

# Big Data & Big Models at BBVA Research

ECB Statistics Day

Jorge Sicilia, Alvaro Ortiz & Tomasa Rodrigo

October 2017

## Index

**01**

Opportunities in the digital era. Big Data at BBVA Research

**02**

Geopolitics, Trade and Spill overs

**03**

Economic & Risk indicators in Real Time

**04**

Text Mining and Sentiment analysis

# 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

... avoiding 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**

## New framework in the digital era...

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



› New availability of data



› 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

## ... 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**

New data may end up changing the way in which economists approach empirical questions and the tools they use to answer them

# Big Data at BBVA Research

## Our work



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

## Our datasets



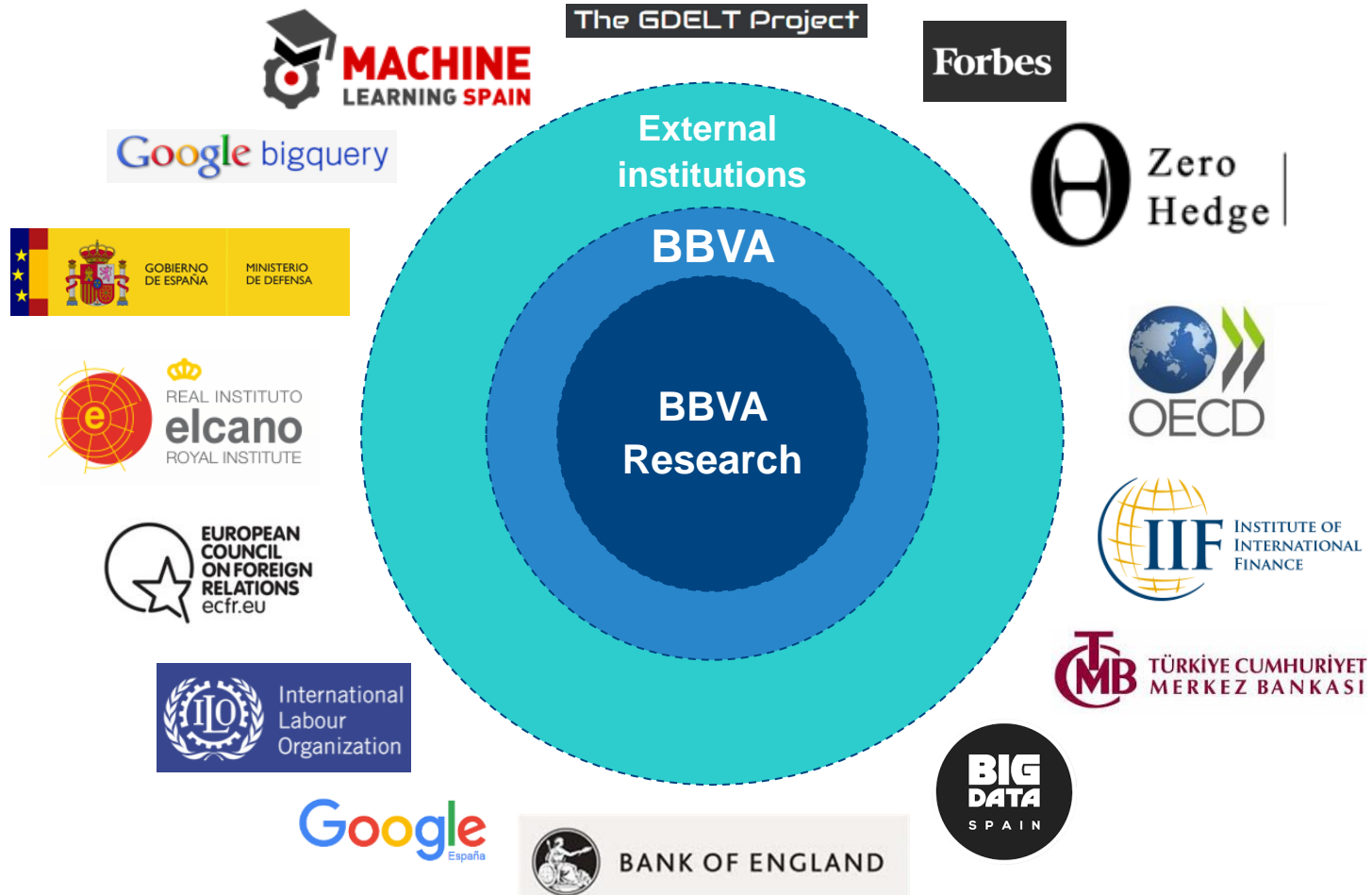
- **Media data** to exploit news intensity, geographic density of events (location intelligence) and emotions across the world (sentiment analysis)
- **BBVA aggregate and anonymized data** from clients digital footprint
- **Data from the web** (Central Banks' reports among others)

## 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

# Internal and external diffusion





# Our working process

## Databases

GDELT  
BBVA data  
Google search  
Web

The GDELT Project  
**BBVA**

## SaaS

BigQuery  
and  
Amazon  
Redshift

 Google BigQuery  
 **amazon**  
REDSHIFT


## Analysis

Clean,  
Aggregate  
transform  
and model  
the data

**BBVA** | Research

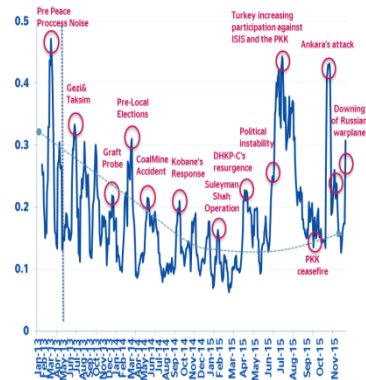
## Visualization

Fuse,  
visualize  
& analyze  
the data

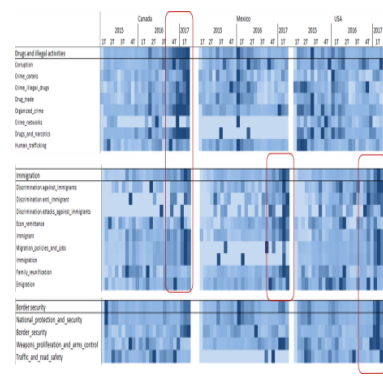
**CARTO**   
 Google Data Studio

# Our products

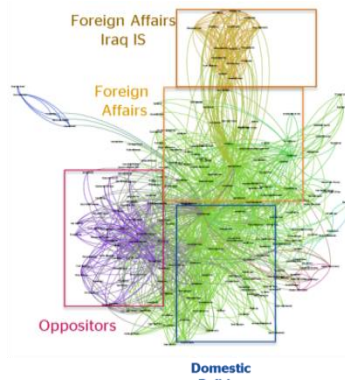
Political, Geopolitical Social Indexes (Political Index)



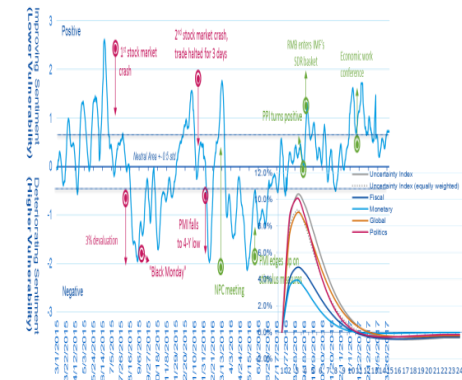
Color Maps NAFTA Topics (Nafta Project)



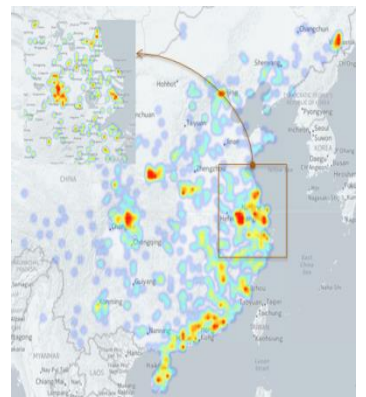
Politics & Financial Networks (Political Networks)



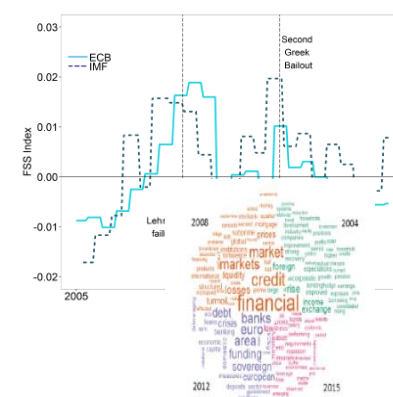
Mix Hard data & Sentiment & VAR models (CBSI and Turkey Sentiment Indexes)



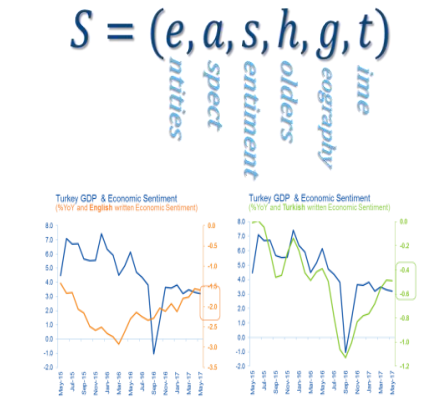
Geographical Analysis Housing Prices (sentiment on Housing Prices)



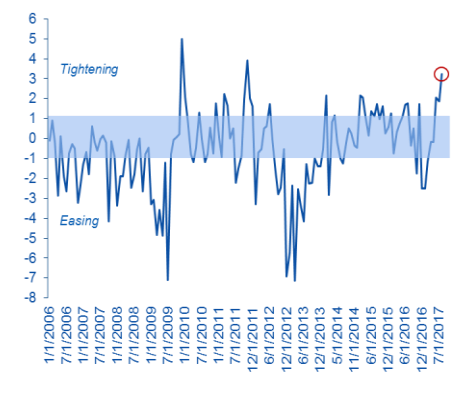
Financial Stability & Macroeprudential (ECB & FED FS index by FED Board)



Measuring Sentiments (sentiment Analysis on Economy and Society)



Monetary & Stability tones by Central Banks



# 02

## Geopolitics, Trade and Spill over analysis

## External databases: GDELT

### Global Database on Events Location and Tone

Open database of human society from every corner of the globe dating back to 1979 ...

... including over 300 events around the world and more than 30000 themes...

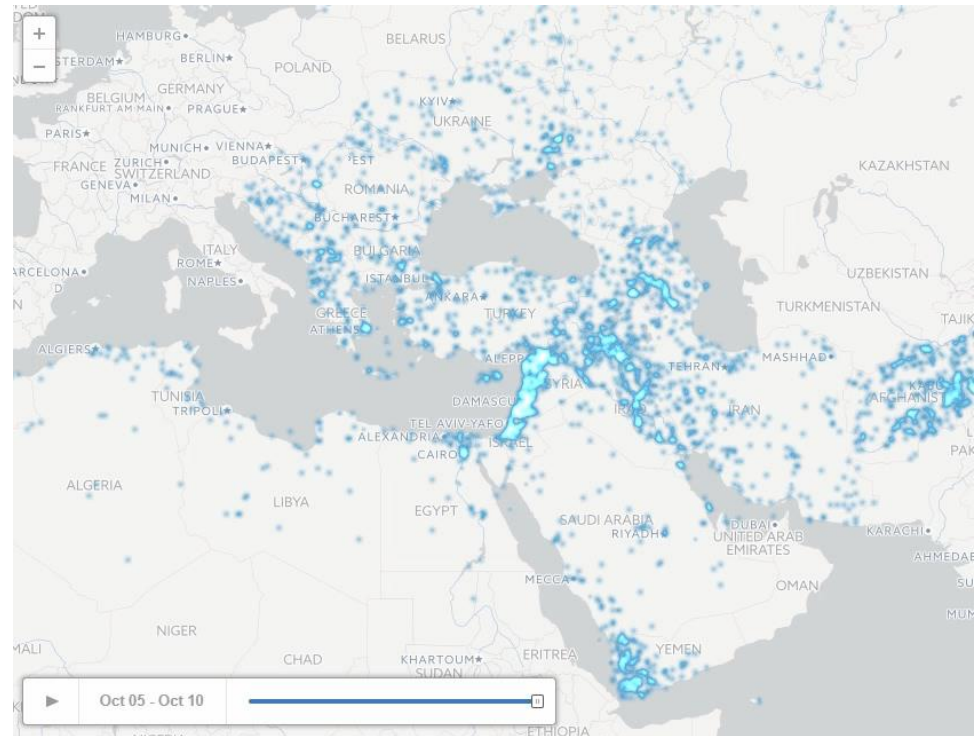
... georeferenced across the entire planet...

...and collecting emotions using some of the most sophisticated algorithms

# Tracking Geopolitics on real time

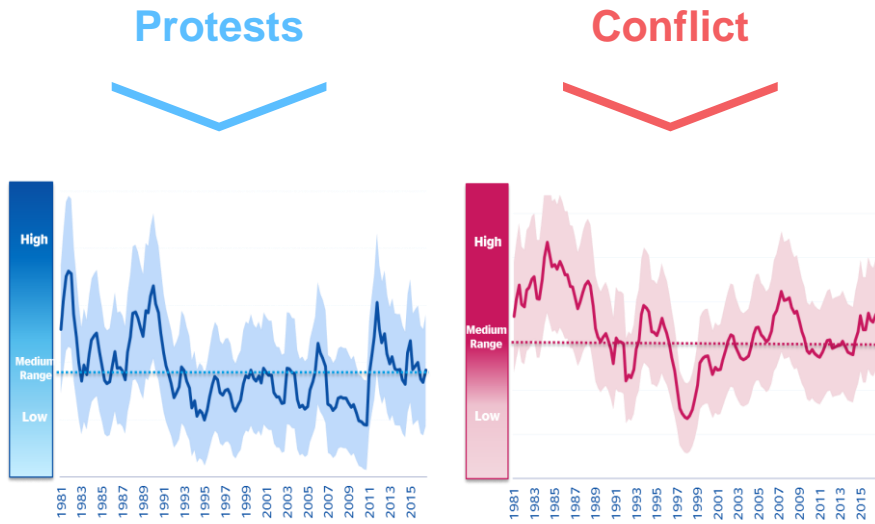
is useful to identify the main hot spots and potential spillovers

