Propuesta de índice de riesgo

Problema

•El **Banco de México** entre sus diferentes actividades, **por medio de encuestas a empresarios** (**y expertos** en el tema), **recopila de manera trimestral información textual a nivel regional** acerca de **sus expectativas sobre la economía mexicana**.

•Ante esto, se tiene como **objetivo principal** la creación de **indicadores que reflejen el impulso o la limitación de la actividad económica** tanto **del país** como **a nivel regional**, basados en datos de texto.

Problema

•Este problema **ya ha sido analizado anteriormente**, extrayendo el sentimiento de cada texto **mediante un Lexicón** para posteriormente crear los indicadores (Miranda 2020), sin embargo, el enfoque que se pretende **en el trabajo actual** es diferente, ya que **consiste en la identificación de tópicos relevantes por métodos no supervisados de aprendizaje máquina** para la generación de los índices.

Métodos para clusters de tópicos

Para encontrar clusters de tópicos, entendiéndolo como grupos de palabras semánticamente relacionadas, hay varios métodos en la literatura, como lo son: **métodos probabilísticos** (LDA), métodos que utilizan bolsa de palabras (BOW o TF-IDF) llamados **métodos de factorización de matrices**, entre otros

Métodos para clusters de tópicos

En este trabajo se emplea un método diferente a los anteriores, que consiste en la **obtención de la representación vectorial de los textos** (embeddings) usando **FastText** ya que tiene la **ventaja** de que se pueden obtener las **representaciones vectoriales de palabras fuera del vocabulario** y posteriormente **clusterizar asociaciones semánticas mediante Fuzzy KMeans**.

Adicionalmente, **la metodología propuesta** en este trabajo tiene la **ventaja** de que **permite rápidamente obtener resultados para todo el periodo de tiempo** con el que se tienen datos (1T11-4T20) **y también para nuevos textos** que se agreguen en el futuro.

Datos

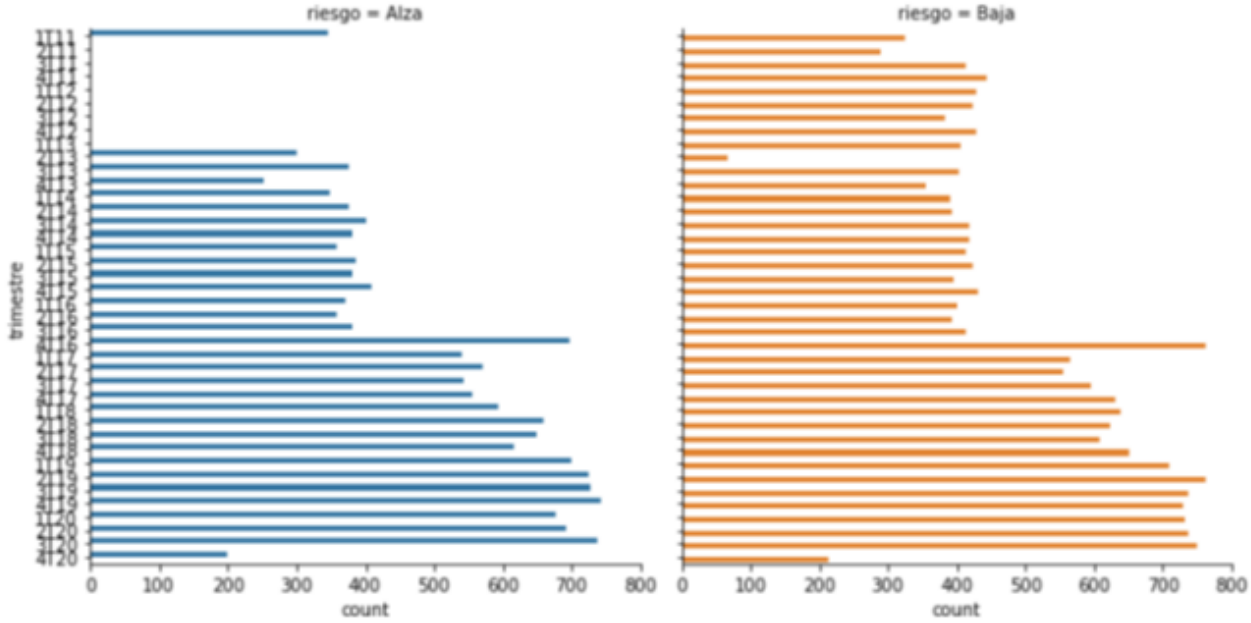


Figura 2.1: Distribución de documentos por trimestre de 2011-2020, según el tipo de riesgo.

Datos

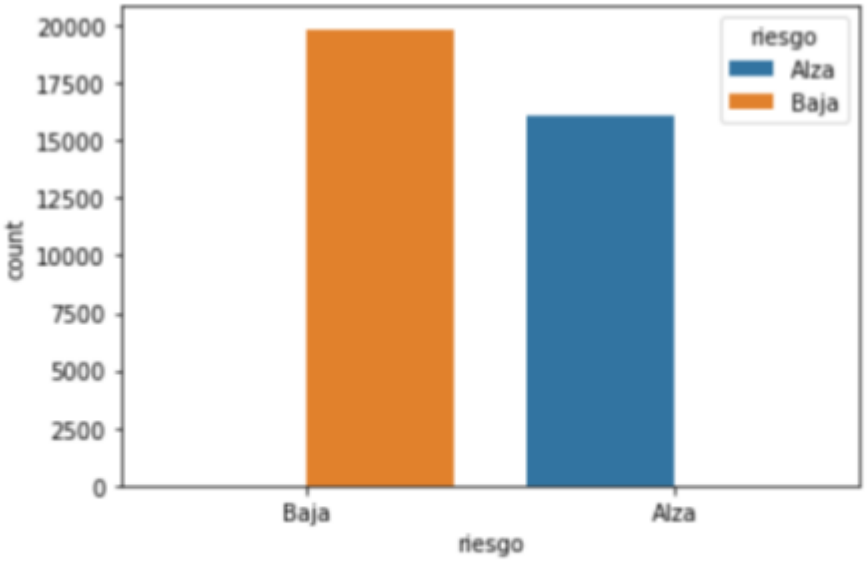


Figura 2.2: Documentos por tipo de riesgo (Baja 55%, Alza 45%). Documentos totales: 35,895.

