

# Big Data y Big Models en BBVA Research

Día de estadísticas del BCE

Jorge Sicilia, Álvaro Ortiz y Tomasa Rodrigo

Octubre de 2017

## Índice

**01**

Oportunidades en la era digital. Big Data en BBVA Research

**02**

Geopolítica, comercio y efectos colaterales

**03**

Indicadores económicos y de riesgo en tiempo real

**04**

Minería de textos y análisis de opiniones

# 01

**Oportunidades en la era digital.  
Big Data en BBVA Research**

## Los datos tradicionales no podían dar respuesta a algunas preguntas relevantes...

 **Conciencia social y la Primavera Árabe**

 **Acontecimientos políticos y reacción social**

 **Desastres naturales y epidemias**

... lo que nos impide medir su impacto económico...  
... en un mundo en el que los riesgos y la  
incertidumbre no paran de crecer



**El uso de las técnicas Big Data y Data Science nos permite  
cuantificar estas tendencias**

## Nuevo marco en la era digital...

*Se necesitan **nuevos enfoques computacionales basados en datos** para permitir que la **nueva era digital** explote las nuevas oportunidades en las que los datos se pueden usar para estudiar el mundo en tiempo real desde el nivel micro hasta el macro.*



› Nueva disponibilidad de datos



› Combinación de datos históricos con datos en tiempo real



› Una infraestructura mejor y más rápida



› Técnicas y algoritmos avanzados de ciencia de datos



› Nuevas respuestas a viejas preguntas



› Mayores capacidades computacionales para hacer frente a una mayor granularidad de los datos

## ...que necesita el desarrollo de nuevas competencias para aprovecharlo



**Hacer las preguntas correctas**



**Desarrollar la gestión de datos y las capacidades de programación para trabajar con conjuntos de datos a gran escala**



**Profundizar en las habilidades estadísticas y econométricas para analizar y tratar con datos de alta dimensión**



**Interpretación de los resultados: resumir, describir y analizar la información**

Los nuevos datos pueden terminar cambiando la forma en que los economistas abordan las preguntas empíricas y las herramientas que utilizan para responderlas

# Big Data en BBVA Research

## Nuestro trabajo



- Analizamos **cuestiones geopolíticas, políticas, sociales y económicas utilizando bases de datos a gran escala y métodos basados en datos cuantitativos** en lugar de introspección cualitativa.

## Nuestros conjuntos de datos



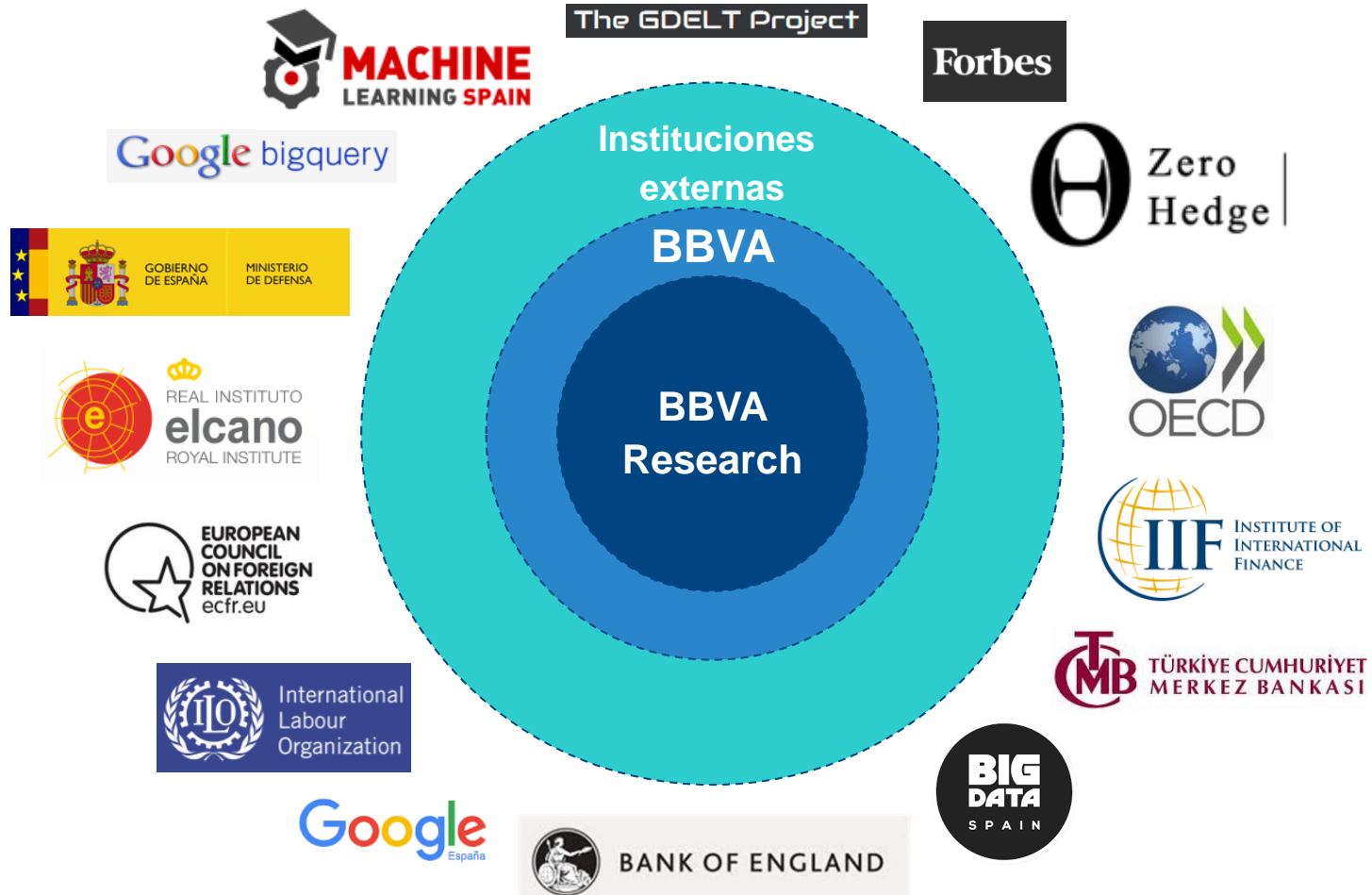
- **Datos de medios** para explorar la intensidad de las noticias, la densidad geográfica de los eventos (inteligencia de ubicación) y las emociones en todo el mundo (análisis de sentimientos)
- **Datos agregados y anonimizados de BBVA** de la huella digital de los clientes
- **Datos de la web** (informes de bancos centrales, entre otros)

## Nuestros resultados



- Estamos en la **frontera de investigación en el área geopolítica y económica** que contribuye a la innovación y aumenta nuestro alcance interno y externo.

# Difusión interna y externa





# Nuestro proceso de trabajo

## Bases de datos

**GDELT**  
Datos de  
BBVA  
Búsqueda de  
Google  
Web

The GDELT Project  
BBVA  
DATA & ANALYTICS

## SaaS

**BigQuery**  
y  
**Amazon**  
**Redshift**

Google BigQuery  
amazon  
REDSHIFT

## Análisis

Limpiar,  
agregar  
transformar  
y modelar  
los datos

BBVA | Research

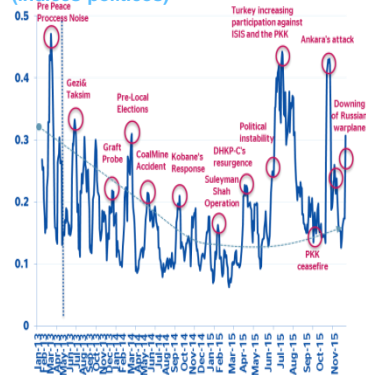
## Visualización

Fusionar,  
visualizar  
y analizar  
los datos

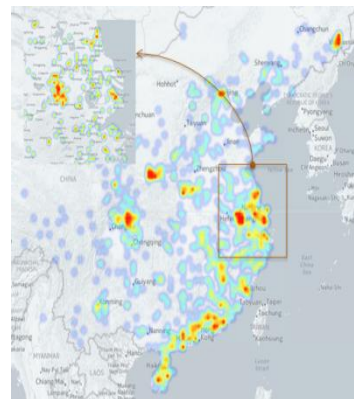
CARTO Gephi  
Google Data Studio

# Nuestros productos

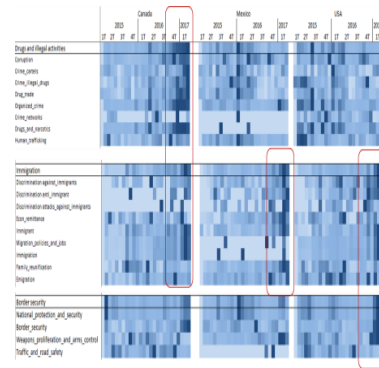
## Índices políticos, geopolíticos y sociales (índices políticos)



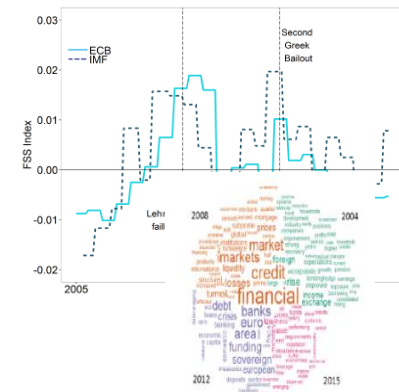
## Análisis geográfico de los precios de la vivienda (Sentimiento sobre los precios de la vivienda)



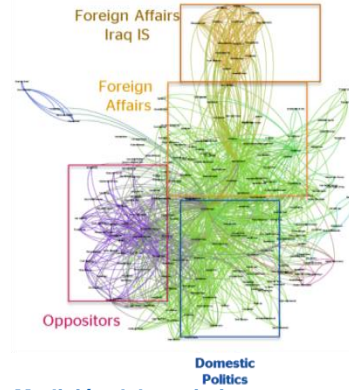
## Mapas de color TLCAN Temas (Proyecto TLCAN)



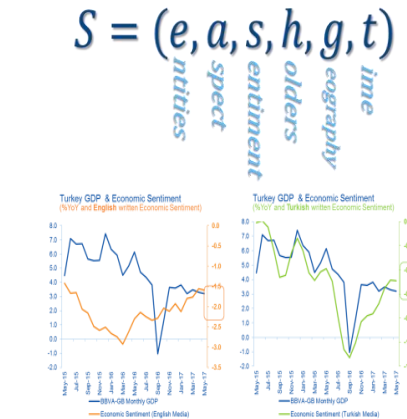
## Estabilidad financiera y macroprudencial (Índice BCE y FED FS por el Consejo de la FED)



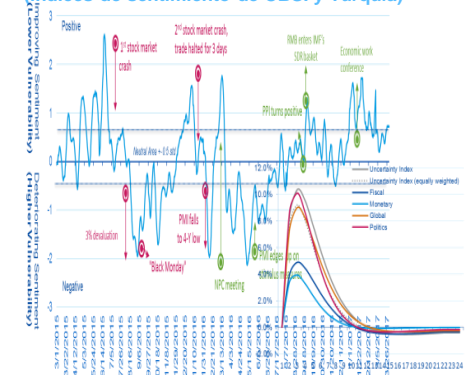
## Redes Políticas y financieras (Redes políticas)



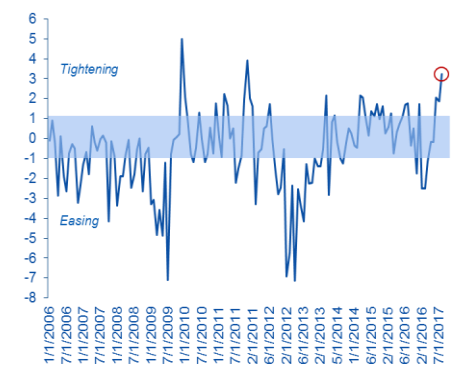
## Medición del sentimiento (Análisis de sentimiento sobre economía y sociedad)



## Mix de datos duros y de sentimiento y modelos VAR (Índices de sentimiento de CBSI y Turquía)



## Tonos monetarios y de estabilidad de los bancos centrales



# 02

## Geopolítica, comercio y efectos colaterales

## Bases de datos externas: GDELT

### Global Database on Events Location and Tone

(Base de datos global sobre eventos, ubicación y tono)

Base de datos abierta de la sociedad humana de todos los rincones del mundo que data de 1979...

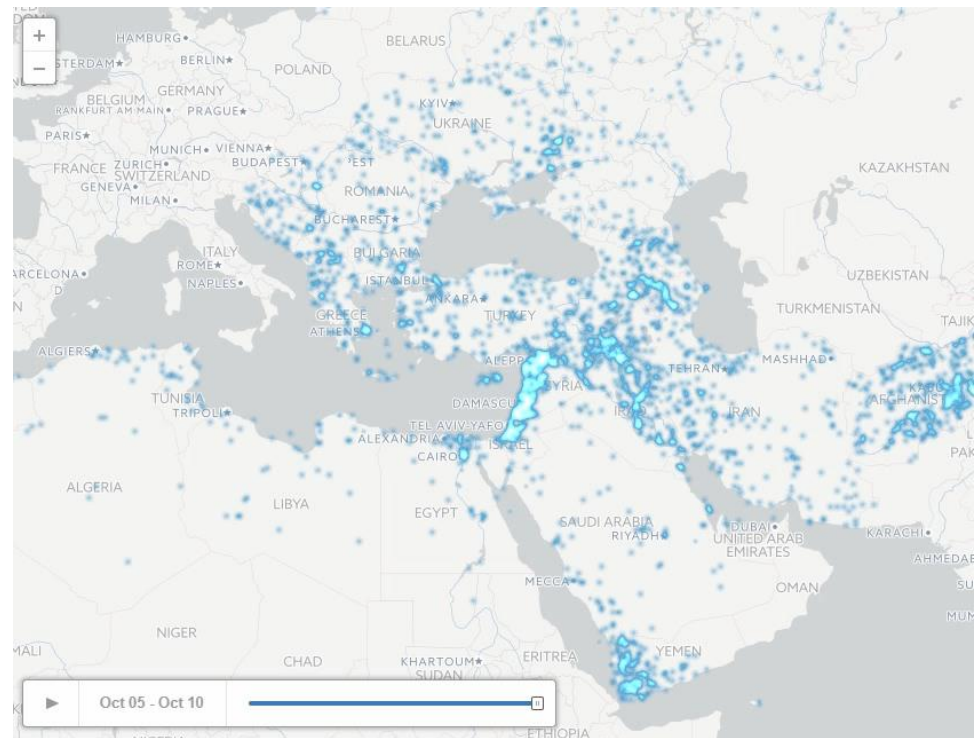
... incluye más de 300 eventos en todo el mundo y más de 30 000 temas...

... por georreferencia en todo el planeta...

... y que recopila emociones usando algunos de los algoritmos más sofisticados.

# El seguimiento de la geopolítica en tiempo real es útil para identificar los principales puntos calientes y los posibles efectos indirectos

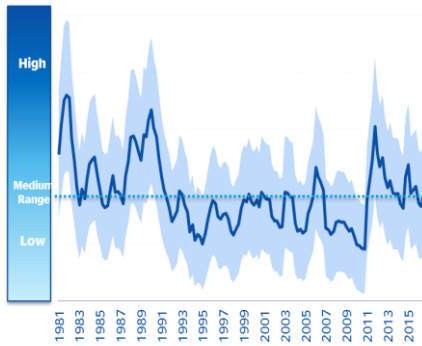
**Mapa de Intensidad de Conflictos 2017**  
(Número de conflictos/Total eventos)



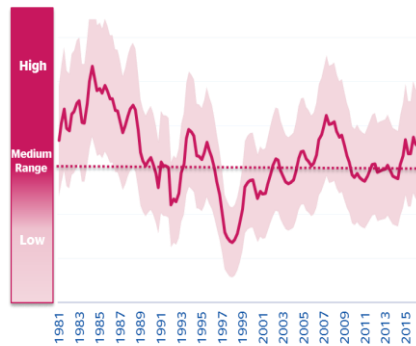
# Desde una perspectiva histórica...