## Conflict Intensity Map 2017 (Number of conflicts/ Total events)

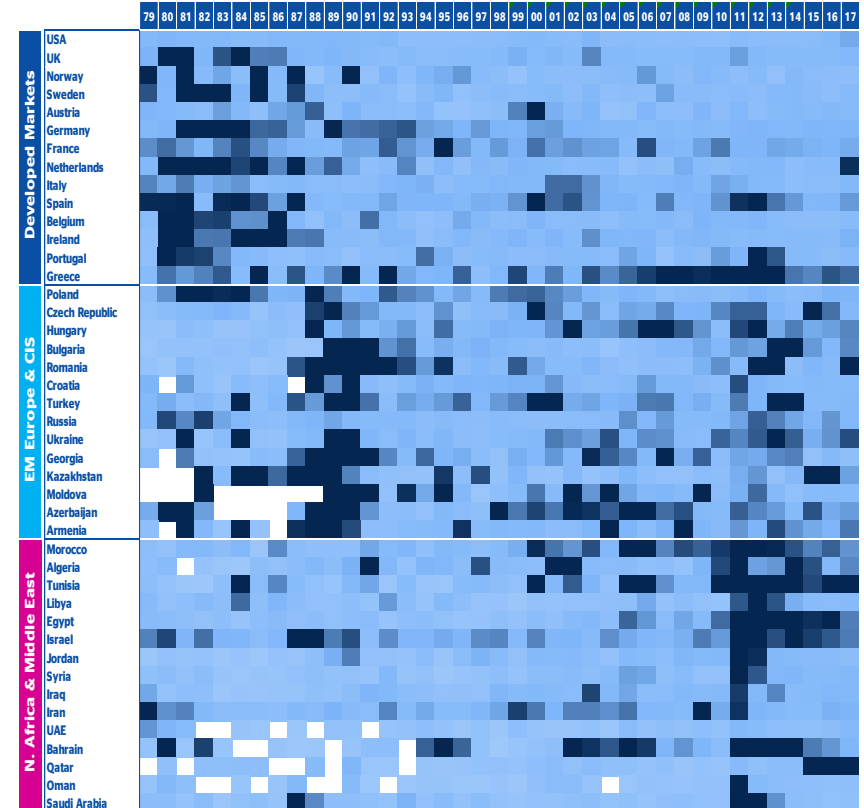


# From an historical perspective...

## BBVA Research World Protest and Conflict Intensity Index 1979-2017



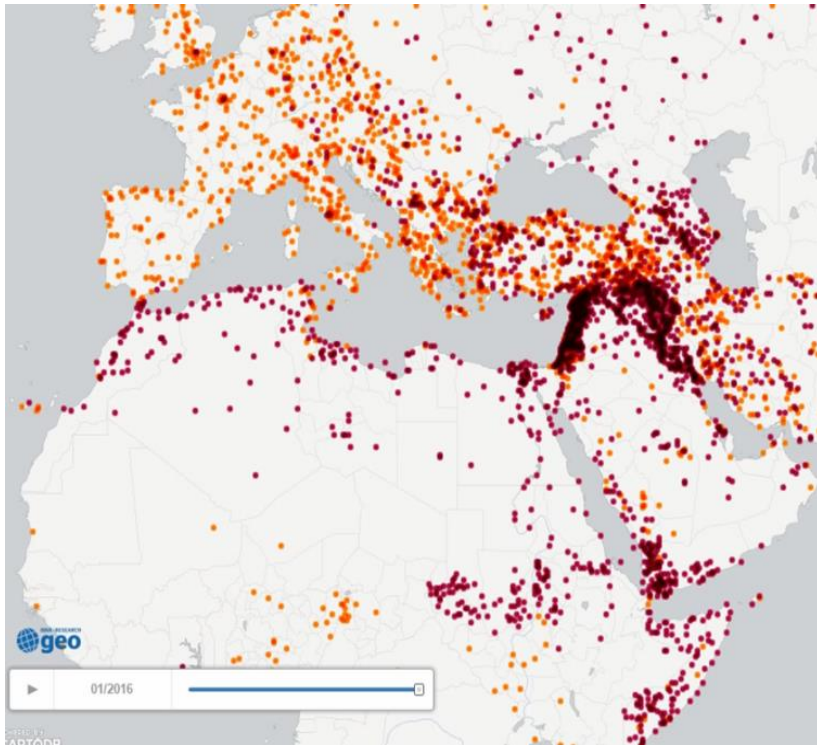
## World Protest Intensity Map 1979- 2017



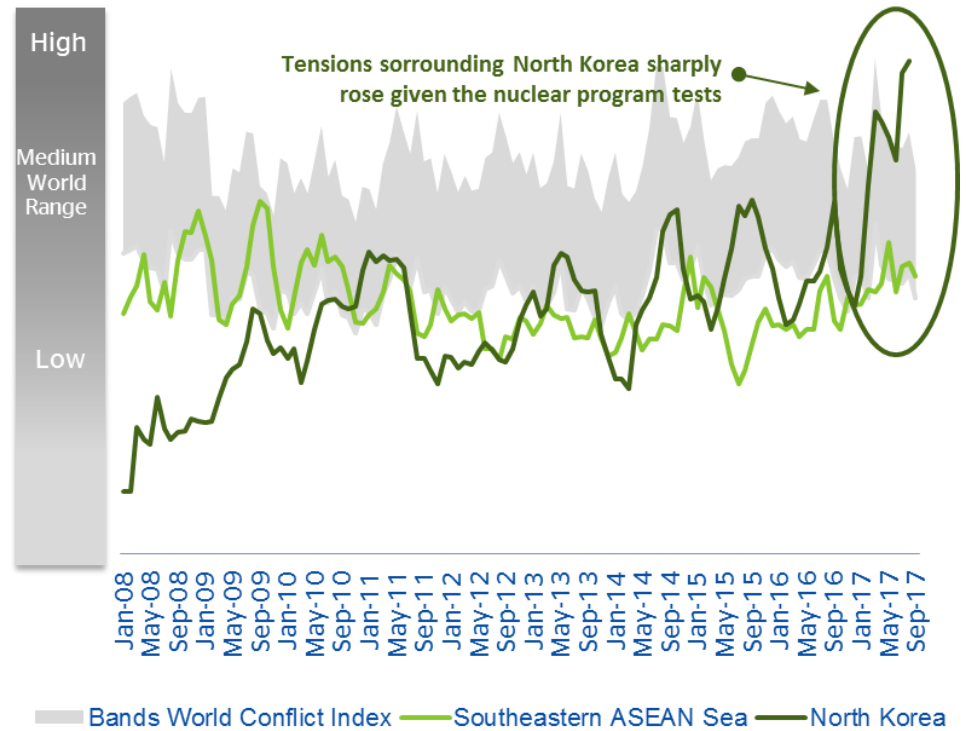
# ...to the main hot spots...

## BBVA Research Refugees Flows Map in 2015-17

Number of media citations about refugees' inflows and outflows



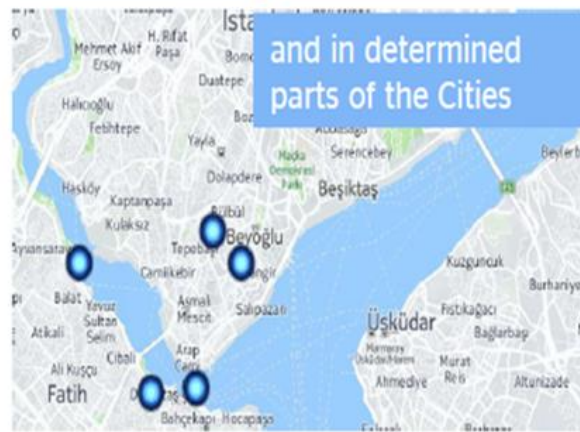
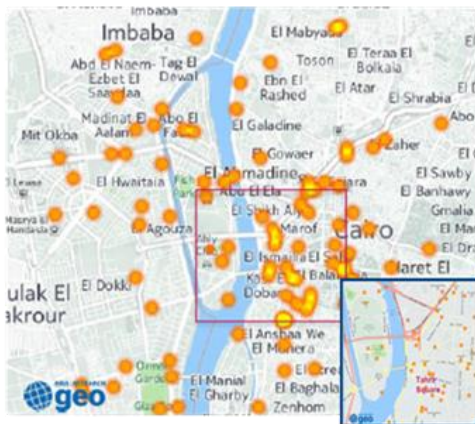
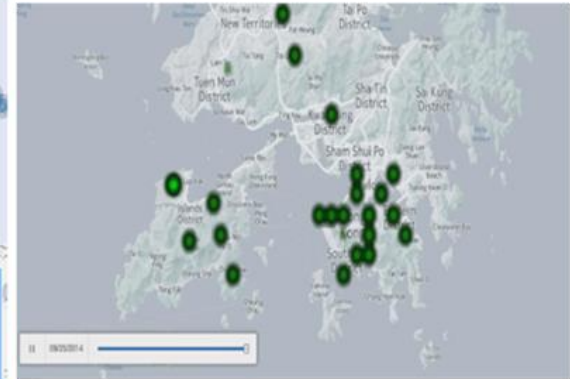
## BBVA Research Asia Conflict Intensity Index 2008-17



# ...at the exact geolocation

## Social unrest events across the world: Cairo, Istanbul and Hong Kong cases

Protest events



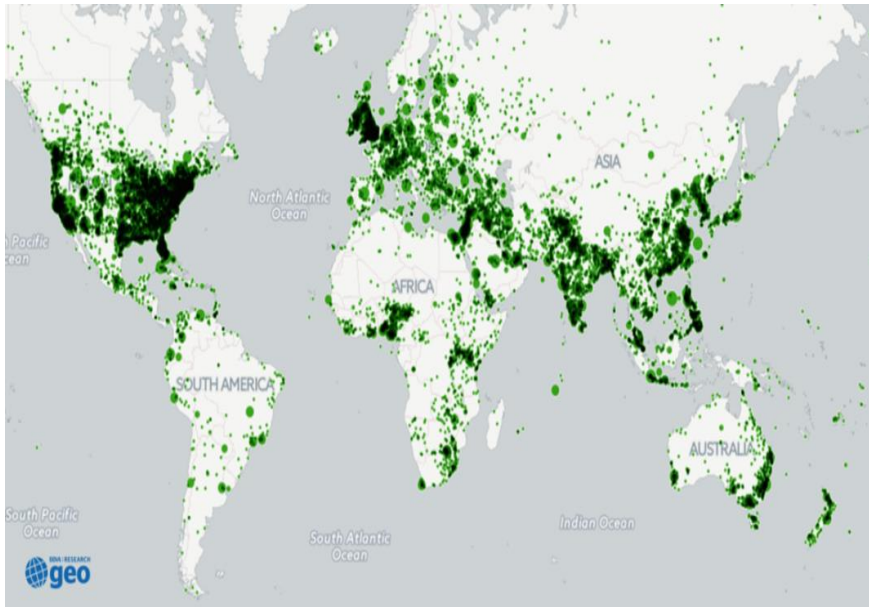


# New threats like cyberattacks can be also monitored

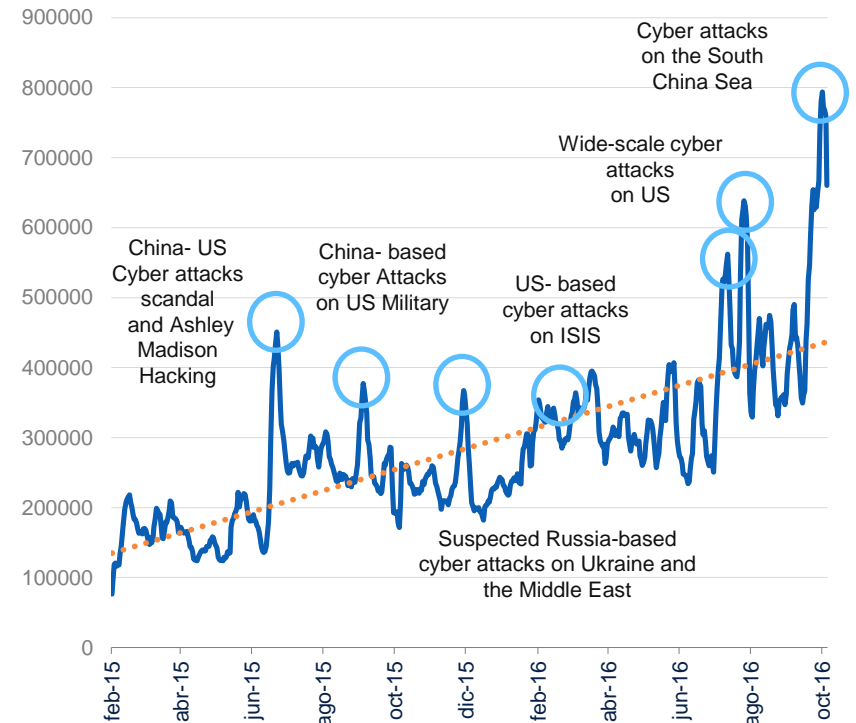
## Cyber-attacks have become one of the main threats in 2015- 2017

(GDELT based indicator of cyber warfare, cyber-attacks, data breaches or another online security issues)

### Media coverage of cyber warfare, cyber-attacks, data breaches and other computer- and online security-related issues around the world 2015-2016