Datos

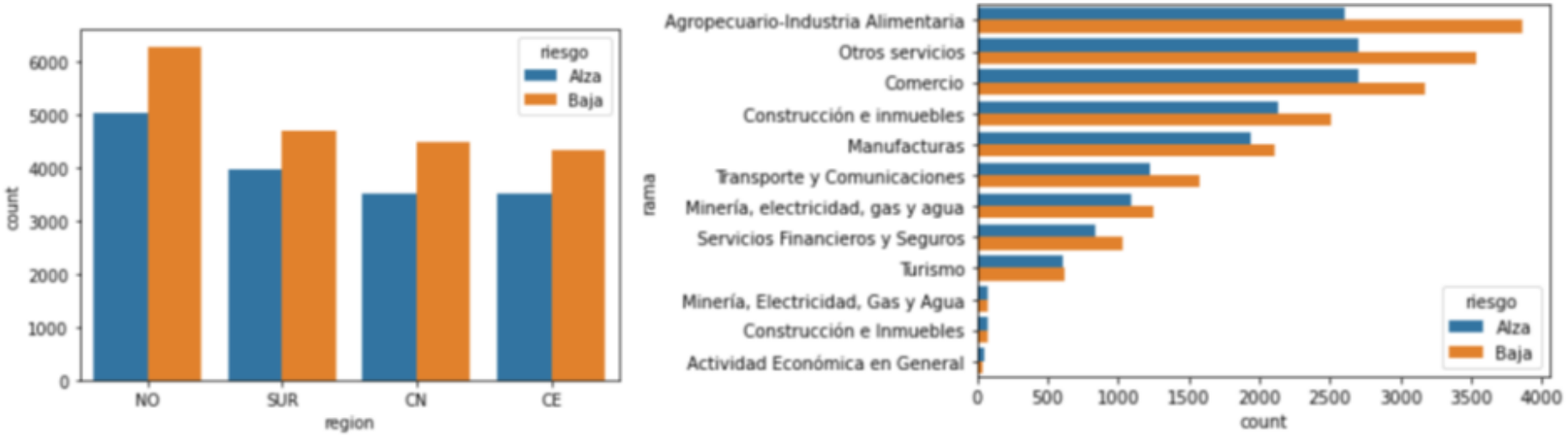


Figura 2.3: a) Documentos por región (NO 31%, CE 22%). b) Documentos por rama (Agropecuario 18%, Actividad económica general 0.2%).

Datos

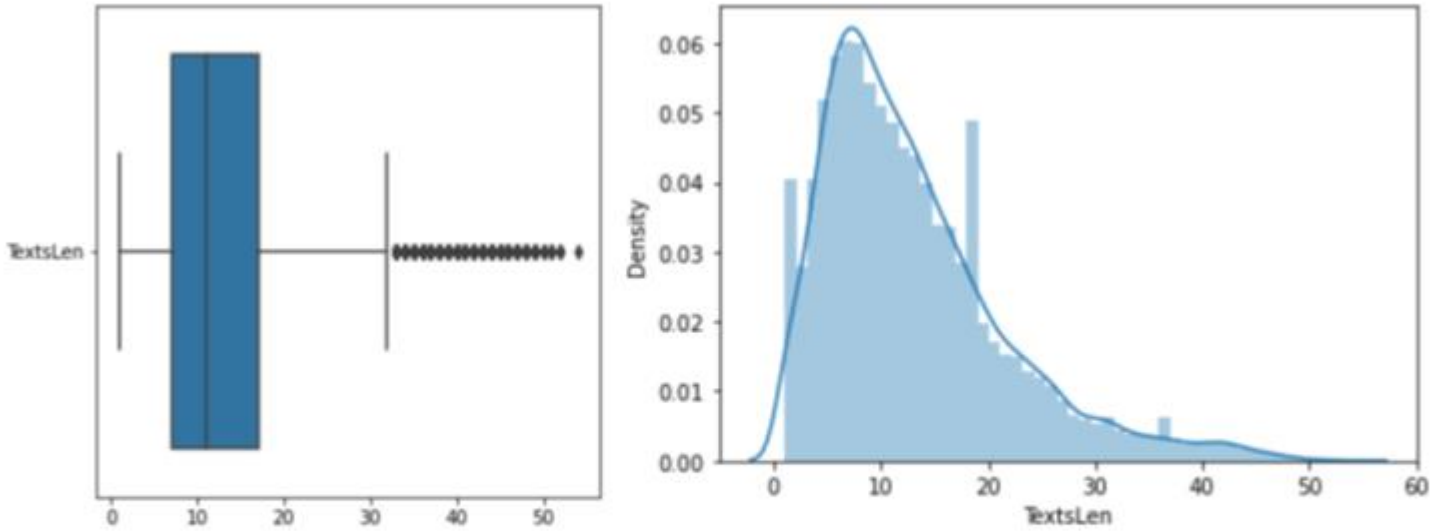


Figura 2.4: Distribución de la longitud de los documentos. Min: 1, 25%: 7, 50%: 11, media: 13, 75%: 17, Max: 54.