## Índice mundial de intensidad de conflictos y protestas de BBVA Research 1979-2017

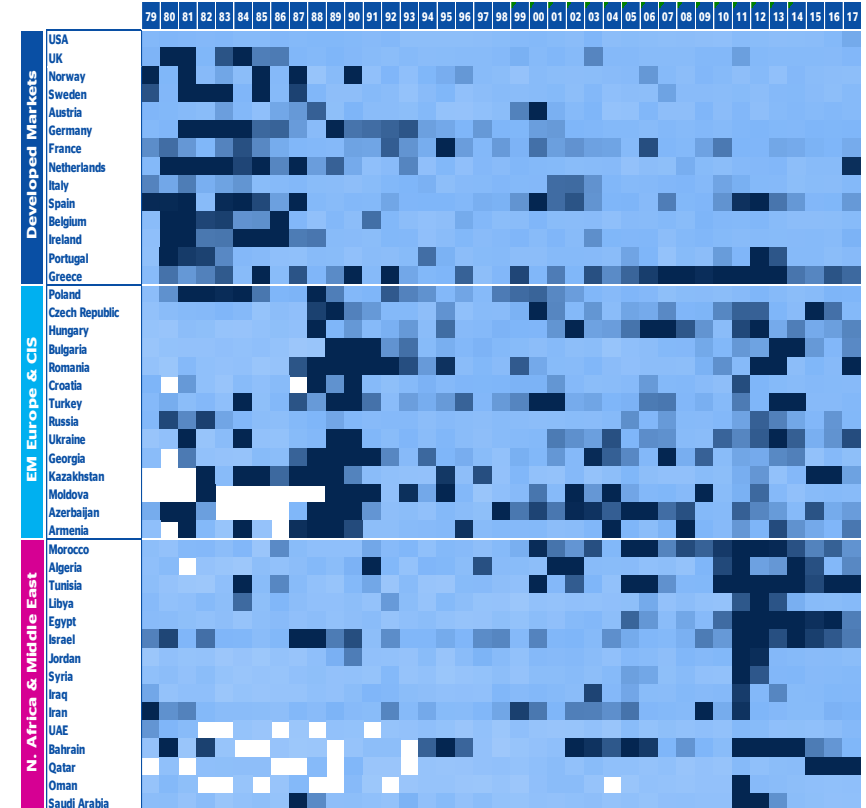
### Protestas



### Conflicto



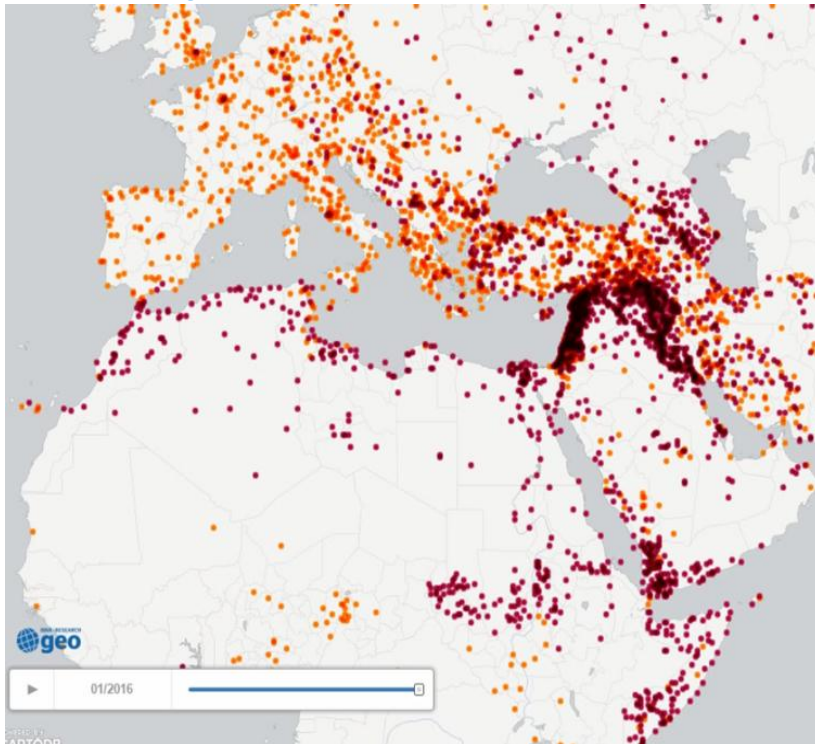
## Mapa mundial de Intensidad de Protestas 1979- 2017



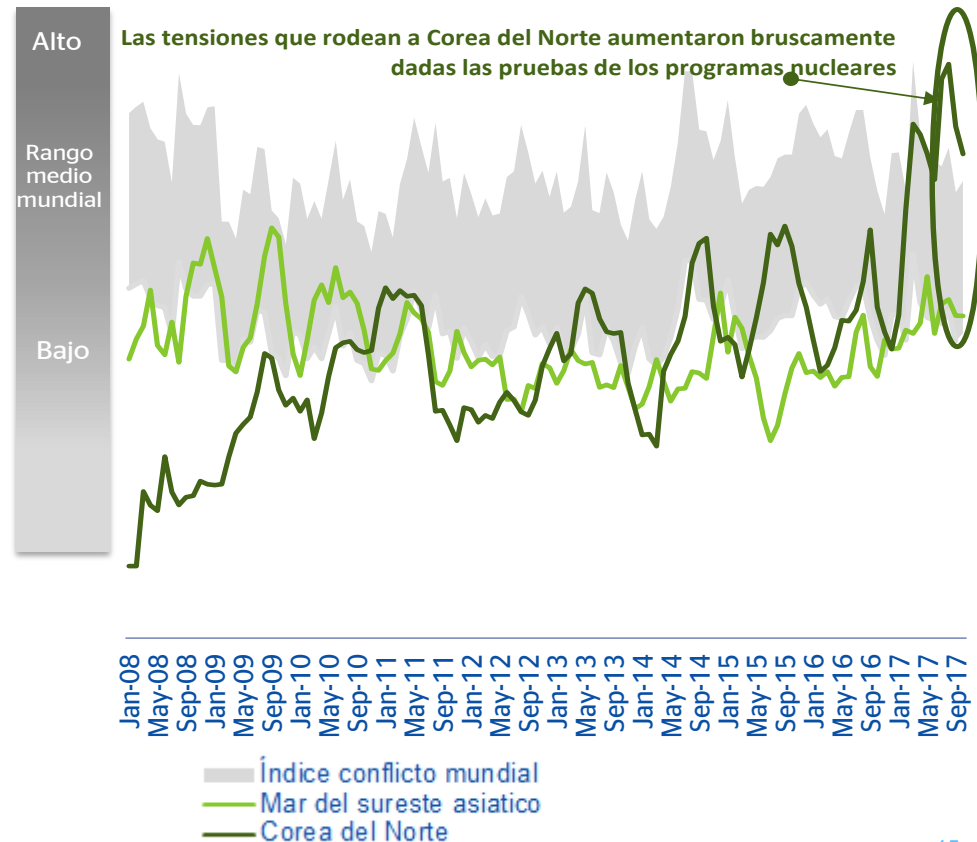
# ...hasta los principales puntos calientes...

## Mapa de flujos de refugiados de BBVA Research en 2015-17

Número de citas de los medios sobre los flujos de entrada y salida de refugiados



## Índice de intensidad de conflictos en Asia 2008-17 de BBVA Research

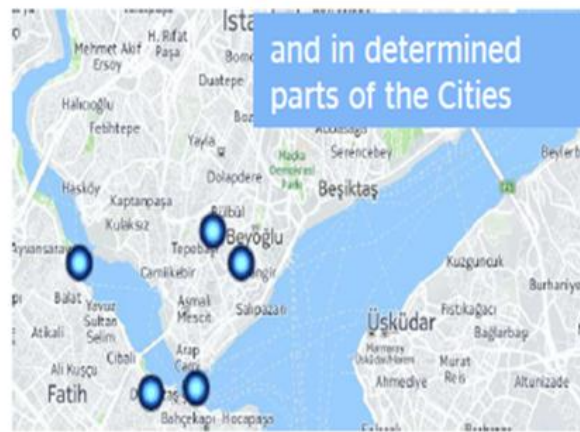
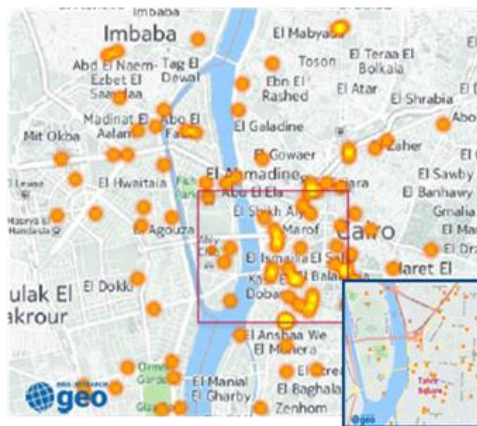
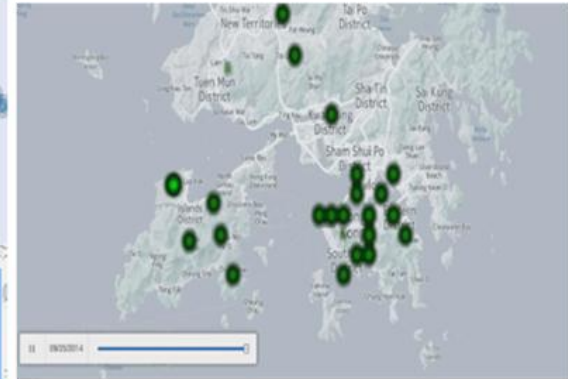


# ...en la geolocalización exacta

## Eventos de disturbios sociales en todo el mundo: casos de El Cairo, Estambul y Hong Kong



Concentrated in the Main Cities...



and in determined parts of the Cities



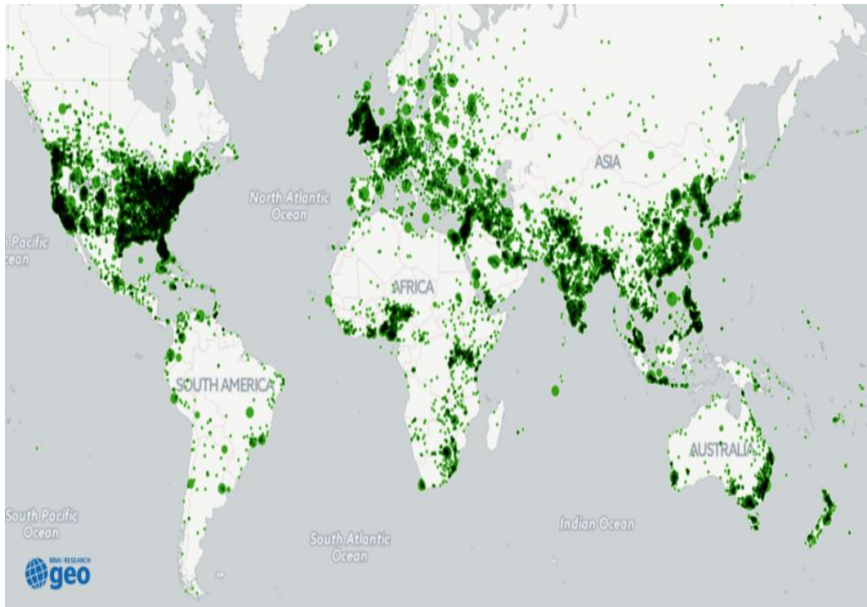


# Las nuevas amenazas como los ataques cibernéticos también se pueden supervisar

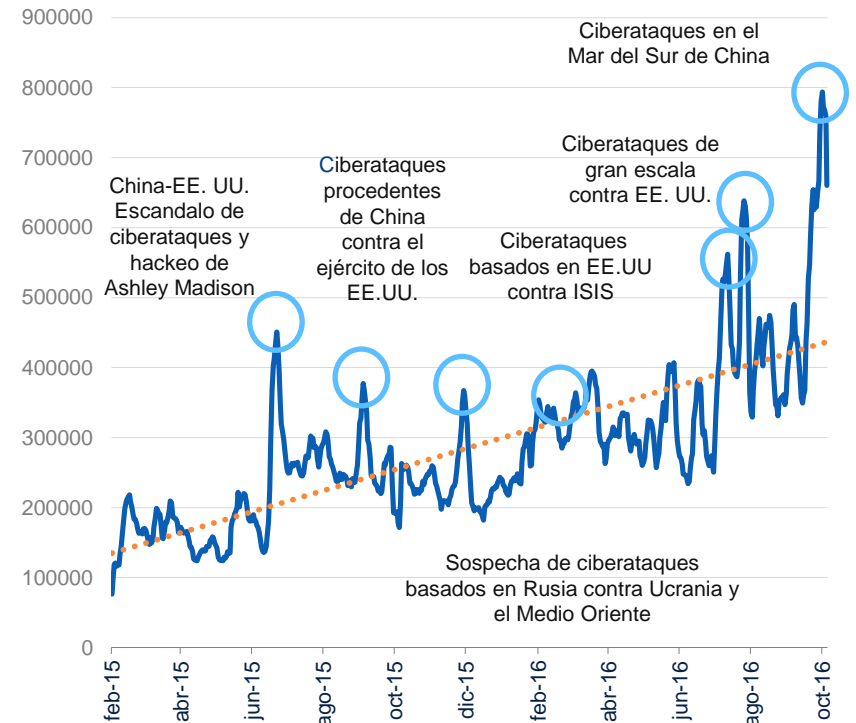
## Los ciberataques se han convertido en una de las principales amenazas en 2015-2017

(Indicador basado en GDELT de ciberguerra, ciberataques, violaciones de datos u otros problemas de seguridad en línea)

**Cobertura de medios de guerra cibernética, ciberataques, violaciones de datos y otros problemas relacionados con la seguridad informática y en línea en todo el mundo 2015-2016**



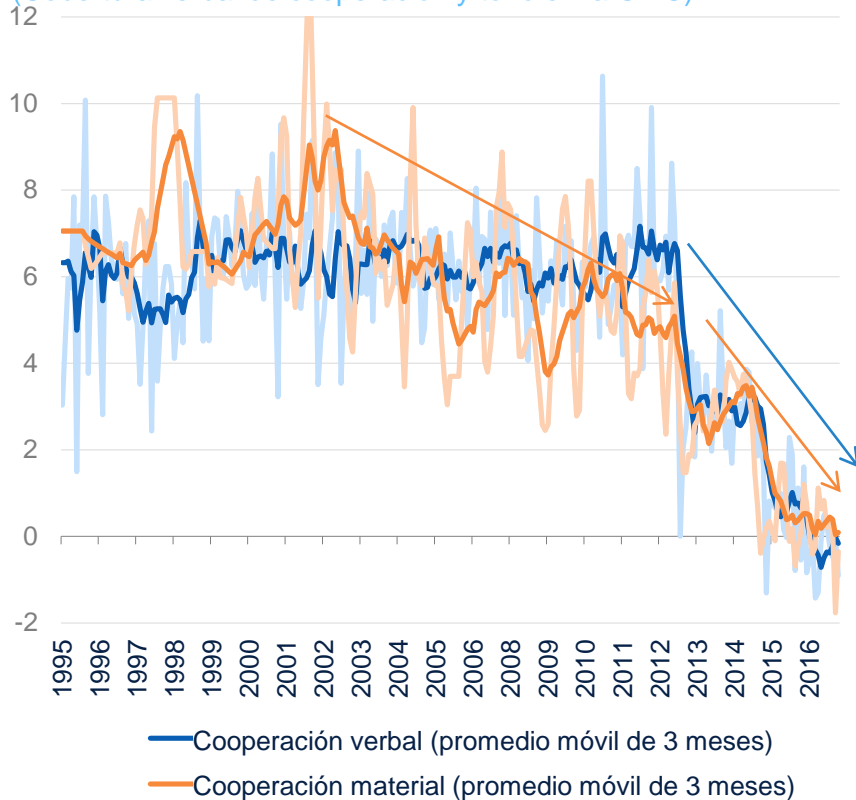
## Cobertura mundial de los ciberataques en 2015-2016