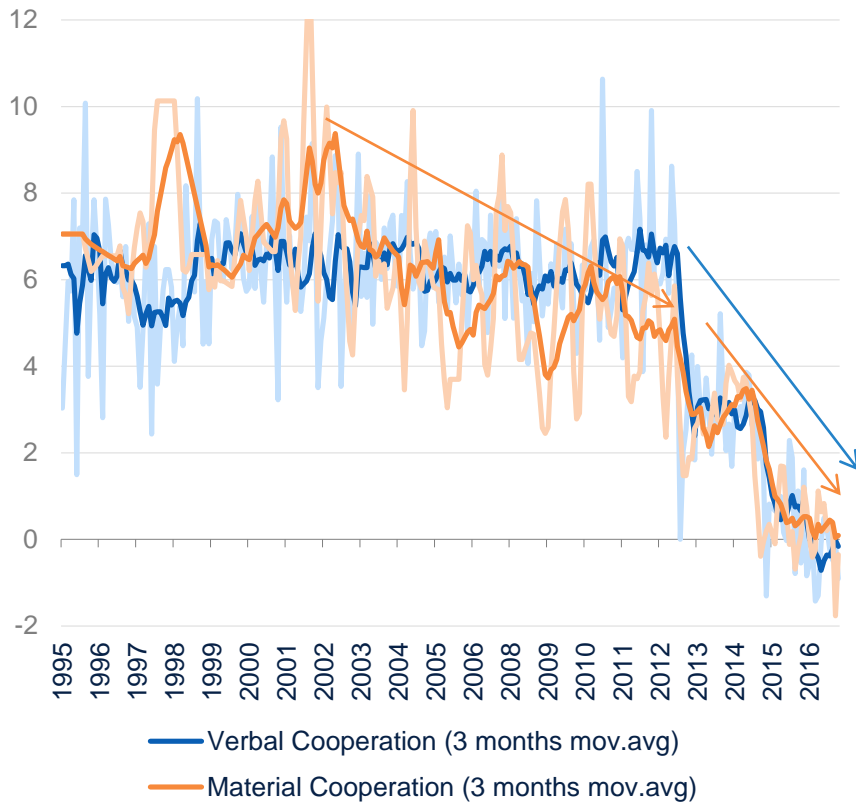


### World Media coverage of cyber-attacks in 2015-2016

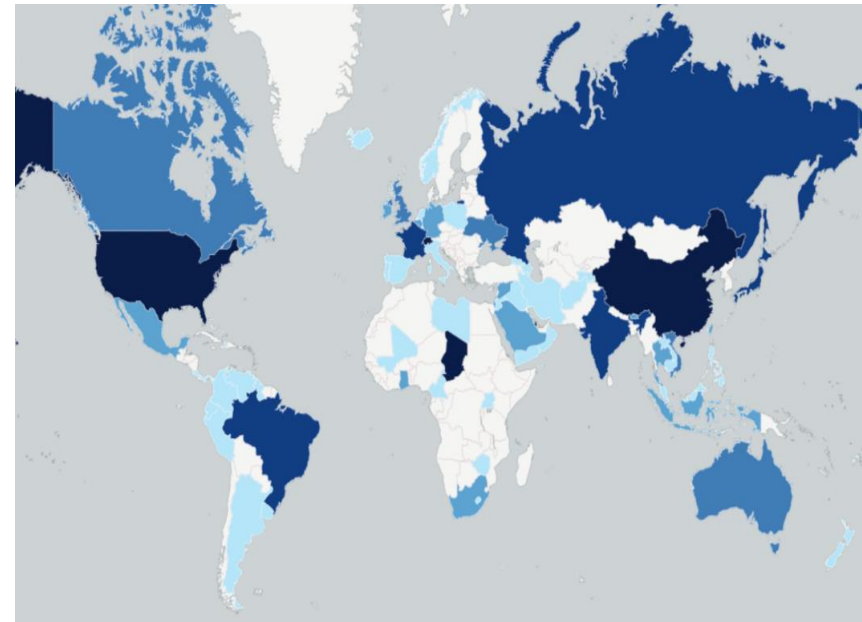


# Thanks to Big Data we can check in real time how is material and verbal support on World Trade...

**BBVA Research World Trade Support Index**  
(Tone & Coverage verbal cooperation at WTO)



**BBVA Research Trade Support Index Changes 2008-17**

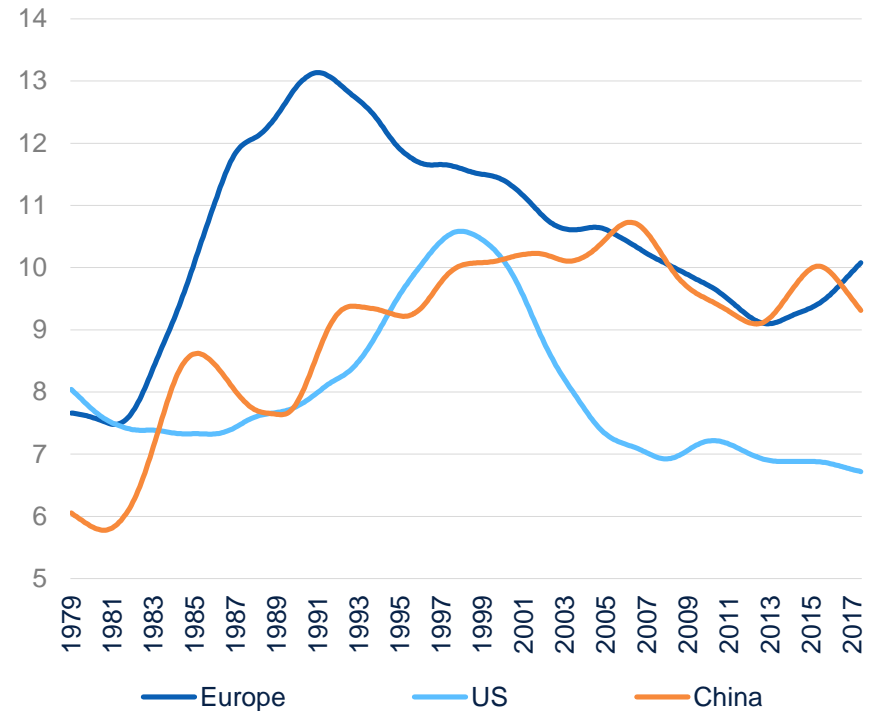
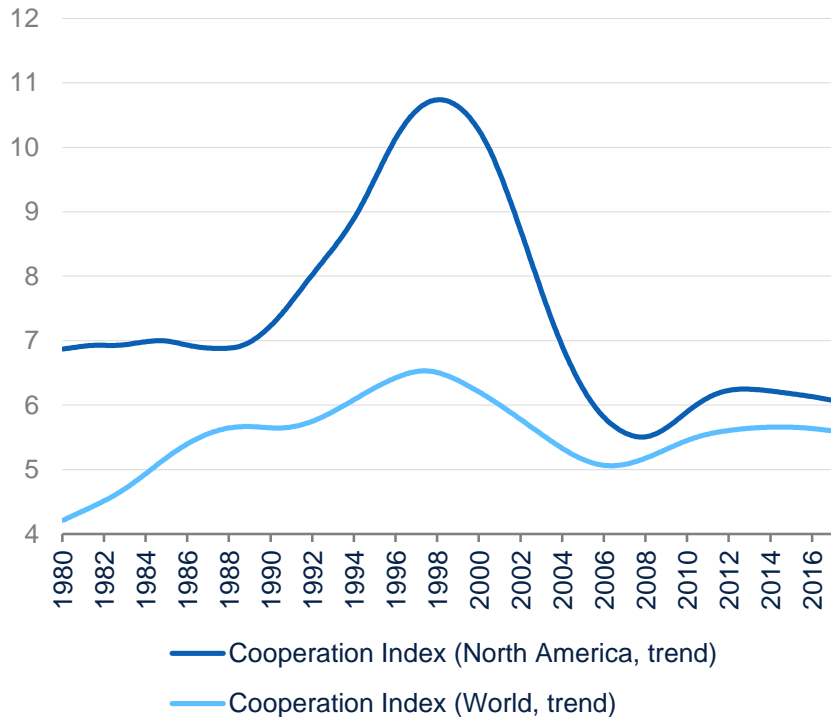


# ... as well as the cooperation index evolution over time of the main world powers

The index is defined as the ratio of the numbers of events of cooperation and demand.

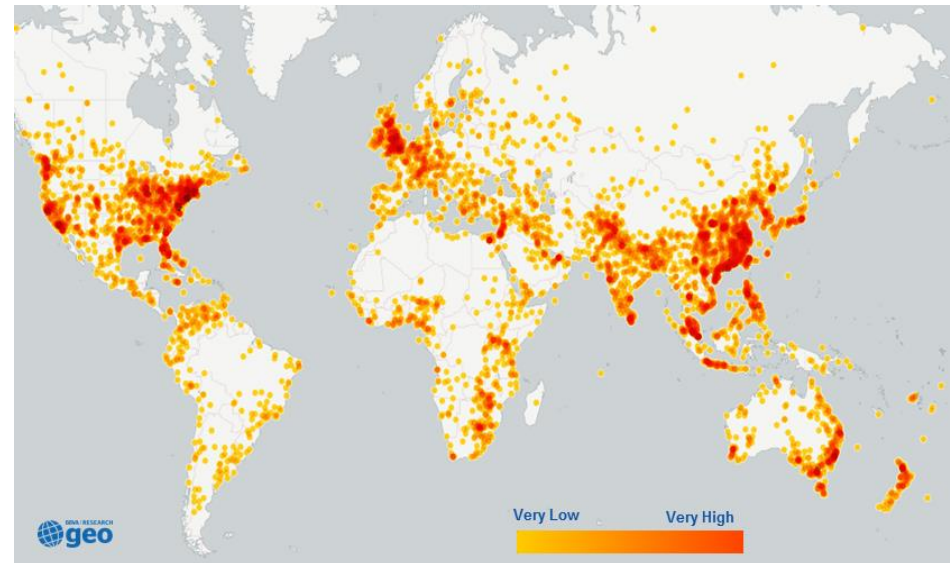
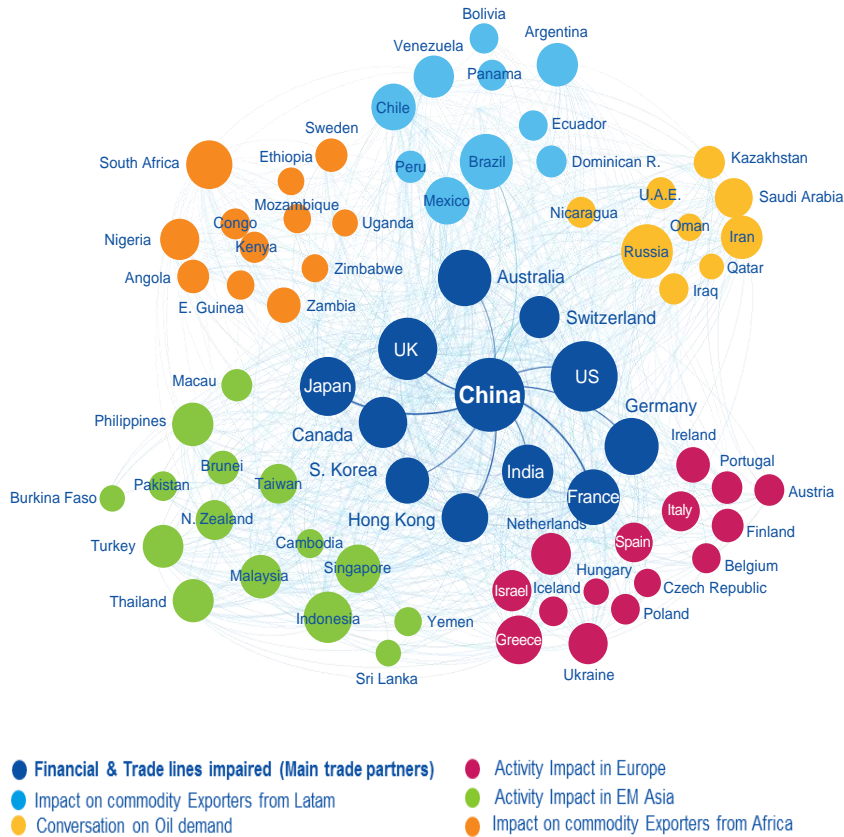
## Trends of the index

(HP filtered)



# Spill Over effects of China's slowdown..

## Chinese slowdown: media perception and country network



# ... or spill overs from trade sanctions on Russia

## Russian Economic Sanctions Network

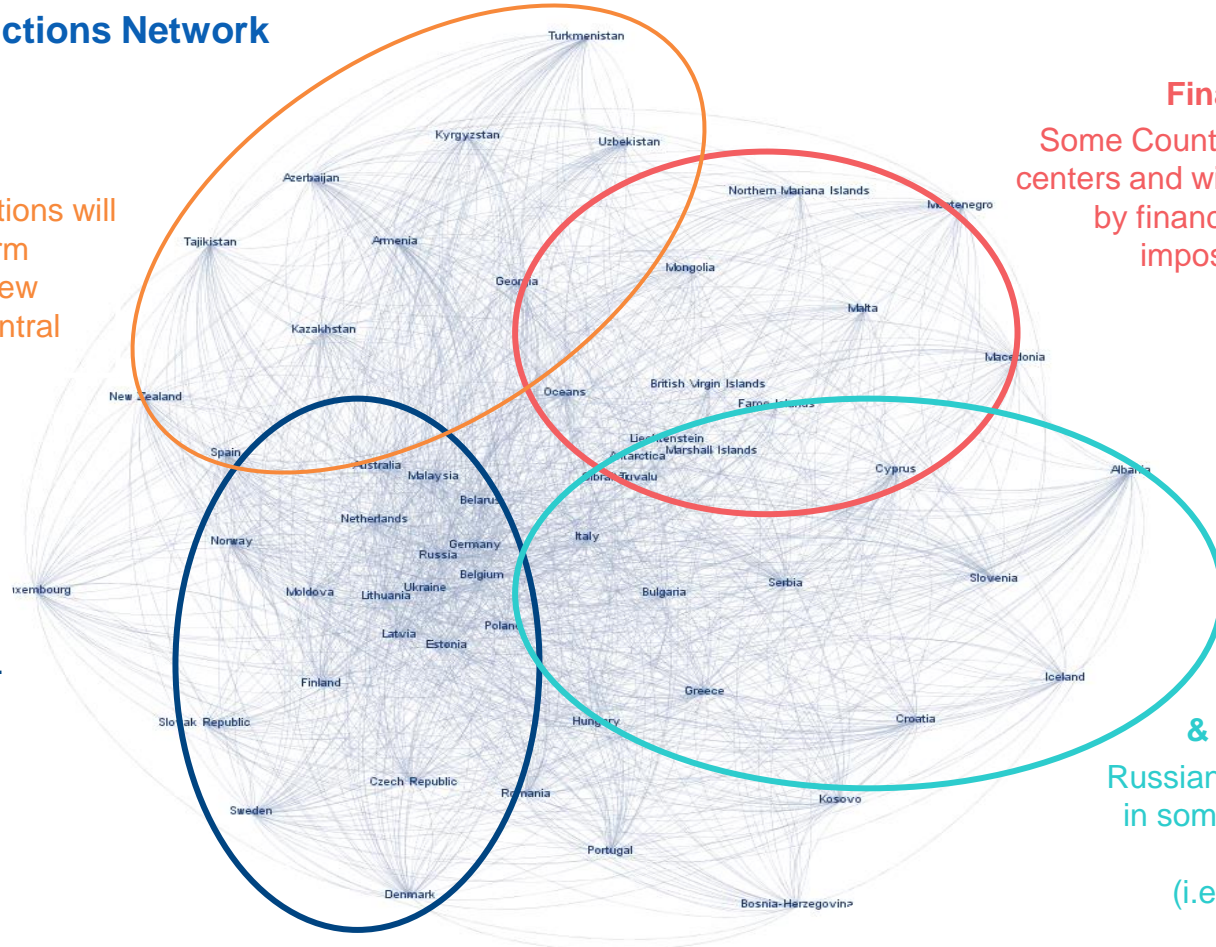
### Central Asia Trade

Technology exchange restrictions will affect to the medium/long term Russian capacity to extract new energy which could affect Central Asia relationships

### Central & Eastern Europe Trade

Trade Effects of commercial Sanctions imposed to Russia will spread to other countries. Particularly traditional Trade Partners in the East

External Demand of some Central Europe Countries (France ,Germany, Italy) will be also affected



### Financial Circle

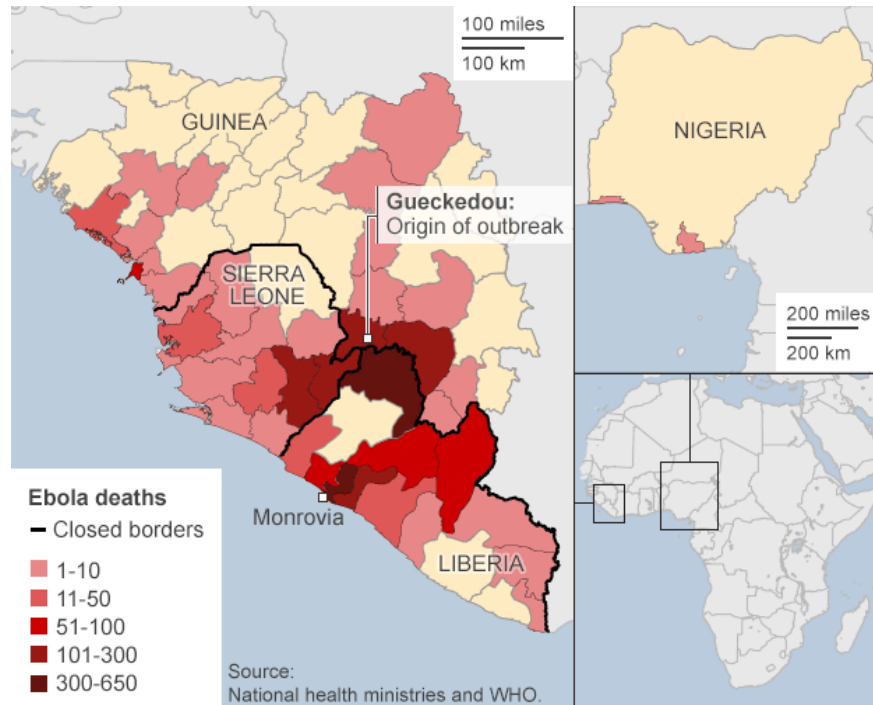
Some Countries, financial centers and will be affected by financial Sanctions imposed to Russia

