BBVA | Research

Big Data at BBVA Research

Big Data Workshop on economics and finance. Bank of Spain

Alvaro Ortiz, Tomasa Rodrigo and Jorge Sicilia

February 2018

BBVA Research

Index

01

Opportunities in the digital era. Big Data at BBVA Research

Traditional data could not answer some relevant questions…

Social awareness and the Arab Spring

Political events and social reaction

Natural disasters and epidemics

… preventing us to measure their economic impact…
… in a world with increasing risks and uncertainty

The use of Big Data and Data science techniques allows us to quantify these trends

BBVA Research

New framework in the digital era…

***Novel data-driven computational approaches** are needed to exploit the new opportunities in the **new digital era** where data can be used to study the world in real time, both at micro or macro level*

- New data availability
- Combination of historical data with real time data
- Better and faster infrastructure
- Advanced data science techniques and algorithms
- New answers to old questions
- Higher computational abilities to face more data granularity

BBVA Research

… which needs the development of new competences to take advantage of it

Making the right questions

Developing the data management and programming capabilities to work with large-scale datasets

Deepening the statistical and econometric skills to analyze and deal with high-dimensional data

Interpreting the results: summarize, describe and analyze the information

BBVA | Research

Big Data at BBVA Research

Our work

- We **analyze geopolitical, political, social and economic issues using large- scale databases and quantitative data-driven methods** rather than just qualitative introspection

Our datasets

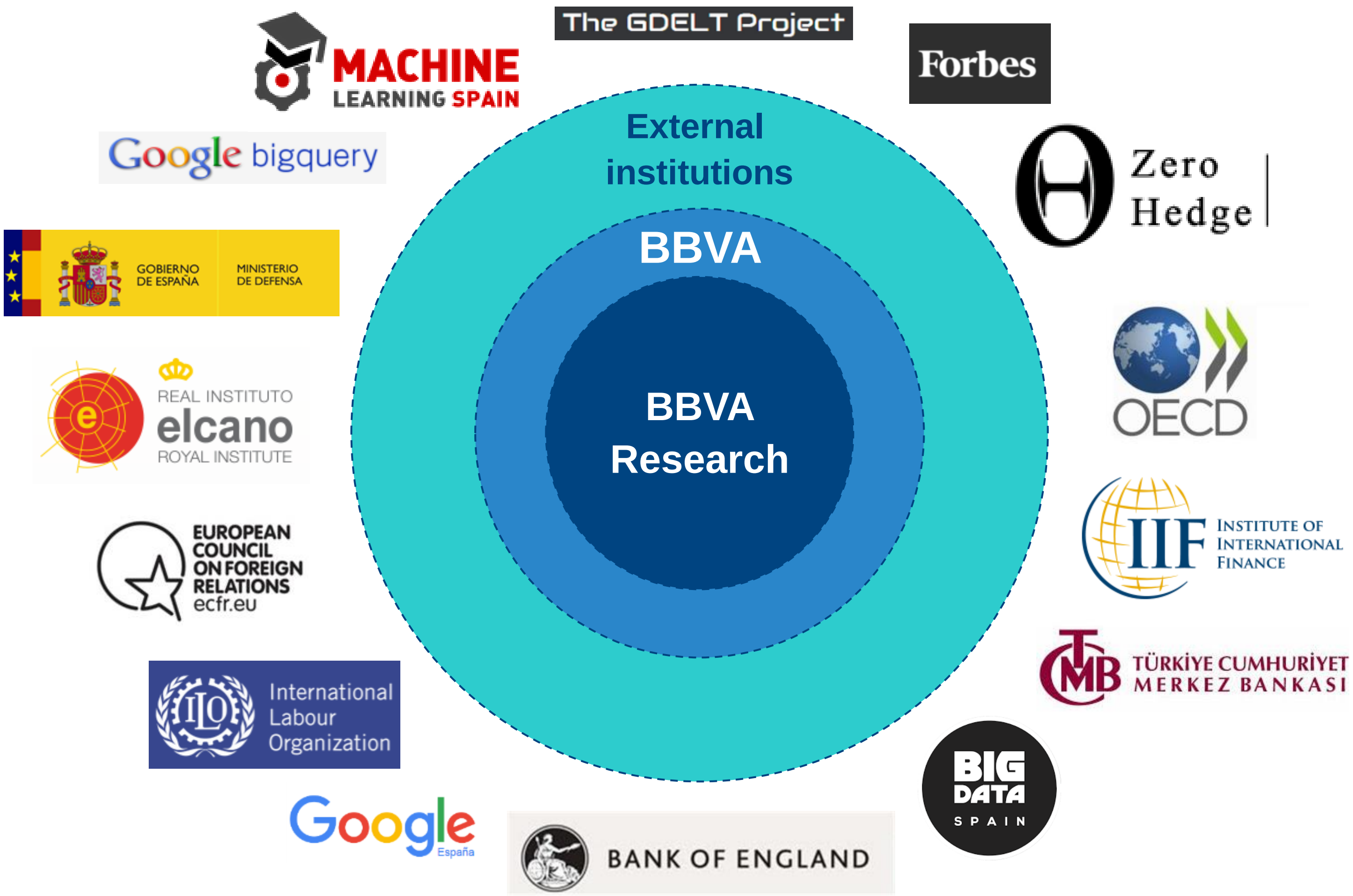
- **Media data** to exploit news intensity, geographic density of events (location intelligence) and emotions across the world (sentiment analysis)
- **BBVA aggregate and anonymized data** from clients digital footprint
- **Data from the web** (Central Banks' reports among others)

Our results

- We are at the **research frontier in the geopolitical and economic area** contributing to the innovation and increasing our internal and external reach

BBVA | Research

Internal and external diffusion



Our working process

Databases	SaaS	Analysis	Visualization
GDELT BBVA data Google search Web	BigQuery and Amazon Redshift	Clean, Aggregate transform and model the data	Fuse, visualize & analyze the data
The GDELT Project BBVA	Google BigQuery amazon REDSHIFT	BBVA Research	CARTO Gephi Google Data Studio

Mean takeaways working with Big Data

It helps us to …

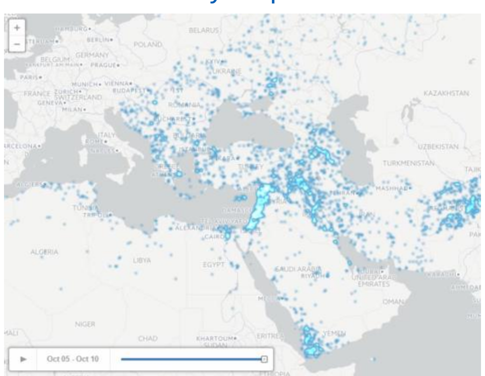
Complement and enrich our traditional databases with high dimensional data:

- Quantifying **new trends** and exploiting new **dimensions**
- Having **timely answers** on the impact of different events
- Improving our models performance at **nowcasting**

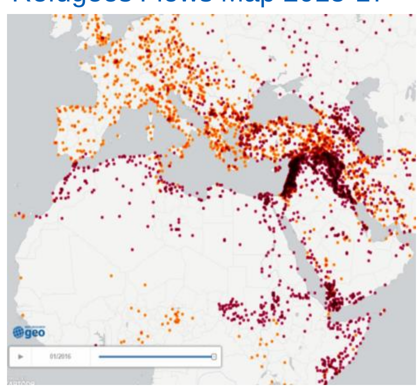
… but still some challenges

- Data challenges: missing data, data sparsity, data quality,…
- There's not enough time horizon to improve our models performance at forecasting

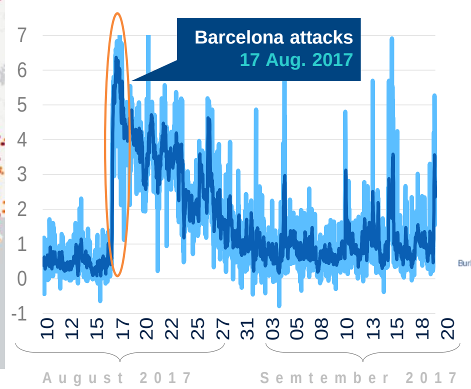
Conflict Intensity Map 2017



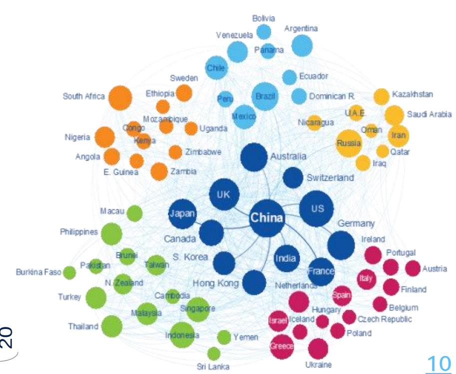
Refugees Flows Map 2015-17



Barcelona Instability index 2017



Chinese slowdown: country network



Data treatment and robustness check

To face with new and high dimensional data

1

Data treatment and analysis:

Data cleaning, missing values, outlier detection, high heterogeneity, sparsity,…

New methodologies to face data challenges: dimensionality reduction, clustering, regularization,…

!

Massive and unstructured datasets:

Importance of making the right questions

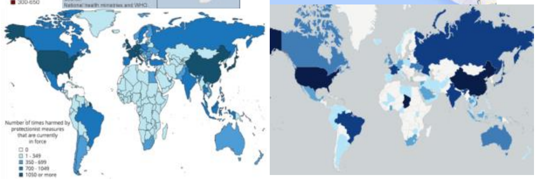
2

Robustness check:

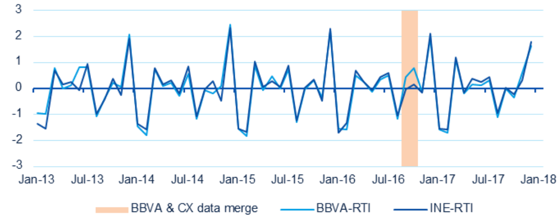
Cross-check of Big Data outcome with traditional data and methodologies

Ebola Outbreak: WHO and GDELT

Protectionism: GTA and GDELT



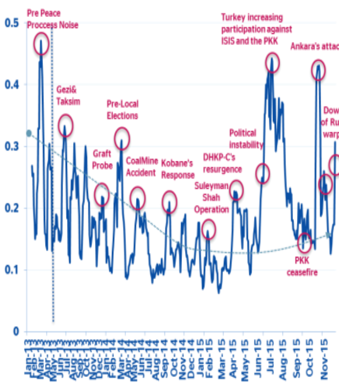
Retail sales: INE and BBVA



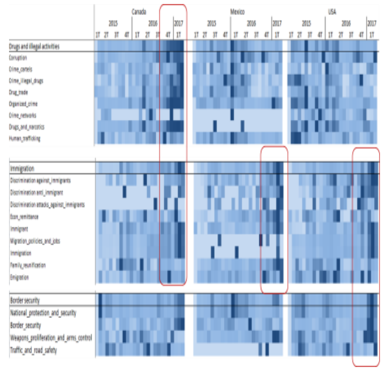
BBVA Research

Our products

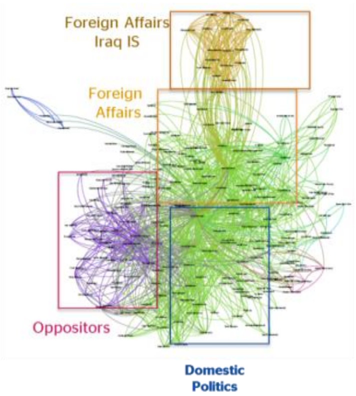
Political, Geopolitical Social Indexes
(Political Indexs)



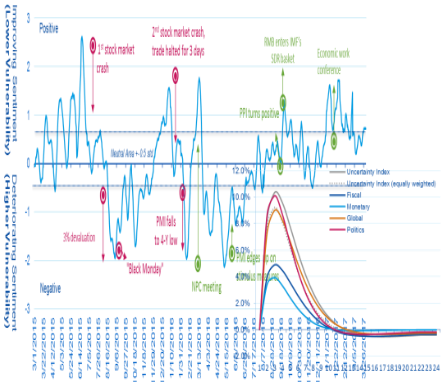
Color Maps NAFTA Topics
(Nafta Project)



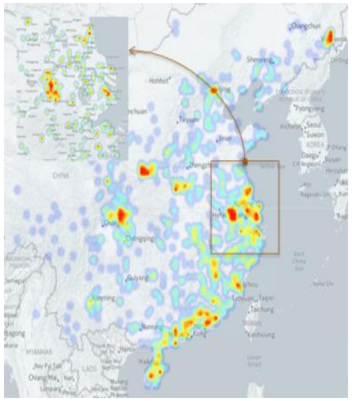
Politics & Financial Networks
(Political Netwoks)



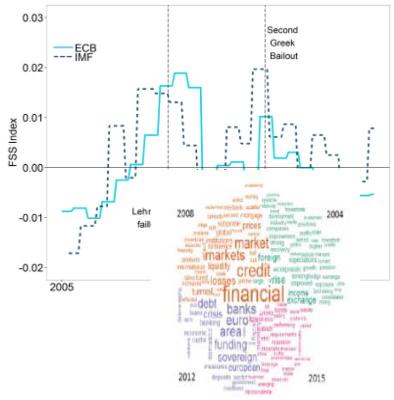
Mix Hard data & Sentiment & VAR models
(CBSI and Turkey Sentiment Indexes)



Geographical Analysis Housing Prices
(sentiment on Housing Prices)

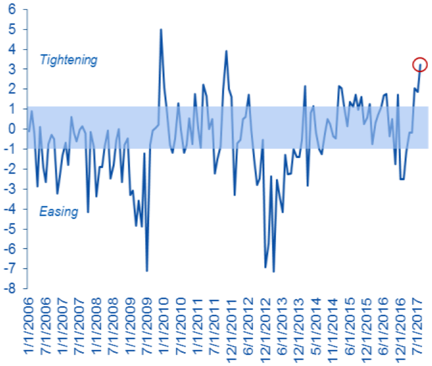


Financial Stability & Macroprudential
(ECB & FED FS index by FED Board)