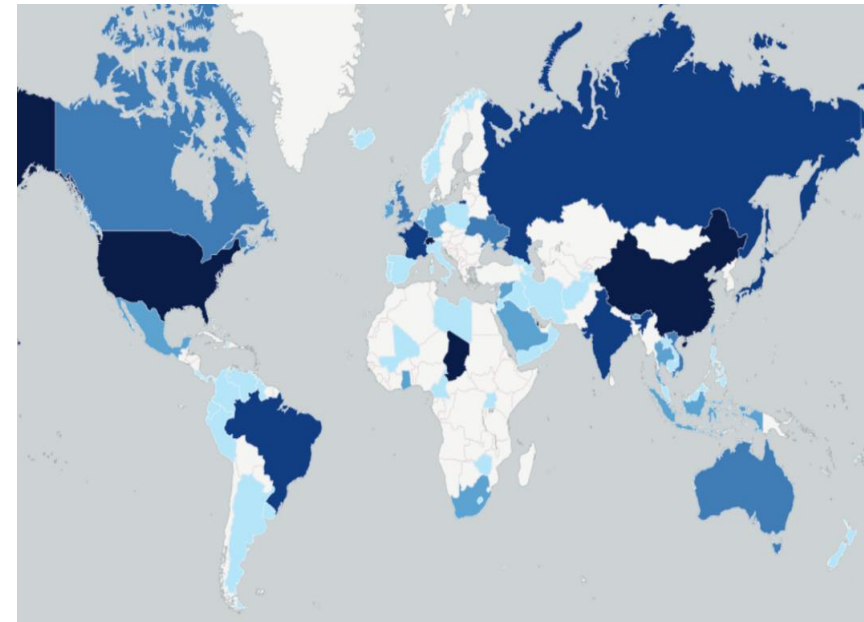
# Gracias a Big Data podemos verificar en tiempo real cómo es el soporte material y verbal en el comercio mundial...

## Índice de soporte del comercio mundial de BBVA Research

(Cobertura verbal de cooperación y tono en la OMC)



## Cambios en el índice de soporte del comercio mundial de BBVA Research 2008-17

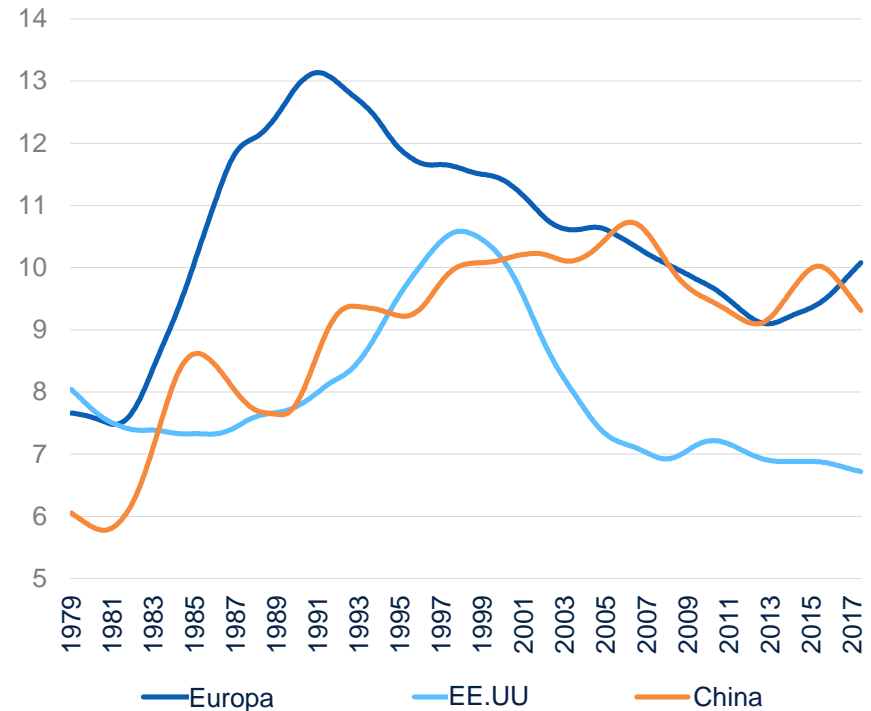
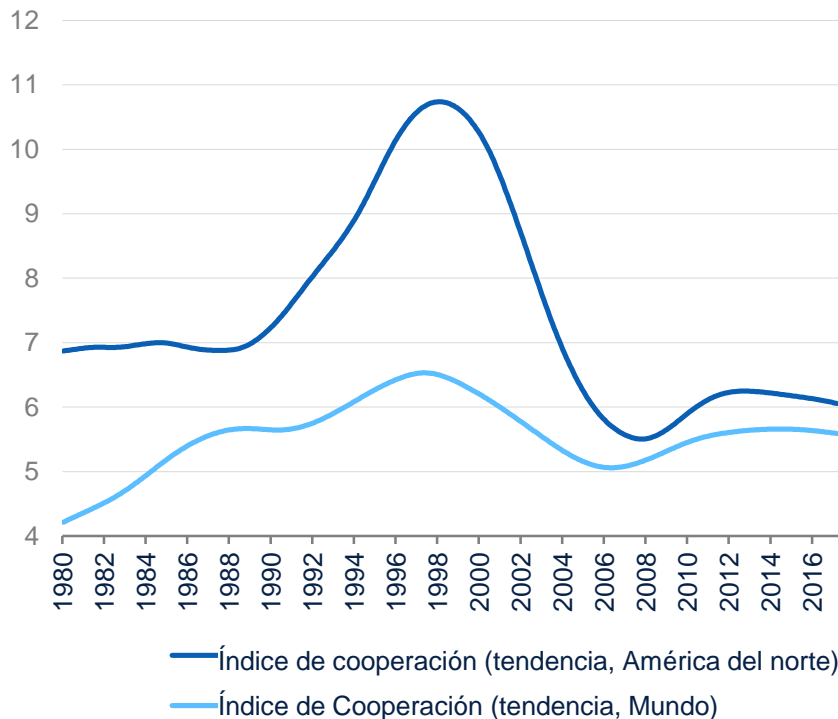


# ...así como la evolución del índice de cooperación en el tiempo de las principales potencias mundiales

El índice se define como la relación del número de eventos de cooperación y demanda.

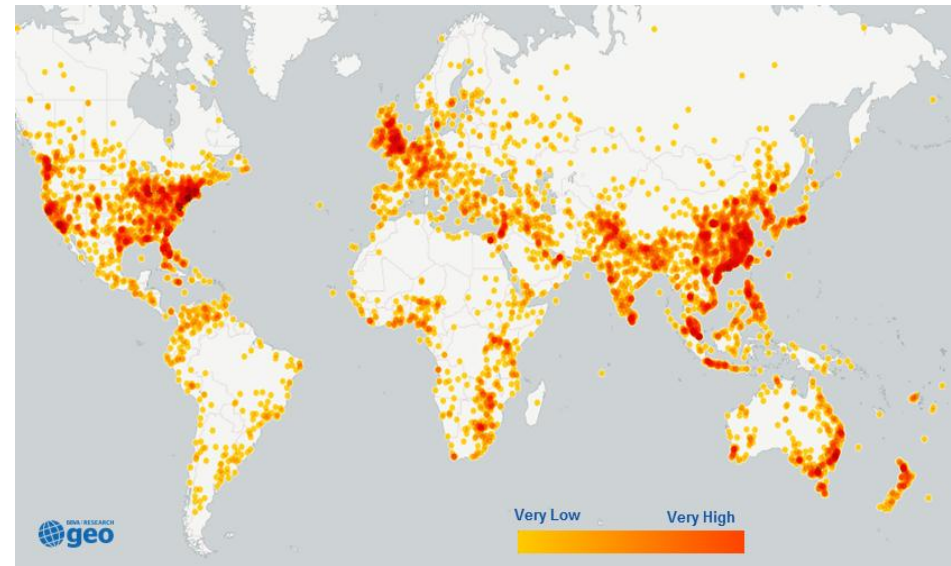
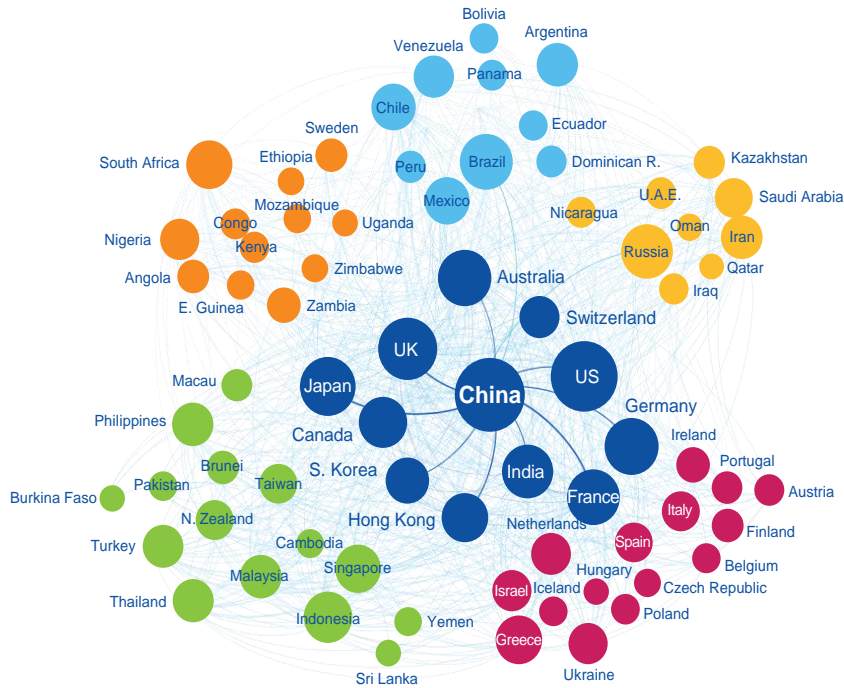
## Tendencias del índice

(Filtro HP)



# Efectos colaterales de la desaceleración de China

## Desaceleración de China: percepción de los medios y red de países



- Financial & Trade lines impaired (Main trade partners)
- Impact on commodity Exporters from Latam
- Conversation on Oil demand
- Activity Impact in Europe
- Activity Impact in EM Asia
- Impact on commodity Exporters from Africa

# ...o efectos colaterales de las sanciones comerciales impuestas a Rusia

## Red de sanciones económicas rusas

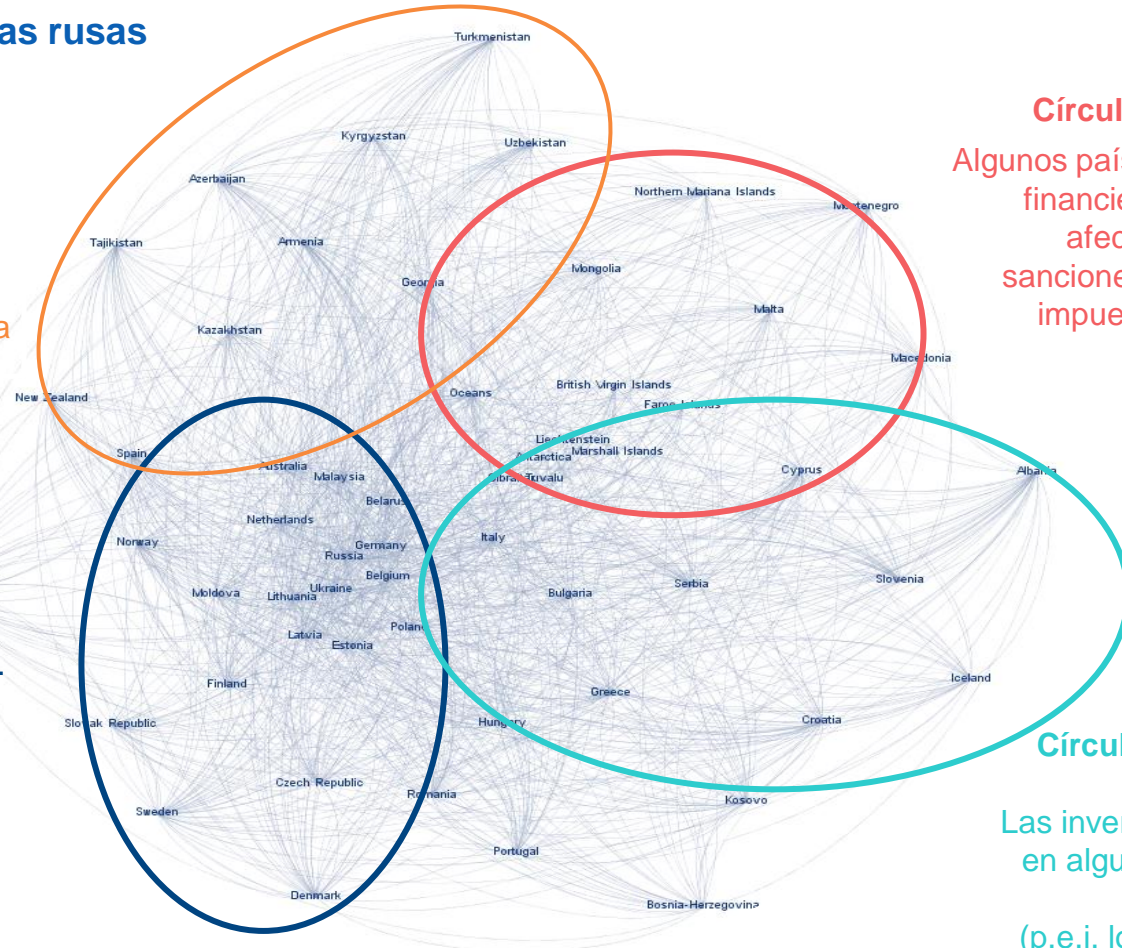
### Comercio de Asia Central

Las restricciones al intercambio de tecnología afectarán a la capacidad rusa a medio/largo plazo para extraer nueva energía, lo que podría afectar a las relaciones de Asia Central.

### Comercio con Europa central y del este

Los efectos comerciales de las sanciones comerciales impuestas a Rusia se extenderán a otros países. En particular, a socios comerciales tradicionales en el este.

La demanda externa de algunos países de Europa Central (Francia, Alemania, Italia) también se verá afectada.



### Círculo financiero

Algunos países y centros financieros se verán afectados por las sanciones financieras impuestas a Rusia.