### Financial & Trade Circle

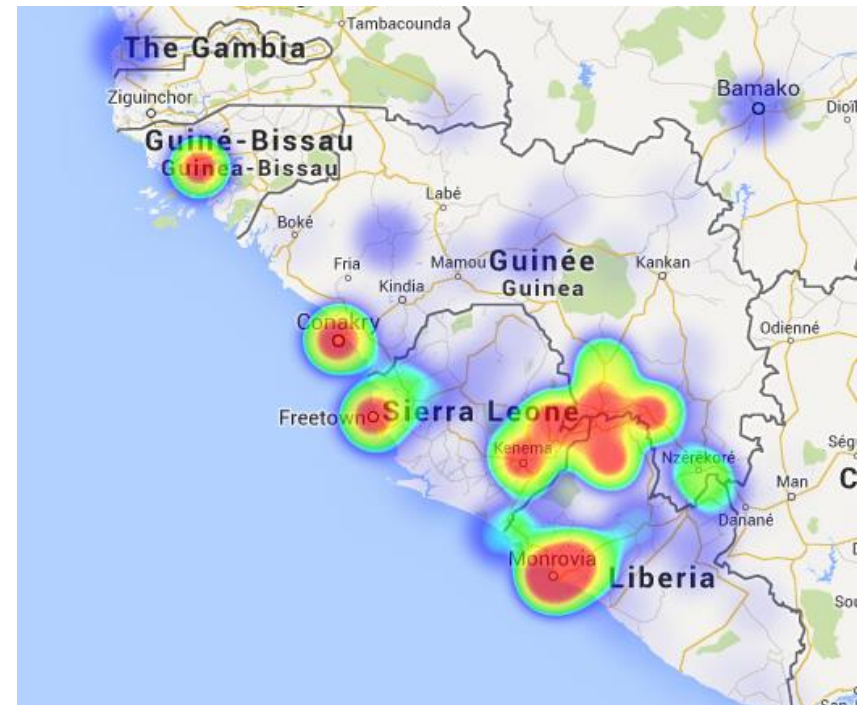
Russian Investments in some regions are huge (i.e the Balkans)

# Robustness checks with official data show a high similarity between the series. From health issues...

**Ebola: Official Debts by the WHO**  
(deaths until mid september)

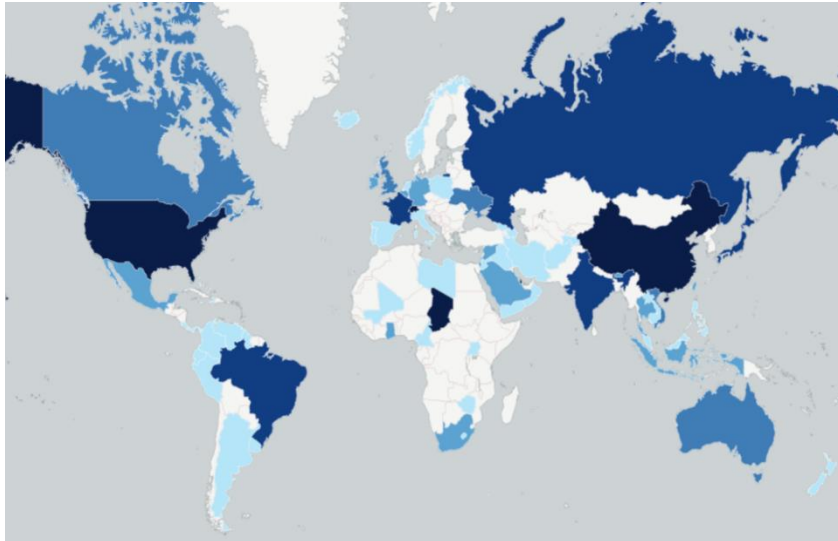


**Ebola: Outbreak according GDELT**  
(deaths until mid september)

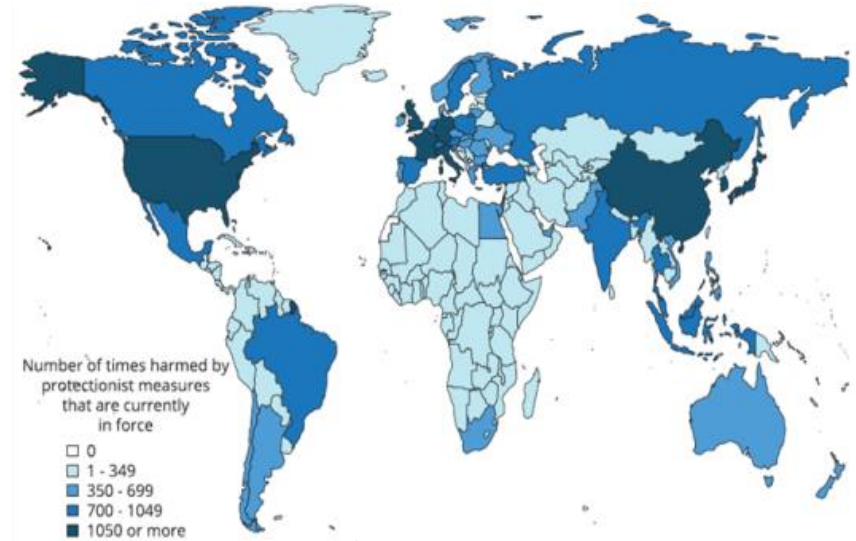


# ... to trade related topics.

## BBVA Research Trade Support Index Changes 2008-17



## The global incidence of protectionism 2008-2015 (global trade alert)

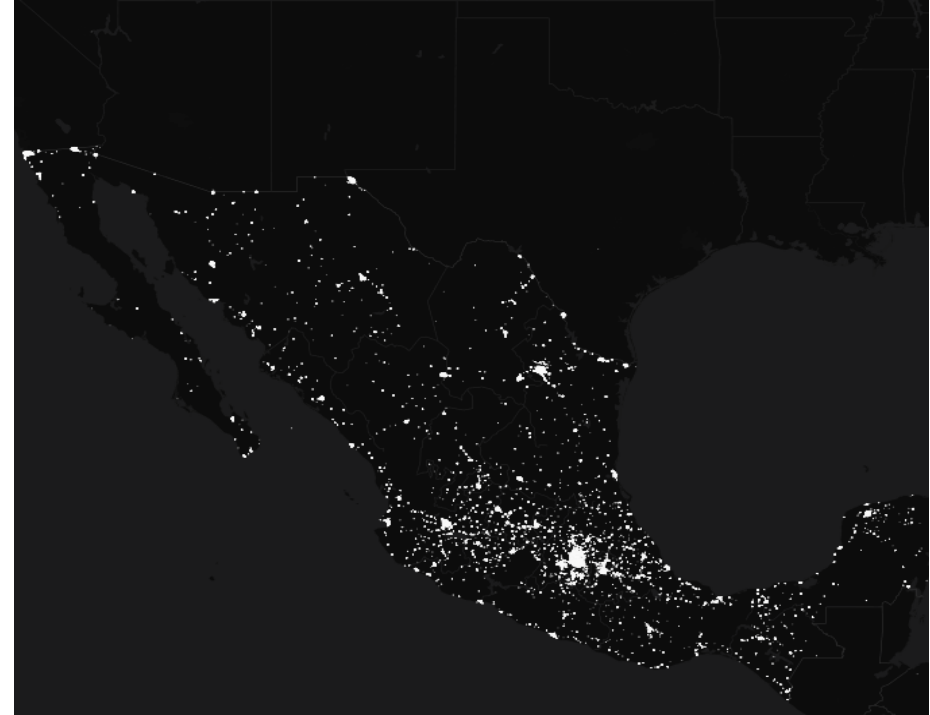


# 03

**Economic & Risk indicators through Transactions,  
Google Searches & International News**



## Internal databases: working with aggregated and anonymized BBVA Data



**710M** card transactions from **1M** PoS, made by **53M** people, representing **€43.000M**

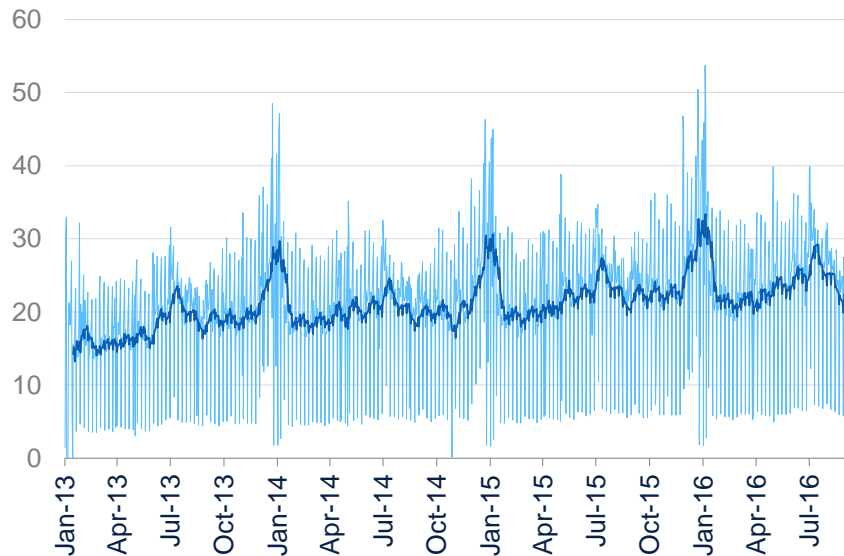
**1.500M** card transactions from **1,1M** PoS, made by **88M** people, representing **€41.000M**

# Using BBVA data, we replicate national figures, gaining frequency...

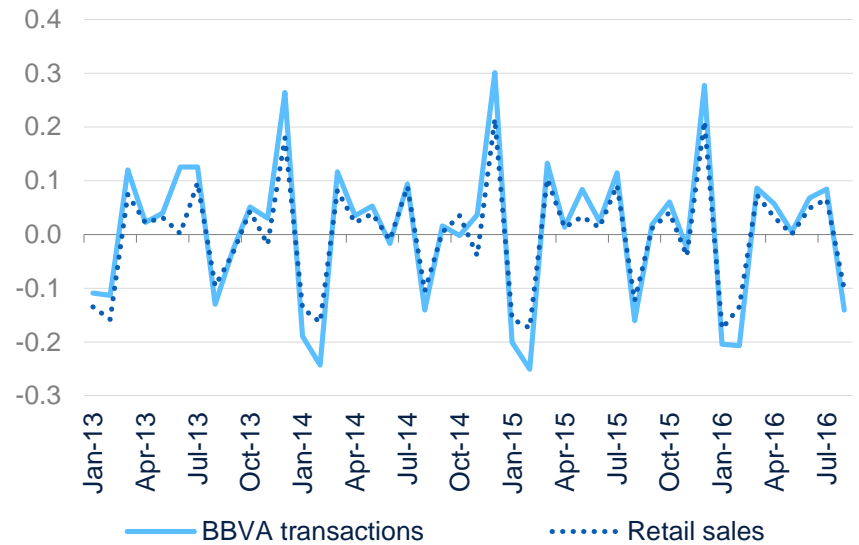
## A “High Definition” Activity Indicator for Spain (and Mexico)

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

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



Comparison Retail Sales-INE and BBVA on monthly basis



### What “HIGH DEFINITION(\*)” means here:

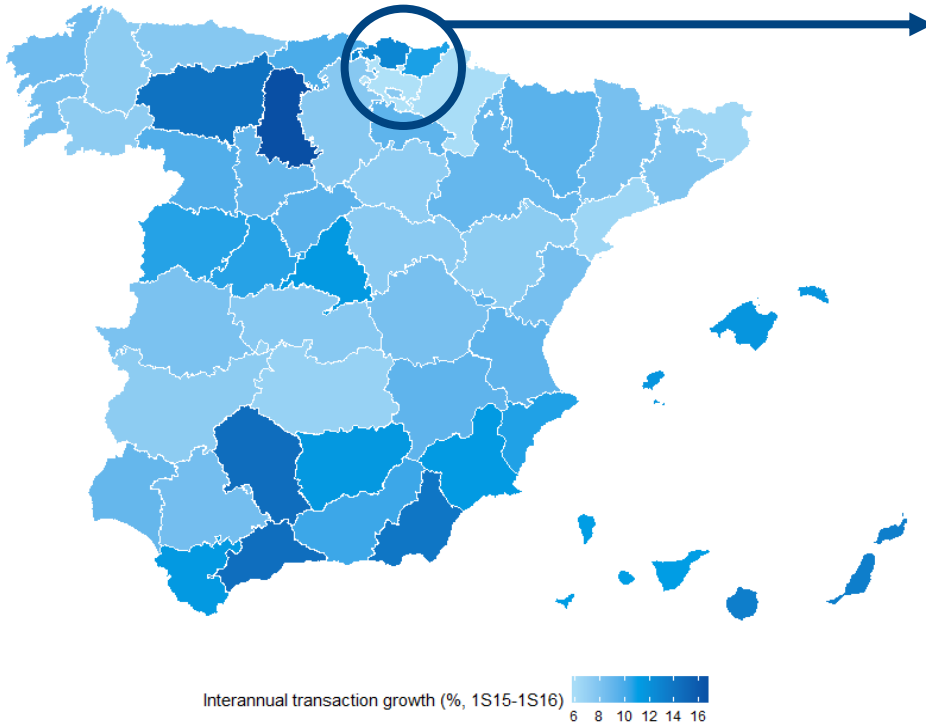
**High granularity:**  
Dynamics down to subnational level

**Ultra High Frequency:**  
Dynamics up to sub-monthly frequency

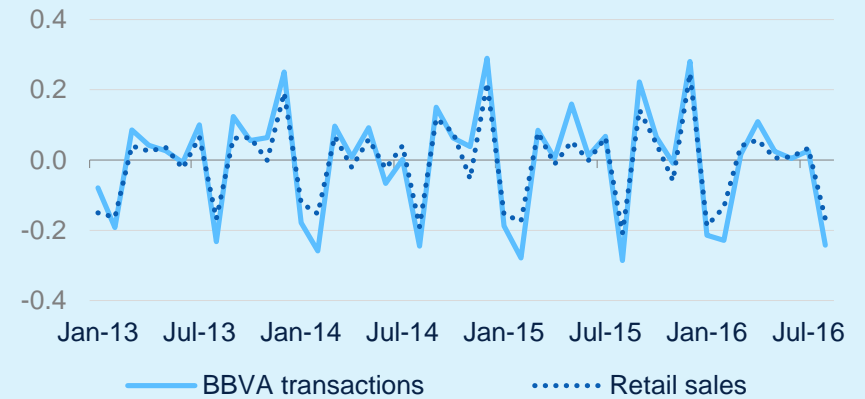
**Multi Dimensional:**  
More detailed socioeconomic features

# ...and granularity, going to regional level

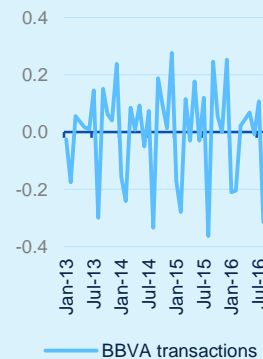
**BBVA transactions 1S15 vs 1S16**  
(% yoy)



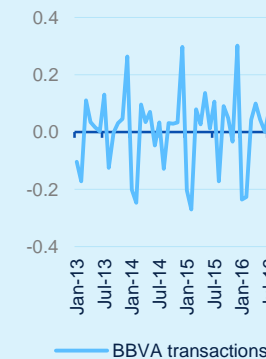
**País Vasco**



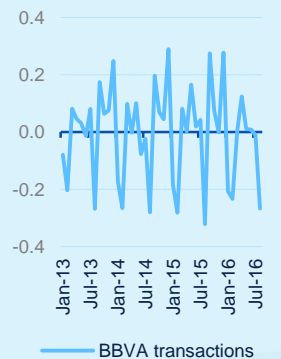
**Álava**



**Guipúzcoa**



**Vizcaya**



## External databases: Google searches database



- ◆ The measurement of **Google queries**, given the **increasing use of internet searches**, has a great potential in predicting future developments
- ◆ **Google Search** provides several features beyond searching for words and are available since **July 2007**
- ◆ The **analysis** of the frequency of search **terms** may indicate the evolution of **economic, social and health trends**