Measuring Sentiments
(sentiment Analysis on Economy and Society)

Monetary & Stability tones by Central Banks



02

Big Data & Big Models: Applying Big Data at BBVA Research

Big Data & Models at BBVA Research: Some Examples

1 **BBVA Data**

(“ Transactions Geo-referenced Data”)

Hard Data

2 **International News (GDELT)**

(Narratives & Sentiment)

Hard + Market + Text Data

3 **Policy Reports**

(Topics , Networks & Sentiment)

Text Data

BBVA Research

1

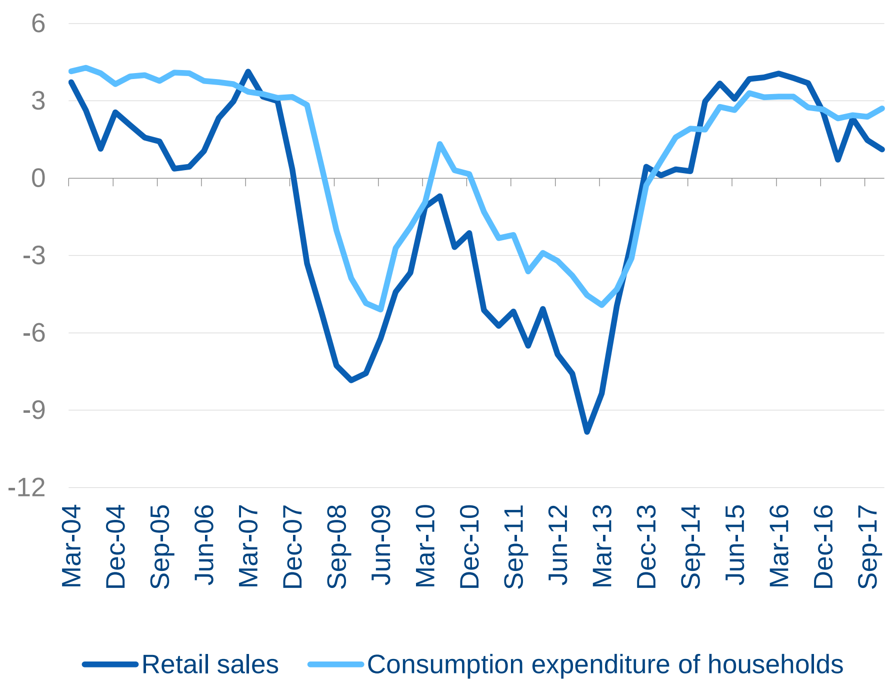
BBVA Data

(“ Transactions Geo-referenced Data”)

Real Time Indicator of Retail Sales (Spain)

Retail trade sector dynamic leads the evolution of the aggregate consumption, which represents a high share of the GDP

Spanish case: Retail sales vs. Consumption expenditure of households
(%, YoY)



- Having accurate estimations on the evolution of the retail trade sector activity is of main importance given it is a key indicator about the current economic situation

Source: BBVA Research and BBVA Data & Analytics from INE data

The Retail Sales Index: Matching - internal and external sources

INTERNAL TAXONOMY - SPAIN

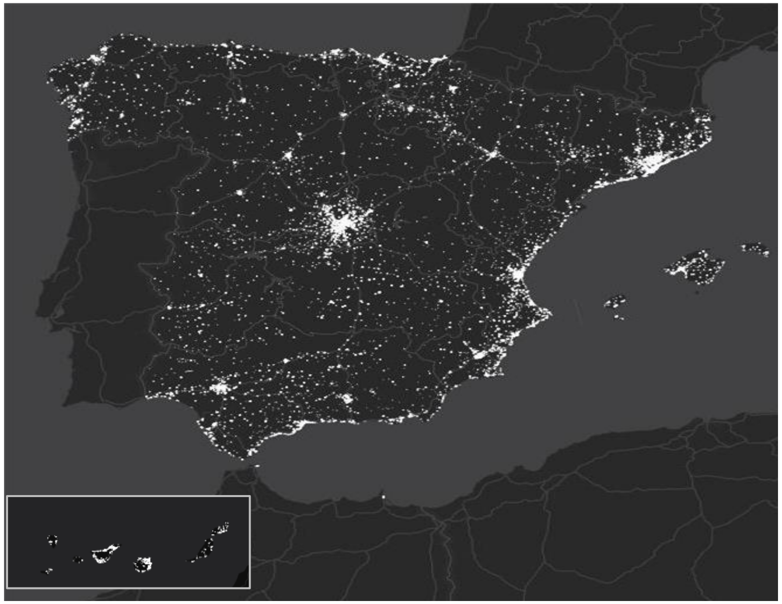
REDSHIFT

CATEGORY (FASHION) → **SUBCATEGORY** (FASHION-BIG) → **RAMO / GIRO** (TEXTILES AND CLOTHING) → **CIF / RFC** (CADENA ZARA) → **FUC / AFILIACIÓN** (ZARA, Gran Vía, Madrid) → **POS ID** (TPV)

EXTERNAL TAXONOMY - SPAIN

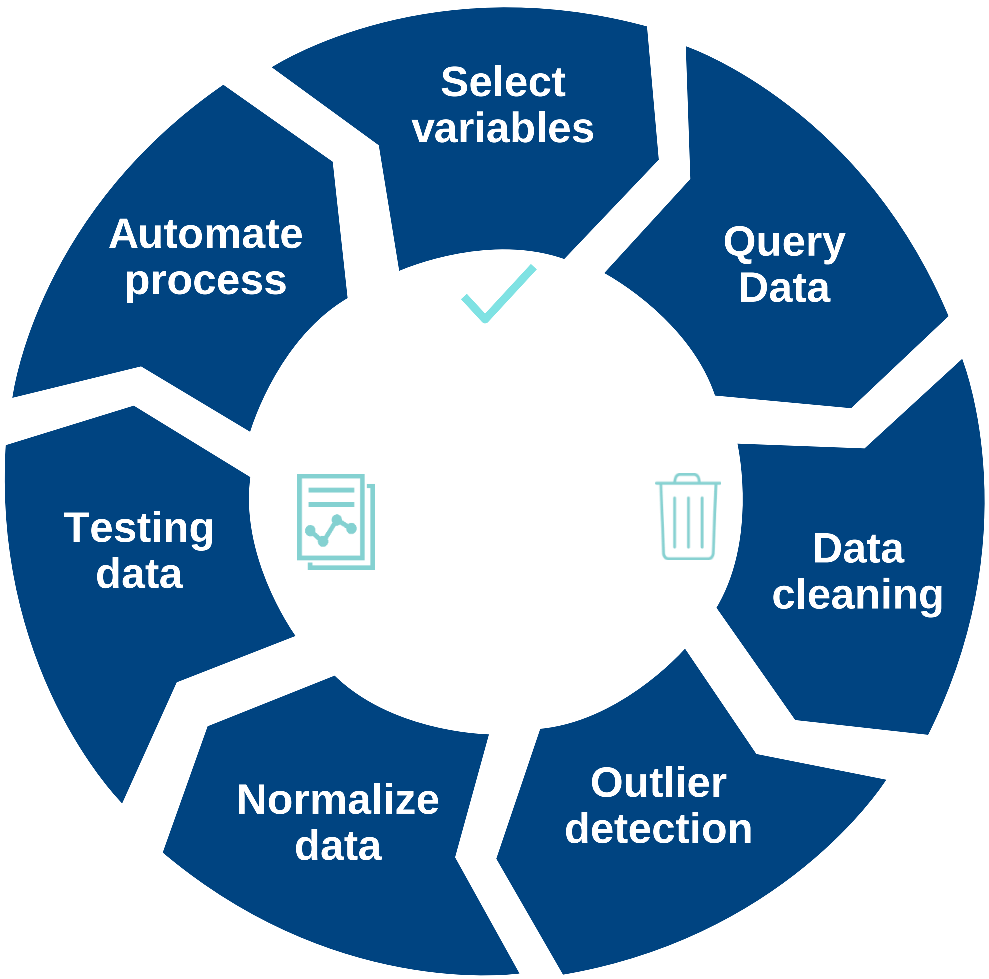
INE: Instituto Nacional de Estadística
5 distribution classes:

- **service stations**
- **single retail stores** (one premise)
- **small chain stores** (2-24 premises & <50 employees)
- **large chain stores** (25 or more premises, and 50 or more employees)
- **department stores** (sales area greater than or equal to 2.500m^2)



More information can be found in the annex

Data extraction, cleaning and transformation



BBVA Research

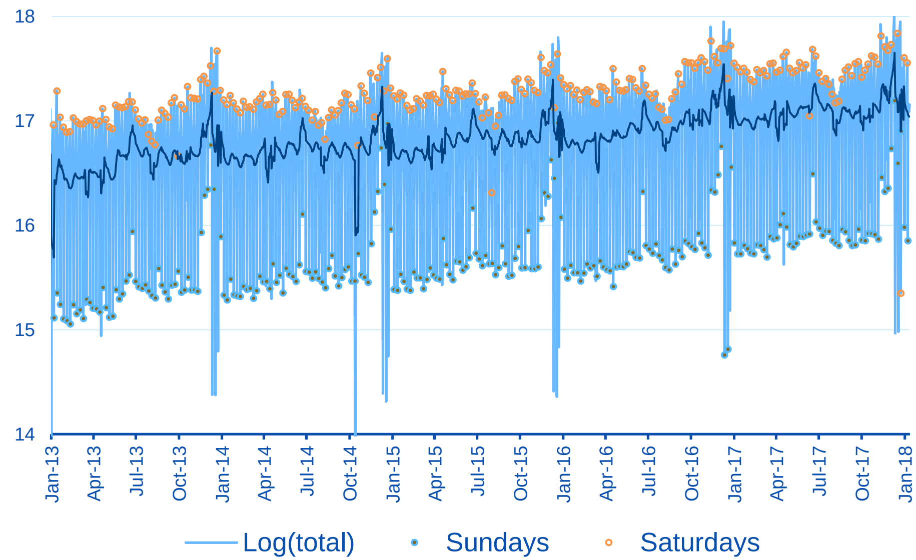
BBVA DATA & ANALYTICS

Using BBVA data, we replicate national figures, gaining frequency…

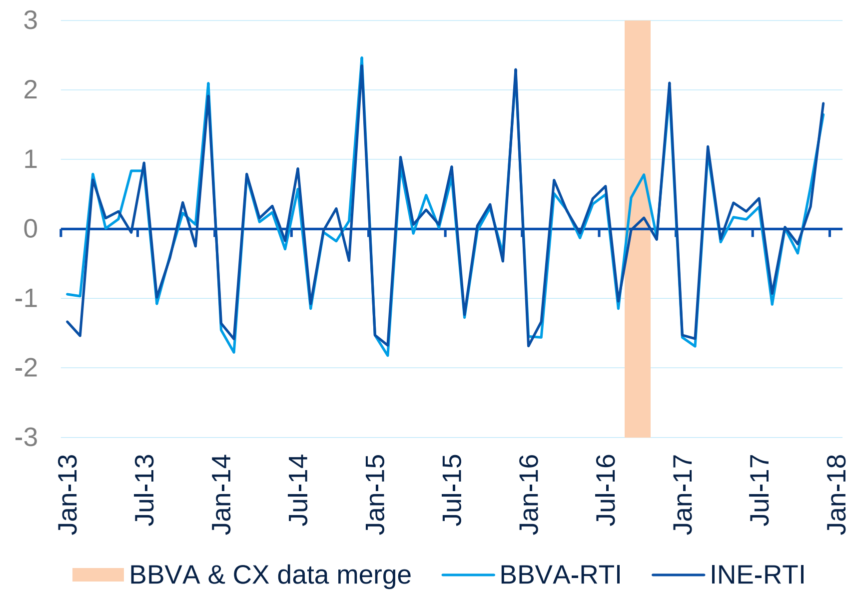
A "High Definition" Retail Sales Index (RTI) for Spain* (and Mexico)

(BBVA consumption indicator for the optimal allocation of BBVA's resources and products)

RTI–BBVA Index, in millions of euros and daily basis



Comparison Retail Sales (RTI) - INE and BBVA on monthly basis (standardized monthly growth)



What "HIGH DEFINITION(*)" means here:

High granularity:
Dynamics down to subnational level

Ultra High Frequency:
Dynamics up to sub-monthly frequency

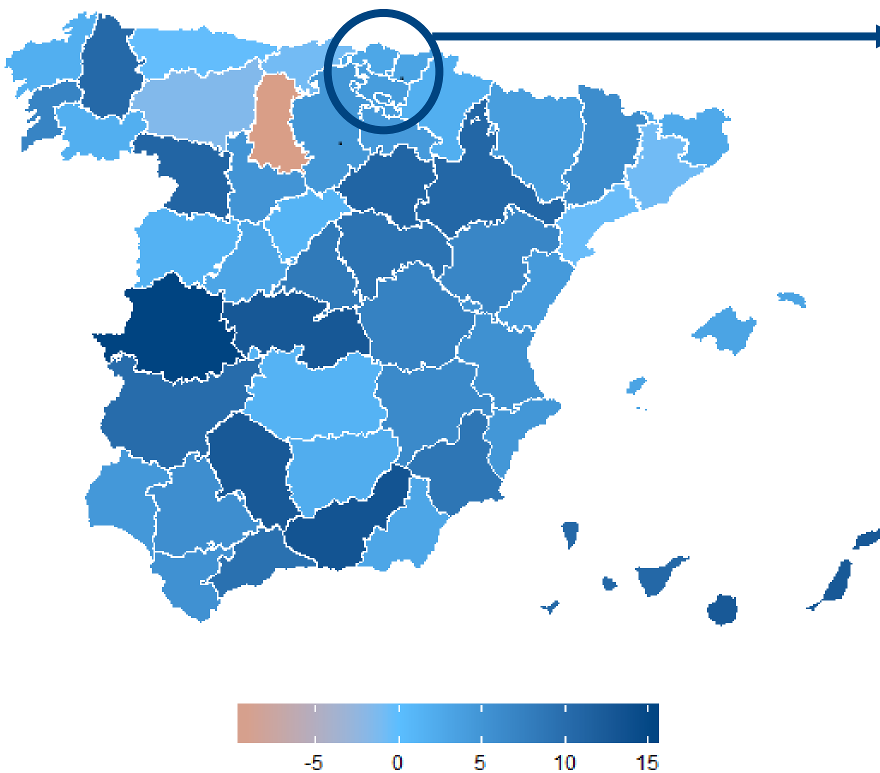
Multi Dimensional:
More detailed socioeconomic features