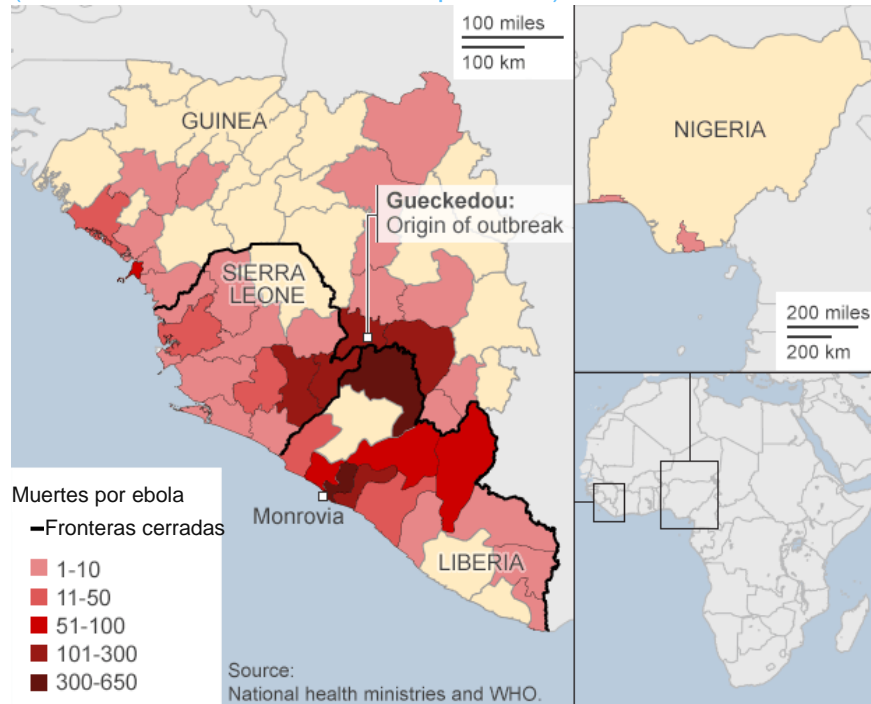
### Círculo financiero y comercial

Las inversiones rusas en algunas regiones son enormes (p.e.j. los Balcanes).

# Los controles de solidez con datos oficiales muestran una gran similitud entre las series. De los problemas de salud...

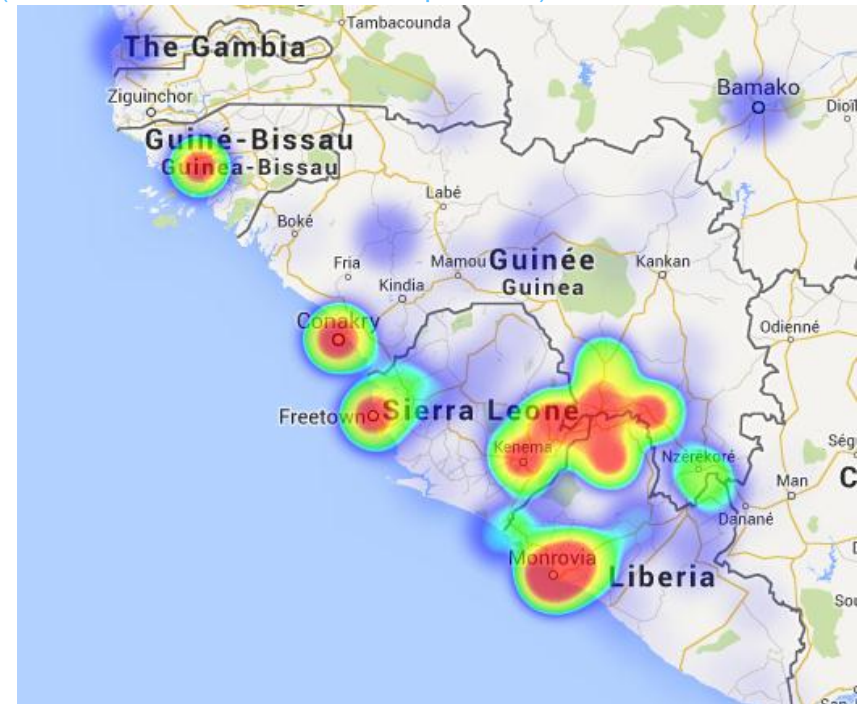
## Ébola: deudas oficiales por la OMS

(muertes hasta mediados de septiembre)



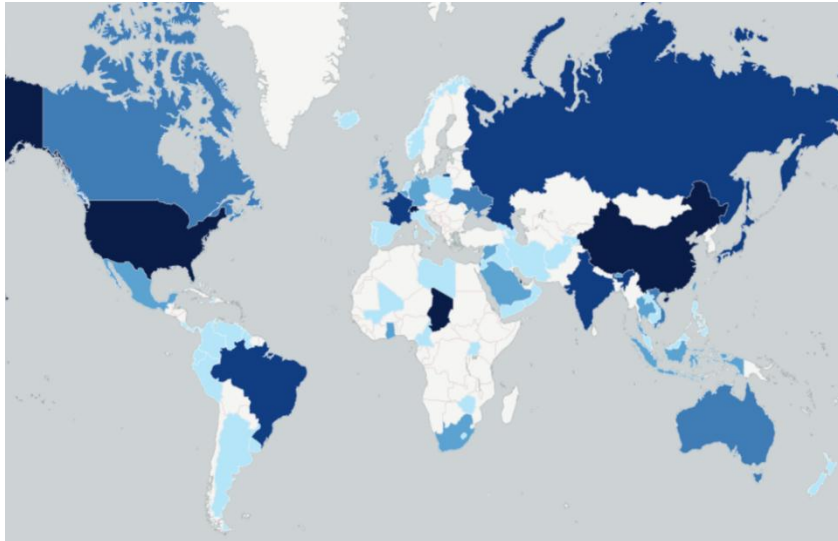
## Ébola: brote según GDELT

(muertes hasta mediados de septiembre)



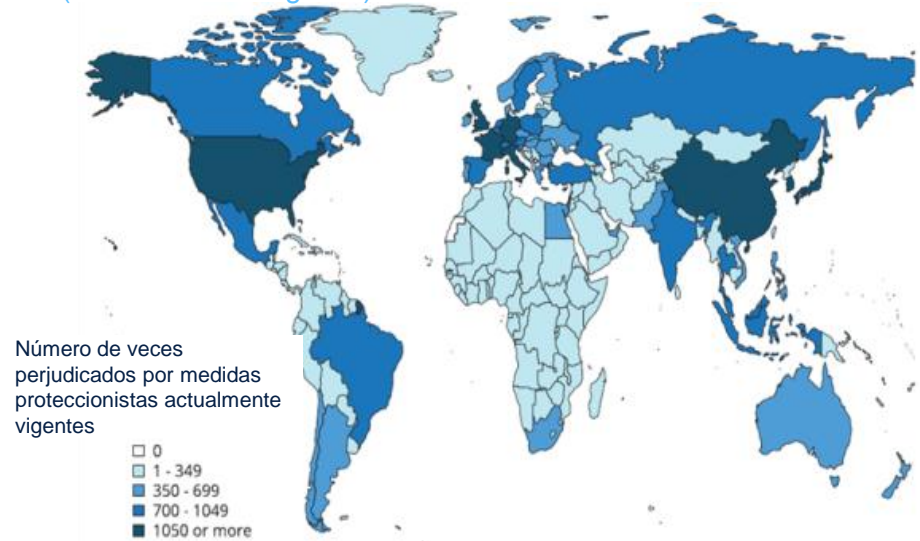
## ...a los temas relacionados con el comercio.

### Cambios en el índice de soporte del comercio mundial de BBVA Research 2008-17



### La incidencia mundial del proteccionismo 2008-2015

(alerta comercial global)



# 03

**Indicadores económicos y de riesgo a través de transacciones, búsquedas de Google y noticias internacionales**



## Bases de datos internas: trabajar con datos de BBVA agregados y anonimizados



**710 millones** de transacciones con tarjeta desde **1 millón** de puntos de venta, hechas por **53 millones** de personas, que representan **43 000 millones de euros**



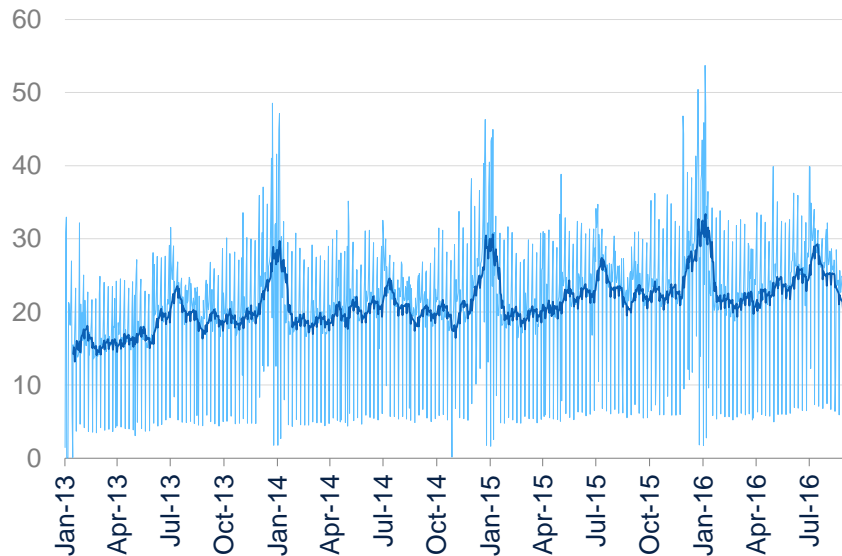
**1 500 millones** de transacciones con tarjeta desde **1,1 millones** de puntos de venta, realizadas por **88 millones** de personas, que representan **41 000 millones de euros**

# Utilizando datos de BBVA, replicamos cifras nacionales, mejorando la frecuencia...

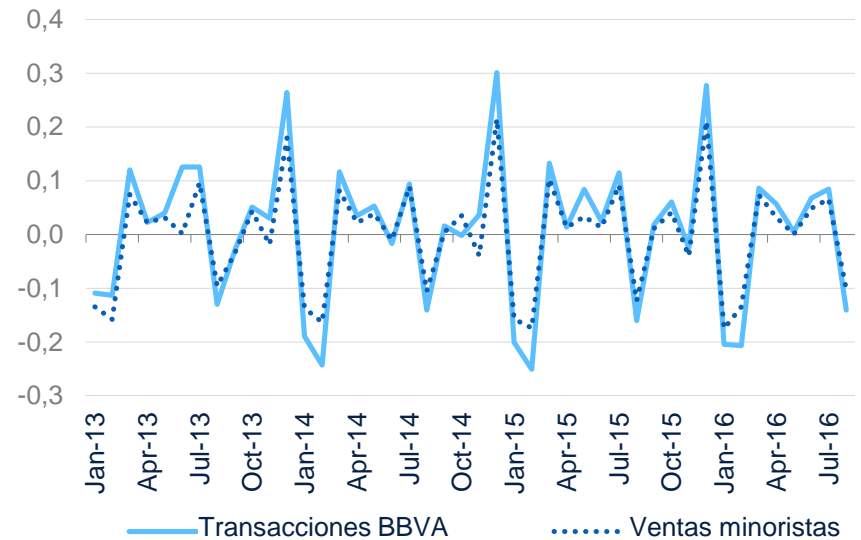
## Un indicador de actividad de "alta definición" para España (y México)

(Indicador de consumo de BBVA para la asignación óptima de los recursos y productos de BBVA)

Índice ICM-BBVA, en millones de euros y a diario



## Comparación de ventas minoristas del INE y BBVA mensualmente



## Qué significa "ALTA DEFINICIÓN (\*)" aquí:

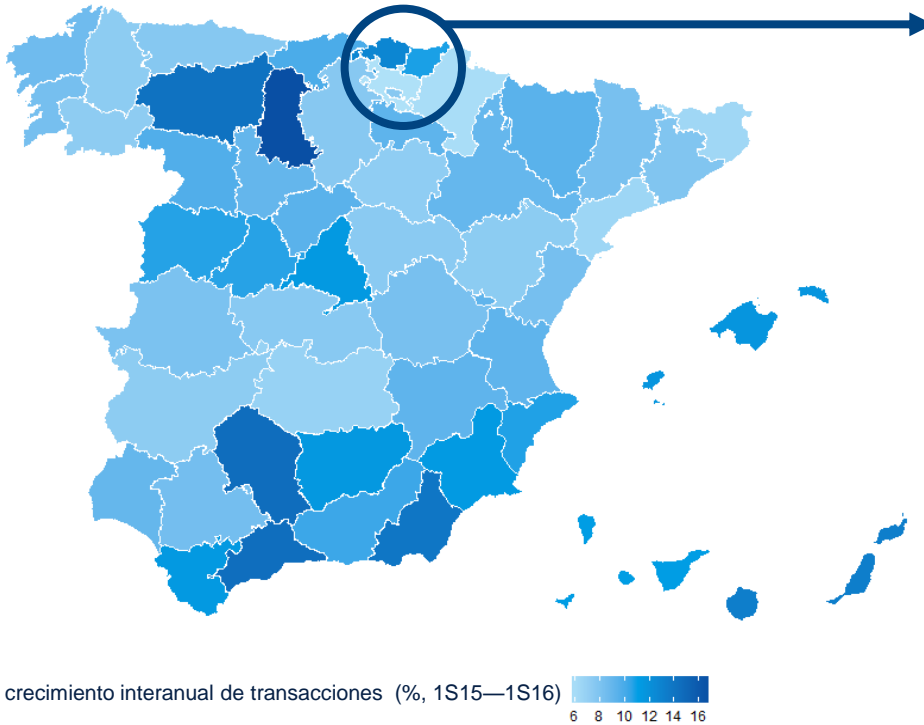
**alta granularidad:**  
dinámica hasta el nivel subnacional

**Frecuencia ultra alta:**  
dinámica hasta frecuencia submensual

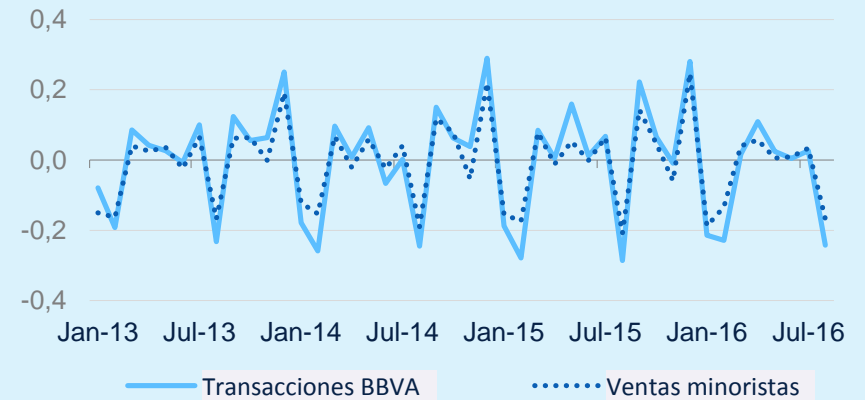
**Multidimensional:**  
características socioeconómicas más detalladas

# ...y granularidad, hasta el nivel regional

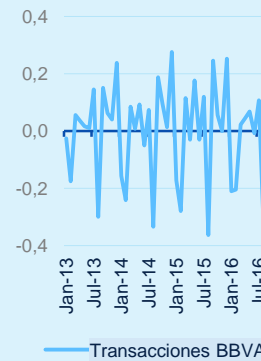
Transacciones de BBVA 1S15 frente a 1S16  
(% a/a)



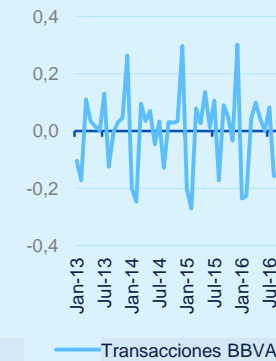
## País Vasco



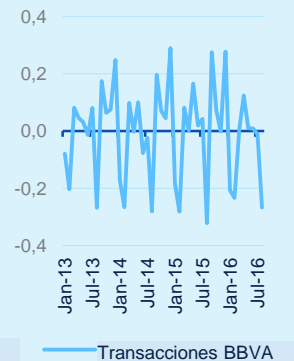
## Álava



## Guipúzcoa



## Vizcaya



## Bases de datos externas: base de datos de búsquedas de Google



- ◆ La medición de las **consultas de Google**, dado el **uso cada vez mayor de las búsquedas en internet**, tiene un gran potencial para predecir la evolución futura de la situación.
- ◆ **Las búsquedas de Google (Google Search)** ofrecen varias funciones más allá de la búsqueda de palabras y están disponibles desde **julio de 2007**.
- ◆ El **análisis** de la frecuencia de los **términos** de búsqueda puede indicar la evolución de **las tendencias económicas, sociales y de salud**.