Conjuntos de tópicos

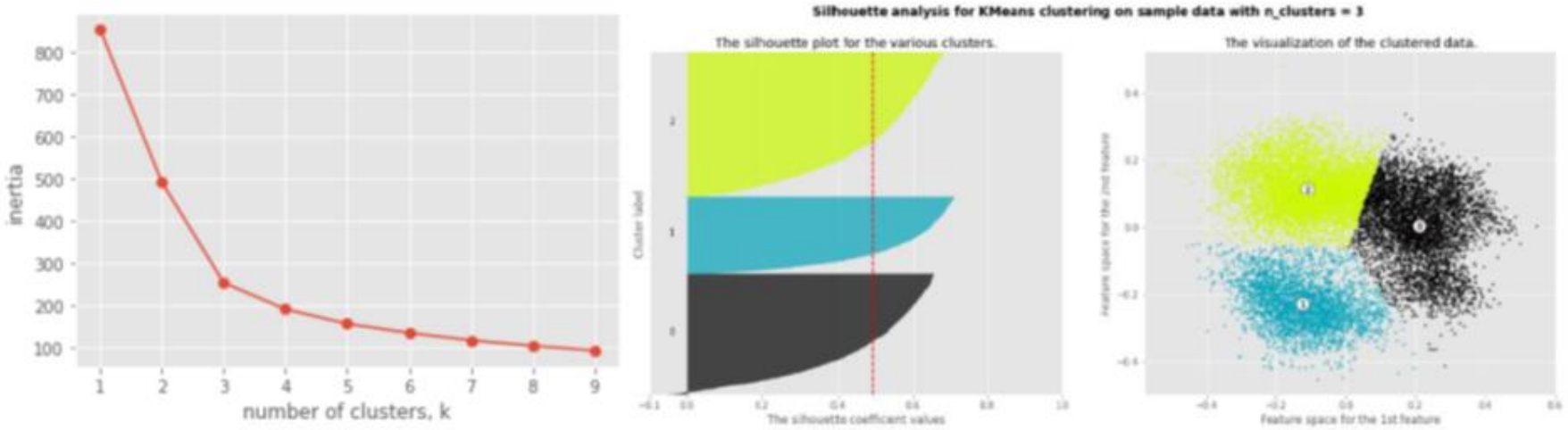


Figura 2.5: a) inercia vs Clústers, punto de cambio en 3. b) Gráfico de silueta, mejor puntuación para 3 clústers: 0.4921.

Conjuntos de tópicos

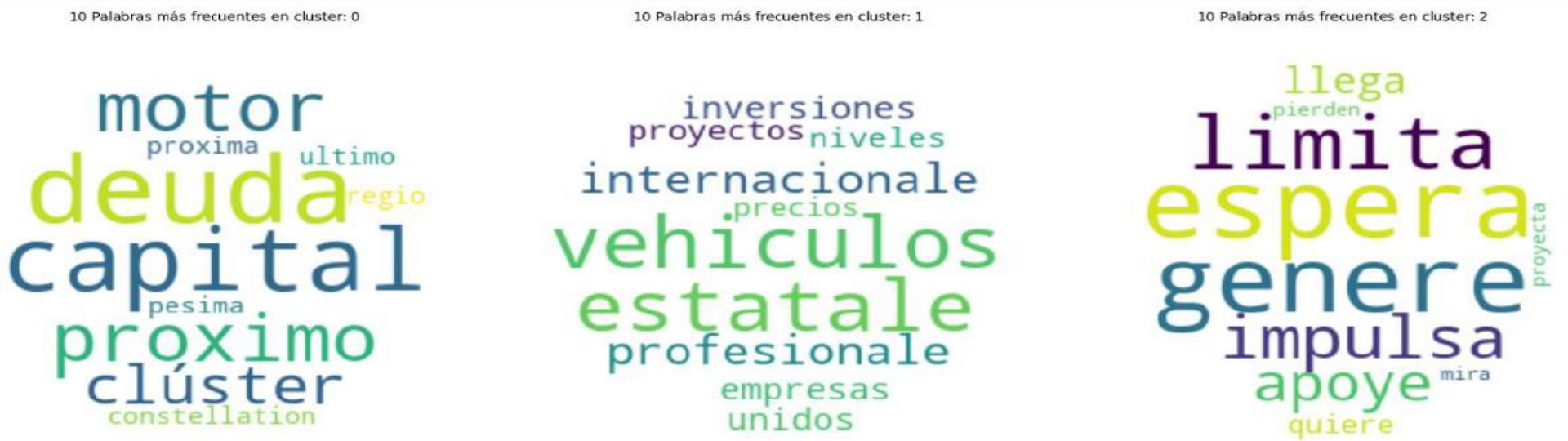


Figura 2.7: Gráficos de nubes de palabras para los clústers: 0, 1, 2.

Conjuntos de tópicos

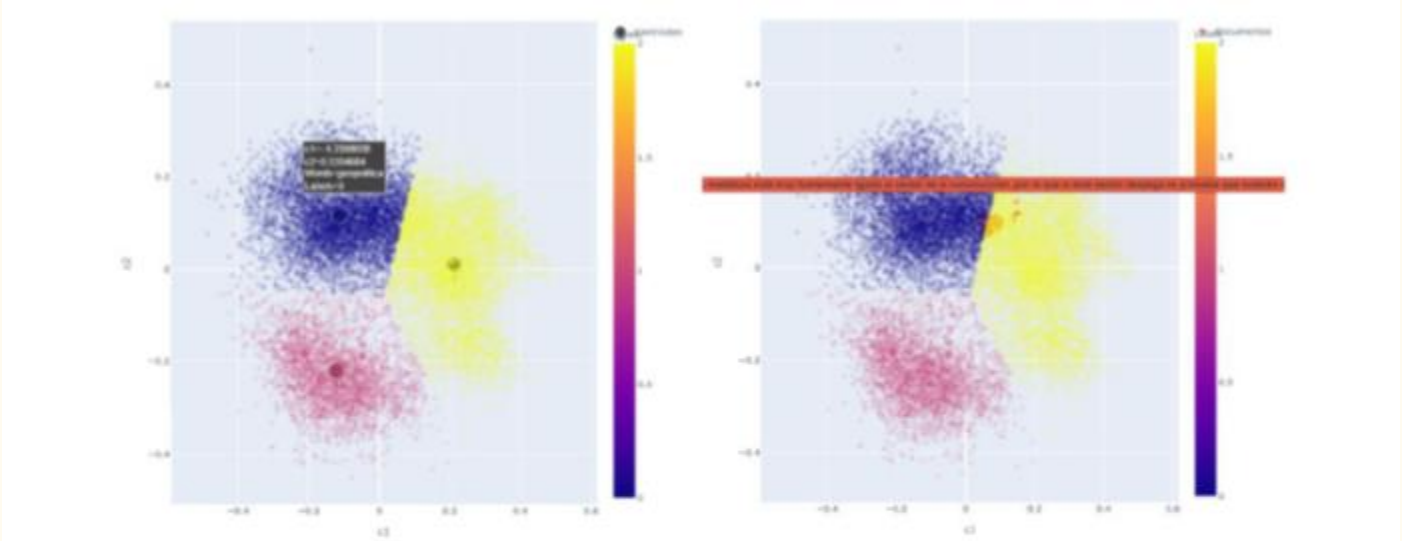
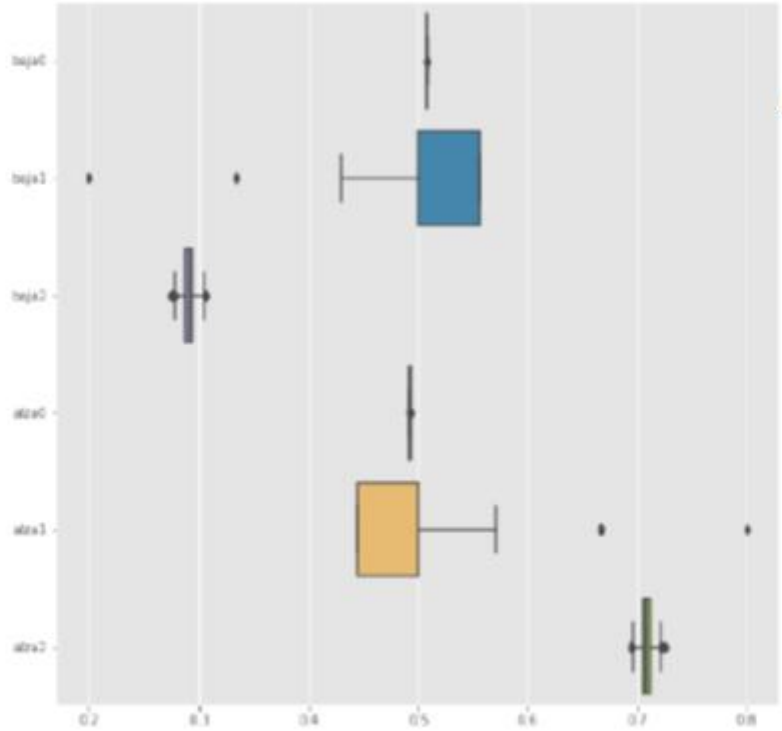