Bodas et al. (2018, Forthcoming). Measuring retail trade through card's transaction data.
Source: BBVA Research and BBVA Data & Analytics

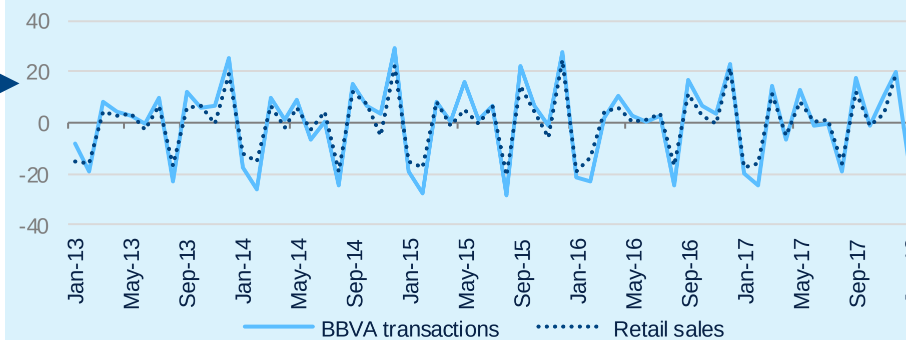
Going further from national figures, the retail sales at regional level …

BBVA transactions December 2017

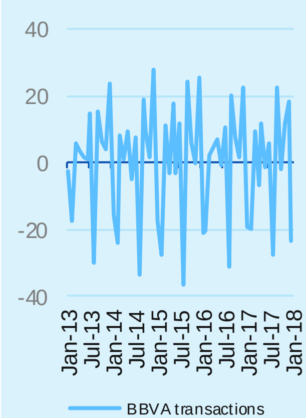
(% yoy)



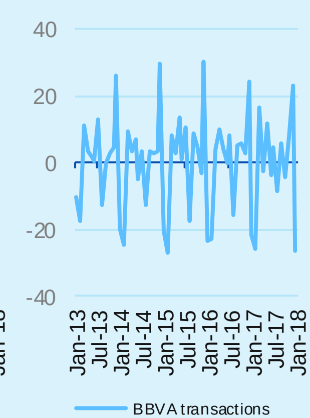
Basque Country



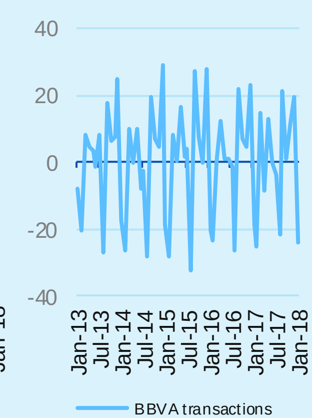
Álava



Vizcaya



Guipúzcoa

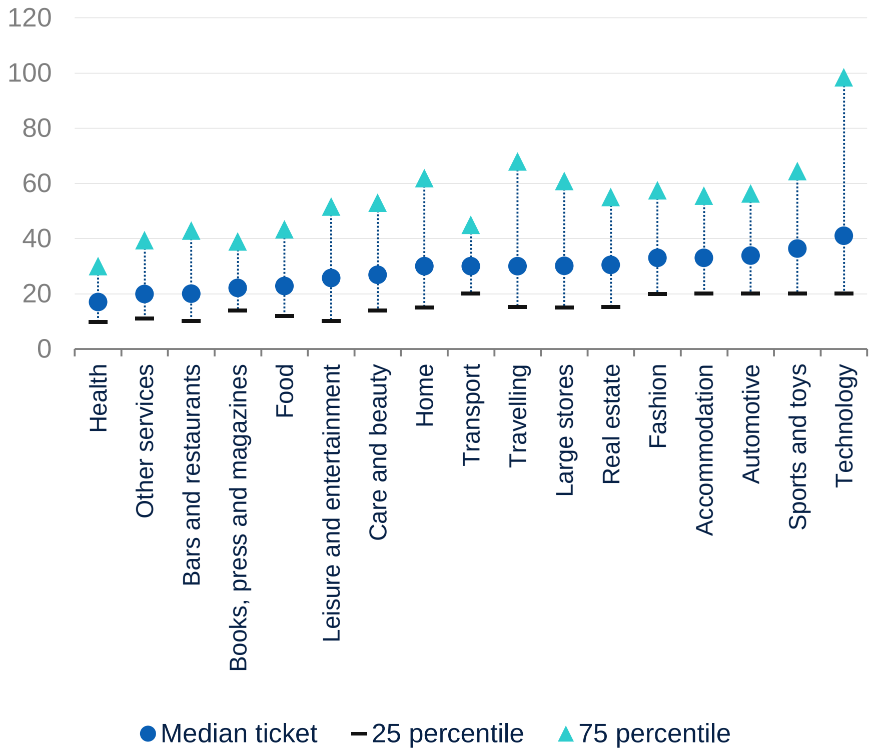


Source: BBVA Research and BBVA Data & Analytics

… and by sector and "size" of activity

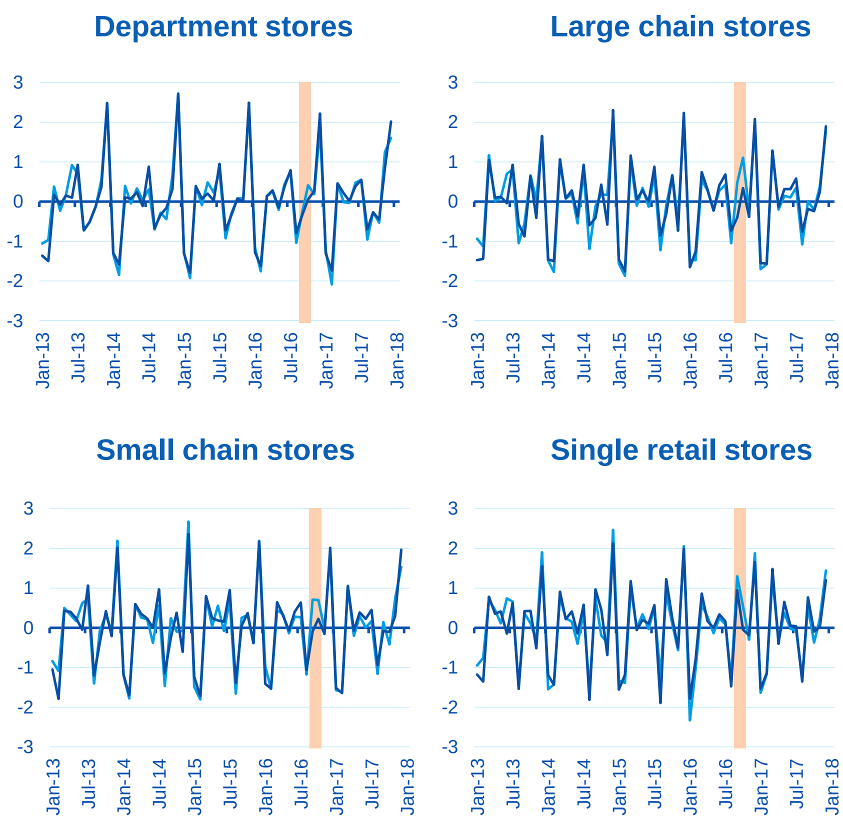
BBVA retail sales index by sector of activity

(median ticket in December 2017)



BBVA retail sales by distribution classes

(standardized monthly growth)



Source: BBVA Research and BBVA Data & Analytics

BBVA Research

International News (GDELT) 2

(Narratives & Sentiment)

Vulnerability Sentiment Indicator (CSVI) (China)

Tracking China Vulnerability in Real Time: Mixing Data and Sentiment

Hard Data Indicators

… provide accurate information… but at lower frequencies and with delays…

Market Indicators

… Real Time but limited Information …. also influenced by global factors…

Sentiment Indicators

… complementing Hard-Soft-Markets in Real Time "sentiment" on Special topics not quotes

BBVA Research

A Balanced set of Information in the Database: Hard Data, Markets and Sentiment

Chinese Vulnerability Sentiment Index (CVSI): components and evolution

China Vulnerability Sentiment Index (CVSI)

SOE Vulnerability Index (SOEI)	Housing Bubble Vulnerability Index (HBI)	Shadow banking Vulnerability Index (SBI)	FX Speculative Pressure Index (FXI)
Principal Components Analysis on each component Tone			
Hard & Financial data			
Total.profits (M) Liabilties (M) 25%	Mortages.loan (M) GICS.Housing.Index (M) Housing.Price (M) New.Construction (M) RealEst.Invest (M) 45%	NPL.Ratio (M) TSF.Aggregate.New Increase (M) Entrusted.Loans (M) Wenzhou.Index (D) WMPs Acceptances (M) 35%	Foreign.Reserves (D) CNY Exchange Rate (D) CNH Exchange Rate (D) HICNHON.Index (D) 40%
Big data (GDELT) indicators in real time			
State_owned_enterprises (D) Resource_misallocs_&policy Failure (D) Resource_misallocs&SOEs (D) Institutional_reform_&_SOEs (D) Industry_policy (D) Industry_laws_and_regulations (D) Local_government_and_SOEs (D) Debt_and_SOEs (D) 75%	Housing_policy_&_institutions (D) Housing_markets (D) Housing_prices (D) Housing_construction (D) Housing_finance (D) Land_reform (D) 55%	Non_bank_financial_institutions (D) Asset_management (D) Bank_capital_adequacy (D) Financial_sector_instability (D) Banking_regulation (D) Infrastructure_funds (D) Financial_vulnerability_&_risks (D) Monetary_&_financial_stability (D) State_financial_institutions (D) 65%	Currency_exchange_rate (D) Currency_reserves (D) Capital_account (D) Macroprudential_policy (D) Exchange_rate_policy (D) Illicit_financial_flows (D) 60%

Source: www.gdelt.org & BBVA Research

BBVA Research

Results show the Importance of Sentiment

Chinese Vulnerability Sentiment Index (CVSI): Weights

SOE Vulnerability

Variable (Type)		Weight
Tota Profits	(HD)	19.63
Institutional Reform & SOEs	(S)	12.37
Debt & SOE	(S)	11.92
Liabilities	(HD)	10.62
Local Goverment & SOE	(S)	9.75
Industry Policy	(S)	9.5
Resource Mis. & P. Failure	(S)	8.18
SOE	(S)	7.15
Industry Laws & Regulation	(S)	5.28
Resource Misaloc. & SOE	(S)	5.61
%of Sentiment in Component	(S)	69.76
% of Variance by 1st PC		63.2
in the CSVI		**29.18**

Housing Bubble

Variable (Type)		Weight
New Construction	(HD)	16.37
Mortgages Loans	(HD)	14.57
Land Reform	(S)	12.62
Housing Price	(HD)	11.6
Housing Construction	(S)	10.59
Housing_Prices	(S)	10.05
Housing Policy & Institutions	(S)	8.94
Housing Finance	(S)	7.83
Housing Markets	(S)	6.71
GICS Housing Index	(M)	0.36
Real State Investment	(HD)	0.31
%of Sentiment in Component	(S)	56.74
% of Variance by 1st PC		65.8
Weight in the CSVI		**26.05**

Shadow Banking

Variable (Type)		Weight
Wenzhou Index	(M)	16.58
WMP Yields	(M)	13.63
Infrastructure_funds	(S)	10.92
NPL ratio	(HD)	9.46
State & Finantial Inst	(S)	8.95
Banking_Regulation	(S)	7.28
Financial Vulnerability	(S)	6.82
Asset_Management	(S)	5.62
Financial Sector Instability	(S)	5.35
Bank Capital Adequacy	(S)	4.6
Non Bank Financial Inst	(S)	4.35
Monetary & F.Stability	(S)	3.54
Acceptances	(HD)	2.22
TSF Aggregate New	(HD)	0.57
Entrusted Loans	(HD)	0.13
%of Sentiment in Component	(S)	57.43
% of Variance by 1st PC		59.5
Weight in the CSVI		**23.12**