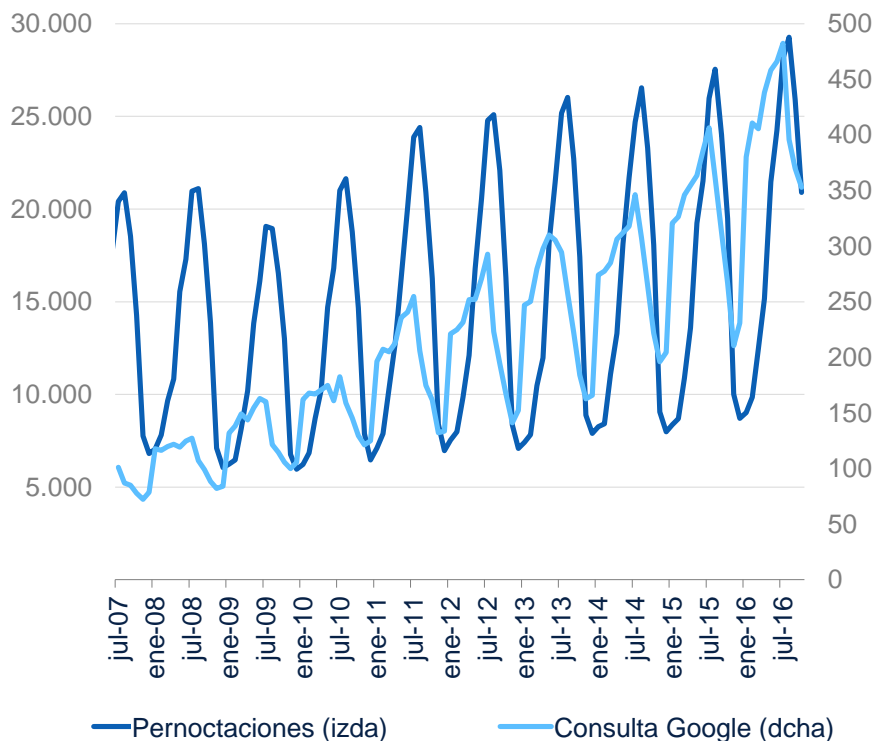
*Ejemplo:* una base de datos con información agregada sobre las **consultas de Google relacionadas con España como destino turístico** desarrollada junto con Google. Las consultas relacionadas con el turismo en Google siguen el mismo patrón estacional que las estadísticas de turismo, anticipándose con uno o dos meses.



# La similitud en la dinámica de las estadísticas oficiales y las consultas de Google nos permiten hacer pronósticos para el turismo español

## Pernoctaciones de turistas no residentes en hoteles y tendencias de búsqueda en Google

(Pernoctaciones en miles, índice de búsquedas = 100, julio de 2007)

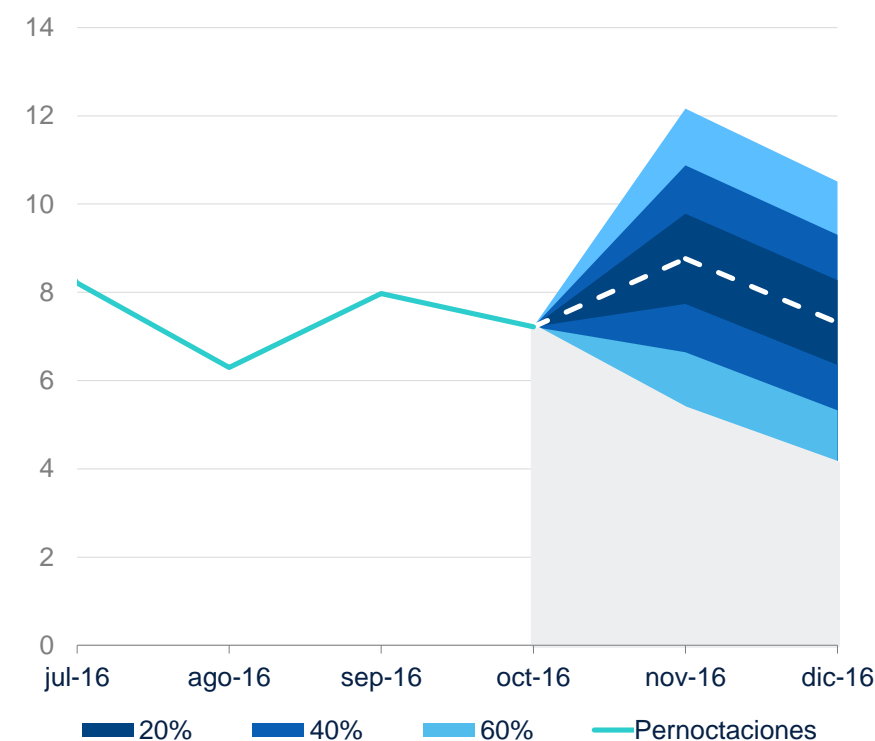


(Más información en el siguiente [enlace](#))

Fuente: BBVA Research, INE y Google

## Pernoctaciones de no residentes en hoteles y previsiones

(% a/a, último pronóstico a 30 de noviembre de 2016)



# El sentimiento de las noticias nos permite elaborar un índice compuesto

## Macroeconomic Sentiment Index (MSI)

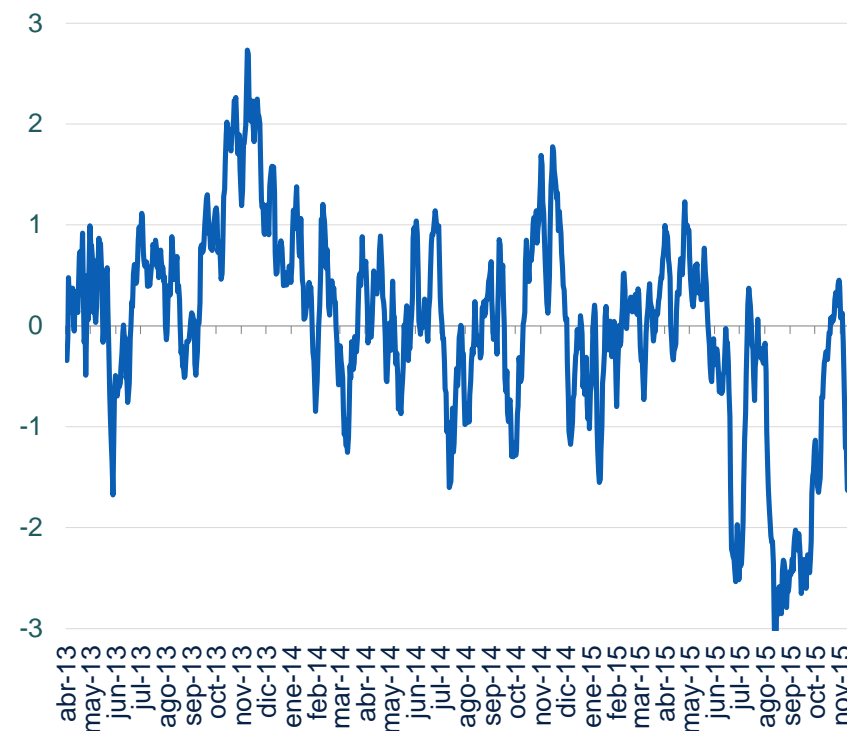
<b>Fiscal Policy Sentiment Index (FPSI)</b>	<b>Monetary Policy Sentiment Index (MPSI)</b>	<b>Political Sentiment Index (PSI)</b>	<b>Global Policy Sentiment Index (GPSI)</b>
---------------------------------------------	-----------------------------------------------	----------------------------------------	---------------------------------------------

### 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_Fiscal Risks, 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
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

## Índice de sentimiento macroeconómico para Turquía

(Evolución del "tono" de los principales temas seguidos)

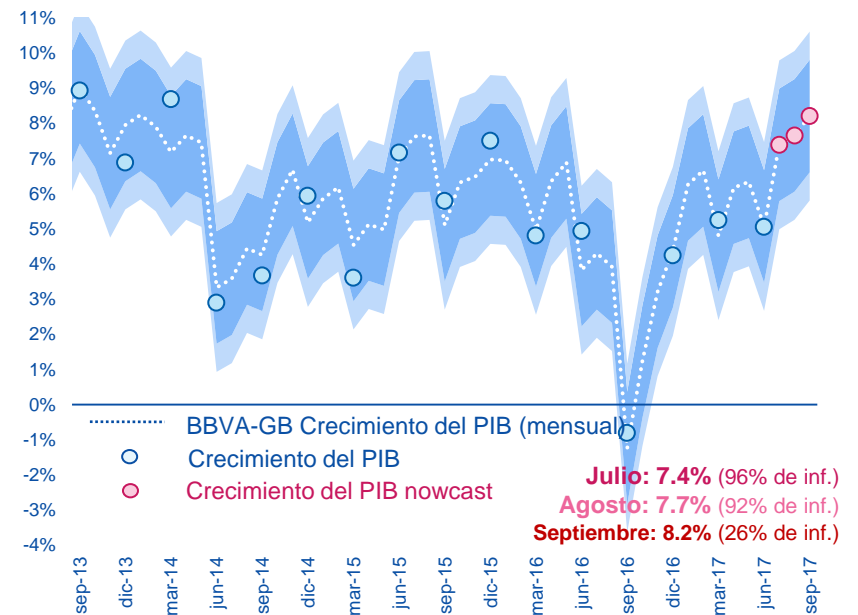


# Podemos usarlo para mejorar nuestros modelos mensuales de PBI... aprovechando las noticias en tiempo real

## Modelo de factor dinámico para el PBI turco Pseudo de errores RMS de muestra

	Mensual	Trimestral	Año
PIB MFD mensual	0,085	0,256	1,024
<b>PIB MFD mensual + Índice MU</b>	<b>0,046</b>	<b>0,139</b>	<b>0,558</b>
PIB MFD mensual + Ponderado MU	0,063	0,190	0,569
PIB MFD mensual + P. Monetaria MU	0,046	0,139	0,556
PIB MFD mensual + Política MU	0,046	0,139	0,556
<b>PIB MFD mensual + P. Fiscal MU</b>	<b>0,046</b>	<b>0,138</b>	<b>0,550</b>
PIB MFD mensual + I. Global MU	0,063	0,188	0,563

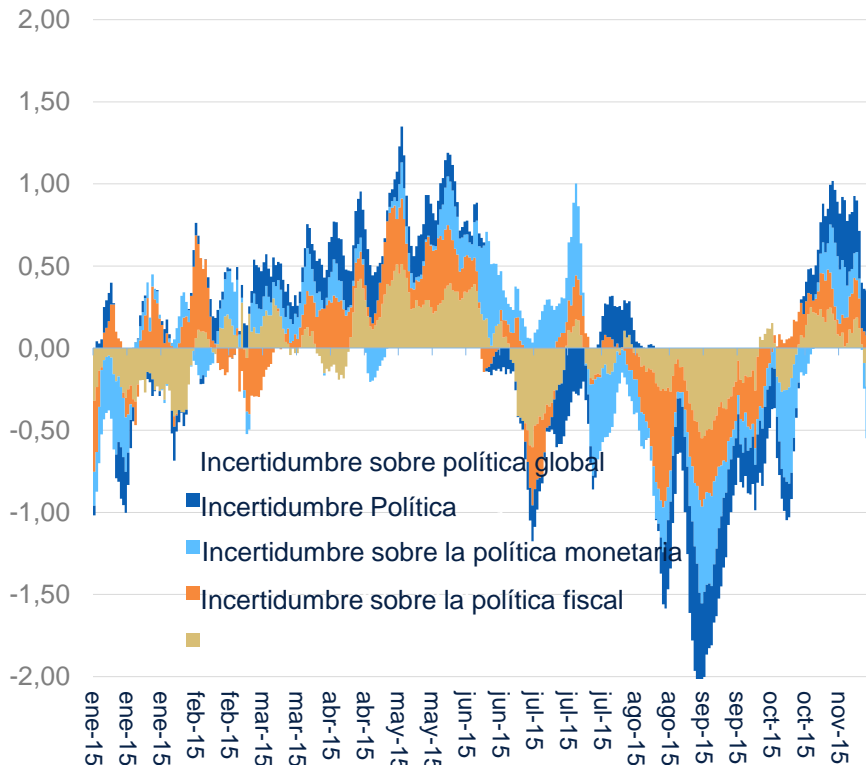
## Indicador mensual de crecimiento del PIB turco y previsiones a corto plazo (% de variación interanual)



# Podemos verificar la evolución a lo largo del tiempo...y cómo los activos financieros responden a diferentes variables de sentimiento...

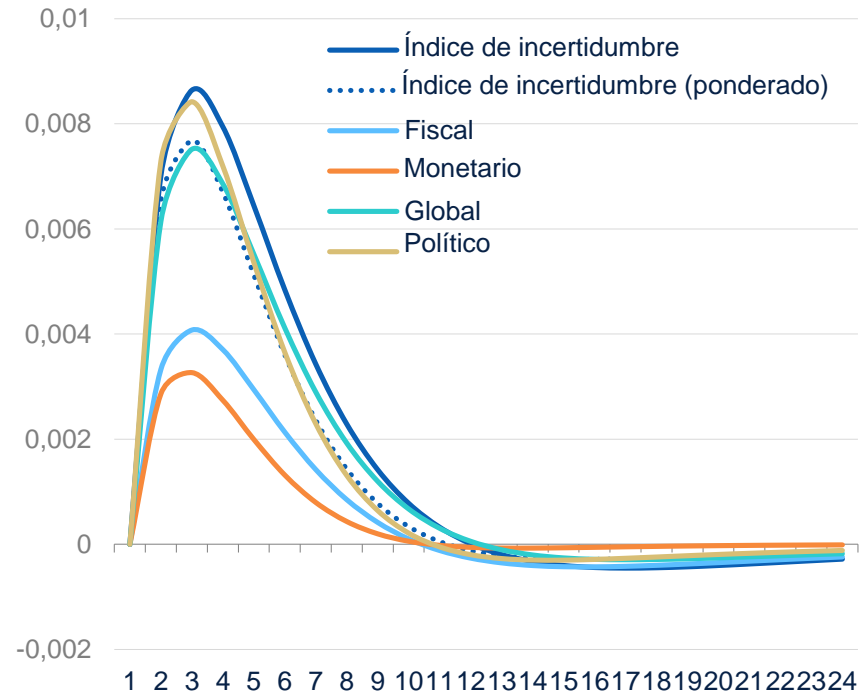
## Turquía: Incertidumbre macroeconómica en 2015

(en desviaciones estándar)



## Turquía: Impulso respuesta del tipo de cambio a los shocks en el sentimiento

(en desviaciones estándar)



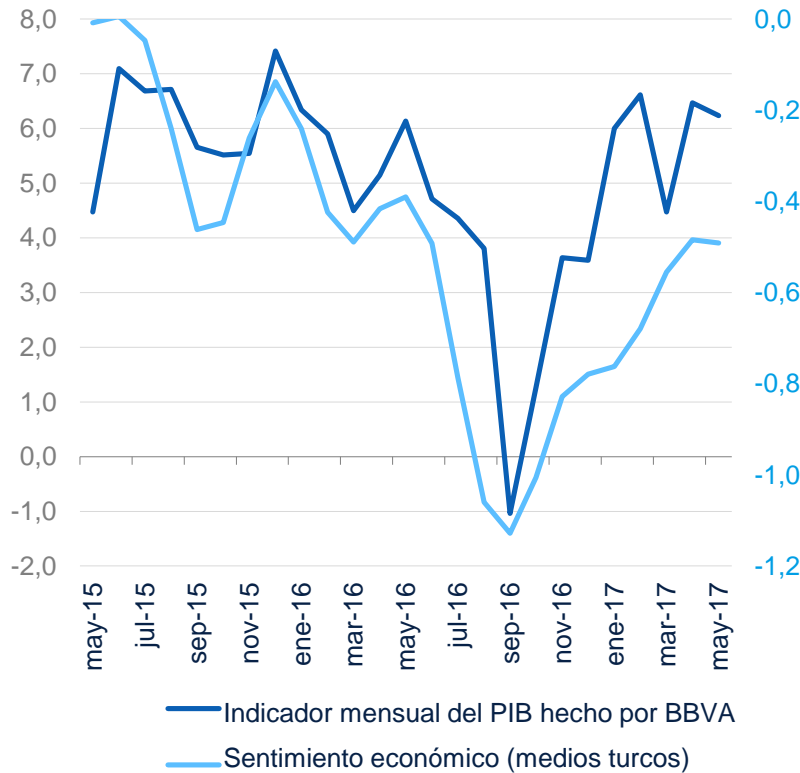
• La respuesta al impulso corresponde a un modelo VAR bayesiano con PIB global, inflación, tasa de interés, PIB local mensual, incertidumbre y tipo de cambio. Se estimó a través del muestreo de Gibbs debido a la restricción de datos.