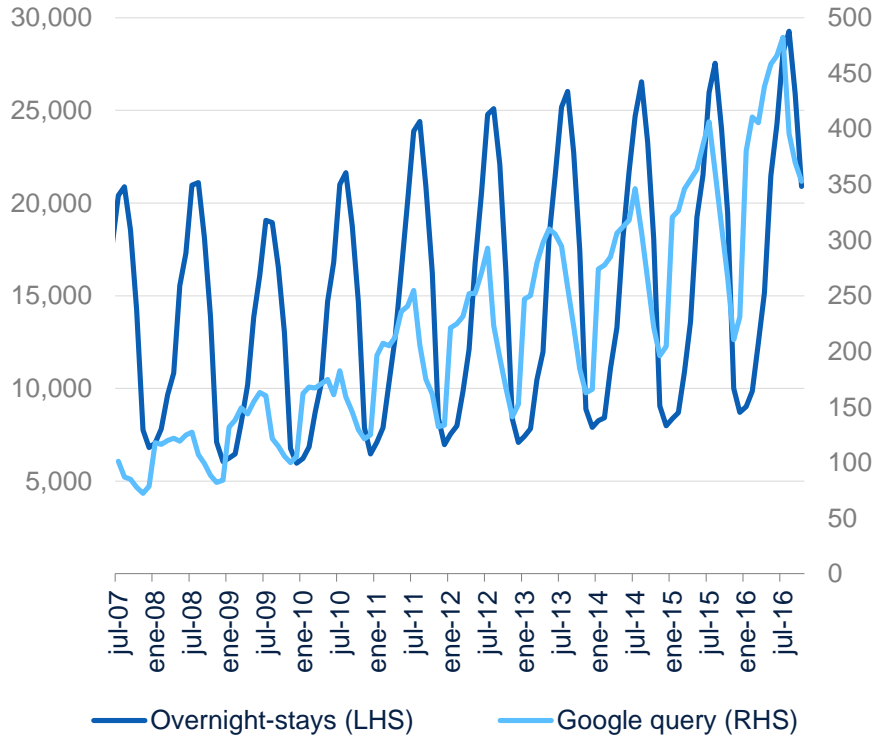
*Example:* a database with aggregate information about **Google queries related to Spain as a tourist destination** has been developed together with Google. Google tourism related queries follow the same seasonal pattern that tourism statistics, anticipating them with one or two months



# Similarity in the dynamics of official statistics and google queries allows us to forecast Spanish tourism

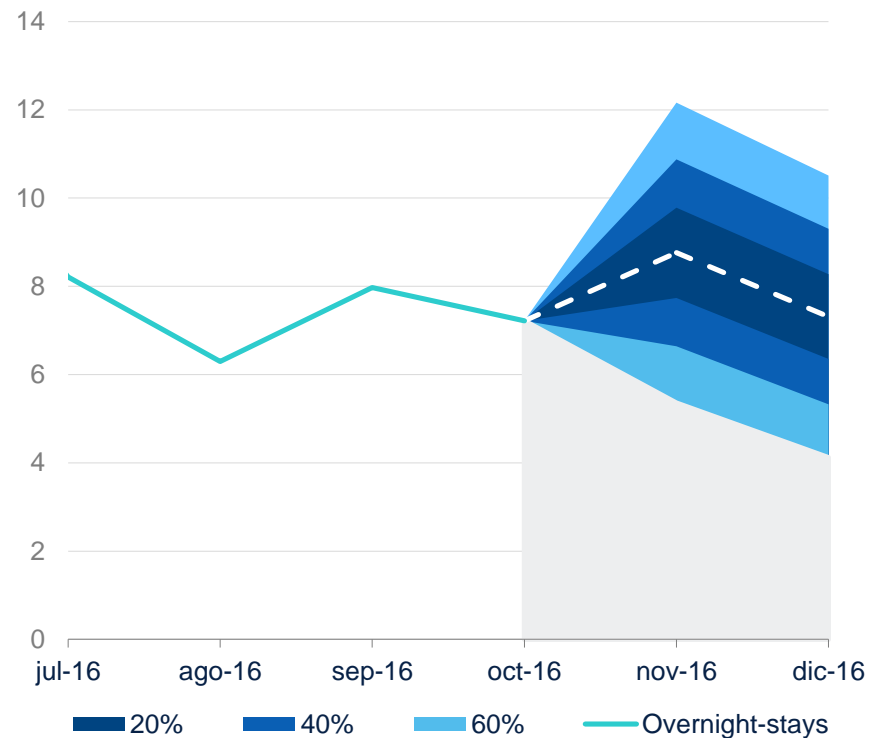
## Overnights of non-resident tourists in hotels and search trends in google

(overnight stays in thousands, searches index = 100, July 2007)



## Overnights of non-resident in hotels and forecasts

(% yoy, latest forecast as of November 30, 2016)



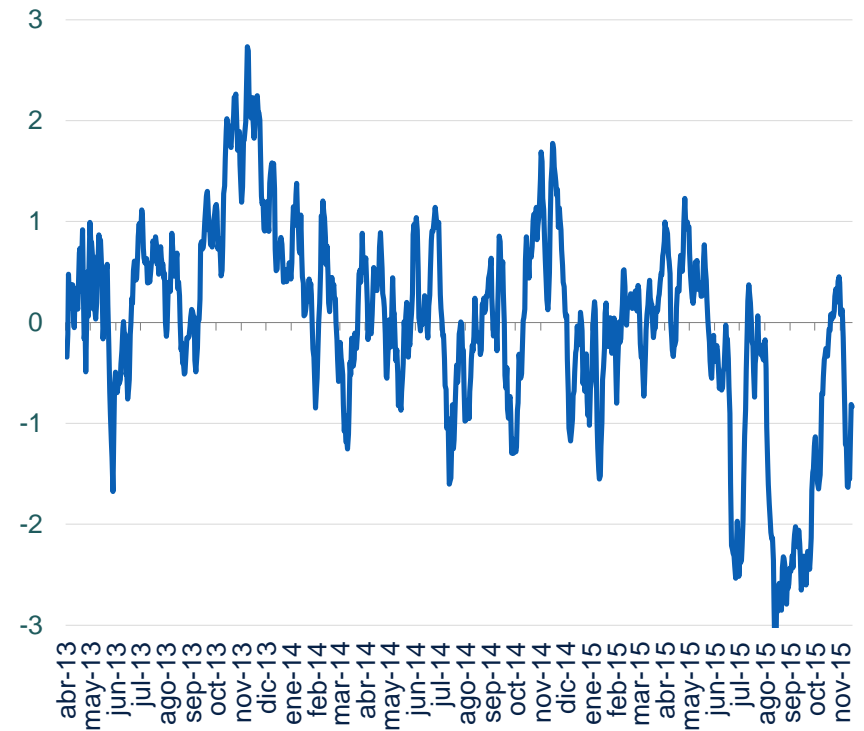
(More information can be found in the following [link](#))

Source: BBVA Research, INE and Google

# The sentiment from News allows us to elaborate a composite index

Macroeconomic Sentiment Index (MSI)			
Fiscal Policy Sentiment Index (FPSI)	Monetary Policy Sentiment Index (MPSI)	Political Sentiment Index (PSI)	Global Policy Sentiment Index (GPSI)
Principal Components Analysis on each component Tone			
Privatization, Austerity, Econ_Debt, Econ_Subsidies, Econ_Taxation, WB_Wages, WB_Fiscal Policy, WB_Investment Policy, WB_Tax and Revenue, WB_Managing Public, WB_Public_Finance, WB_Debt_WB_FiscalRisks, WB_Tax Credit and Subsidies, WB_Fiscal Policy & Jobs, WB_Tax Expenditures, WB_Taxation, WB_Fiscal Cconsolidation	Econ_Interest rates, Econ_cost of living, Econ_currency exchange rate, Econ_Currency reserves, WB_Inflation, WB_Monetary Policy, WB_Exchange rate policy, WB_Central Banks, Fuel Prices	Gov_Reform, General Government, Elections, Democracy, Political_Party, Political Turmoil	Monetary Policy_US, WB_CentralBanks_US, WB_Interest rates US, WB_Monetary Policy_EU, WB_CentralBanks_EU, WB_Interest rates EU, WB_Monetary Policy_CH, WB_CentralBanks_CH, WB_Interest rates CH

**Macroeconomic Sentiment Index for Turkey**  
(Evolution of the “Tone” of main followed themes)

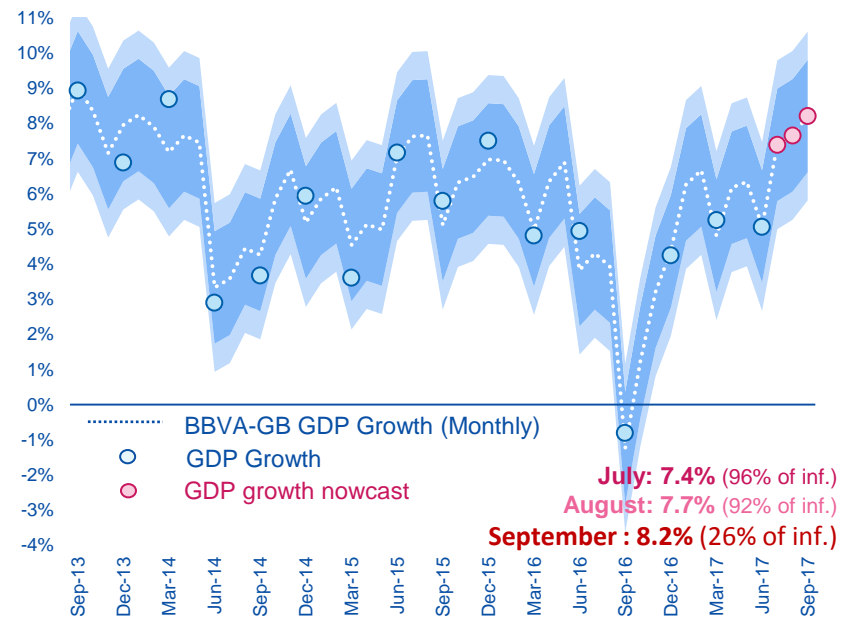


# We can use it to improve our Monthly GDP models...taking advantage of Real Time News

## Dynamic Factor Model for Turkish GDP Pseudo Out of Sample RMS errors

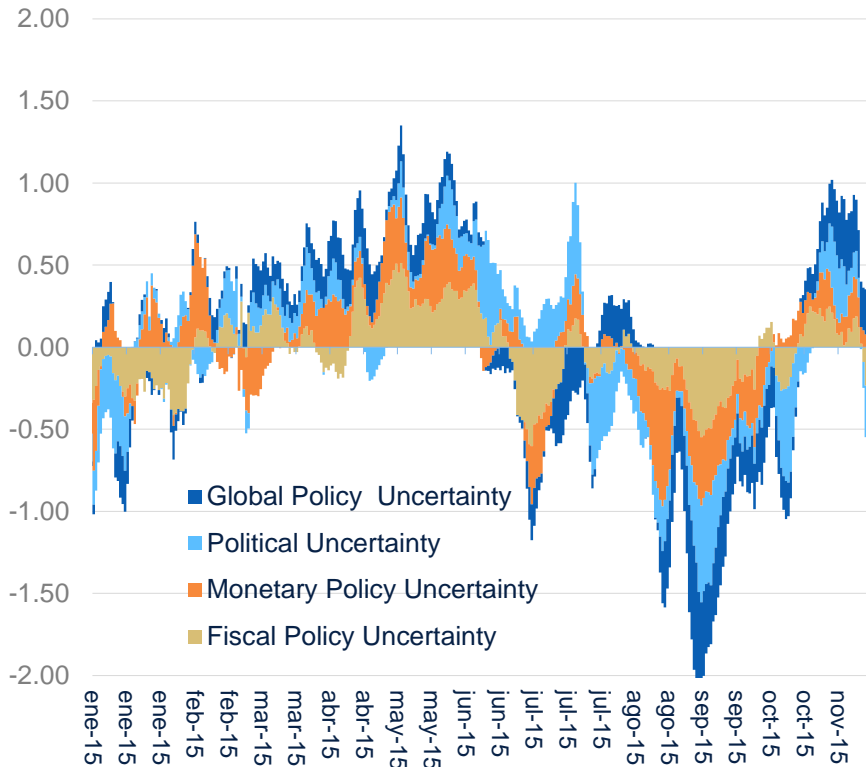
	Monthly	Quarterly	Year
Monthly DFM GDP	0.085	0.256	1.024
<b>Monthly DFM GDP + MU Index</b>	<b>0.046</b>	<b>0.139</b>	<b>0.558</b>
Monthly DFM GDP + MU Weighted	0.063	0.190	0.569
Monthly DFM GDP + MU Monetary P.	0.046	0.139	0.556
Monthly DFM GDP + MU Politics	0.046	0.139	0.556
<b>Monthly DFM GDP + MU Fiscal .P</b>	<b>0.046</b>	<b>0.138</b>	<b>0.550</b>
Monthly DFM GDP + MU Global I	0.063	0.188	0.563