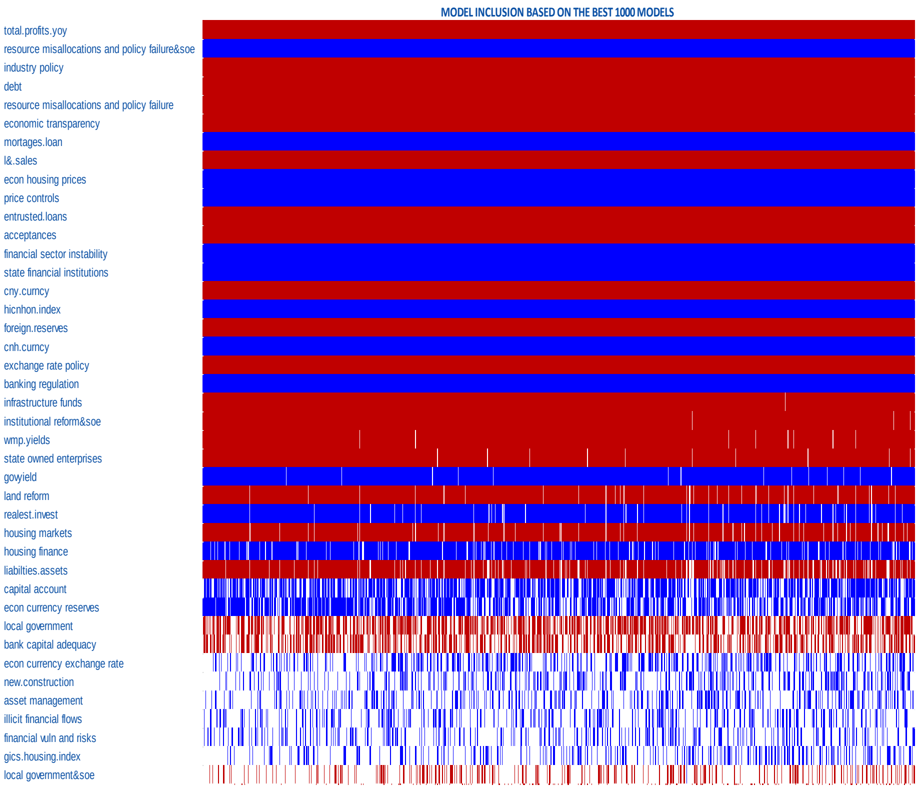
FX Speculative Pressure

Variable (Type)		Weight
Currency Exchange Rate	(S)	19.94
Exchange Rate Policy	(S)	17.95
Macroprudential_Policy	(S)	15.33
Hinchon Index	(M)	13.84
CNY Currency	(M)	11.92
Capital Account	(S)	10.05
CNH Currency	(M)	8.73
Illicit Financial Flows	(S)	1.58
Foreign Reserves	(S)	0.6
Currency Reserves	(S)	0.06
%of Sentiment in Component	(S)	48.27
% of Variance by 1st PC		78.99
Weight in the CSVI		**21.64**

Source: BBVA Research. Ortiz et al. (2017). Tracking Chinese Vulnerability in Real Time Using Big Data

Bayesian Model Averaging Robustness check confirms the relevance of Sentiment in explaining Market Risk proxies

Bayesian Model Averaging Results
(PIP Inclusion after 1000 models)



Bayesian Model Averaging Results: X with PIP>50%
(PIP Inclusion after 1000 models)

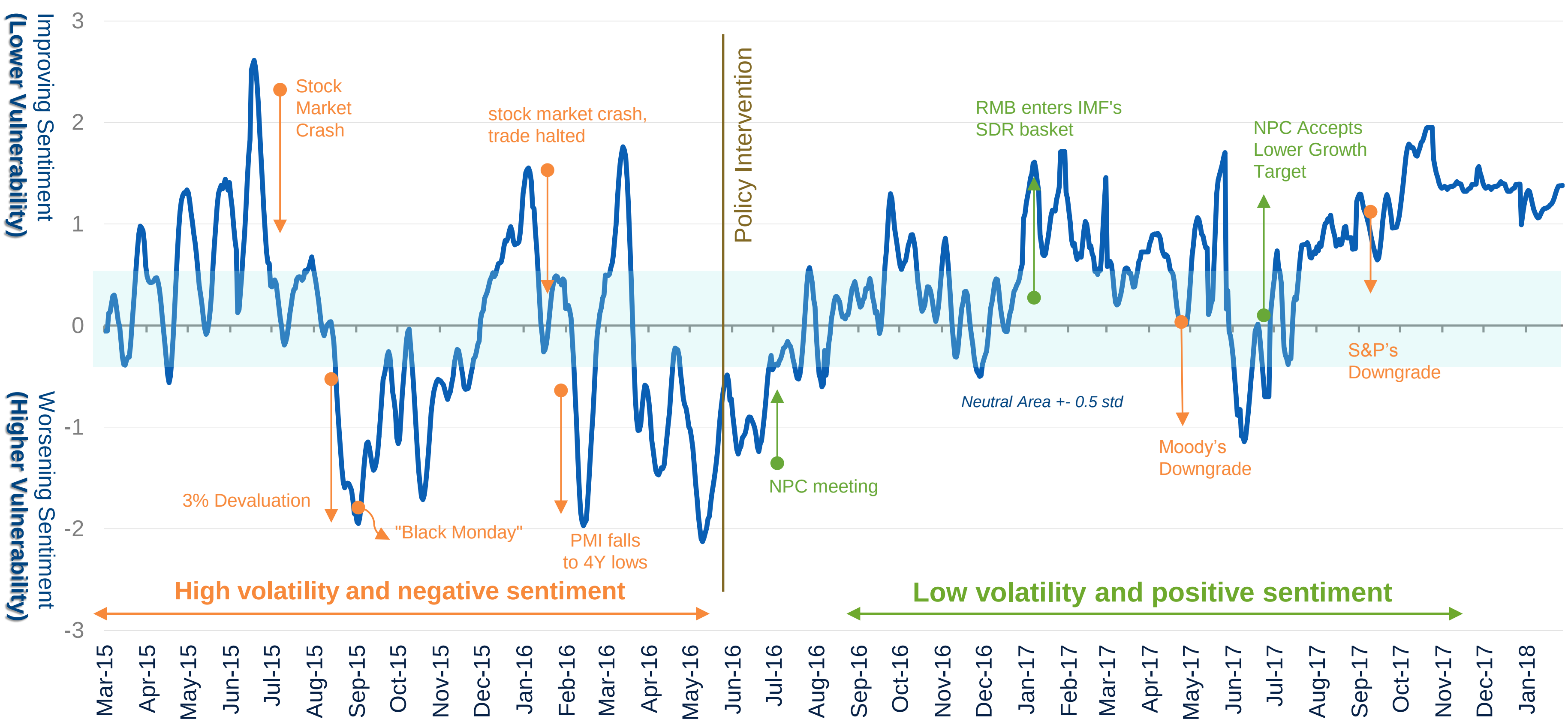
	Dependent variable: CDS Spread	Type of data	Component	PIP	Posterior Mean	Posterior standard deviation SD
1	Total profits	Hard Data	SOE	1.000	-0.358	0.042
2	Institutional reform&SOE	Sentiment	SOE	1.000	-0.089	0.019
3	Industry policy	Sentiment	SOE	1.000	-0.112	0.015
4	Economic transparency	Sentiment	Shadow Banking	1.000	-0.069	0.015
5	mortages loan	Hard Data	Housing bubble	1.000	1.014	0.105
6	housing finance	Sentiment	Housing bubble	1.000	0.086	0.016
7	NPL ratio	Financial	Shadow Banking	1.000	-0.976	0.130
8	Acceptances	Financial	Shadow Banking	1.000	-0.264	0.053
9	**CNY curncy**	Financial	FX	1.000	-0.811	0.086
10	Econ currency reserves	Sentiment	FX	1.000	0.129	0.020
11	Exchange rate policy	Sentiment	FX	1.000	-0.113	0.017
12	Ilicit financial flows	Sentiment	FX	1.000	0.063	0.015
13	Industry laws and regulations	Sentiment	SOE	0.995	0.062	0.016
14	Hicnhon index	Financial	FX	0.994	0.080	0.022
15	State owned enterprises	Sentiment	SOE	0.953	-0.055	0.020
16	Econ housing prices	Sentiment	Housing bubble	0.927	0.058	0.025
17	Financial sector instability	Sentiment	Shadow Banking	0.908	0.089	0.038
18	Wenzhou index	Financial	Shadow Banking	0.860	0.071	0.040
19	Asset managment	Sentiment	Shadow Banking	0.817	0.042	0.025
20	Banking regulation	Sentiment	Shadow Banking	0.791	0.032	0.020
21	Macroprudential policy	Sentiment	FX	0.716	-0.030	0.023
22	Housing policy and institutions	Sentiment	Housing bubble	0.706	0.030	0.023
23	New construction	Hard Data	Housing bubble	0.697	0.045	0.035
24	Entrusted loans	Financial	Shadow Banking	0.674	-0.072	0.058
25	Real Estate Investment	Hard Data	Housing bubble	0.655	0.055	0.048
26	Non bank Financial institutions	Sentiment	Shadow Banking	0.655	-0.029	0.025

Source: BBVA Research

The index shows that Vulnerability sentiment has been improved since the authorities implemented some policies

Chinese Vulnerability Sentiment Index (CVSI)

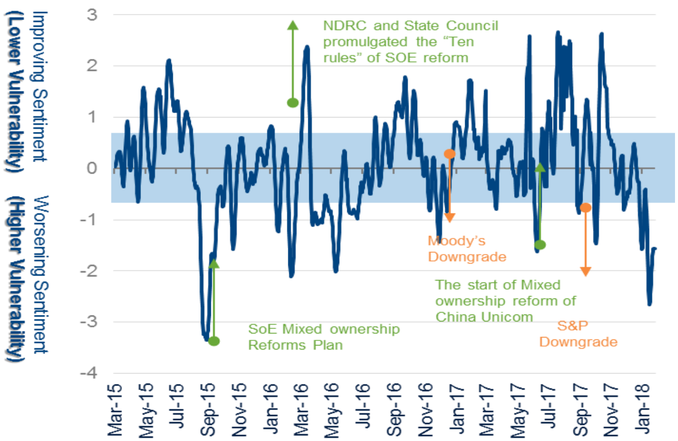
(Evolution of the “Tone” or “Sentiment”. Lower values indicate a deterioration of sentiment and higher vulnerability)