Fuente: BBVA Research

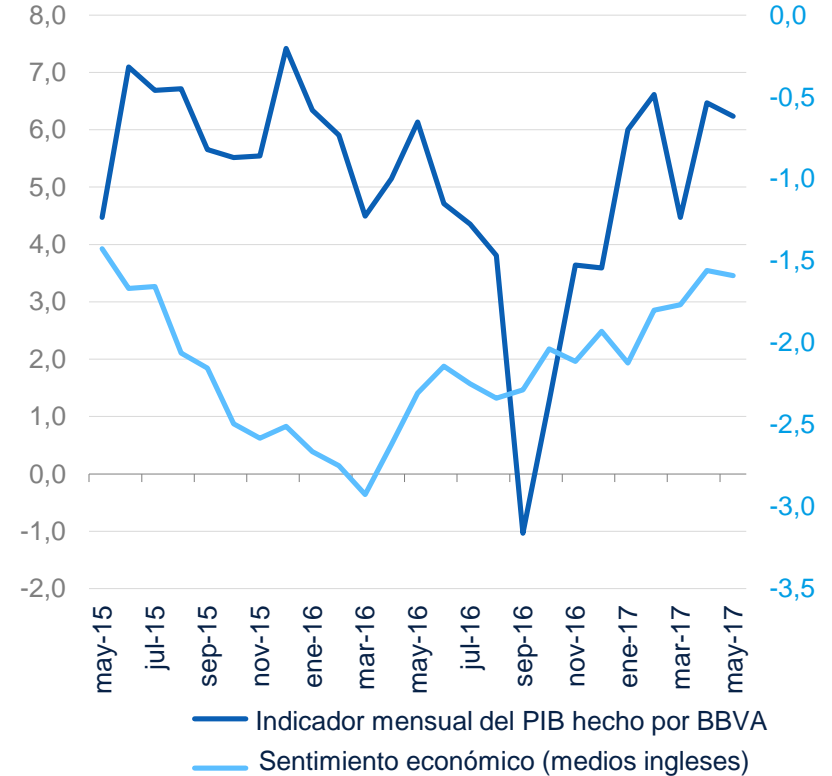


# ...o para analizar la importancia de las narrativas y el sesgo del lenguaje: y sí, importan...

**PIB de Turquía y sentimiento económico**  
(%a/a y sentimiento económico escrito en turco)

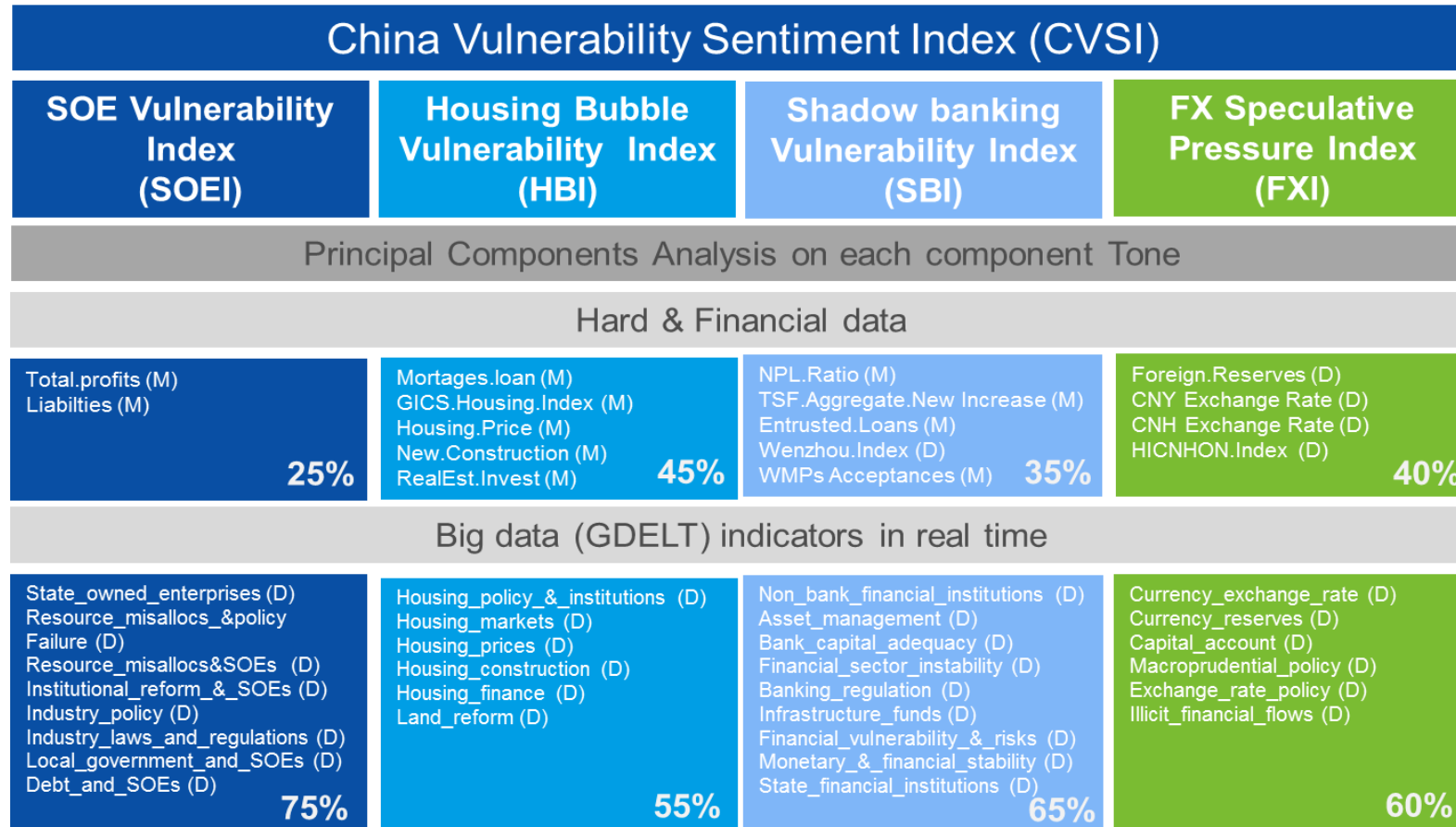


**PIB de Turquía y sentimiento económico**  
(%a/a y sentimiento económico escrito en inglés)



# No se trata solo del sentimiento económico... sino también de complementar los datos oficiales...

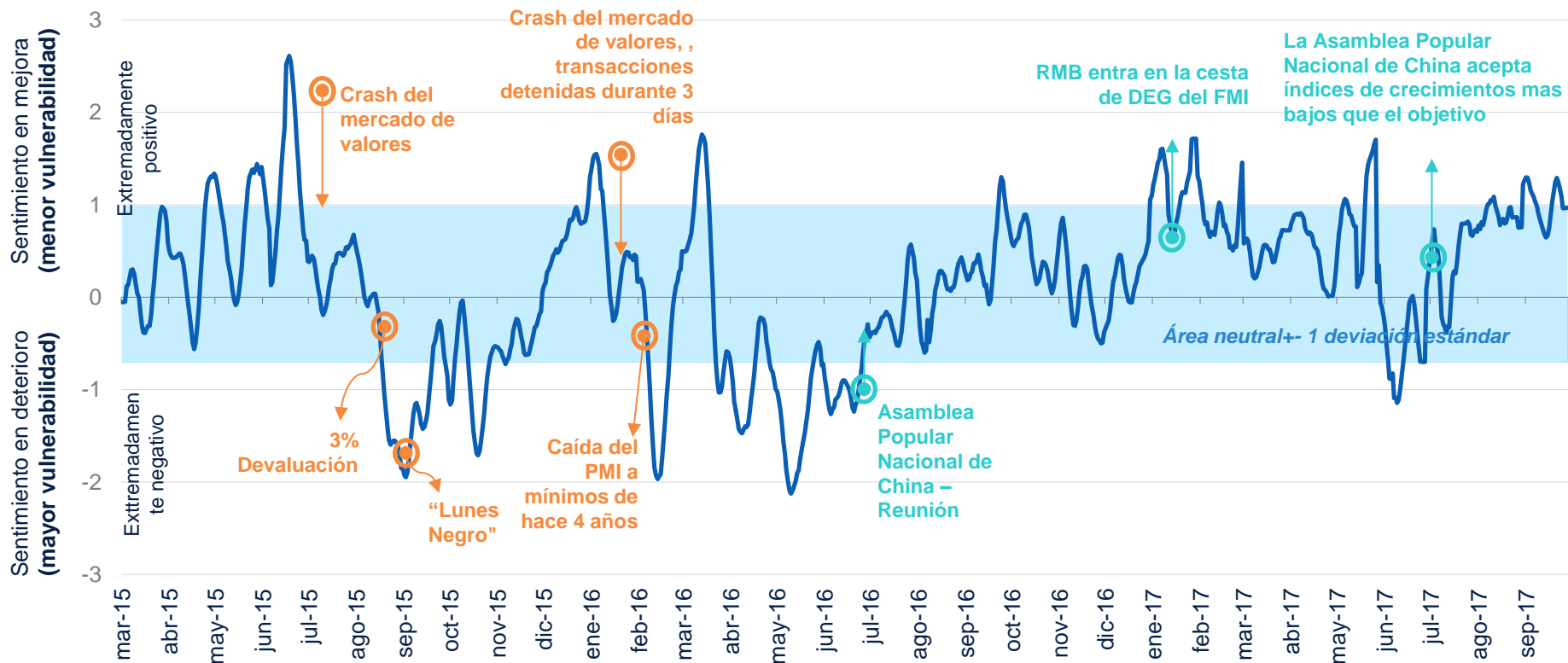
## Índice de Sentimiento de Vulnerabilidad Chino (CVSI): componentes y evolución



# ...para rastrear los riesgos en tiempo real...

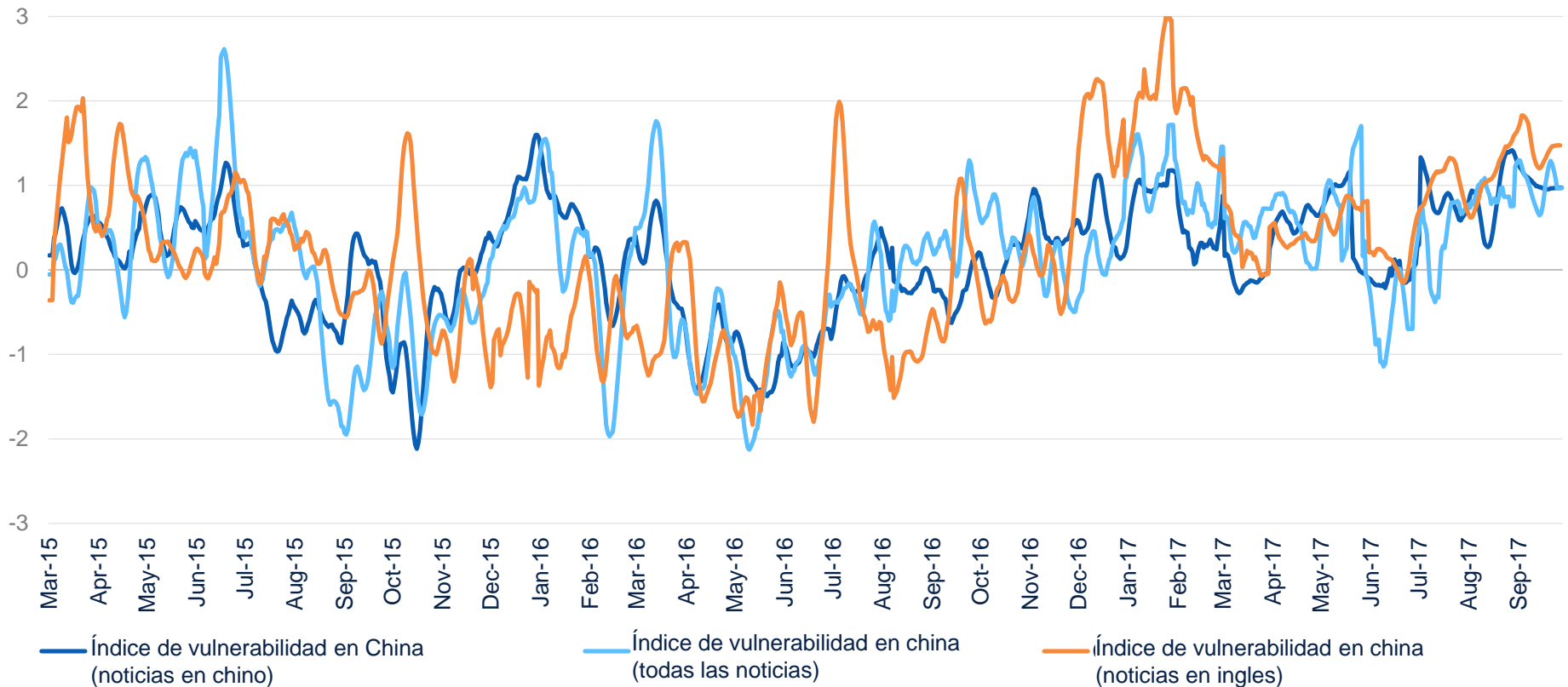
## Índice de Sentimiento de Vulnerabilidad Chino (CVSI)

(Evolución del "tono" de los principales temas seguidos sobre la vulnerabilidad en China. Los valores más bajos indican un deterioro del sentimiento y una mayor vulnerabilidad)



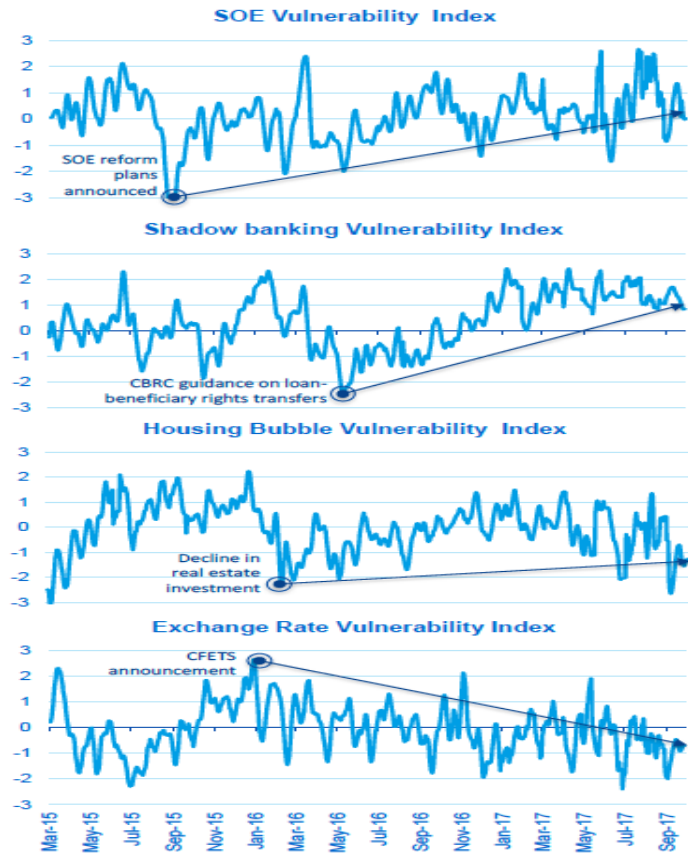
# ...desenmarañando los efectos del lenguaje de los medios...