## Monthly Turkish GDP Growth Indicator & Nowcast (YoY Change, %)

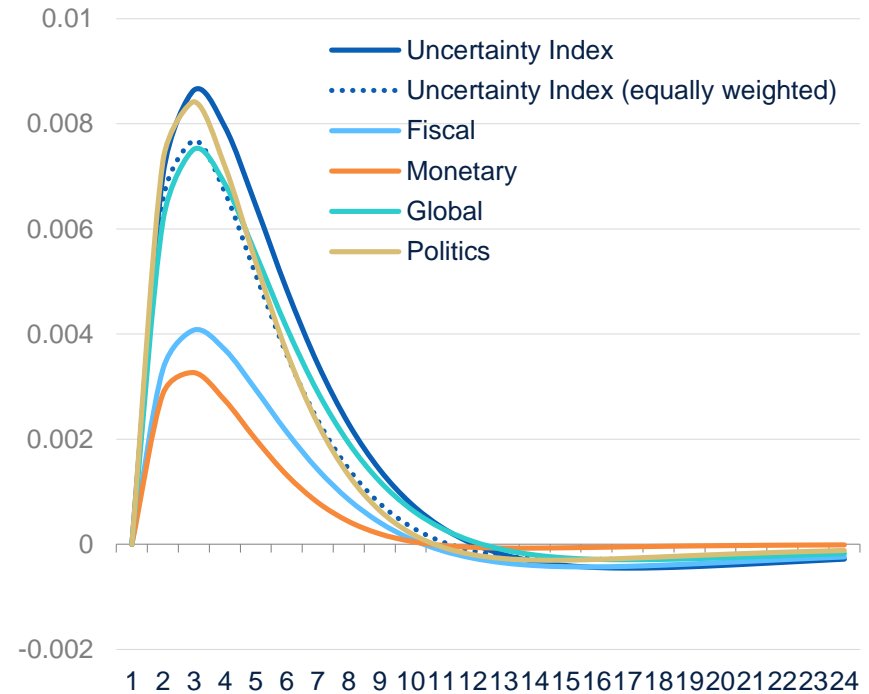


# We can check the evolution over time...and how the financial assets response to different sentiment variables...

**Turkey: Macroeconomic Uncertainty in 215**  
(in Standard Deviations)



**Turkey: Impulse Response of Exchange rate to shocks in sentiment**  
(in Standard Deviations)

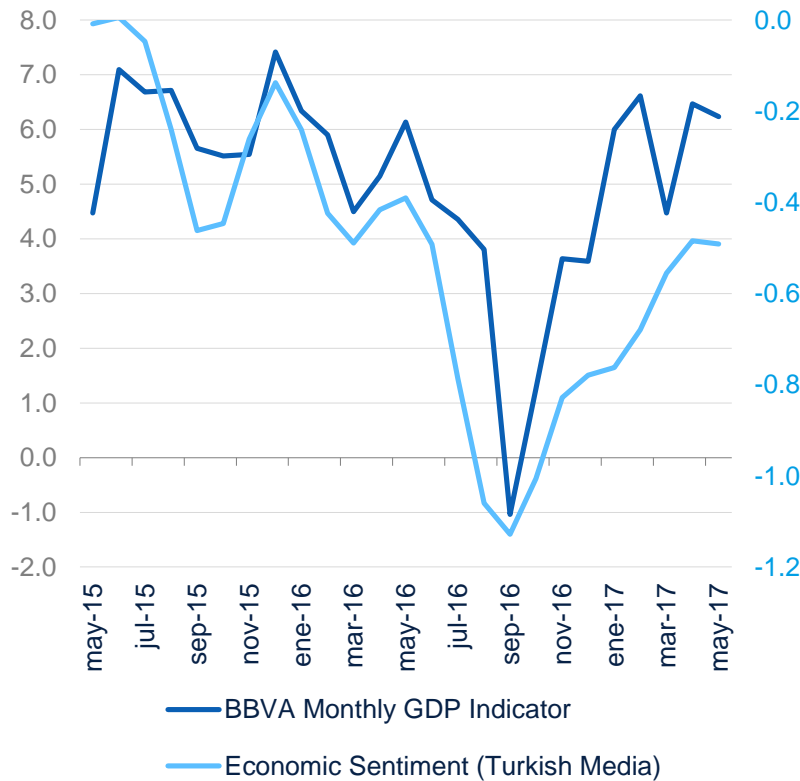


\* The impulse response correspond to a Bayesian VAR model  
With Global GDP, Inflation, Interest rate, Monthly local GDP , Uncertainty and Exchange rate . It was estimated through Gibbs Sampling due to restriction on data  
Source: BBVA Research

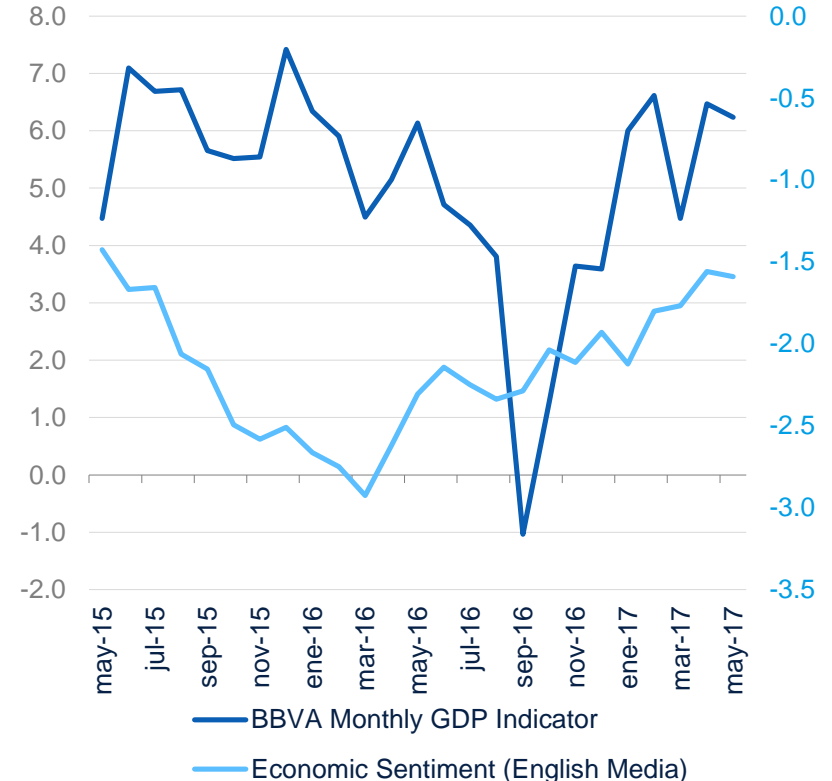


# ... or to analyze the importance of narratives and language bias: and yes, they matter...

**Turkey GDP & Economic Sentiment**  
(%YoY and Turkish written Economic Sentiment)



**Turkey GDP & Economic Sentiment**  
(%YoY and English written Economic Sentiment)



# It is not only about Economic Sentiment... but also about complementing official data...

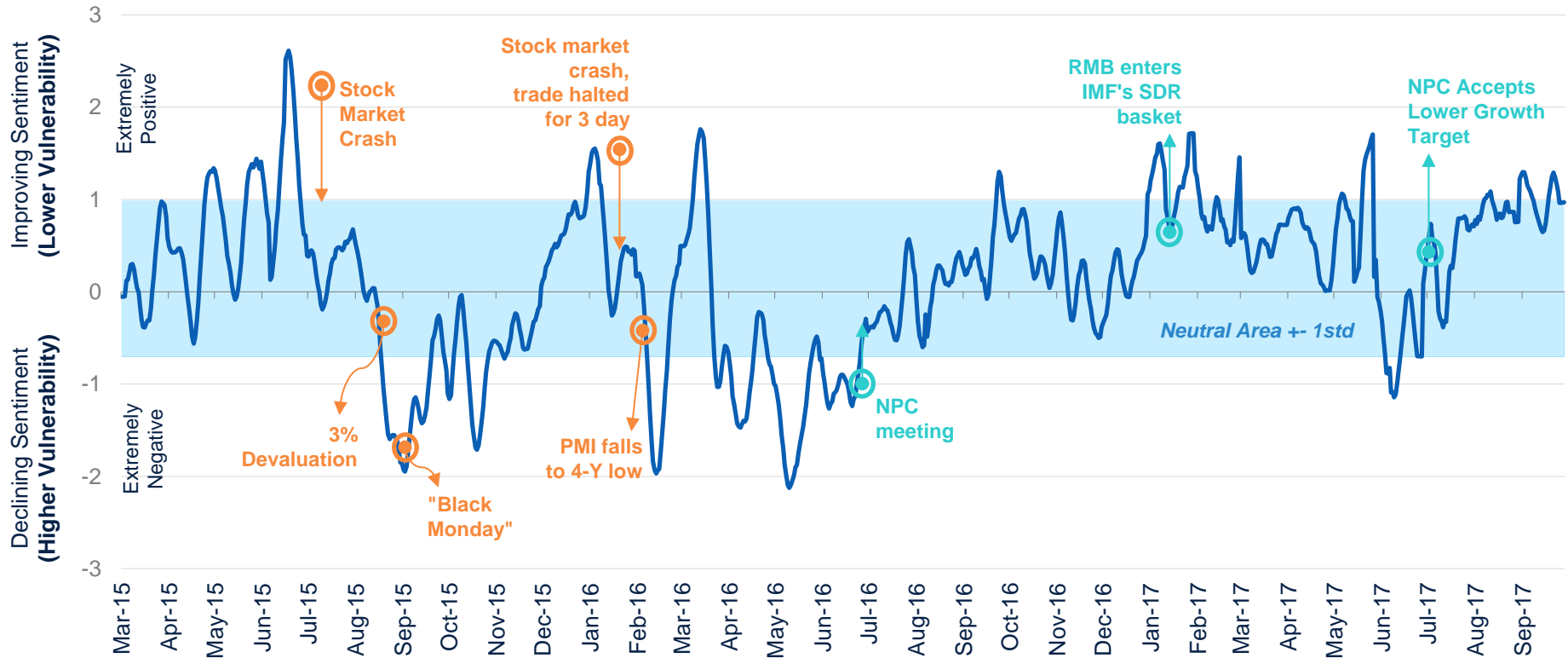
## Chinese Vulnerability Sentiment Index (CVSI): components and evolution

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) Liabilities (M)  <b>25%</b>	Mortgages.loan (M) GICS.Housing.Index (M) Housing.Price (M) New.Construction (M) RealEst.Invest (M)  <b>45%</b>	NPL.Ratio (M) TSF.Aggregate.New Increase (M) Entrusted.Loans (M) Wenzhou.Index (D) WMPs Acceptances (M)  <b>35%</b>	Foreign.Reserves (D) CNY Exchange Rate (D) CNH Exchange Rate (D) HICNHON.Index (D)  <b>40%</b>
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)  <b>75%</b>	Housing_policy_&_institutions (D) Housing_markets (D) Housing_prices (D) Housing_construction (D) Housing_finance (D) Land_reform (D)  <b>55%</b>	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)  <b>65%</b>	Currency_exchange_rate (D) Currency_reserves (D) Capital_account (D) Macroprudential_policy (D) Exchange_rate_policy (D) Illicit_financial_flows (D)  <b>60%</b>

# ...to track Risks in Real Time...

## Chinese Vulnerability Sentiment Index (CVSI)

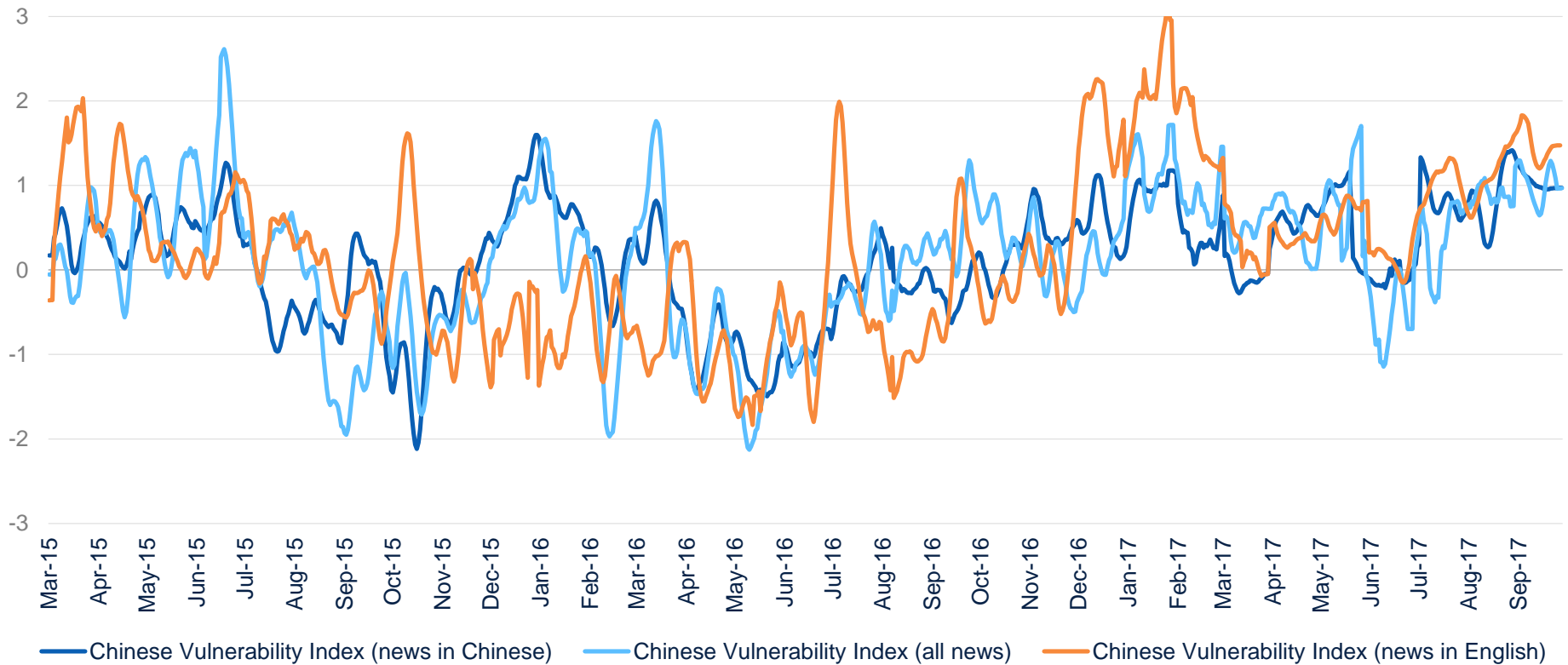
(Evolution of the "Tone" of main followed themes about vulnerability in China. Lower values indicate a deterioration of sentiment and higher vulnerability)



# ...disentangling media language effects...

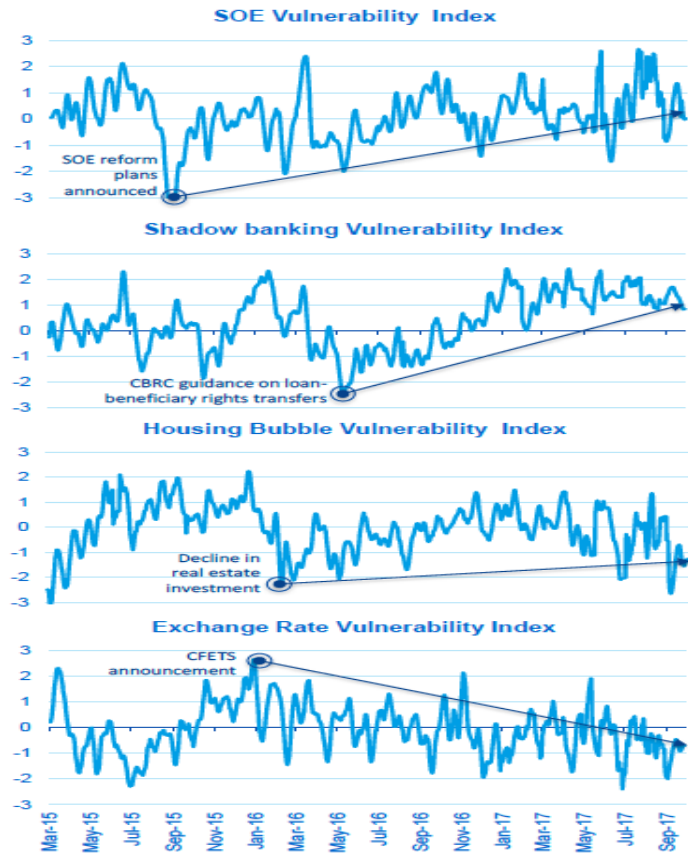
## Chinese Vulnerability Sentiment Index by media language: total, Chinese and English

(Evolution of the “Tone” of main followed themes about vulnerability in China. Lower values indicate a deterioration of sentiment and higher vulnerability)

