.77 *	0.68 **	0.73 **	0.78 *	0.78 *
Hard and payment cards (am)[d]	0.62 **	0.69 **	0.71 **	0.53 ***	0.69 **	0.52 ***	0.79 *	0.86	0.84
Hard, payment cards (am)[d] and soft	0.65 **	0.67 **	0.67 **	0.68 ***	0.74 **	0.59 ***	0.83 *	0.89	0.92
Hard and soft	0.68 **	0.66 **	0.66 **	0.77 **	0.75 **	0.69 **	0.91	0.94	1.02
Hard and Google (durables)	0.77 **	0.78 **	0.76 **	0.74 ***	0.83	0.78 *	0.85	0.91	0.90

Notes: The asterisks denote the Diebold–Mariano test results for the null hypothesis of equal forecast accuracy of two forecast methods. A squared loss function is used. The number in each cell represents the loss differential of the method in its horizontal line as compared to the quarterly random walk alternative. * (**) [***] denotes rejection of the null hypothesis at the 10% (5%) [1%] level. [a]Nowcast/forecast errors computed as the difference from the first released vintage of private consumption data. Forecasts are generated recursively over the moving window 2008Q1 (m1) to 2017Q4 (m3). [b]Social security registrations; Retail Trade Index; Activity Services Index. [c]PMI Services; Consumer Confidence Index. [d]Aggregate of payment cards via POS and ATMs. [e]Stock market volatility (IBEX); Economic Policy Uncertainty (EPU) index. [f]Combination of the results of 30 models, including models in which the indicators of each block are included separately, models that include the quantitative block and each other block, and versions of all the previous models but including lags of the variables

4 Conclusions

Central banks use structured data (micro and macro) to monitor and forecast economic activity. Recent technological developments have unveiled the potential of exploiting new sources of data to enhance the economic and statistical analyses of CBs. These sources are typically more granular and available at a higher frequency than traditional ones and cover structured (e.g., credit card transactions) and unstructured (e.g., newspaper articles, social media posts) sources. They pose significant challenges from the data management and storage and security and confidentiality points of view. In addition, new sources of data can provide timely information, which is extremely powerful in forecasting. However, they may entail econometric problems. For instance, in many cases they are not linked to the target variables by a causal relationship but rather reflect the same phenomena they aim to measure (for instance, credit card transactions are correlated with—and do not cause—consumption). Nevertheless, a causal relationship exists in specific cases, e.g., uncertainty shocks affect economic activity.

In this chapter, we first discussed the advantages and challenges that CBs face in using new sources of data to carry out their functions. In addition, we described a few successful case studies in which new data sources (mainly text data from newspapers, Google Trends data, and credit card data) have been incorporated into a CBs' functioning to improve its economic and forecasting analyses.

References

1. Aguilar, P., Ghirelli, C., Pacce, M., & Urtasun, A. (2020). *Can news help to measure economic sentiment? An application in Covid-19 times*. Working Papers 2027, Banco de España.
2. Ahir, H., Bloom, N., & Furceri, D. (2019). *The world uncertainty index*. Working Paper 19–027, Stanford Institute for Economic Policy Research.
3. Aprigliano, V., Ardizzi, G., & Monteforte, L. (2017). *Using the payment system data to forecast the Italian GDP*. Working paper No. 1098, Bank of Italy.
4. Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics, 131*(4), 1593–1636.
5. Banco de España (2017). *Survey of household finances, 2014: Methods, results and changes since 2011*. Analytical Article No. 1/2017, Bank of Spain, January.
6. Banco de España (2018). *Central balance sheet data office*. Annual results of non-financial corporations 2017. https://www.bde.es/bde/en/areas/cenbal/
7. Barrow, D. K., & Crone, S. F. (2016). A comparison of AdaBoost algorithms for time series forecast combination. *International Journal of Forecasting, 32*(4), 1103–1119.
8. Bhattarai, S., Chatterjee, A., & Park, W. Y. (2019). Global spillover effects of US uncertainty. *Journal of Monetary Economics, 114*, 71–89. https://doi.org/10.1016/j.jmoneco.2019.05.008
9. Biljanovska, N., Grigoli, F., & Hengge, M. (2017). *Fear thy neighbor: Spillovers from economic policy uncertainty*. Working Paper No. 17/240, International Monetary Fund.
10. Binette, A., & Tchebotarev, D. (2019). *Canada's monetary policy report: If text could speak, what would it say?* Staff Analytical Note 2019–5, Bank of Canada.
11. Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