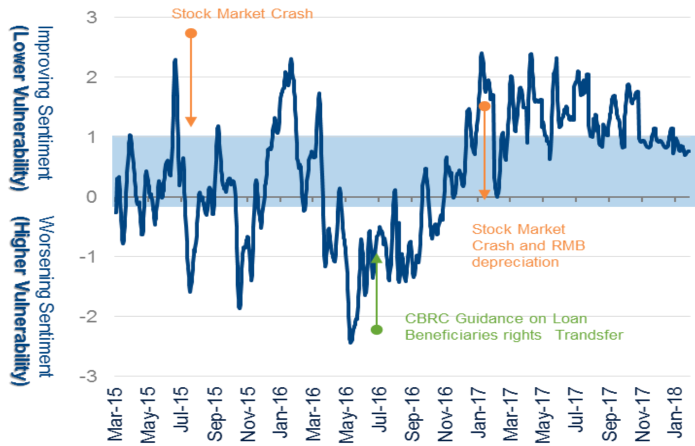
Source: www.gdelt.org & BBVA Research

The performance of the components has not been uniform…

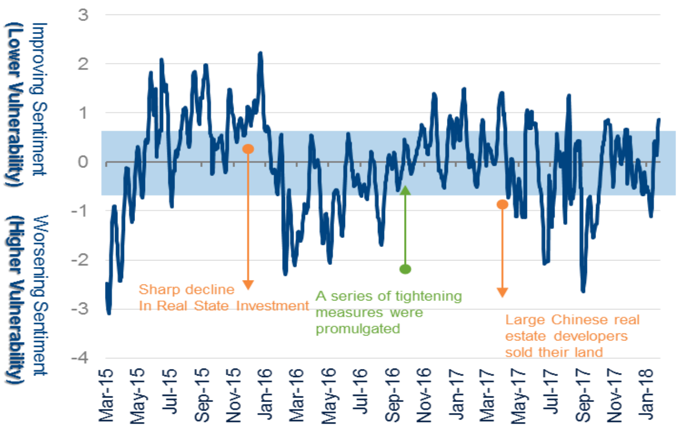
CVSI: SOEs Component



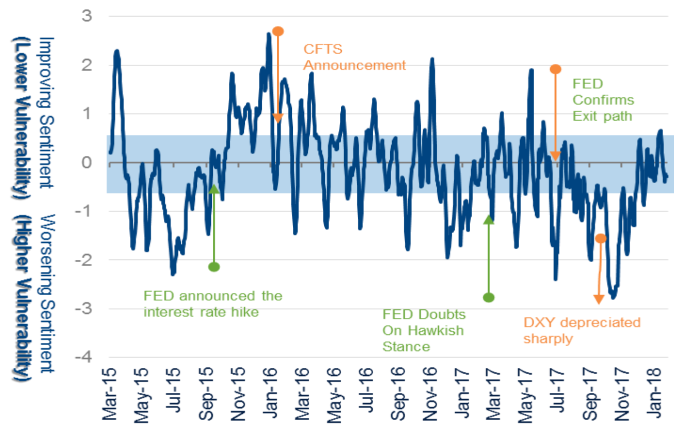
CVSI: Shadow Banking Component



CVSI: Real State Component



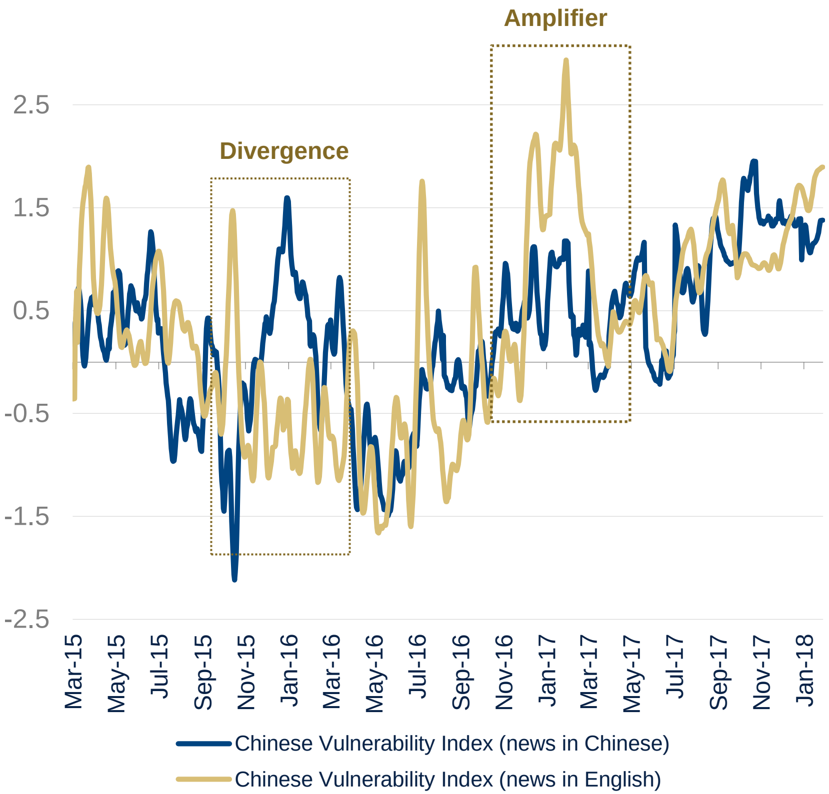
CVSI: External Component



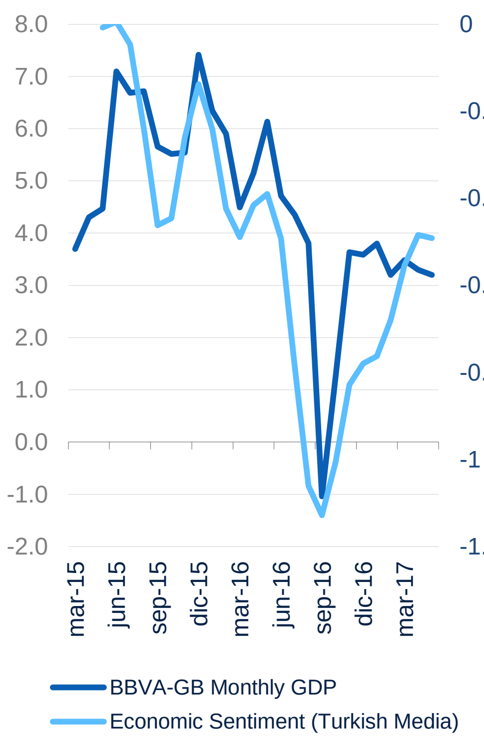
BBVA | Research

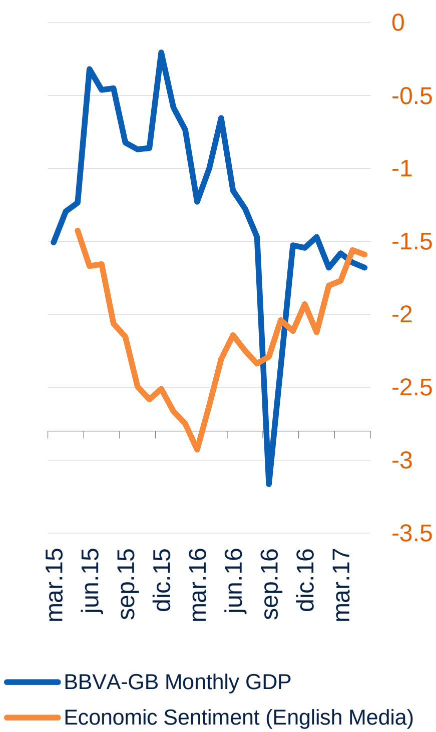
The role of language: "What" & "Where" you read matters

China CSVI: English vs Chinese
(index)



Turkey Econ. Sentiment: English vs Turkish News
(% YoY and index)

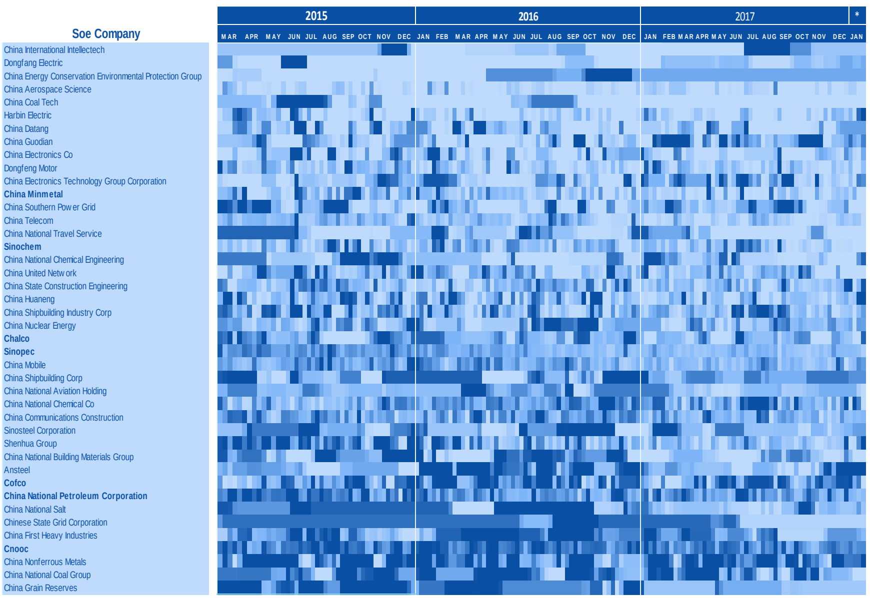




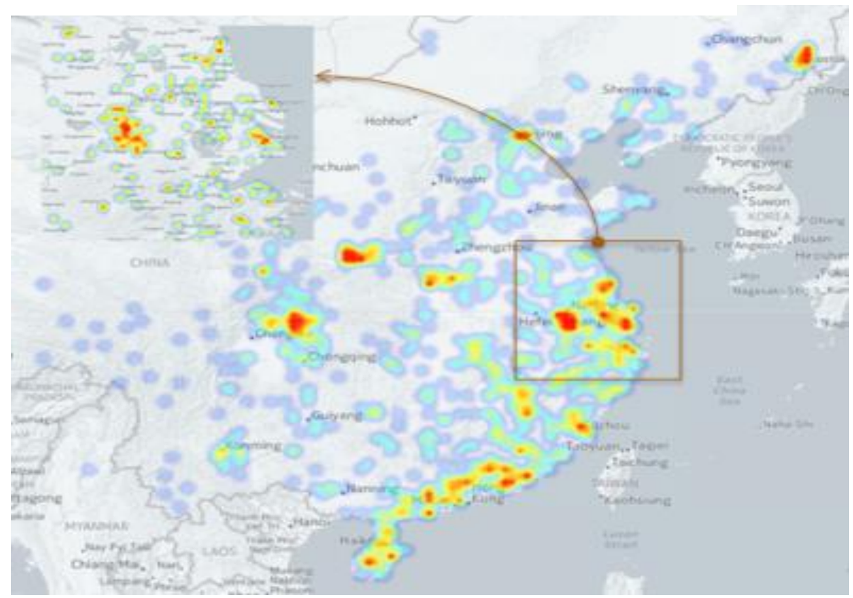
Source: www.gdelt.org & BBVA Research

Some of the tools developed in the analysis can be used to track vulnerability in a very granular way…

China SOE Sentiment Diffusion Map
(sentiment on SOE)



Geographical Analysis Housing Prices
(sentiment on Housing Prices)



Source: www.gdelt.org & BBVA Research

Policy Reports

3

(Topics , Networks & Sentiment)

Monetary Policy Topics & Sentiment (Turkey)

BBVA | Research

Text as Data: "Natural Language Processing (NLP) & Text Mining" for analysis of the Central Bank of Turkey

- We use communication reports of the Central Bank of the Republic of Turkey CBRT from 2006 to October 2017
- We Analyze **"What" the CBRT is talking about** through Latent Dirichlet Allocation (LDA) and Dynamic Topic Models (DTM)
- We apply **network analysis** to understand Monetary Policy Complexity
- We check **"How" the Central Bank talks** by using Sentiment Analysis (Dictionary Assisted)
- We design some analytical tools to understand the Monetary Policy of Turkey through the official documents

External databases: web scrapping and NPL techniques

Information extraction

- Documents
- Web pages

Pre-Processing and text parsing

- Extract words
- Identify parts of speech
- Tokenization and multi-word tokens
- Stopword Removal
- Stemming
- Case-folding

Transformation

- Text filtering
- Indexing to quantify text in lists of term counts
- Create the Document-term matrix
- Weighting matrix
- Factorization (SVD)

Text mining and NPL

- Analysis and Machine learning
- Topics extraction (LDA)
- Clustering
- Modelling (STM and DTM)

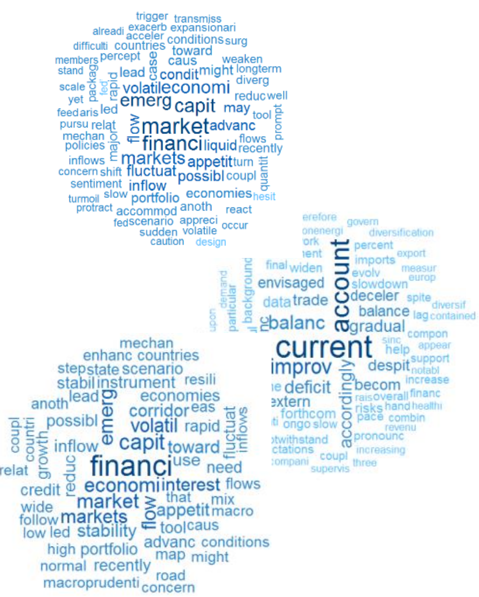
Sentiment analysis

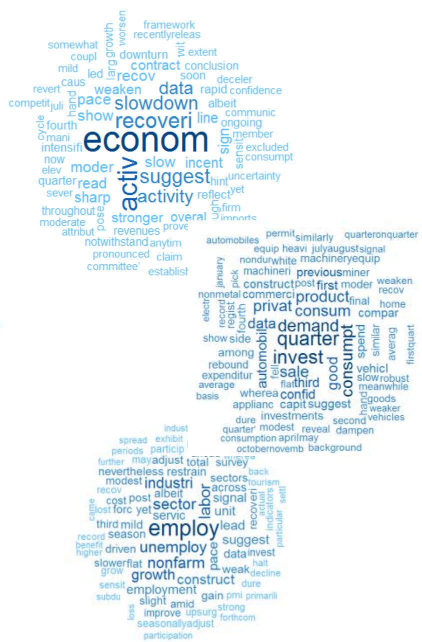
- Apply sentiment dictionaries
- Semantic analysis and classification
- Clustering

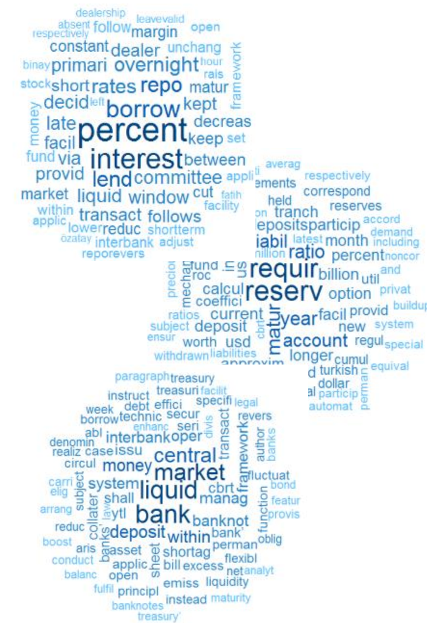
More information can be found in the annex

First we identify the topics: Word clouds will help us to understand and identify topics… here there is a big room for the Researcher

Global Flows	Activity	Inflation	Monetary Policy



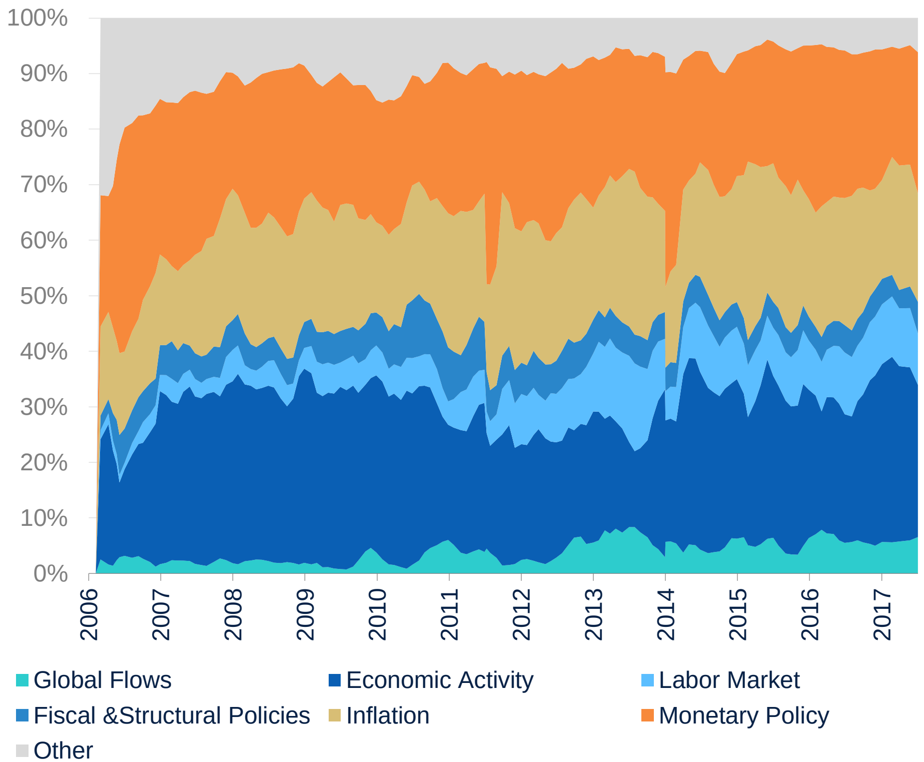




Each word cloud represents the probability distribution of words within a given topic. The size of the word and the color indicates its probability of occurring within that topic

What's the CBRT talking about? We aggregate topics in groups… to see the "dynamics" of Central Bank Communication over time…

Central Bank of Turkey Topics Evolution
(in % of total)