Figura 2.6: a) Clústers de tópicos únicos en el vocabulario (14,475) con sus respectivos centroides, Técnica de clustering empelada: Fuzzy KMeans. b) proyección de los documentos de un trimestre sobre los clústers de tópicos, se asigna el cluster del centroide más cercano al texto.

Conjuntos de tópicos



$$P(B/C_i) = I_{b,i} / N_i$$

Figura 2.8: Frecuencias relativas de los riesgos a la baja o al alza por cada clúster (0,1,2).

Indicadores

$$P(C_i)_Q = {}^{I_{i,Q}}\!/_{N_Q} \tag{1.2}$$

Donde, $\left(P(C_i)_Q\right)$ es la frecuencia relativa para el clúster $(i=0,1,2)$ en el trimestre (Q), $\left(N_Q\right)$es el número de textos en el trimestre y $\left(I_{i,Q}\right)$ es el número de textos clasificados en el clúster en el trimestre.

$$Risk\_index = P(B)_Q = \sum_{i=0}^{2} P(C_i)_Q \, P(B/C_i) \tag{1.3}$$

Donde $P(B)_Q$ es la probabilidad de que para ese trimestre sea riesgo a la baja.

Indicador, General

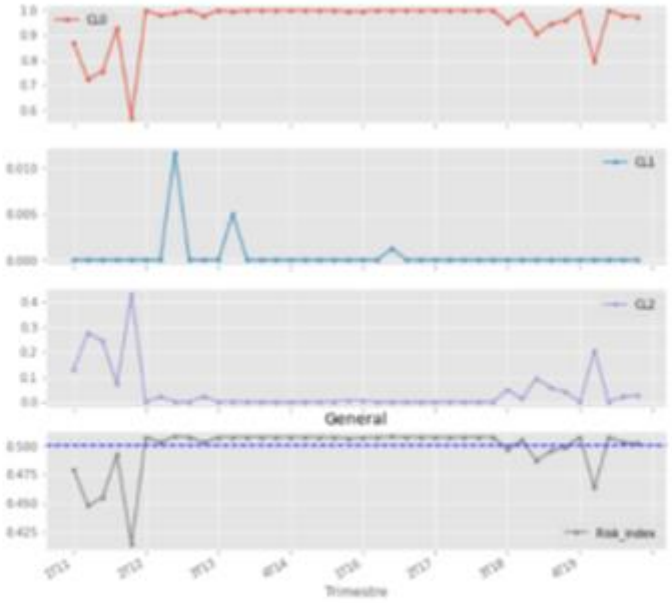


Figura 3.2: Indicador de riesgo con sus respectivas componentes para el periodo de 1T11 a 4T20.

Indicador, General

Figura 3.1: Gráficos de nubes de palabras, con las 50 palabras más representativas en el cluster 0, para el trimestre a) 3T12, b) 3T18.

Indicador, General

Figura 3.3: Gráfico de nube de palabras, para las 50 palabras más frecuentes en el a) clúster 0 y b) clúster 2 para el trimestre 1T20.