12. Bloom, N. (2014). Fluctuations in uncertainty. *Journal of Economic Perspectives, 28*(2), 153–76.
13. Bodas, D., García, J., Murillo, J., Pacce, M., Rodrigo, T., Ruiz, P., et al. (2018). *Measuring retail trade using card transactional data*. Working Paper No. 18/03, BBVA Research.
14. Broeders, D., & J. Prenio (2018). *Innovative technology in financial supervision (Suptech): The experience of early users*. Financial Stability Institute Insights on Policy Implementation, Working paper No. 9, Bank for International Settlements, July.
15. Bulut, L. (2018). Google Trends and the forecasting performance of exchange rate models. *Journal of Forecasting, 37*(3), 303–315.
16. Cakmakli, C., Demircan, H., & Altug, S. (2018). *Modeling coincident and leading financial indicators for nowcasting and forecasting recessions: A unified approach*. Discussion Paper No. 13171, Center for Research in Economics and Statistics.
17. Camacho, M., & Perez-Quiros, G. (2011). Spain-sting: Spain short-term indicator of growth. *The Manchester School, 79*,594–616.
18. Carlsen, M., & Storgaard, P. E. (2010). *Dankort payments as a timely indicator of retail sales in Denmark*. Working paper No. 66, Bank of Denmark.
19. Carriére-Swallow Y., & Labbé, F. (2013). Nowcasting with google trends in an emerging market. *Journal of Forecasting, 32*(4), 289–298.
20. Hinge, D., & Šilytė, K. (2019). *Big data in central banks: 2019 survey results*. Central Banking, Article No. 4508326. https://www.centralbanking.com/central-banks/economics/data/4508326/big-data-in-central-banks-2019-survey-results
21. Chakraborty, C., & Joseph, A. (2017). *Machine learning at central banks*. Working paper No. 674, Bank of England.
22. Chetty, R., Friedman, J., & Rockoff, J. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *The American Economic Review, 104*(9), 2633–2679.
23. Colombo, V. (2013). Economic policy uncertainty in the us: Does it matter for the euro area? *Economics Letters, 121*(1), 39–42.
24. D'Amuri F., & Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment, *International Journal of Forecasting, 33*(4), 801–816.
25. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv:1810.04805v2.
26. Diaz-Sobrino, N., Ghirelli, C., Hurtado, S., Perez, J. J., & Urtasun, A. (2020), *The narrative about the economy as a shadow forecast: an analysis using bank of Spain quarterly reports*. Bank of Spain. Working Paper No. 2042. https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosTrabajo/20/Files/dt2042e.pdf
27. Duarte, C., Rodrigues, P. M., & 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.
28. Einav, L., & Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy, 14*, 1–24.
29. Elliott, G., Granger, C. W. J., & Timmermann, A. (Eds.). (2006). *Handbook of economic forecasting*. Holland, Amsterdam: Elsevier.
30. Esuli, A., & Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006* (pp. 417–422).
31. Farné, M., & Vouldis, A. T. (2018) *A methodology for automatised outlier detection in high-dimensional datasets: An application to Euro area banks' supervisory data*. Working Paper No. 2171, European Central Bank.
32. Fernández, A. (2019). *Artificial intelligence in financial services*. Analytical Articles, Economic Bulletin No. 2/2019, Bank of Spain.
33. Fernández-Villaverde, J., Hurtado, S., & Nuño, G. (2019). *Financial frictions and the wealth distribution*. Working Paper No. 26302, National Bureau of Economic Research.