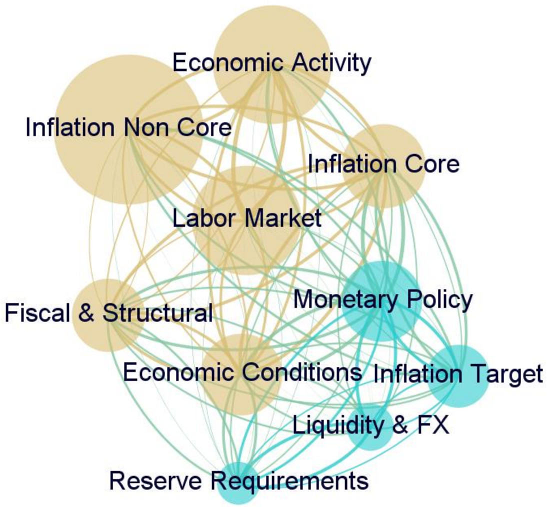


- **Monetary Policy maintains its share but increases in "Stress" periods**
- **Inflation remains stable**
- **Discussions on Structural Policies remain at minimum***
- **Employment issues increased Relevance since the crisis**
- **Economic Activity discussion has returned to the fore**
- **The Global Capital flows are increasingly important**

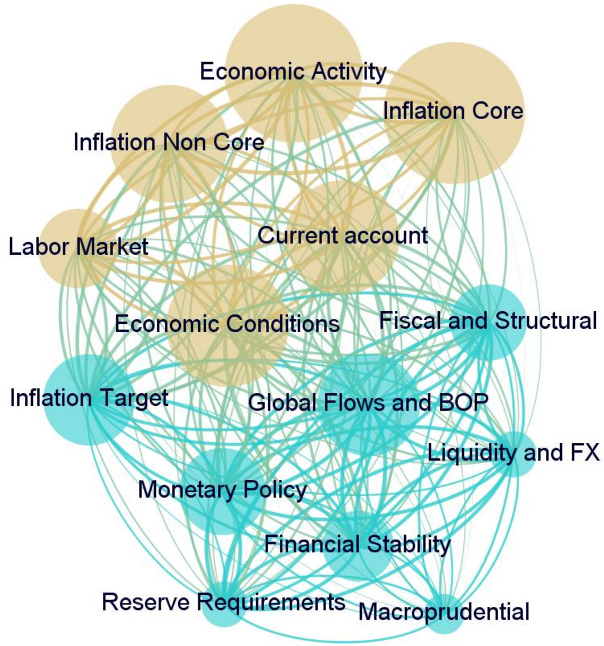
Source: BBVA Research. Ortiz et al (2017). How do the EM Central Bank talk? A Big Data approach to the Central Bank of Turkey

Networks can be very useful to understand how the Central Banks elaborate their strategy and the interconnectedness of topics

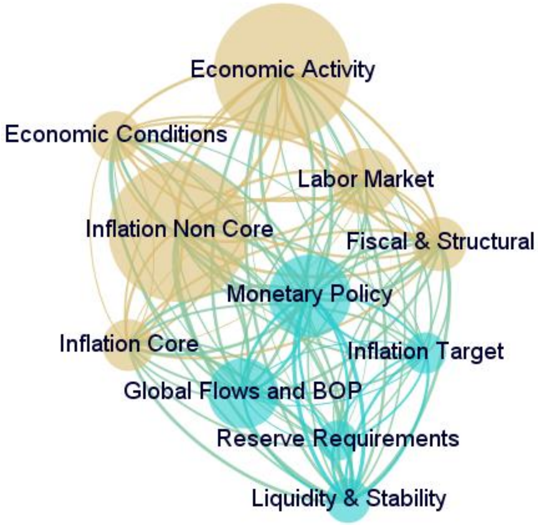
Full Fledge Inflation Target
2006-09



The global financial crisis
2010-15



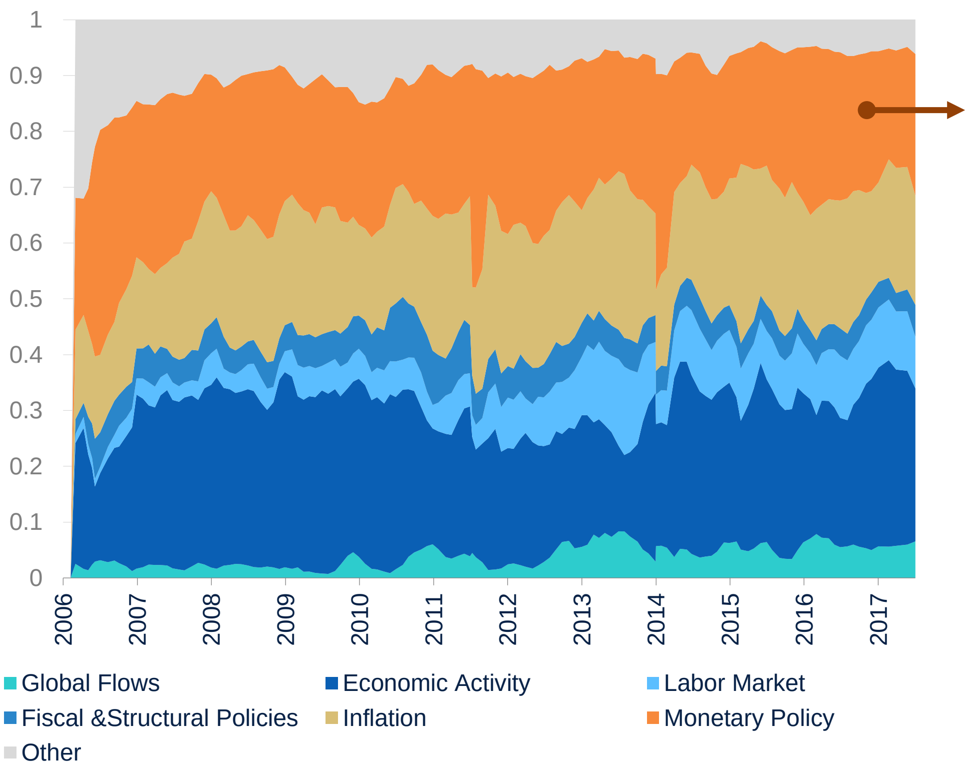
In search of price stability
2016-17



The network of the estimated and correlated topics using STM. The nodes in the graph represent the identified topics. Node size is proportional to the number of words in the corpus devoted to each topic (weight). Node color indicates clusters using a community detection algorithm called modularity developed by Blondel et al (2008). Topics for which labeling is Unknown are removed from the graph in the interest of visual clarity. Edges represent words that are common to the topics they connect (co-ocurrence of words between topics). Edge width is proportional to the strength of this co-ocurrence between topics.

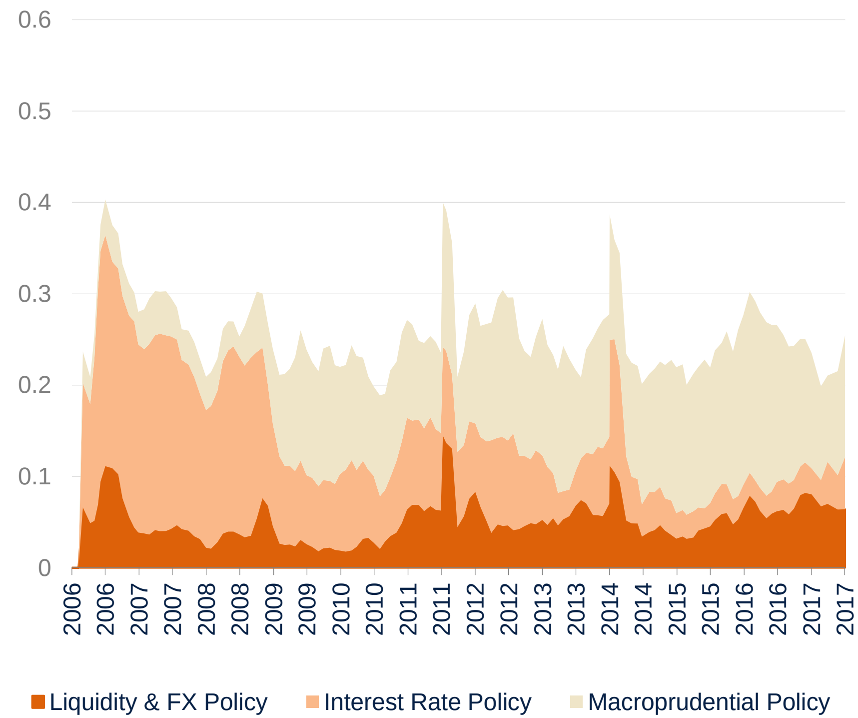
Disentangling communication policy by topic: the monetary policy case

Central Bank Of Turkey: Evolution of Topics



Monetary Policy Topics Distribution

(% of Total)



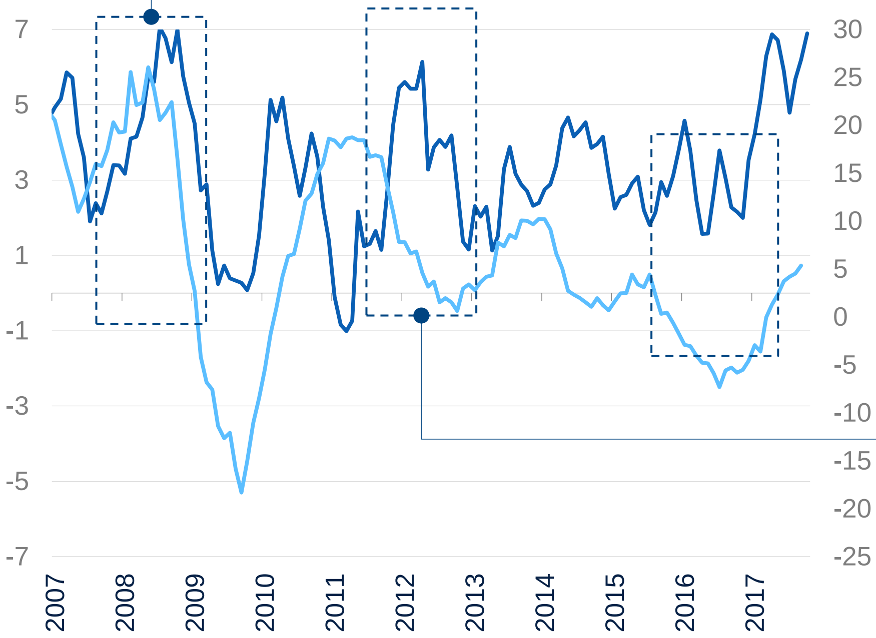
Source: BBVA Research

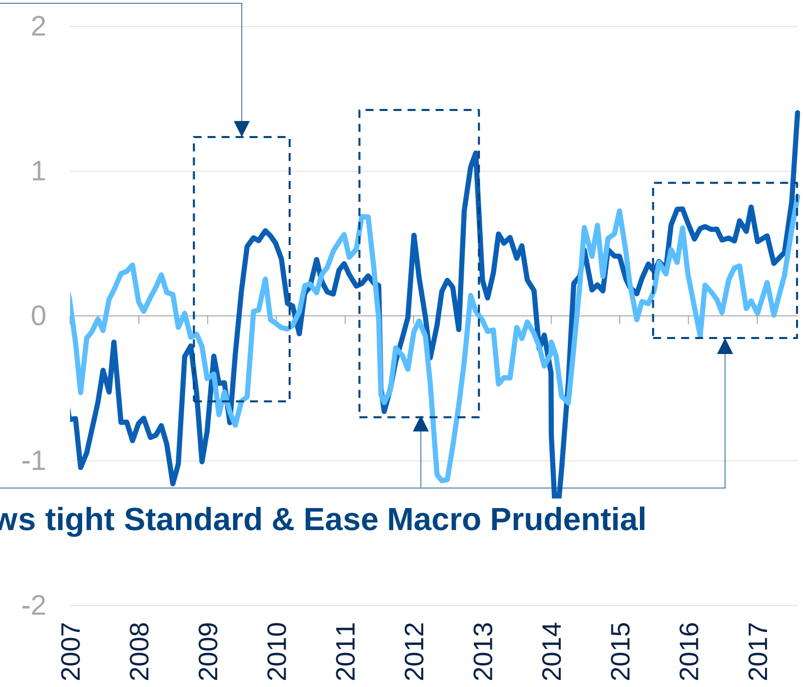
Multiple targets lead to different Policies…

Deviation from target (reference): Inflation & Credit
(FX adj. Loans YoY ninus 15% and inflation minus 5%)

Standard Monetary & Macroprudencial policies
(Sentiments)

Requires Tight Standard & Macro Prudential





Allows tight Standard & Ease Macro Prudential

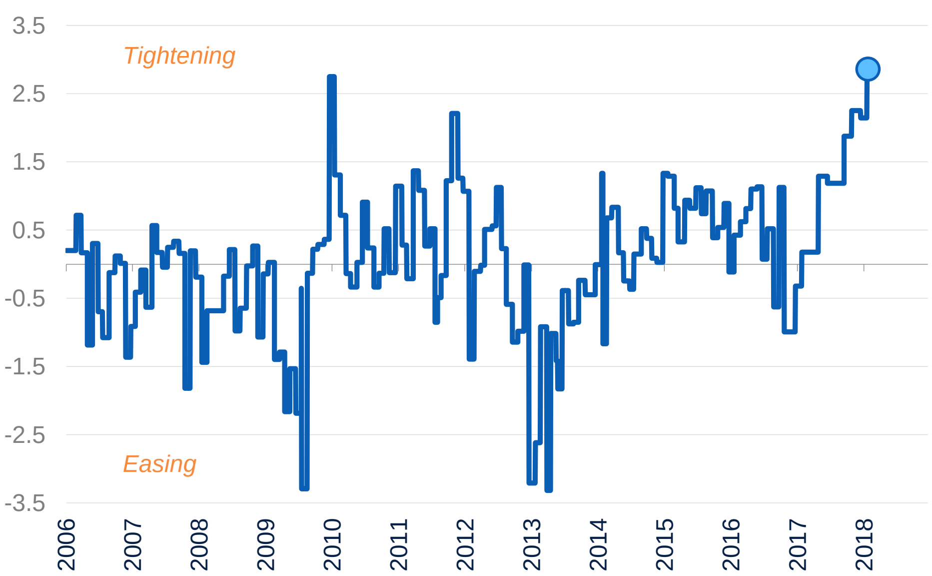
Source: BBVA Research

Through Sentiment Analysis we can check "how" the CBRT is talking and obtain some assessment of the monetary policy stance… (they can be different depending on the documents)