Indicador, General

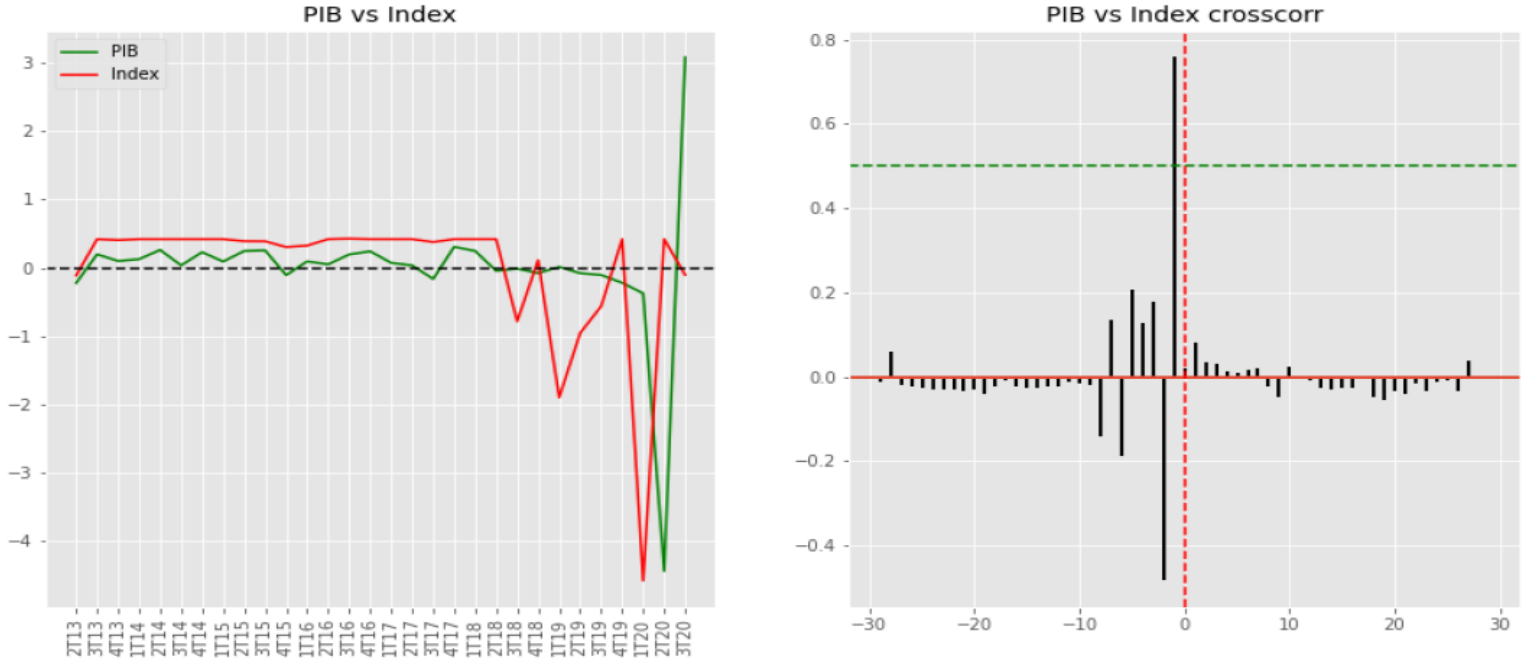


Figura 3.4: a) Indicador general contra el **PIB** del 2T13 al 3T20, b) correlación cruzada para el indicador y el **PIB**.

Indicador, Regional

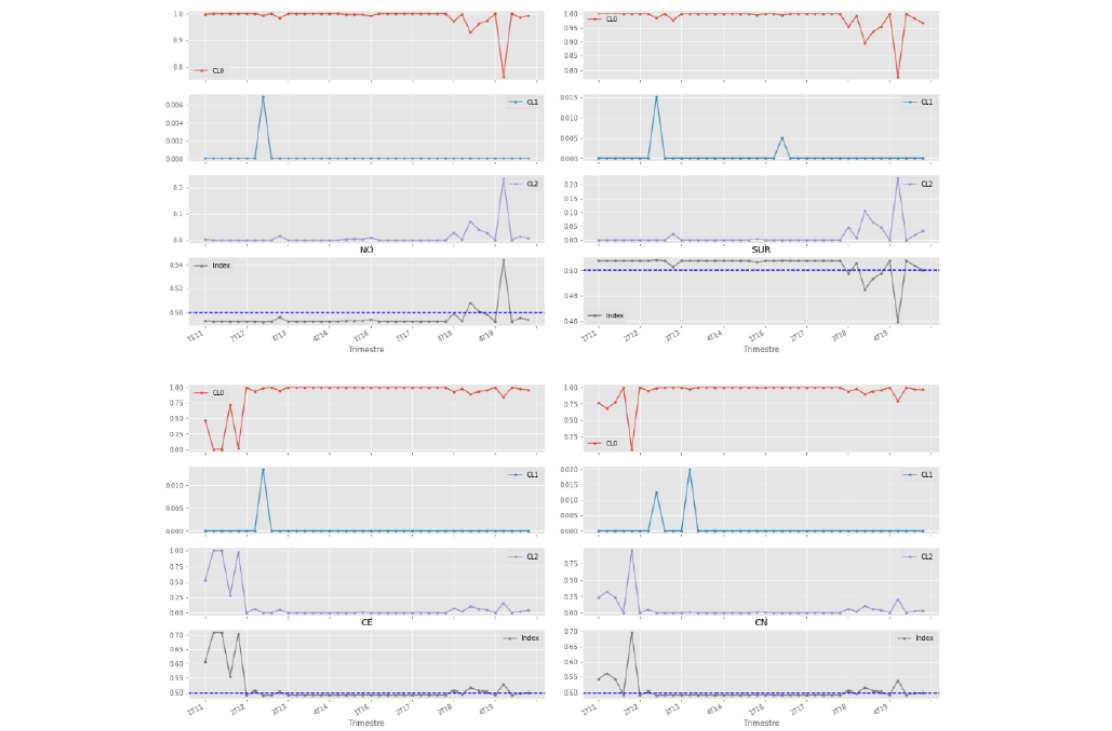


Figura 3.5: Indicadores regionales con sus respectivas componentes para el periodo de 1T11 a 4T20, para todas las regiones.