34. Ferrara, L., & Simoni, A. (2019). *When are Google data useful to nowcast GDP? An approach via pre-selection and shrinkage*. Working paper No. 2019–04, Center for Research in Economics and Statistics.
35. Fontaine, I., Didier, L., & Razafindravaosolonirina, J. (2017). Foreign policy uncertainty shocks and US macroeconomic activity: Evidence from China. *Economics Letters, 155*, 121–125.
36. Galbraith, J. W., & Tkacz, G. (2015). *Nowcasting GDP with electronic payments data*. Working Paper No. 10, Statistics Paper Series, European Central Bank.
37. Ghirelli, C., Peñalosa, J., Pérez, J. J., & Urtasun, A. (2019a). *Some implications of new data sources for economic analysis and official statistics*. Economic Bulletin. Bank of Spain. May 2019.
38. Ghirelli, C., Pérez, J. J., & Urtasun, A. (2019). A new economic policy uncertainty index for Spain. *Economics Letters, 182*, 64–67.
39. Ghirelli, C., Pérez, J. J., & Urtasun, A. (2020). *Economic policy uncertainty in Latin America: measurement using Spanish newspapers and economic spillovers*. Working Papers 2024, Bank of Spain. https://ideas.repec.org/p/bde/wpaper/2024.html
40. Gil, M., Pérez, J. J., Sánchez, A. J., & Urtasun, A. (2018). *Nowcasting private consumption: Traditional indicators, uncertainty measures, credit cards and some internet data*. Working Paper No. 1842, Bank of Spain.
41. Gómez, V., & Maravall, A. (1996). *Programs TRAMO and SEATS: Instructions for the user*, Working paper No. 9628, Bank of Spain.
42. Götz, T. B., & Knetsch, T. A. (2019). Google data in bridge equation models for German GDP. *International Journal of Forecasting, 35*(1), 45–66.
43. Hammer, C. L., Kostroch, D. C., & Quirós, G. (2017). *Big data: Potential, challenges and statistical implications*, IMF Staff Discussion Note, 17/06, Washington, DC, USA: International Monetary Fund.
44. Hardy, A., Hyslop, S., Booth, K., Robards, B., Aryal, J., Gretzel, U., et al. (2017). Tracking tourists' travel with smartphone-based GPS technology: A methodological discussion. *Information Technology and Tourism, 17*(3), 255–274.
45. Harvey, A., & Chung, C. (2000). Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistical Society, Series A: Statistics in Society, 163*(3), 303–309.
46. Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 168–177). New York, NY, USA: ACM.
47. Kapetanios, G., & Papailias, F. (2018). *Big data & macroeconomic nowcasting: Methodological review*. ESCoE Discussion Paper 2018-12, Economic Statistics Centre of Excellence.
48. Lacroix R. (2019). The Bank of France datalake. In Bank for International Settlements (Ed.), IFC Bulletins chapters, *The use of big data analytics and artificial intelligence in central banking* (vol. 50). Basel: Bank for International Settlements.
49. Loberto, M., Luciani, A., & Pangallo, M. (2018). *The potential of big housing data: An application to the Italian real-estate market*. Working paper No. 1171, Bank of Italy.
50. Meinen, P., & Roehe, O. (2017). On measuring uncertainty and its impact on investment: Cross-country evidence from the euro area. *European Economic Review, 92*, 161–179.
51. Menéndez, Á., & Mulino, M. (2018). *Results of non-financial corporations in the first half of 2018*. Economic Bulletin No. 3/2018, Bank of Spain.
52. Molina-González, M. D., Martínez-Cámara, E., Martín-Valdivia, M.-T., & Perea-Ortega, J. M. (2013). Semantic orientation for polarity classification in Spanish reviews. *Expert Systems with Applications, 40*(18), 7250–7257.
53. Mueller, H., & Rauh, C. (2018). Reading between the lines: Prediction of political violence using newspaper text. *American Political Science Review, 112*(2), 358–375.
54. Nyman R., Kapadia, S., Tuckett, D., Gregory, D., Ormerod, P. & Smith, R. (2018). *News and narratives in financial systems: exploiting big data for systemic risk assessment*. Staff Working Paper No. 704, Bank of England.

55. Petropoulos, A., Siakoulis, V., Stavroulakis, E., & Klamargias, A. (2019). A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting. In Bank for International Settlements (Ed.), *IFC Bulletins chapters, The use of big data analytics and artificial intelligence in central banking* (vol. 50). Basel: Bank for International Settlements.
56. Pew Research Center (2012). *Assessing the Representativeness of Public Opinion Surveys*. Mimeo. See: https://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/
57. Reforgiato Recupero, D., Presutti, V., Consoli, S., Gangemi, A., & Nuzzolese, A. G. (2015). Sentilo: Frame-based sentiment analysis. *Cognitive Computation, 7*(2), 211–225.
58. Thorsrud, L.A. (2020). Words are the new numbers: A newsy coincident index of business cycles. *Journal of Business and Economic Statistics. 38*(2), 393–409.
59. Trung, N. B. (2019). The spillover effect of the US uncertainty on emerging economies: A panel VAR approach. *Applied Economics Letters, 26*(3), 210–216.
60. Xu, C., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data cleaning: Overview and emerging challenges. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, 26-June-2016* (pp. 2201–2206).
61. Vosen, S., & Schmidt, T. (2011). Forecasting private consumption: Survey-based indicators vs. Google trends. *Journal of Forecasting, 30*(6), 565–578.