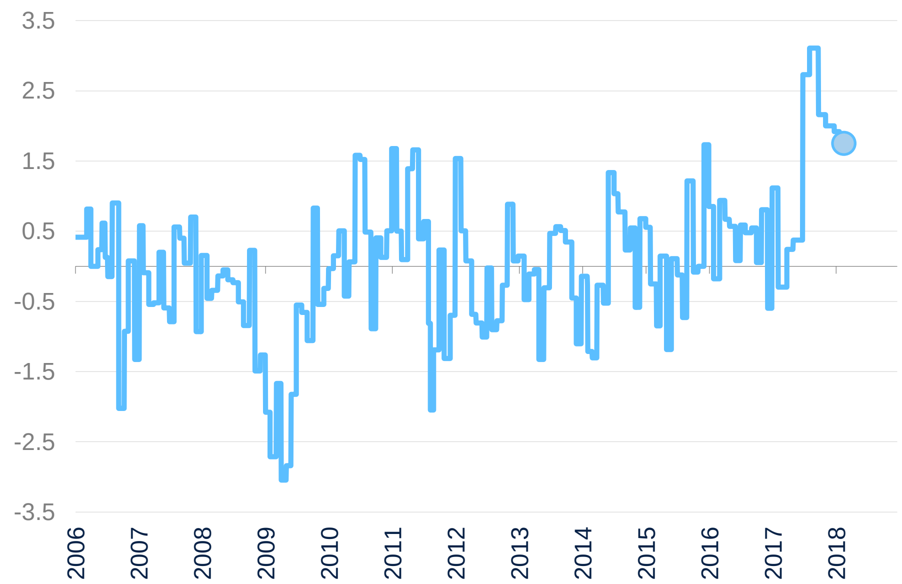
Central Bank of Turkey: Monetary Policy Sentiment

(Standardized, estimated through Big Data LDA and STM Techniques from Minutes & Statements)

Monetary Policy "Statements"



Monetary Policy "Minutes"



Source: Iglesias, J, Ortiz, A & Rodrigo, T (2007)
Source: BBVA Research

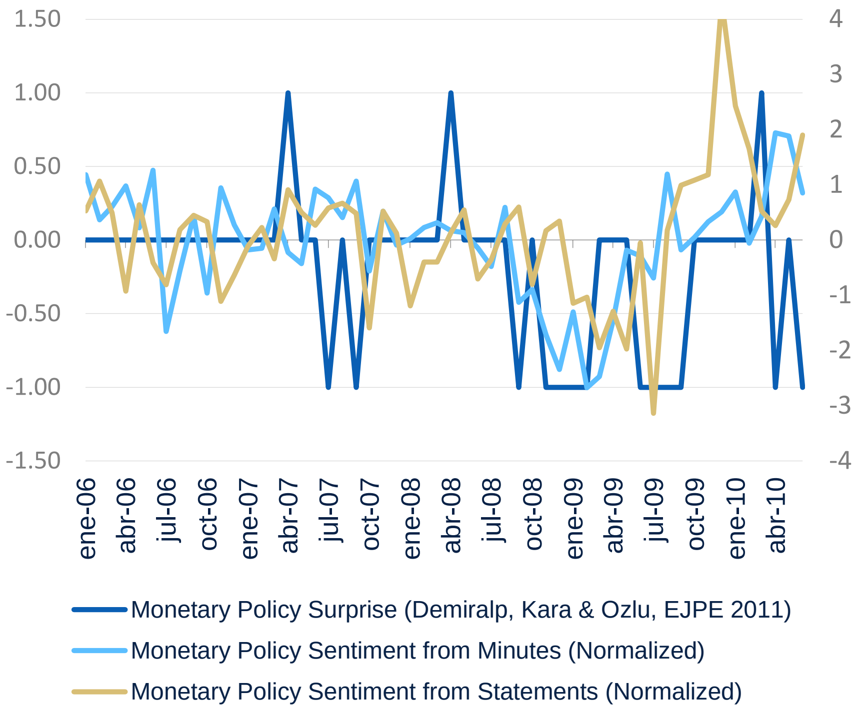
A more formal Statement…

More extensive and analytical…

We can check whether the sentiment affects analysts (and test the Machines & Dictionary methods compare with Experts analysis)

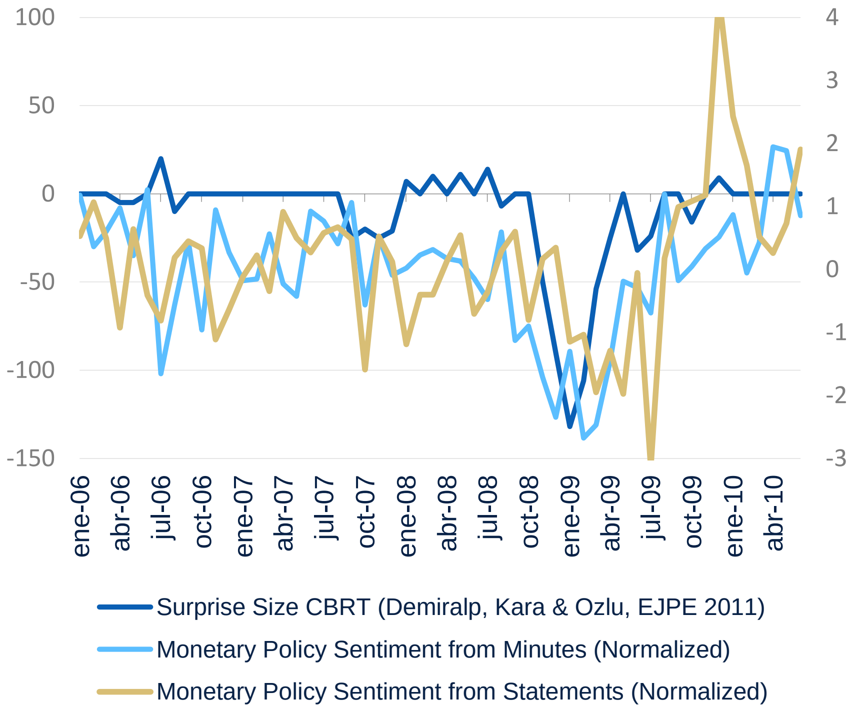
Monetary Policy: Experts vs Algorithms

(Sentiments fron LDA Algorithm and MP Surprises by Demiralp et Al1=Hawkish, 0= Neutral, -1=Dovish)



Experts vs Algorithms: Size of Surprises & Sentiments

(Sentiments fron LDA Algorithm and MP Surprises by Demiralp et Al)



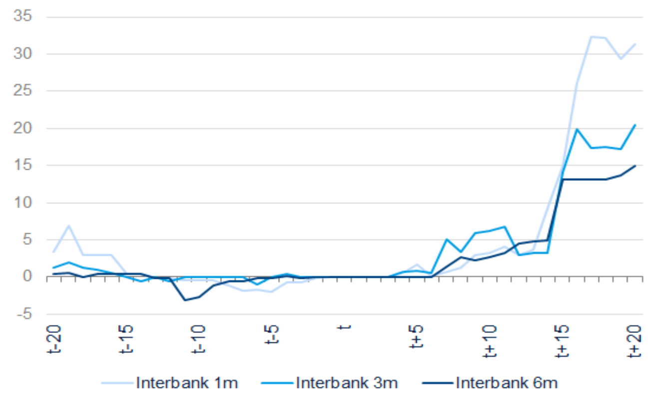
Source: BBVA Research

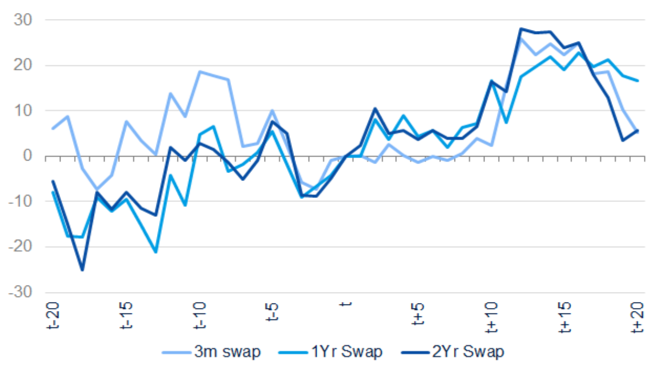
A good test is whether Monetary Policy Sentiment can affect markets (a necessary condition to affect the MTM)

Sentiment and the Markets: Response to Tighten and Easing

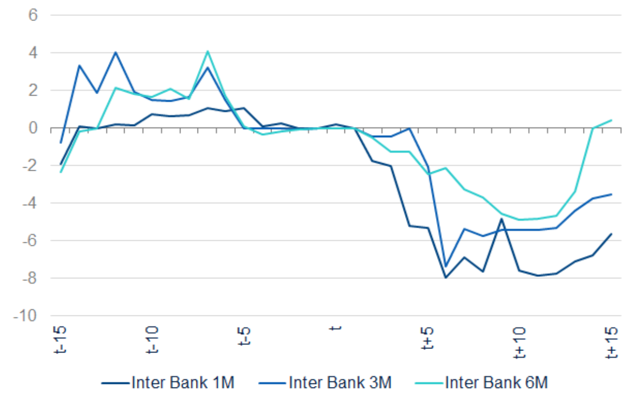
(Changes interbank deposits and Swap rates after monetary policy sentiments changes Higher than 1 STD)

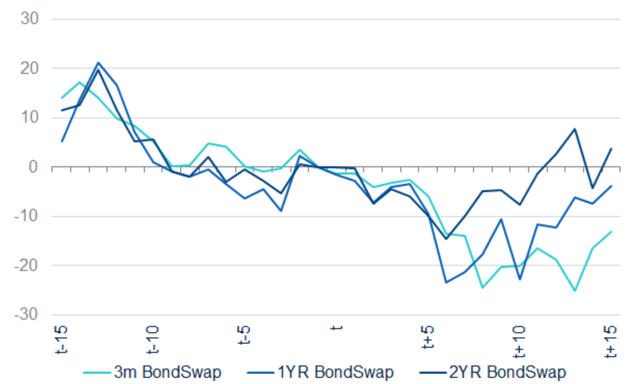
Tigthening





Easing





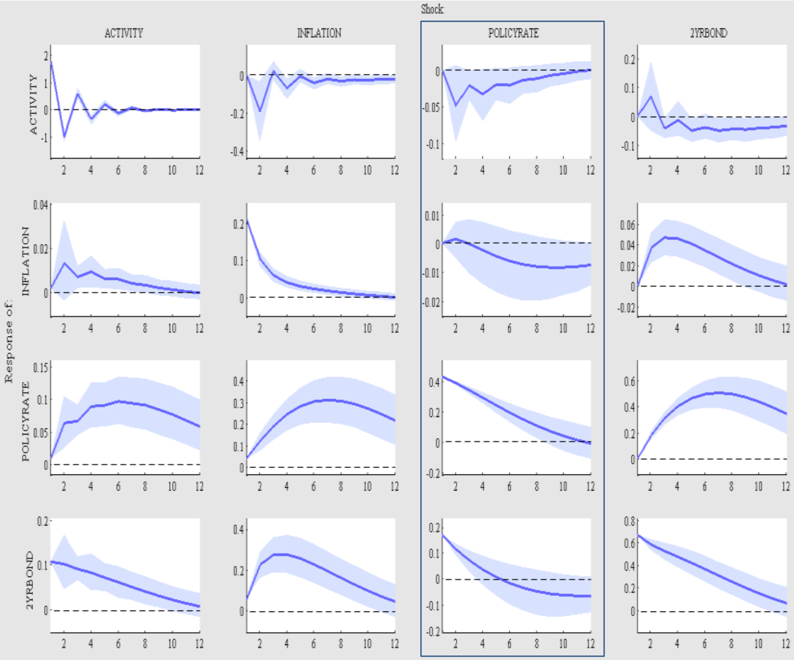
Source: BBVA Research

But at the end “words” need to be complement by “deeds” to affect the real economy an prices

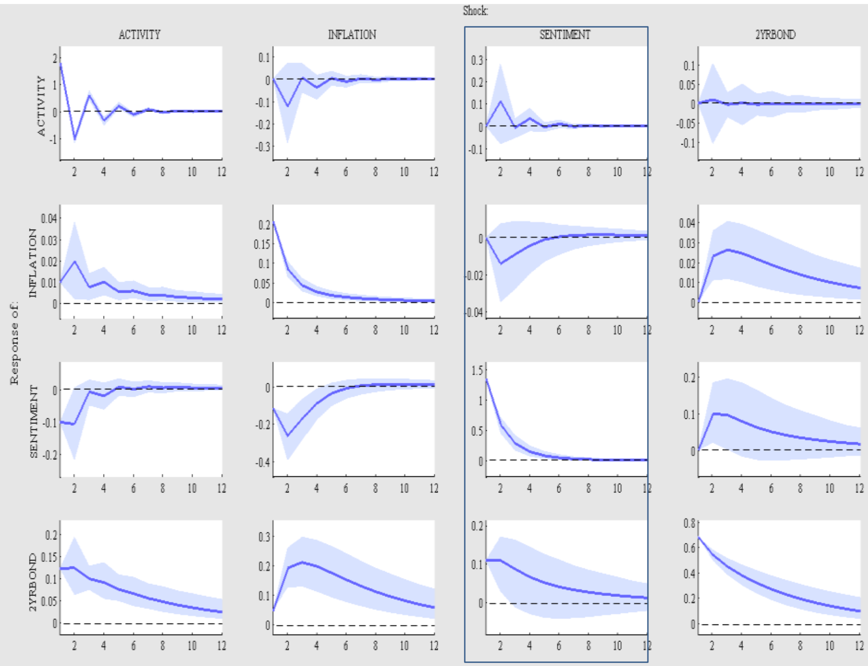
Monetary Policy Sentiment Effects: Standard Monetary Policy vs Sentiment

(Changes interbank deposits and Swap rates after monetary policy sentiments changes Higher than 1 STD)

Monetary Policy Rate Shock



Monetary Policy “Sentiment” Shock



Source: BBVA Research

BBVA Research

ANNEX

External sources: the case of Spain

The **Retail Trade Index** is a business cycle indicator which shows the **monthly activity** of the retail sector (**turnover**)

Spain: Retail Trade Survey (RTS)

Population scope: stores whose main activity is registered in **Division 47 of the NACE-2009**, which includes the following groups:

- **Retail sale in non-specialized establishments** (supermarkets, deparment stores, etc.)
- **Retail sale in specialized establishments** (food, beverages and tobacco; fuel; IT equipment and communications; personal goods, such as fabric, clothing and footwear; household items, such as textiles, hardware, electrical appliances and furniture; cultural and recreational items, such as books, newspapers and software; pharmaceutical products; etc.)
- **Retail trade not carried out in establishments** (**eCommerce**, home delivery, vending machines, etc.)