Indicador, Regional: CE

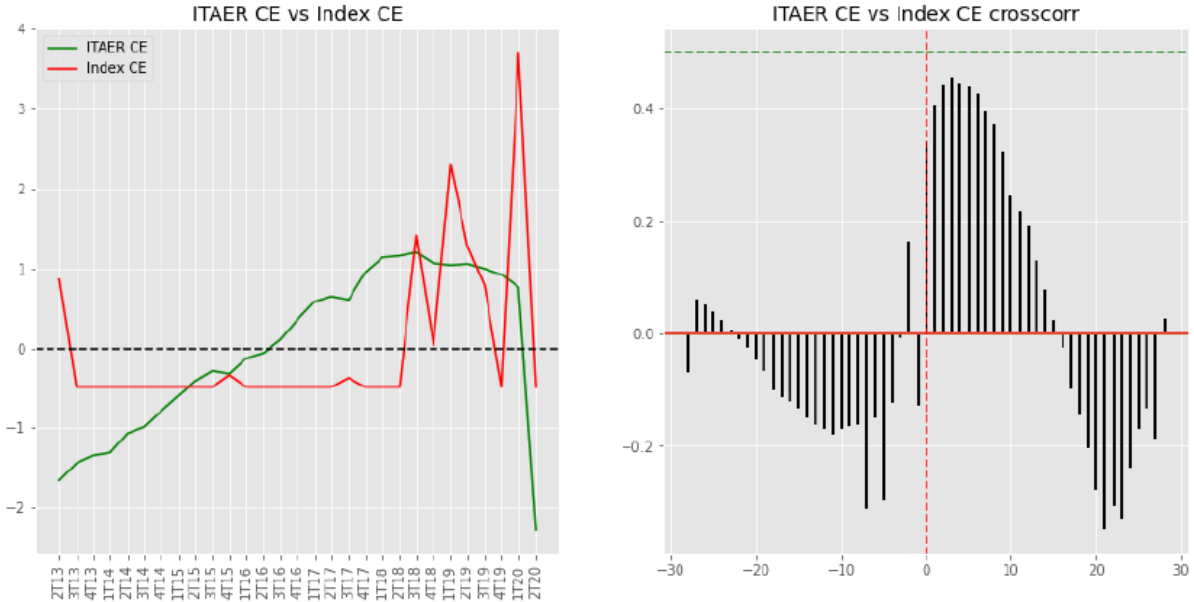
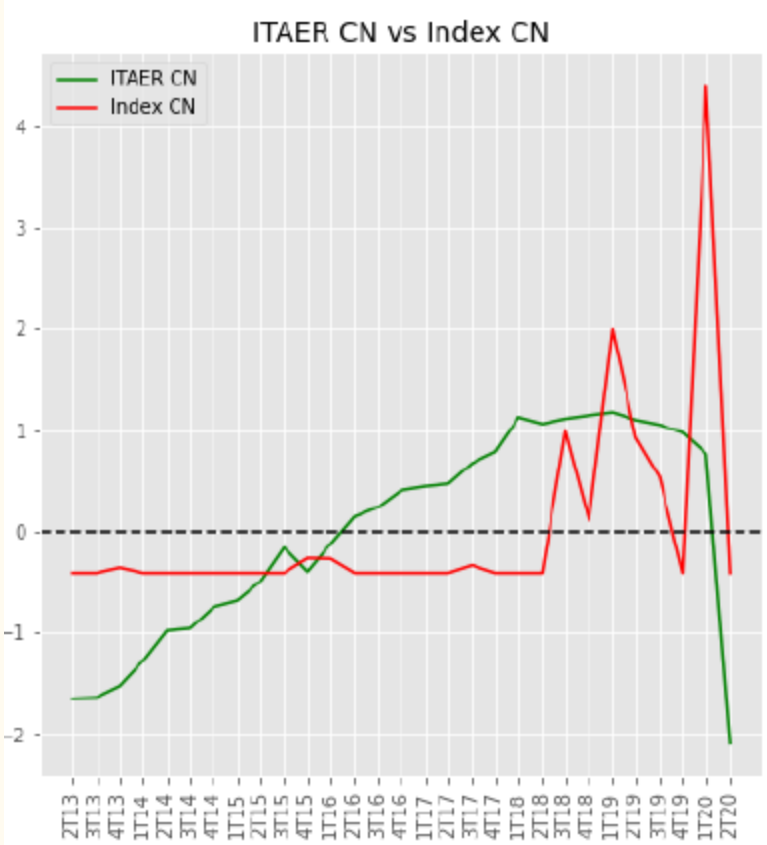
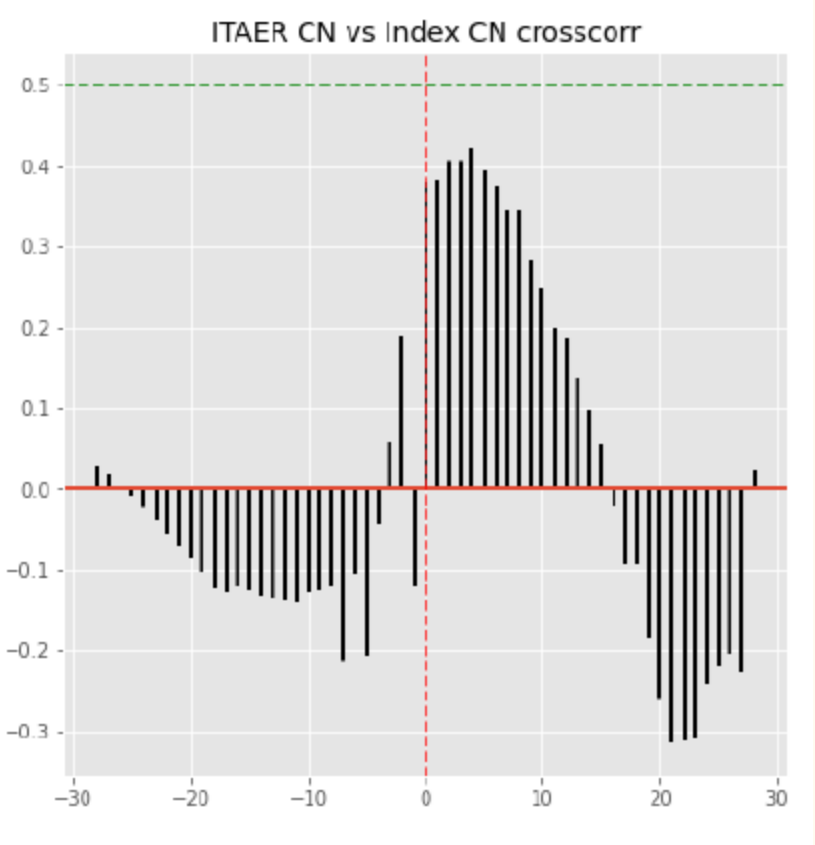