**Índice de Sentimiento de Vulnerabilidad en China por idioma de los medios: total, chino e inglés**  
 (Evolución del "tono" de los principales temas seguidos sobre la vulnerabilidad en China. Los valores más bajos indican un deterioro del sentimiento y una mayor vulnerabilidad)

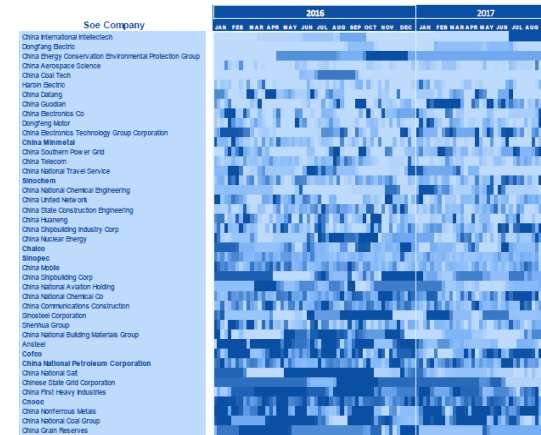


# ...y analizado los riesgos en un alto grado de granularidad

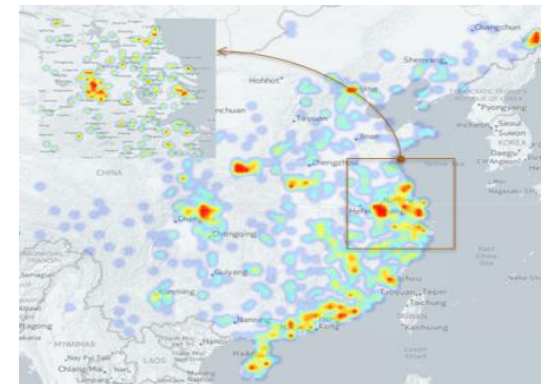
## Componentes del Índice de Sentimiento de Vulnerabilidad Chino (CVSI)



## Mapa de empresas estatales de China (sentimiento sobre empresas estatales)

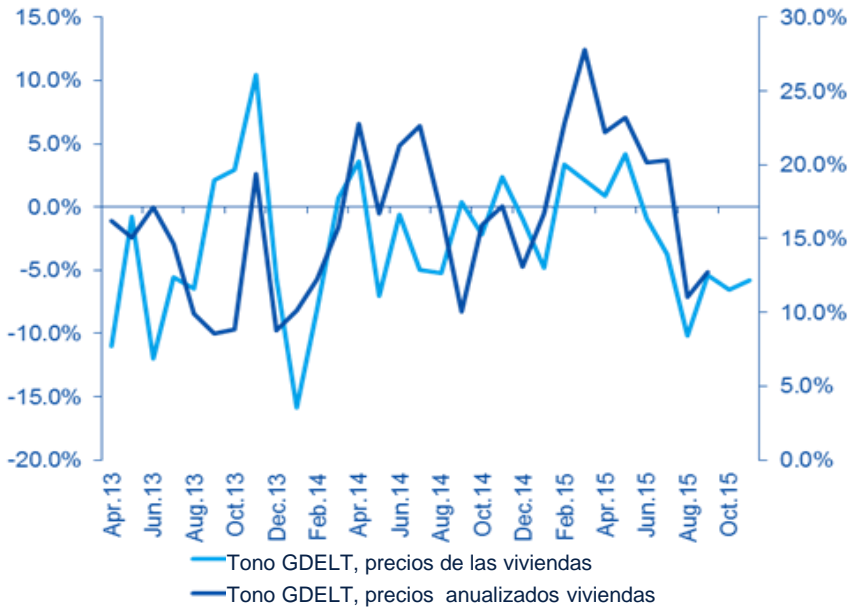


## Análisis geográfico Precios de la vivienda (sentimiento sobre los precios de la vivienda)

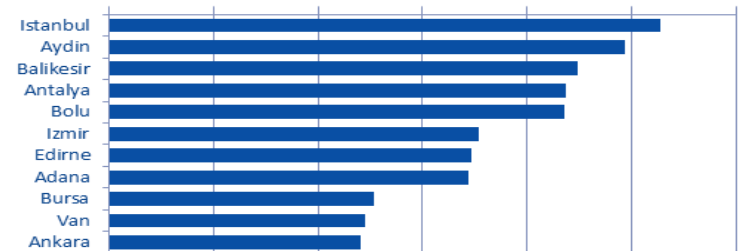
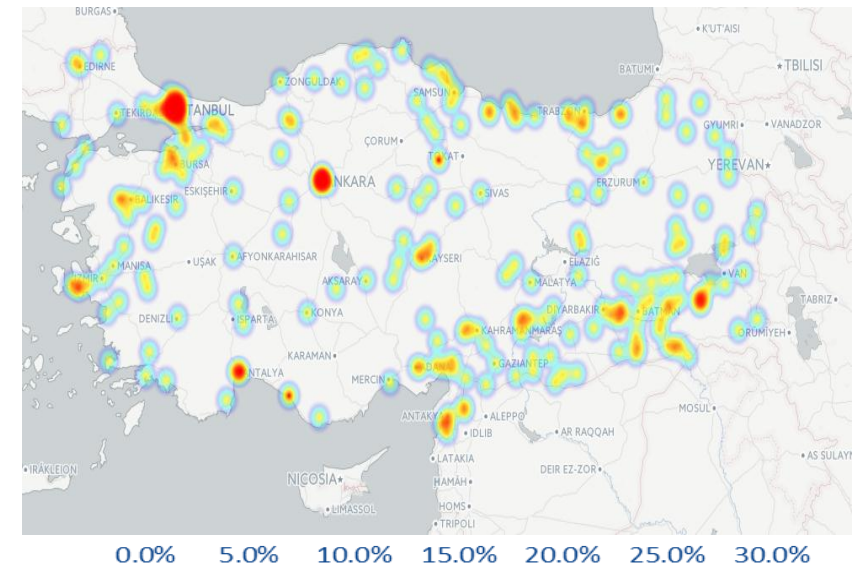


# Las previsiones a corto plazo (nowcast) de los precios de la vivienda también son un aspecto prometedor del Big Data

**Precios de la vivienda y tono de los precios de la vivienda en Turquía**  
(Azul oscuro: tono más negativo)



**Distribución geográfica del tono de los precios de la vivienda 2015**  
(Azul oscuro: tono más negativo)



# 04

## Minería de textos y análisis de opiniones

# Bases de datos externas: técnicas de web scrapping y NPL

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

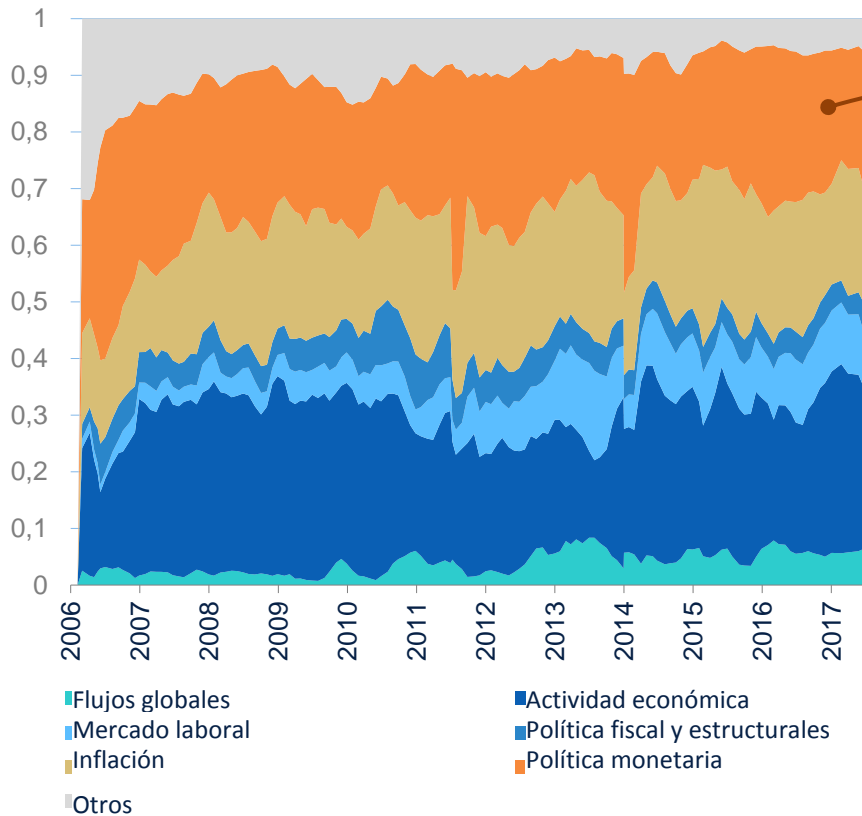




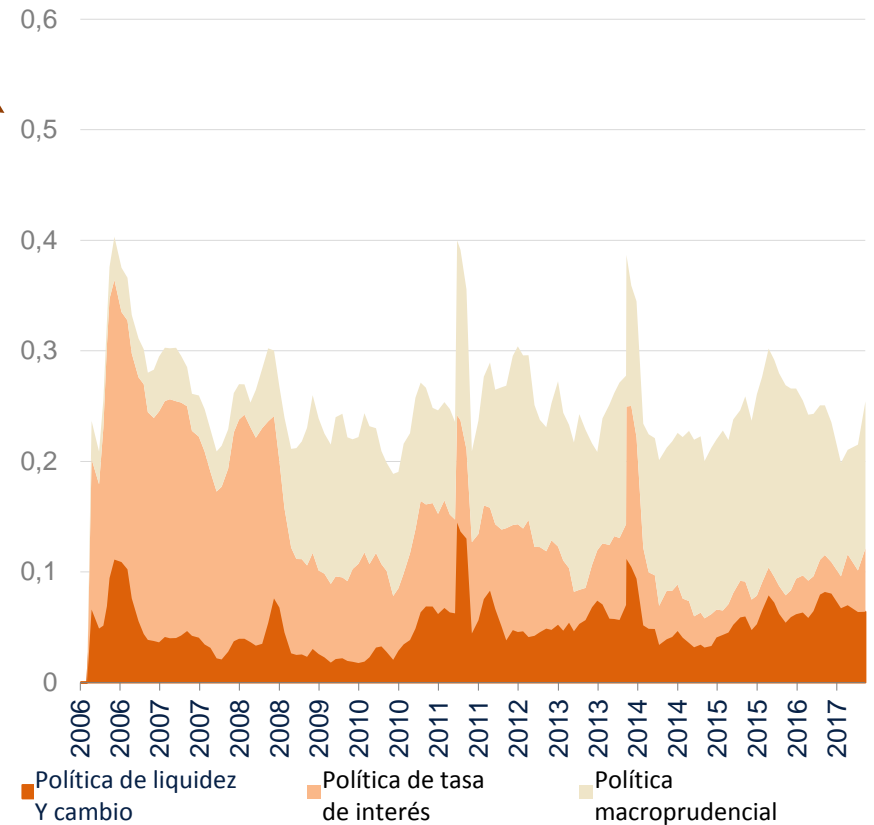



# ...y podemos verificar "de lo que habla el Banco Central"...

**Banco Central de Turquía: Evolución de los temas**

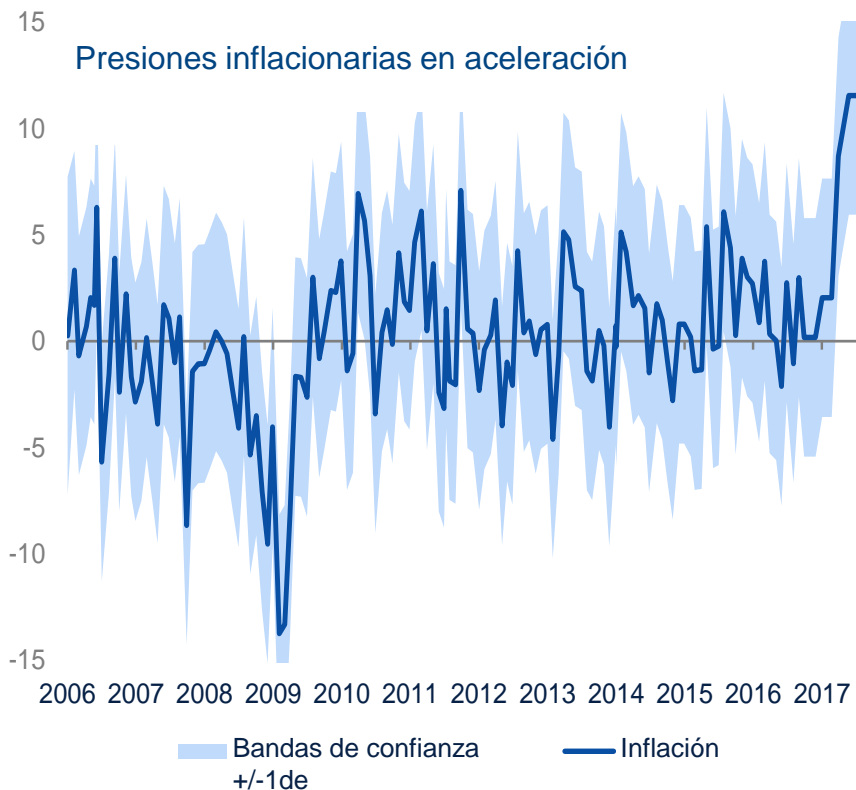


**Distribución de los temas de política monetaria (% del total)**

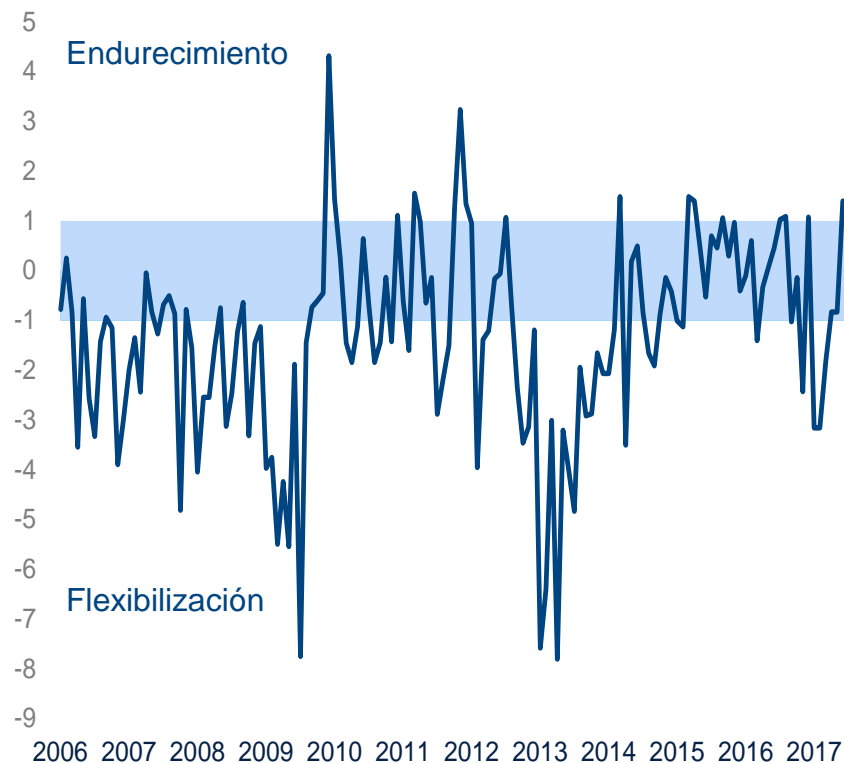


# ...así como el sentimiento del tema y la postura de los informes de CB...

## Sentimiento del Banco Central sobre la inflación (Técnicas estandarizadas de Big Data LDA aplicadas a las actas y declaraciones)