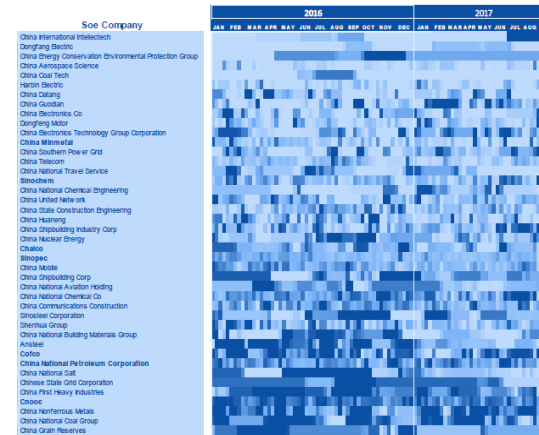


# ...and analyzing risks at a high degree of granularity

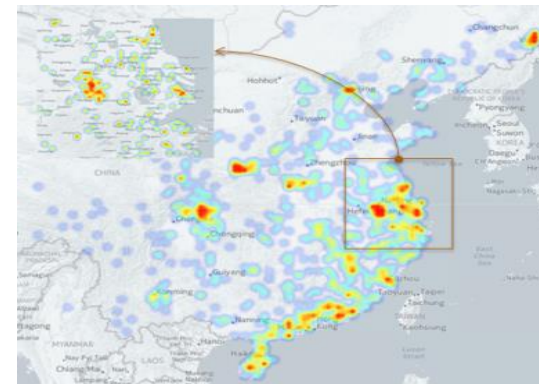
## Chinese Vulnerability Sentiment Index Components (CVSI)



## China SOE Map (sentiment on SOE)

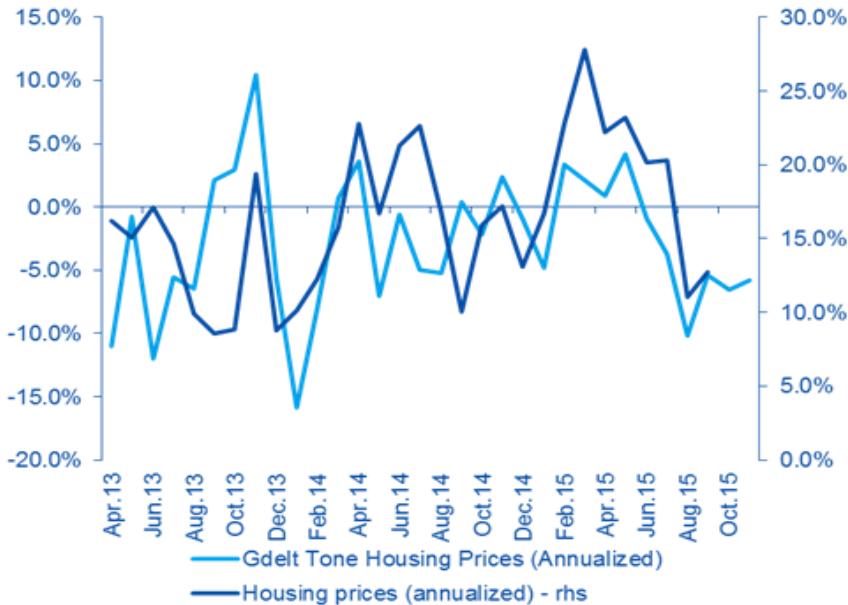


## Geographical Analysis Housing Prices (sentiment on Housing Prices)

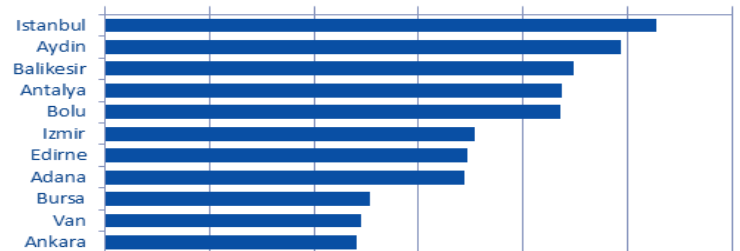
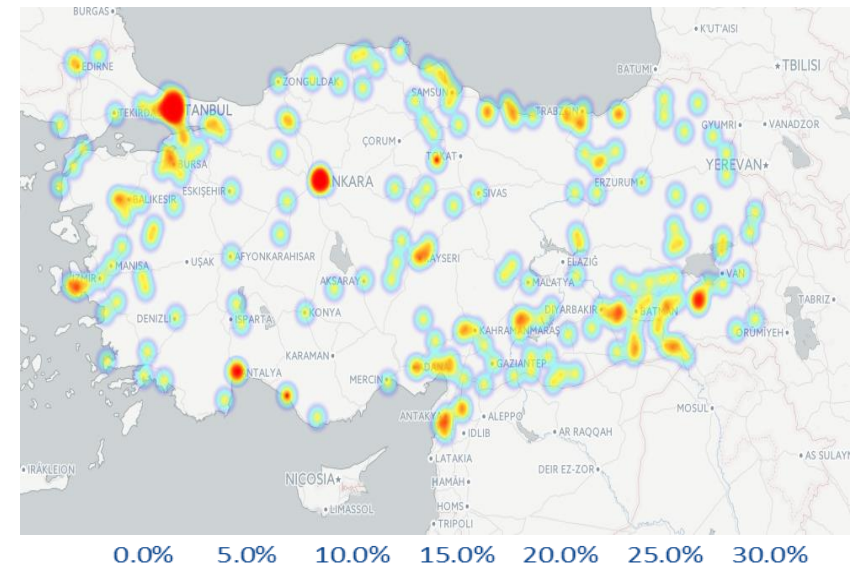


# Housing Prices nowcasting is also a promising aspect of Big Data

**Turkey House Prices Tone and Housing Prices**  
 (Dark Blue: more negative tone)



**Geographical distribution of House Prices Tone 2015**  
 (Dark Blue: more negative tone)

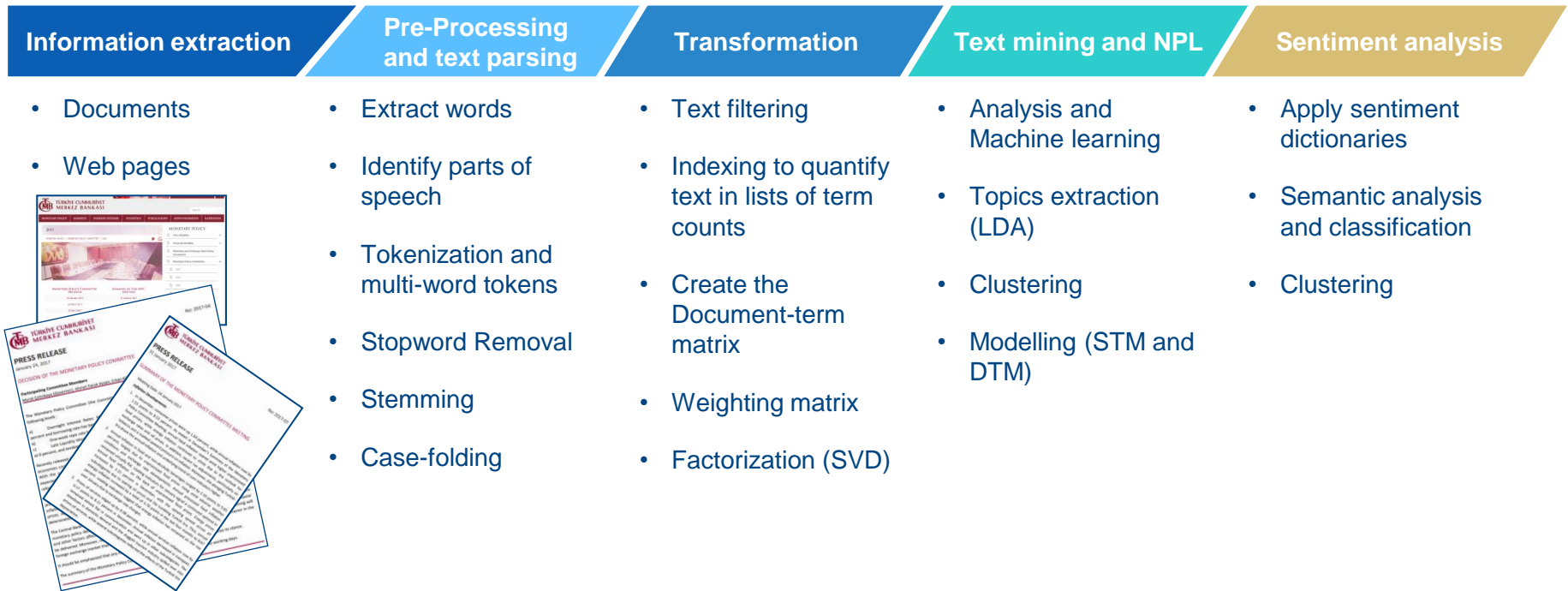


# 04

## **Text Mining and Sentiment analysis**

# External databases: web scrapping and NPL techniques

Text mining makes information extraction from huge volumes of data easier and structures the information as important facts, key terms or persons.

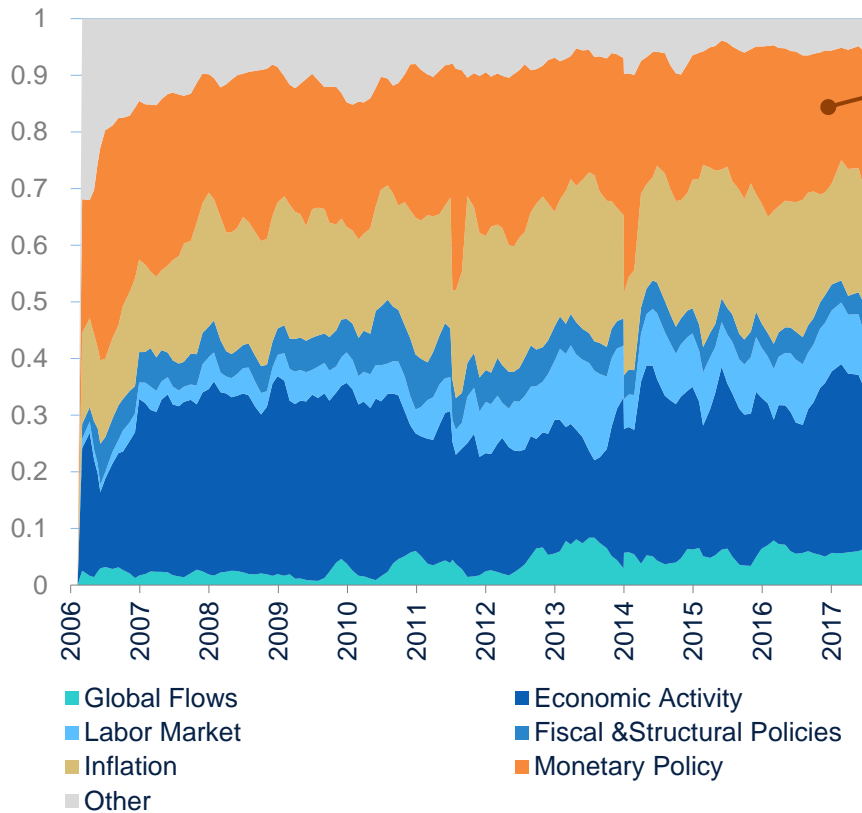





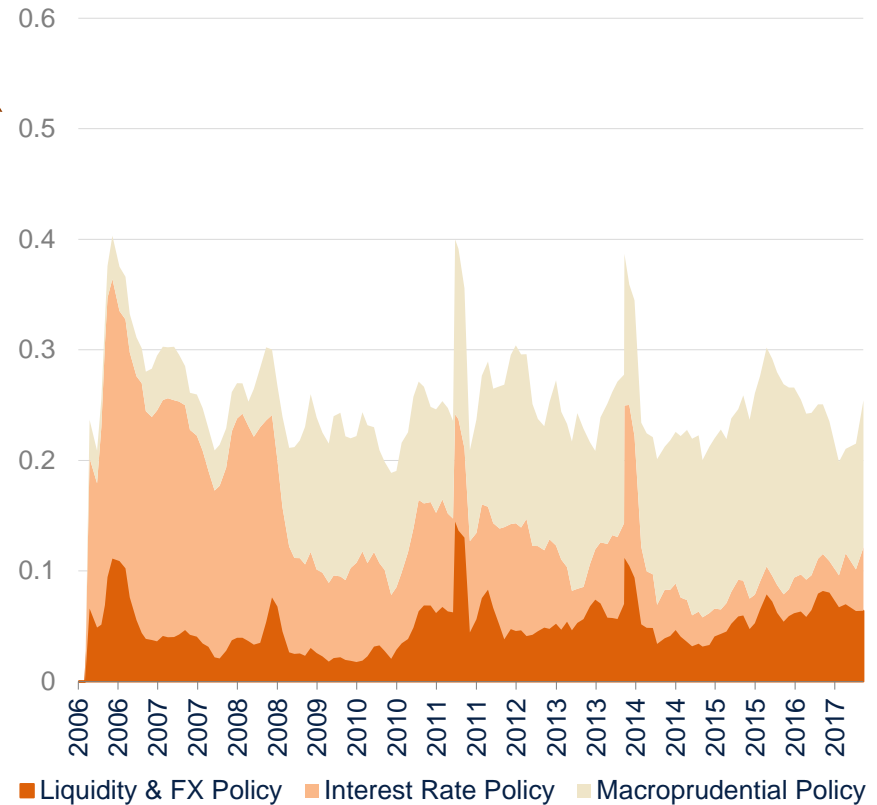


... and we can check “what the Central Bank is talking about”...

Central Bank Of Turkey: Evolution of Topics

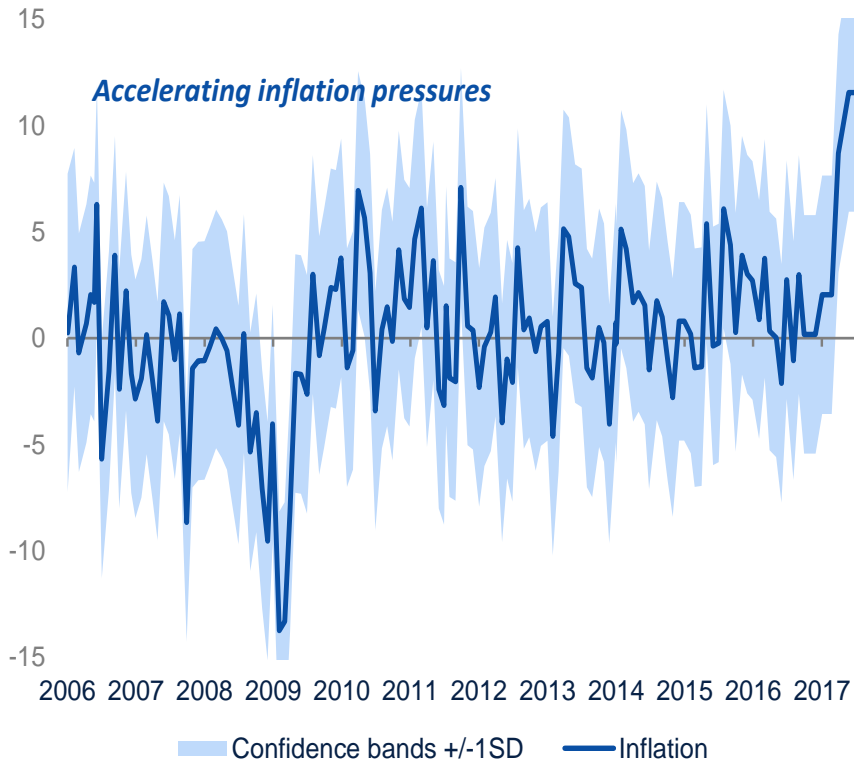


Monetary Policy Topics Distribution (% of Total)