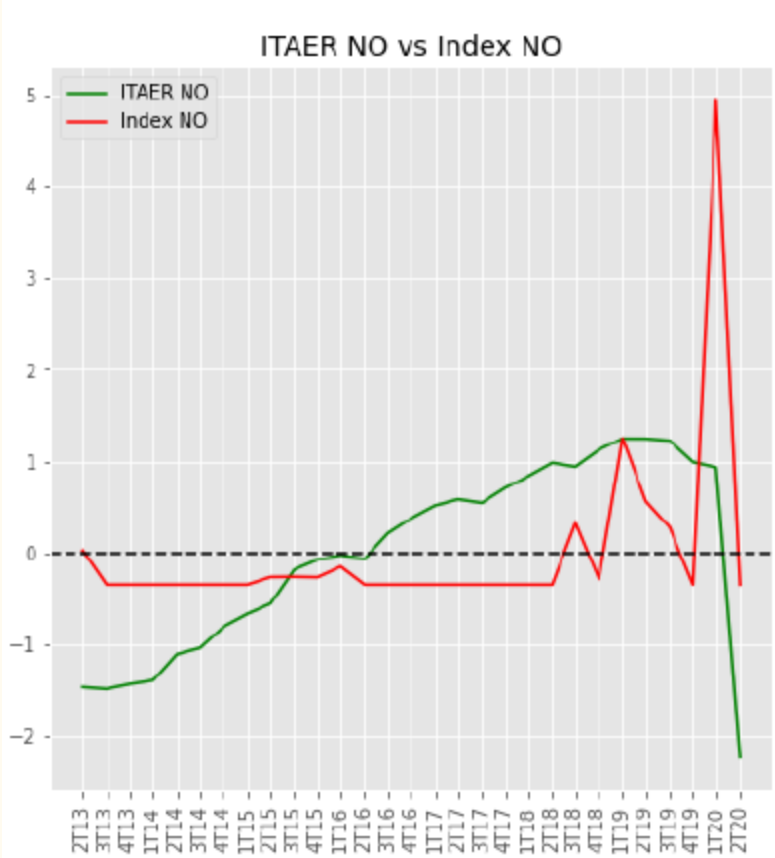
Figura 3.6: a) Indicadores regionales contra el **ITAER** del 2T13 al 2T20, b) correlación cruzada para el indicador y el **ITAER**.