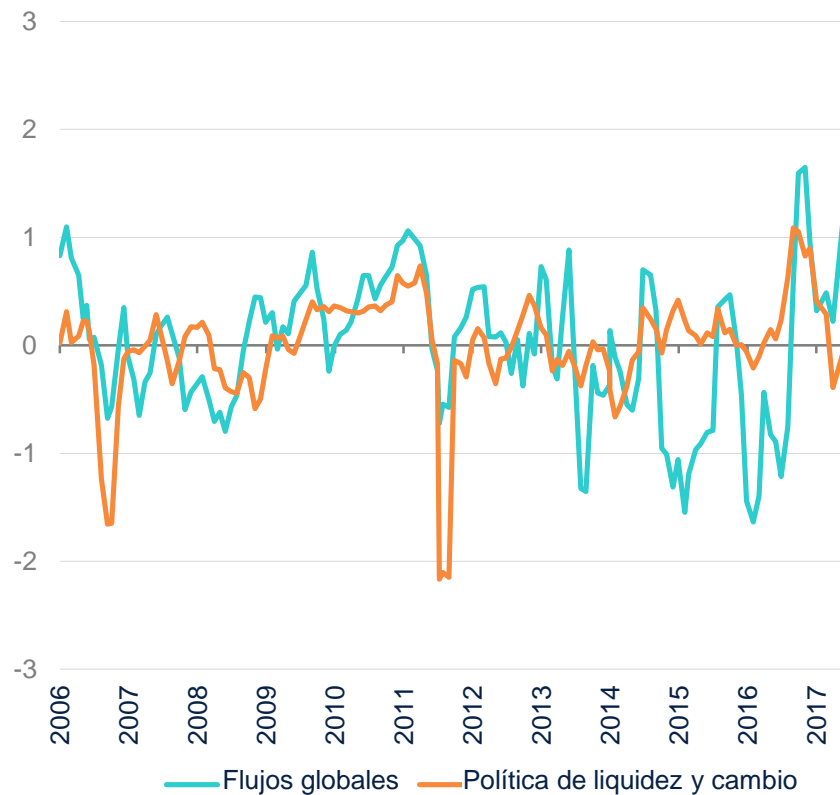
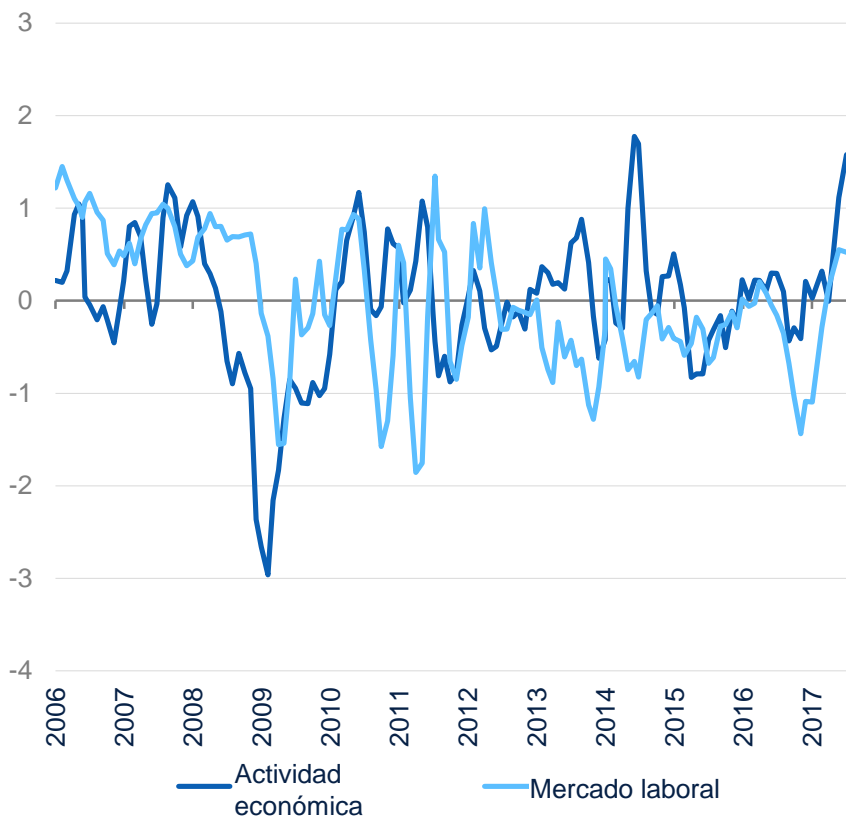


## Sentimiento de la política monetaria (Estandarizado, estimado mediante técnicas de Big Data LDA a partir de actas y declaraciones)



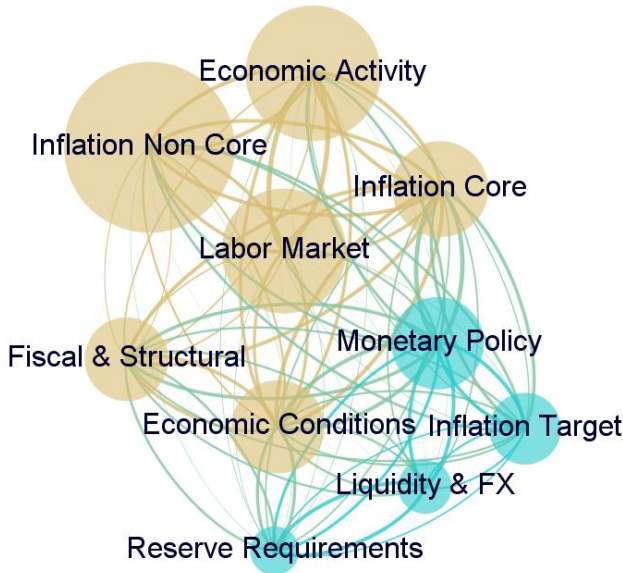
# Lo que cambia con el tiempo...según las técnicas de minería de textos y aprendizaje automático...

Evolución del sentimiento de los temas en los informes de BC en 2006-17.

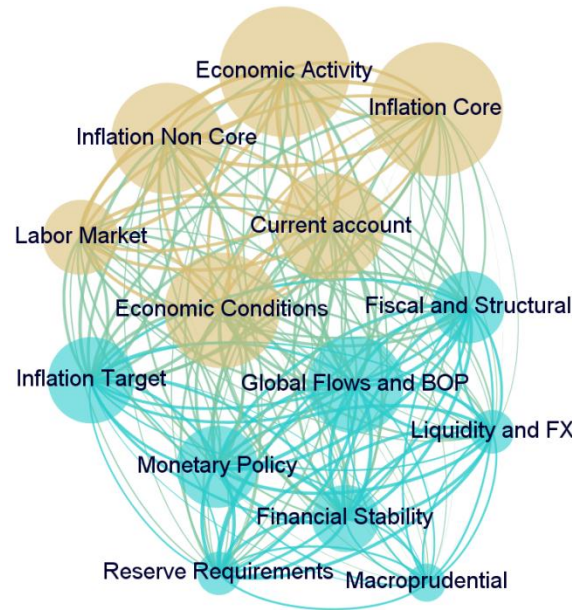


# ...así como las relaciones entre los temas y su evolución en el tiempo mediante el análisis de redes de temas

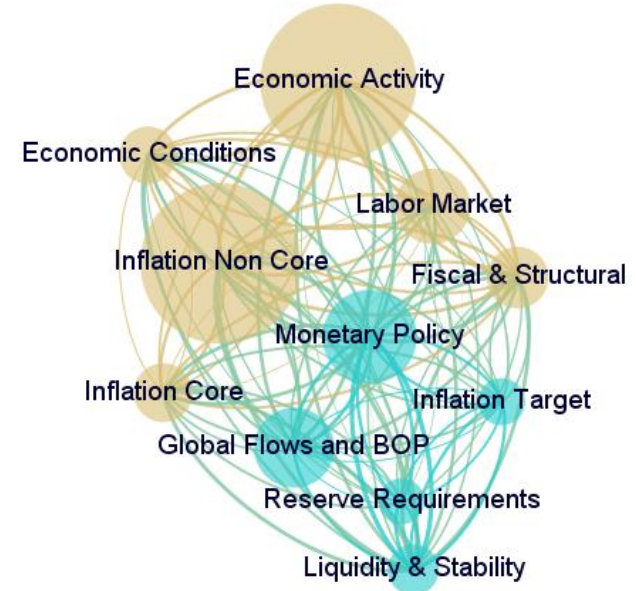
**Red de temas 2006-09:  
la meta de inflación**



**Red de temas 2010-15:  
el período de la crisis financiera mundial**



**Red de temas 2016-17:  
en busca de la estabilidad de precios**



La red de los temas estimados y correlacionados usando STM. Los nodos en el gráfico representan los temas identificados. El tamaño del nodo es proporcional al número de palabras en el corpus dedicado a cada tema (peso). El color del nodo indica clústers que usan un algoritmo de detección de comunidad llamado modularidad desarrollado por Blondel et al (2008). Los temas cuyo etiquetado se desconoce se eliminan del gráfico por razones de claridad visual. Los bordes representan palabras que son comunes a los temas que conectan (coocurrencia de palabras entre los temas). El ancho del borde es proporcional a la fuerza de esta coocurrencia entre los temas.

# ANEXO



## Minería de texto y NPL: preprocesamiento y transformación

- ◆ Los documentos se definen como párrafos.
- ◆ Se excluyen los documentos con menos de 200 caracteres (títulos, secciones de contenido...)
- ◆ A continuación, se derivan las palabras (reducir una palabra a su raíz semántica) para generar tokens.
- ◆ La selección de características se realiza en los tokens: las palabras vacías comunes y las palabras con una longitud de 3 o menos caracteres se eliminan y las palabras restantes se derivan. Los tokens se filtran en función de un índice de frecuencia de término – frecuencia inversa de documento (tf.idf) (Manning y Schutze 1999); se eliminan las palabras del cuantil más bajo. Este esquema de indexación es una combinación de un índice de frecuencia de término (tf) y un índice de frecuencia de documento (df). tf es solo el recuento de una palabra dada en un documento, la media tf se usa para construir el índice final. df es la cantidad de documentos que contienen una palabra dada. Entonces, el tf.idf usado para filtrar palabras es:

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

- ◆ donde i indexa términos y j documentos. Este índice otorga un gran peso a las palabras frecuentes a través del componente tf, pero si una palabra es muy frecuente a través del corpus; su peso se reduce a través del componente idf. El objetivo de este procedimiento de filtrado es eliminar palabras muy poco frecuentes y muy frecuentes para eliminar palabras con bajo contenido semántico.



## Algoritmos de aprendizaje automático en texto: LDA, STM y DTM

- ◆ La **asignación latente de Dirichlet** (LDA) (Blei, Ng y Jordan 2003) es un modelo bayesiano con una distribución previa con respecto a las probabilidades de mezcla específicas del documento, donde el recuento de términos dentro de los documentos se distribuye de manera independiente e idéntica dada una distribución previa de Dirichlet.
- ◆ Para introducir dependencias temporales en el proceso de generación de datos, utilizamos el **modelo de tema dinámico** (DTM), una particularización de los **Modelos de tema estructural** (STM) donde cada período de tiempo tiene un modelo de tema separado y los períodos de tiempo están vinculados a través de parámetros que evolucionan con fluidez.
- ◆ STM (Roberts et al. 2016) introduce de manera explícita covariables en un modelo de temas que nos permite valorar el impacto de las covariantes a nivel de documento sobre el contenido del tema y la prevalencia como parte del modelo del tema en sí mismo.
- ◆ El proceso para generar palabras individuales es el mismo que para el LDA simple. Sin embargo, ambos objetos pueden depender de conjuntos potencialmente diferentes de covariables a nivel de documento: prevalencia de temas (cada documento tiene atributos  $P$  que pueden afectar a la probabilidad de discutir el tema  $k$ ) y el contenido temático (cada documento tiene un atributo categórico de nivel  $A$  que afecta a la probabilidad de discutir el término  $v$  en general y de analizarlo dentro del tema  $k$ ). La generación de los términos  $k$  y  $d$  se realiza mediante regresión logística multinomial

## Análisis del sentimiento en el texto: enfoque léxico

- ◆ Empleamos los métodos Lexicon utilizando el **diccionario Loughran-McDonald** (Loughran McDonald 2009), un diccionario creado específicamente para analizar textos financieros y el **diccionario de la FED para la estabilidad financiera** (Correa et al, 2017)
- ◆ Usando las palabras negativas y positivas de este diccionario, el "tono" promedio de un documento dado se calcula del siguiente modo:

$$\text{Tono medio} = 100 * \frac{\sum \text{Palabras positivas} - \sum \text{Palabras negativas}}{\sum \text{Palabras totales}}$$

- ◆ la puntuación oscila entre -100 (extremadamente negativo) y +100 (extremadamente positivo), pero los valores comunes oscilan entre -10 y +10, donde 0 indica neutral.
- ◆ Para crear los **índices de sentimiento** finales, usamos la mezcla de temas que **combina los métodos de diccionario** con el resultado de **LDA** para recuentos de palabras de peso por tema, siguiendo el enfoque propuesto por Hansen y McMahon (2015). Esto permite generar diferentes medidas de sentimiento a partir de un conjunto de texto, y enfocar ese sentimiento en los temas de interés.

## Metodología de impacto causal

- ◆ Para medir el impacto de los ataques sobre el rendimiento del comercio en la ciudad de Barcelona se ha utilizado un modelo bayesiano de series temporales ([aquí el documento de referencia](#)). Este modelo se basa en la comparación del comportamiento observado en una serie temporal objetivo, a partir de la fecha del evento analizado, con una predicción de los valores esperados de que no se haya producido. Para crear esta serie contrafáctica, usamos un conjunto de series de control no afectadas por el evento
- ◆ En este caso particular, la serie temporal utilizada corresponde al gasto diario con tarjeta de crédito en el comercio físico. El período cubierto por la serie va desde el 1 de enero de 2015 hasta el 24 de septiembre de 2017, fijando la fecha del evento el 17 de agosto de 2017. La serie objetivo es el gasto registrado en la ciudad de Barcelona y la serie de control corresponde al resto de municipios españoles con mayor correlación con Barcelona en el período anterior
- ◆ Por lo tanto, la predicción contrafáctica se obtiene mediante un proceso de inferencia bayesiana en el cual cada uno de los componentes de las series temporales objetivas (tendencias, estacionalidad, ciclos...) se aproxima utilizando el conjunto de series de control. Una vez hecho esto, se combinan para obtener las probabilidades a priori de la serie objetivo
- ◆ La metodología utiliza el método de la cadena Monte Carlo Markov para simular distribuciones posteriores. Esto permite no solo generar un valor esperado para cada uno de los días posteriores al evento, sino también permitir intervalos de confianza para determinar si las diferencias entre las series observadas y pronosticadas (crecimiento y decremento) podrían haber ocurrido incluso si el evento no se produce o si estadísticamente no se justifican sin el evento. En este análisis, se ha considerado estadísticamente demostrado que una diferencia se debe al ataque cuando su valor se encuentra en el 1 % final de la distribución de probabilidad calculada

# Big Data y Big Models en BBVA Research

Día de estadísticas del BCE

Jorge Sicilia, Álvaro Ortiz y Tomasa Rodrigo

Octubre de 2017