Sale of motor vehicles, Foodservice, hospitality industry, financial services, etc., are not included in RTS!

Sample: 12,500 stores (Random stratified sampling <50 employees + exhaustive>=50)

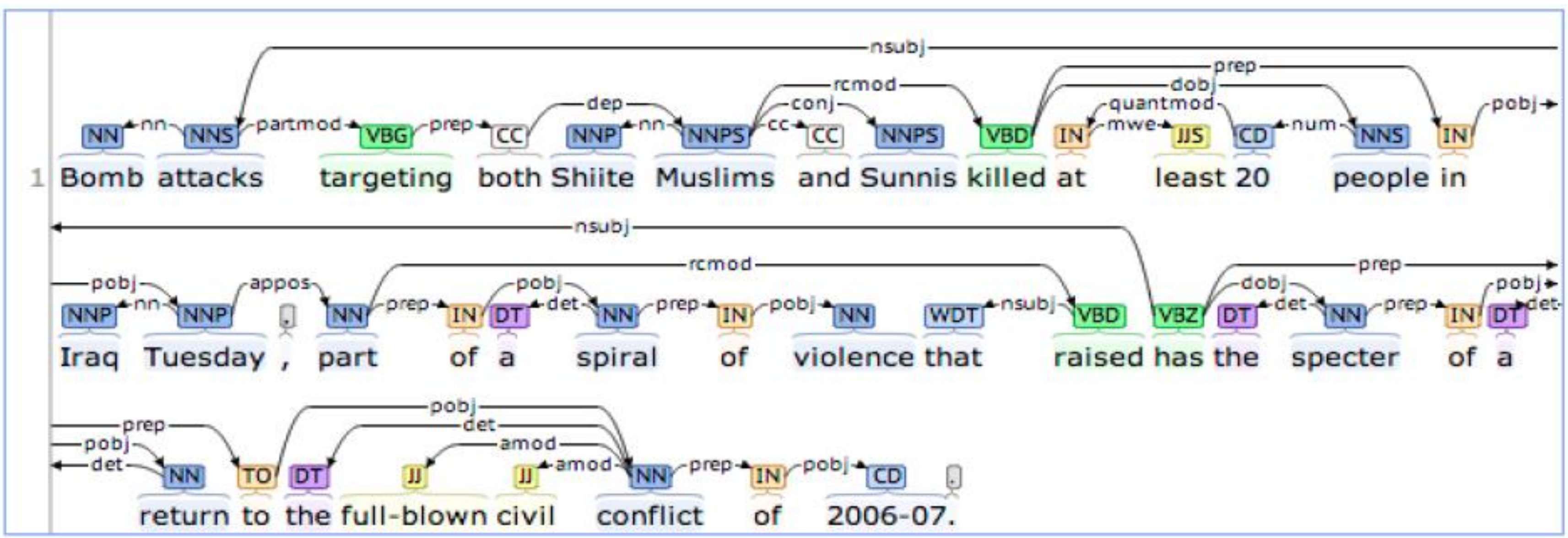
Dissemination: AA. CC. OR 5 distribution classes:

- **service stations,**
- **single retail stores (one premises),**
- **small chain stores (2-24 premises & <50 employees),**
- **large chain stores (25 or more premises, and 50 or more employees)**
- **department stores (sales area greater than or equal to 2500 m^2)**

Emotional indicator and coding system in GDELT

Average Tone**:** GDELT uses more than 40 tonal dictionaries to build a score ranging from -100 (extremely negative) to +100 (extremely positive) for each piece of news, with common values ranging between -10 (negative) and +10 (positive), with 0 indicating neutral tone. A neutral sentiment can be the result of a neutral language or a balancing of some extreme positive sentiments compensated by negative ones. The sentiment variable is based on the balance between the percentage of all words in the article having a positive and negative emotional connotation within an article divided by the total number of words included the article

PETRARCH coding system example:



BBVA Research

A two step procedure to extract common vulnerability factor from Hard data, Markets and News Sentiment

1st Step Estimation: Components → 2nd Step Estimation: Index

1st Step Estimation: Components

(1) $\text{SOEI} = \gamma_1 x_1 + \gamma_2 x_2 + \ldots + \gamma_{10} x_{10} + \epsilon_1$

(2) $\text{HBI} = \delta_1 y_1 + \delta_2 y_2 + \ldots + \delta_{11} y_{11} + \epsilon_2$

(3) $SBI = \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_{15} z_{15} + \epsilon_3$

(4) $FXI = \rho_1 v_1 + \rho_2 v_2 + \ldots + \rho_{10} v_{10} + \epsilon_4$

with $\gamma_i, \delta_i, \beta_i, \rho_i$ being the weight of every variable in the first principal component

2nd Step Estimation: Index

(6) $CSVI = \mu_1 \text{SOE} + \mu_2 \text{HB} + \mu_3 \text{SB} + \mu_4 \text{FX} + \varepsilon$

with $\mu_1, \mu_2, \mu_3, \mu_4$ being the weight of every component in the first principal component of the four components

Text mining and NPL: pre-proccesing and transformation

- Documents are defined as paragraphs.
- Documents with less than 200 characters are excluded (titles, contents sections,…)
- Then words are stemmed (reduce a word to their semantic root) to generate tokens
- Feature selection is conducted on the tokens: common stopwords and words with length 3 or less are removed and the remaining words are stemmed. Tokens are filtered out based on a term-frequency-inverse-document-frequency (tf.idf) index (Manning and Schutze 1999), words of the lowest quantile are removed. This indexing scheme is combined of a term-frequency index (tf) and a document frequency index (df). tf is just the count of a given word in a document, mean tf is used to construct the final index. df is the number of documents that contain a given word Then, the tf.idf used to filter words out is:

$$tf.idf_i = mean(tf_{ij}) * log_2\left(\frac{N}{df_i}\right)$$

- where i indexes terms and j documents. This index gives high weight to frequent words through the tf component, but if a word is very prevalent through the corpus; its weight is reduced through the idf component. The aim of this filtering procedure is to remove very unfrequent as well as very frequent words, to remove words with low semantic content

Machine learning algorithms on text: LDA, STM and DTM

- Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) is a Bayesian model with a prior distribution on the document-specific mixing probabilities where the count of terms within documents are independent and identically distributed given a Dirichlet prior distribution

- To introduce time-series dependencies into the data generating process, we use the dynamic topic model (DTM), a particularization of the Structural Topic Models (STM) where each time period has a separate topic model and time periods are linked via smoothly evolving parameters

- STM (Roberts et. al. 2016) explicitly introduces covariates into a topic model allowing us to estimate the impact of document-level covariates on topic content and prevalence as part of the topic model itself,

- The process for generating individual words is the same as for plain LDA. However both objects can depend on potentially different sets of document-level covariates: Topic Prevalence (each document has P attributes that can affect the likelihood of discussing topic k) and Topic Content (each document has an A-level categorical attribute that affects the likelihood of discussing term v overall, and of discussing it within topic k. The generation of the k and d terms is via multinomial logistic regression

“Parsing” through (LDA): Some Basics

- Words (Tokens): basic unit of discrete data. Represented as an unit vector with a single 1 entry, and 0 in the remainder, this vector ha as many entries as total words under analysis.
- Stop Words: “A”, “the” very frequent but don´t add value in term
- Document: sequence of N words
- Corpus: a collection of documents
- Document-Term-Matrix: matrix where each row is the sum of all the words in a given document. As such we have documents in the rows, words in the columns, and each entry in the matrix is the number of occurences of a word in a given document

The Latent Dirichlet Allocation (LDA) Model

- Latent Dirichlet Allocation (LDA) (Blei et al. 2003) is a generative probabilistic (hierarchical Bayesian) model of a corpus. Documents are represented as mixtures over latent topics, where each topic is characterized by a distribution over words.

- Simplified corpus generative process:

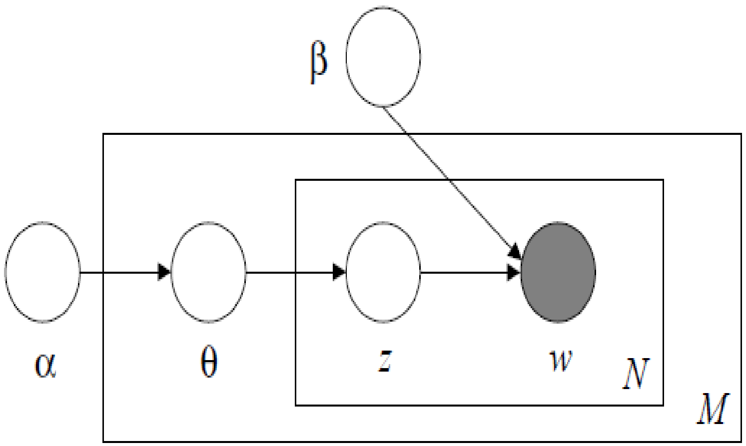
$N \sim Poisson(\zeta)$

$\theta \sim Dirichlet(\alpha)$

for each word w_n

topic $z_n \sim Multinomial(\zeta)$

$w_n \sim p(w_n | z_n, \beta)$



Source: Blei et al. 2003

- Bag of Words assumption: Order of the words is not importat , only the occurence is relevant. This assumption is inherited, as LDA is an extension of the Latent Semantic Indexing algorithm (an SVD on the Document-Term-Matrix)

- Words are conditionally Independent and Identically distributed: Needed when working with latent mixture of distributions, following de Finneti's theorem (*exchangeable observations are conditionally independent given some latent variable to which an epistemic probability distribution would then be assigned*)

Extending the LDA: The Dynamic Topic Model

- Structural Topic Model (Roberts et al. 2016) extends the LDA algorithm such that metadata (covariates) can affect the topic distribution. This allows us to introduce time series dependencies, estimating what is known as a Dynamic Topic Model

- Topics can depend on 2 classes of covariates:

    - Topic Prevalence (each document has P attributes that can affect the likelihood of discussing topic k)

    - Topic Content (each document has an A-level categorical attribute that affects the likelihood of discussing term v overall, and of discussing it within topic k)

Extending the LDA: Structural Topic Model & The Dynamic Topic Model

- Structural Topic Model (Roberts et al. 2016) extends the LDA algorithm such that metadata (covariates) can affect the topic distribution. This allows us to introduce time series dependencies, estimating what is known as a Dynamic Topic Model (i.e Topics Change over time). Topics can depend on 2 classes of covariates:
    - Topic Prevalence (each document has P attributes that can affect the likelihood of discussing topic k)
    - Topic Content (each document has an A-level categorical attribute that affects the likelihood of discussing term v overall, and of discussing it within topic k)
- Dynamic Topic Models generative process:

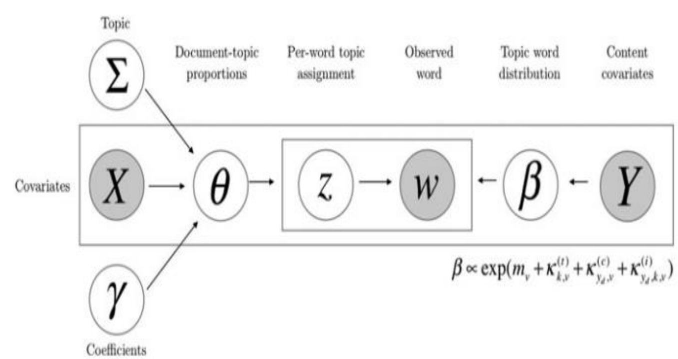
were Γ is a *Px(K – 1)* matrix of prevalence coefficients, *d* indexes documents, *n* indexes words within documents and *k* indexes the latent topics

$$\gamma_k \sim Normal(0, \sigma_k^2 I_p)$$

$$\theta_d \sim LogisticNormal\ (\Gamma' x_d', \Sigma)$$

$$z_{d,n} \sim Mutinomial\ (\theta)$$

$$w_{d,n} \sim Multinomial(B z_{d,n})$$



Source: Roberts et al. 2016

Sentiment analysis on text: lexicon approach

- We rely on Lexicon methods using the **Loughran-McDonald dictionary** (Loughran McDonald 2009), a created dictionary specifically to analyze financial texts and the **FED dictionary for financial stability** (Correa et al, 2017)

- Using the negative and positive words of this dictionary, the average "tone" of a given document is computed by:

$$\text{Average tone} = 100 * \frac{\sum Positive\ words\ - \sum Negative\ words}{\sum Total\ words}$$

- The score ranges from -100 (extremely negative) to +100 (extremely positive) but common values range between -10 and +10, with 0 indicating neutral

- To build the final sentiment indices, we use the topic mixture that combines dictionary methods with the output of LDA to weight word counts by topic, following the approach proposed by Hansen and McMahon (2015). This allows generating different sentiment measures from a set of text, and focusing that sentiment on the topics of interest

BBVA | Research

Big Data at BBVA Research

Big Data Workshop on economics and finance. Bank of Spain

Alvaro Ortiz, Tomasa Rodrigo and Jorge Sicilia

February 2018