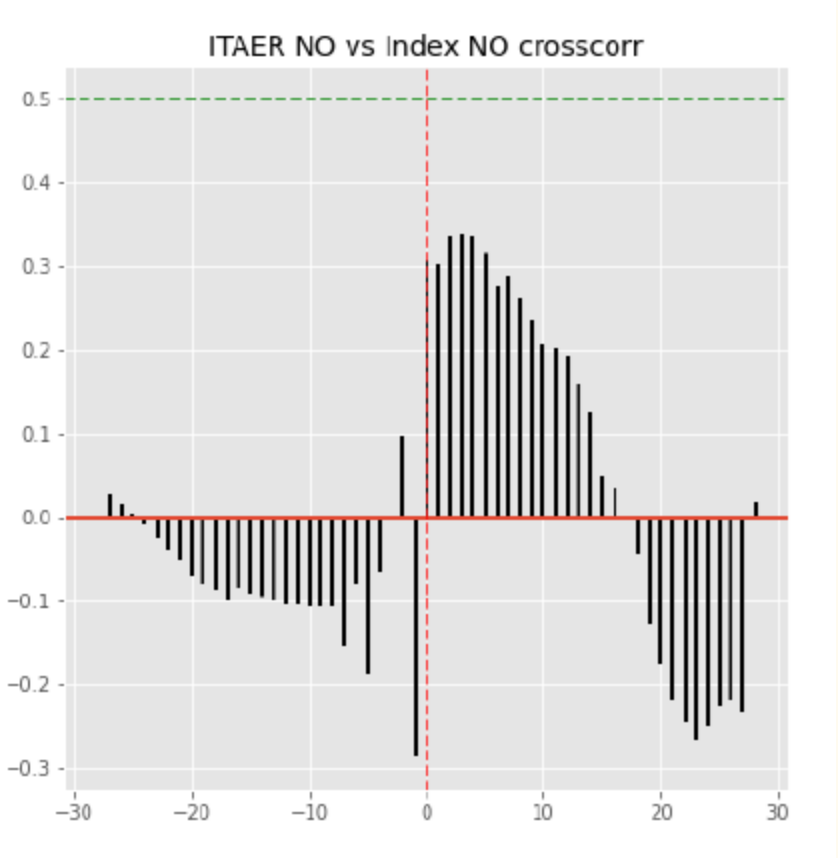
Indicador, Regional: CN



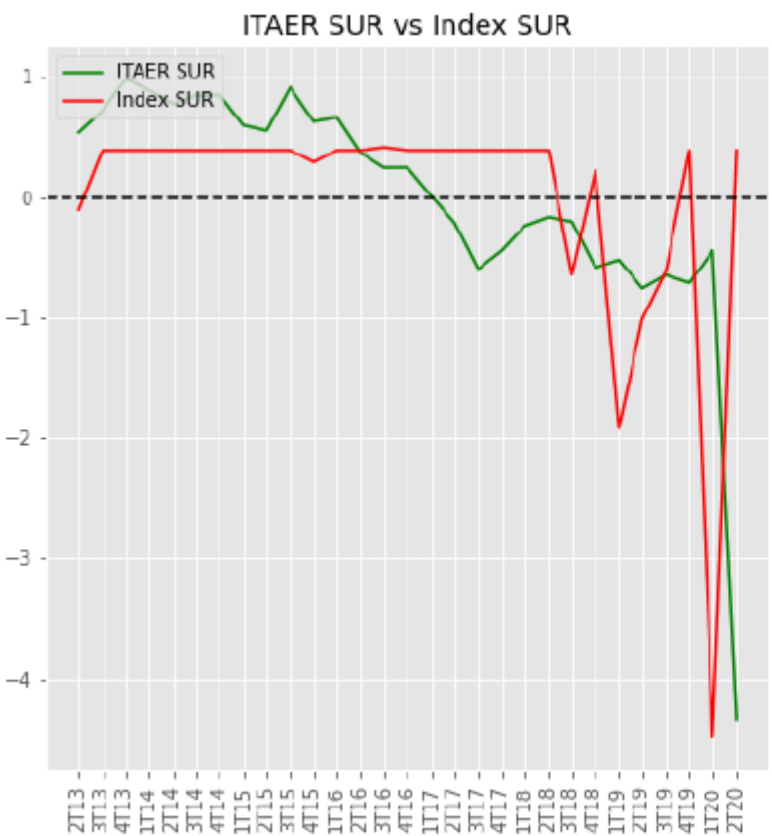


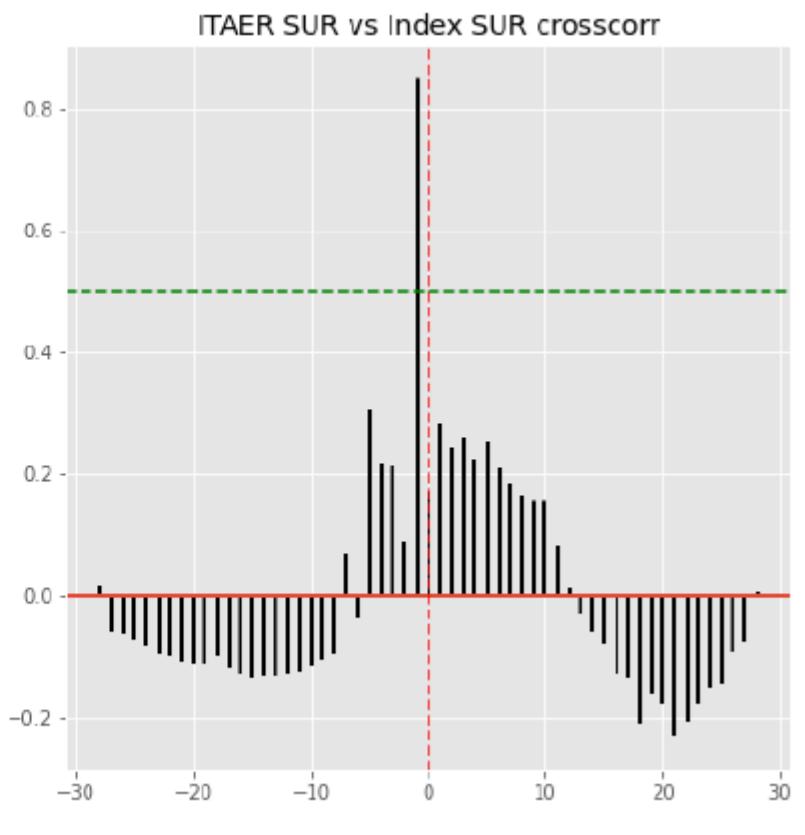
Indicador, Regional: NO



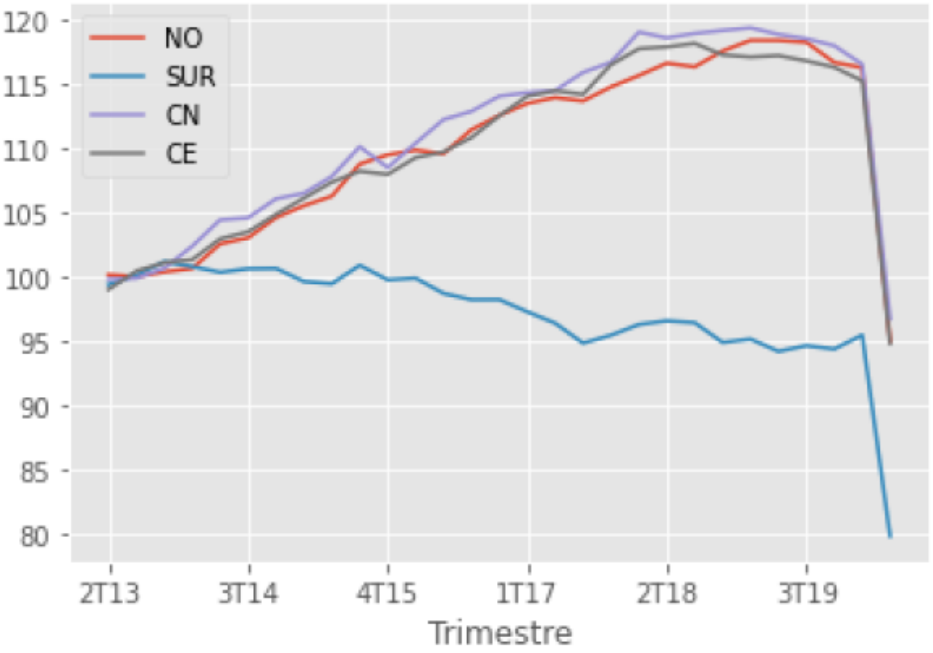


Indicador, Regional: SUR





Indicador, Regional



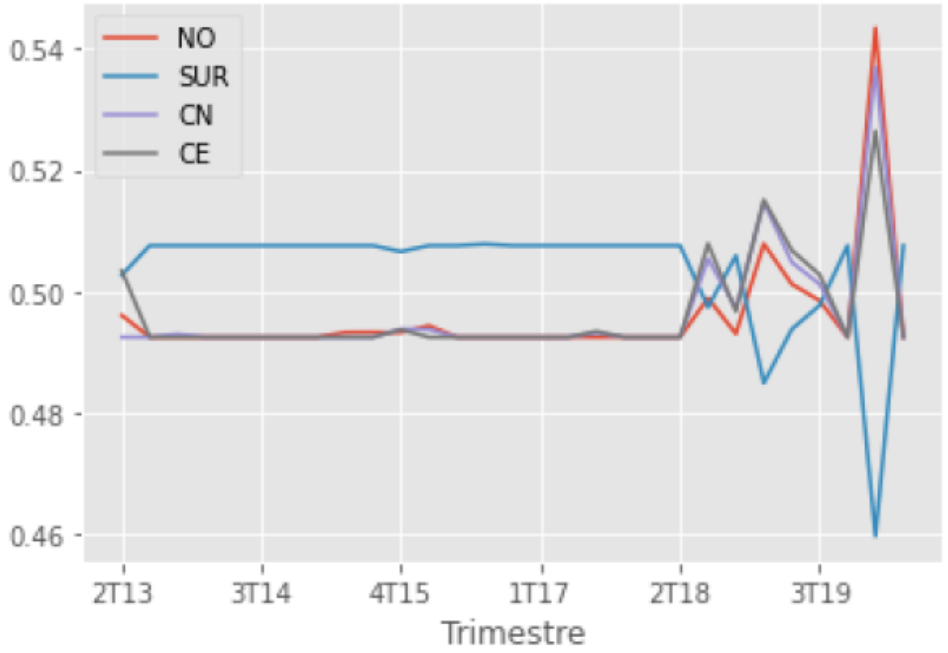


Figura 3.7: a) ITAER para cada región. b) Índice regional