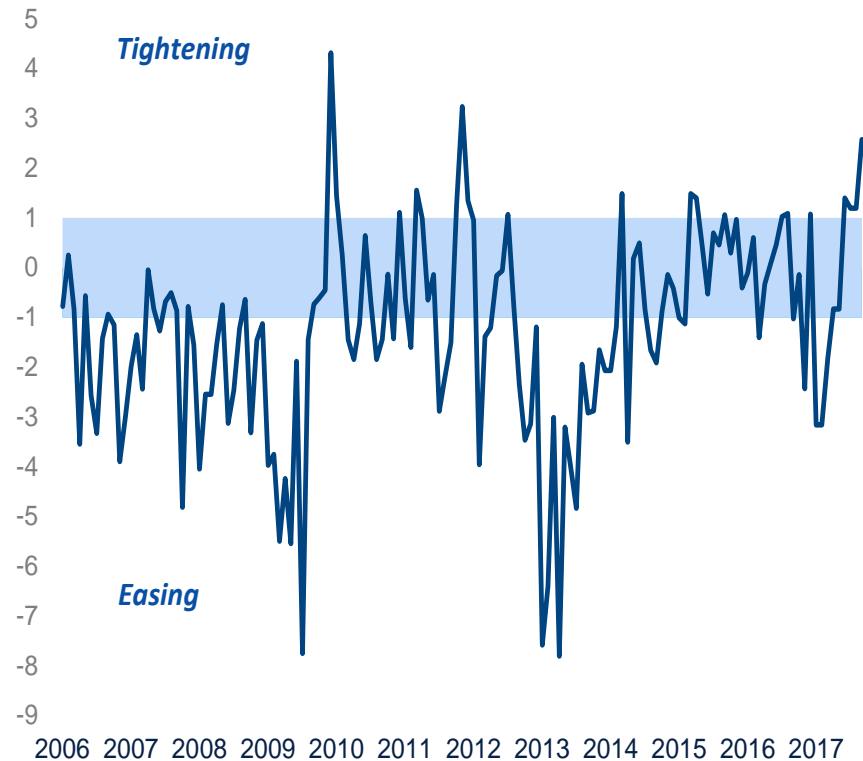


# ... as well as the topic sentiment and the stance of CB reports...

**Central Bank Sentiment on Inflation**  
(Standardized, Big Data LDA Techniques applied to Minutes & statements)

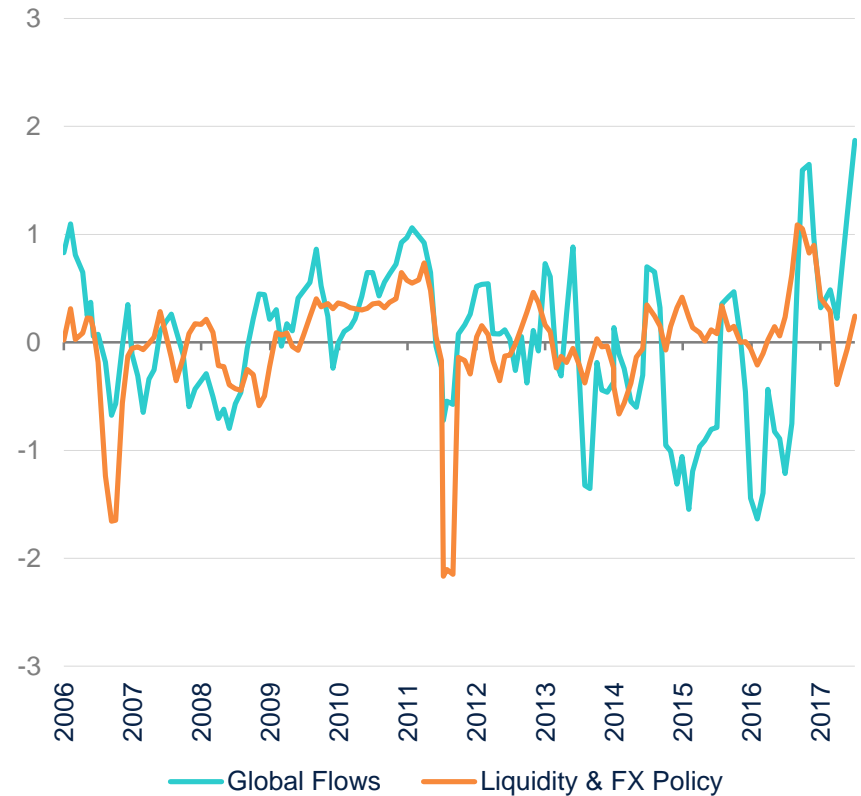
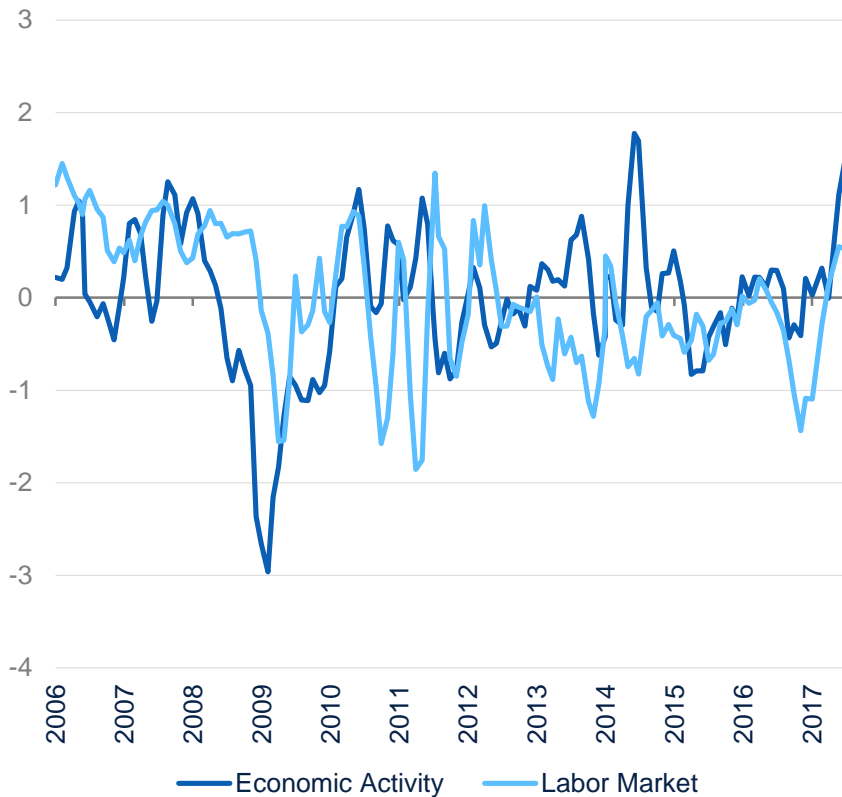


**Monetary Policy Sentiment**  
(Standardized, estimated through Big Data LDA Techniques from Minutes & Statements)



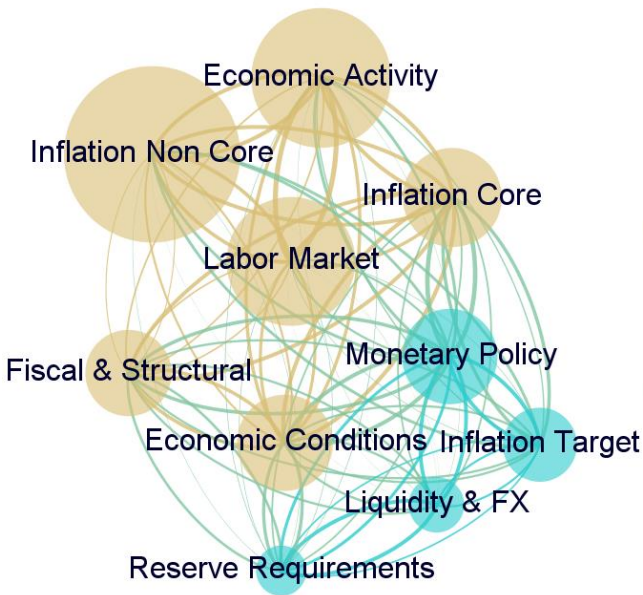
# Which is changing over time... according to text mining and machine learning techniques...

Sentiment evolution of Topics in CB reports in 2006-17.

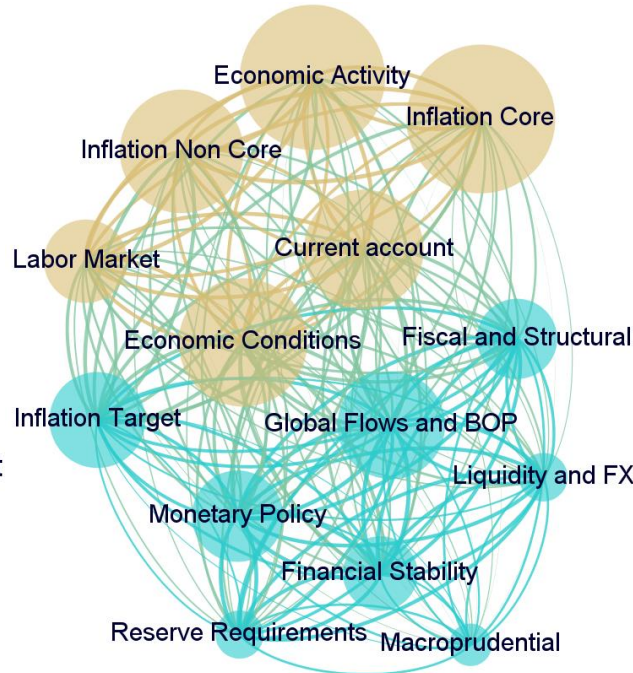


# ...as well as the relationships between topics and their evolution over time using topic network analysis

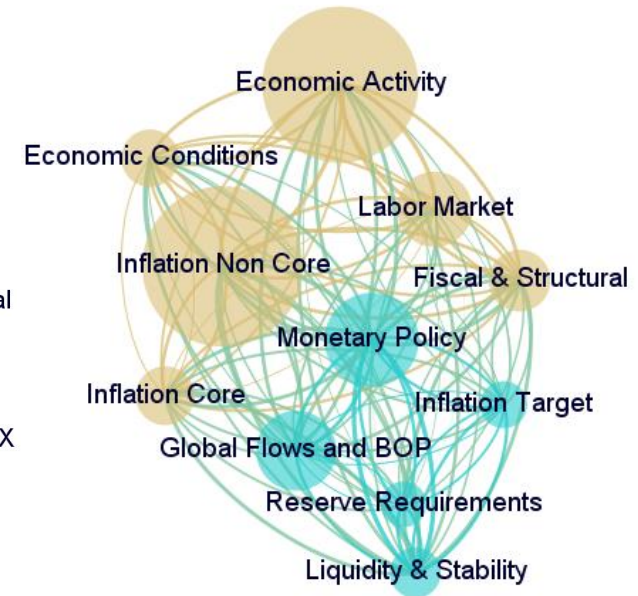
**Topic network 2006-09: the inflation Target**



**Topic network 2010-15: the global financial crisis period**



**Topic network 2016-17: in search of price stability**



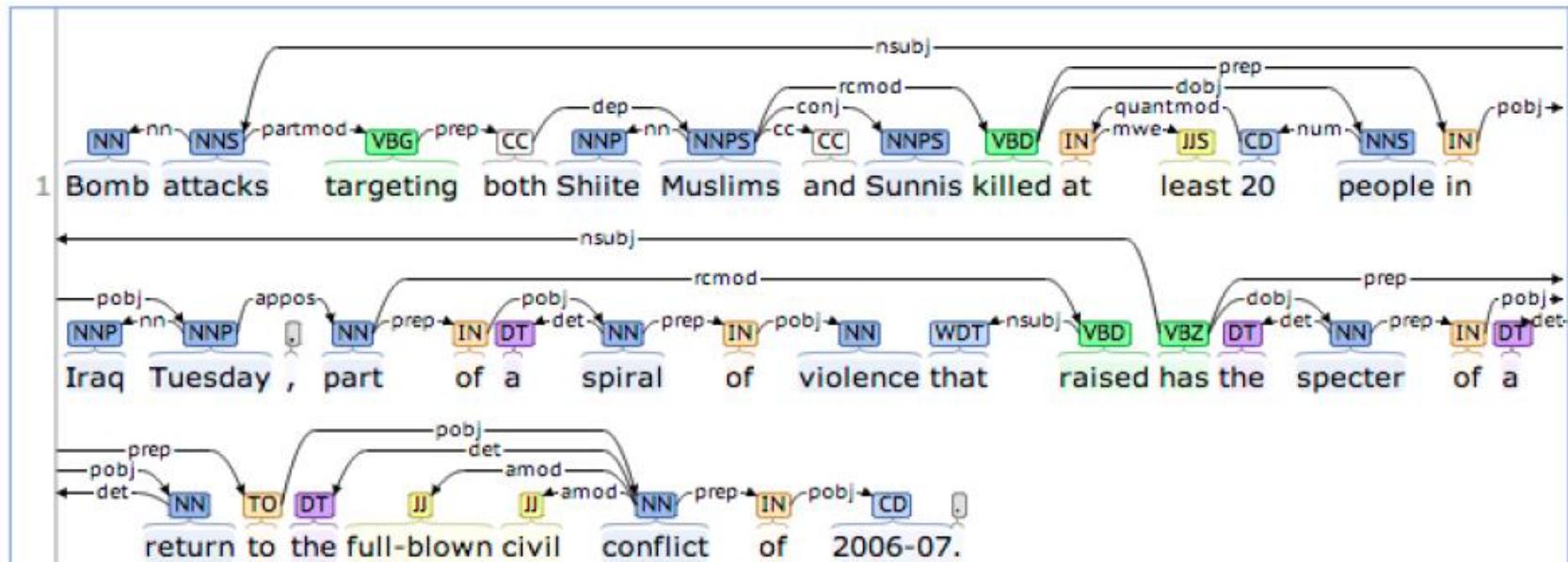
The network of the estimated and correlated topics using STM. The nodes in the graph represent the identified topics. Node size is proportional to the number of words in the corpus devoted to each topic (weight). Node color indicates clusters using a community detection algorithm called modularity developed 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-occurrence of words between topics). Edge width is proportional to the strength of this co-occurrence between topics.

# ANNEX

# 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:



# Text mining and NPL: pre-processing 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 Schütze 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

## 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:

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

- ◆ 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

# Causal impact methodology

- ◆ To measure the impact of the attacks on the performance of commerce in the city of Barcelona has been used a Bayesian model of time series ([here the reference paper](#)). This model is based on the comparison of the observed behavior in a target time series, from the date of the analyzed event, with a prediction of the expected values of not having occurred. To construct this counterfactual series we use a set of control series not affected by the event
- ◆ In this particular case, the used time series corresponds to the daily expenditure with credit card in physical commerce. The covered period by the series goes from January 1, 2015 to September 24, 2017, setting the date of the event on August 17, 2017. The target series is the recorded expenditure in the city of Barcelona and the control series corresponds to the rest of Spanish municipalities with highest correlation with Barcelona in the previous period
- ◆ Thus, the counterfactual prediction is obtained by a Bayesian inference process in which each of the components of the objective time series (trends, seasonality, cycles ...) is approximated using the set of control series. Once this is done, they are combined to obtain the a priori probabilities of the target series
- ◆ The methodology uses the Monte Carlo Markov chain method to simulate a posterior distributions. This allows not only to generate an expected value for each of the days after the event, but also to allow confidence intervals to determine if the differences between the observed and predicted series (growth and decrement) could have occurred even if the event doesn't occur or if they are statistically not justified without the event. In this analysis it has been considered statistically demonstrated that a difference is due to the attack when its value is in the final 1% of the calculated probability distribution.

# Big Data & Big Models at BBVA Research

ECB Statistics Day

Jorge Sicilia, Alvaro Ortiz & Tomasa Rodrigo

October 2017