UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

'Nowcasting' de indicadores económicos combinando series de Google Trends

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: Cabria Rodriguez, Minerva

Tutor/a: Doménech i de Soria, Josep

CURSO ACADÉMICO: 2022/2023

Resumen

Los indicadores económicos suelen publicarse con retraso con respecto al periodo temporal al que hacen referencia. Por su parte, Google Trends suministra información actualizada en tiempo real sobre la popularidad de los términos de búsqueda. Algunos de estos términos están estrechamente relacionados con ciertos indicadores económicos, lo que permite la creación de modelos para realizar "nowcasting" de dichos indicadores. Es decir, para estimar en tiempo real la evolución de los indicadores, mejorando así la capacidad de diagnóstico de la situación económica del país en el mismo momento. Con este TFG se pretende comparar diversos métodos estadísticos y econométricos para realizar estas predicciones, y así evaluar su capacidad para hacer 'nowcasting' de los indicadores económicos escogidos. Esto se hará mediante la combinación de series de Google Trends y evaluando posteriormente los modelos creados desde distintas perspectivas.

Palabras clave: indicadores económicos, Google Trends, nowcasting, series temporales, regresión, análisis de componentes principales

Abstract

Economic indicators are often published with a delay with respect to the time period to which they refer. On its part, Google Trends provides instant updated information about the popularity of search terms. As some of these terms are correlated with certain economic indicators, it is possible to create models to perform nowcasting of these indicators. That is, to estimate in real time the evolution of the indicators, thus improving the ability to diagnose the economic situation of the country at the same time. The aim of this ungraduate tesis project is to compare different statistical and econometric methods to make these predictions, and thus evaluate their capacity to do 'nowcasting' of the chosen economic indicators. This will be done by combining Google Trends series and subsequently evaluating the models created from different perspectives.

Keywords: economic indicators, Google Trends, nowcasting, temporal series, regression, principal component analysis

Tabla de contenidos

1. INTRODUCCIÓN

1.1. Resumen

La herramienta Google Trends es un medio en línea desarrollado por Google que proporciona información acerca de la notoriedad de las investigaciones realizadas por los usuarios. En los últimos años, el uso de estas series temporales para realizar predicciones económicas y sociales ha aumentado significativamente. Las investigaciones realizadas en este campo han demostrado de manera satisfactoria que la inclusión de los datos de Google Trends en los modelos de predicción de indicadores económicos mejora los resultados obtenidos en países altamente desarrollados, como Alemania y los Estados Unidos de América.

Uno de los pasos previos antes de enriquecer un modelo de predicción con datos de GT es seleccionar qué palabras clave anticipan el comportamiento de un indicador económico. Trabajos previos han abordado esta selección de forma intuitiva y sin una metodología clara. En este contexto, el presente Trabajo Fin de Grado (TFG) pretende evaluar distintos métodos para seleccionar y combinar series de GT asociadas a palabras clave para predecir algunos indicadores económicos a tiempo real. Además, se pretende validar la relación existente entre Google Trends y los indicadores económicos en España.

Así pues, la innovación de este trabajo reside en la construcción y evaluación de modelos *nowcasting* con series temporales de Google Trends, que sean capaces de predecir ESI, desempleo e inflación, con una selección automática de las palabras clave más relevantes en cada momento.

La metodología se basa en previas investigaciones científicas que tratan sobre el uso de modelos de *nowcasting* y, en particular, del uso de Google Trends. Las bases de datos de los indicadores se descargan directamente del Eurostat, Instituto Nacional de Estadística y Servicio de Empleo Público Estatal y las series temporales de GT se adquieren con el paquete 'trendecon' de Rstudio. Con los datos de GT se entrenan los modelos estadísticos y econométricos como Stepwise regression, PCA combinado con regresión lineal y el método Lasso para obtener la estimación de los indicadores. Posteriormente, se realizan dos tipos de validación, la primera de ellas dentro de la muestra y la segunda fuera de la muestra, con los datos reales de los indicadores. Esta validación se efectúa con métricas como el error cuadrático medio, el error medio absoluto, error porcentual absoluto medio y R al cuadrado.

1.2. Motivación

En primer lugar, se resalta la utilidad de las herramientas que aportan información sobre el comportamiento de los usuarios en la web, como es el caso de Google Trends. Estas herramientas permiten entender la manera en que actúan los usuarios a través de la huella digital que generan. Debido al uso diario de las TIC, la cantidad de datos que los usuarios depositan online es cada vez mayor y más variada. Dicha información, resulta particularmente importante si se trata del uso del buscador de Google en actividades cotidianas como; explorar productos en sitios web de compras específicos, la búsqueda de información sobre el clima económico en España, la indagación de ofertas laborales o la consulta diaria de noticias en diferentes temas, entre otros. Es evidente que el uso de Internet para recopilar información se ha vuelto crucial en la actualidad.

Por lo comentado en el párrafo anterior, tanto el concepto de sentimiento económico, medido por el ESI, como el desempleo y la inflación, están directamente relacionados con las series proporcionadas por GT, las cuales miden la popularidad de búsqueda de palabras clave relacionadas con dichos indicadores económicos. Es por ello que, como principal motivación se quiere probar la efectividad de GT para estimar en tiempo real estos indicadores.

Por tanto, se tienen dos fuentes de datos: los que ofrece GT y los datos reales de los indicadores ESI, desempleo e inflación, calculados por las instituciones oficiales. Ambas fuentes recolectan datos de los usuarios de manera diferente, los cuales pueden ser útiles en el ámbito económico. Dada la situación, se plantea la posibilidad de que la evolución de los datos de ambas fuentes esté correlacionada, lo que favorecería el cálculo de los indicadores económicos adelantados.

La relación entre los indicadores económicos y Google Trends ya ha sido estudiada en numerosas ocasiones, como bien se comenta en el apartado 2.7.1. Cabe destacar que, otra de las motivaciones es evitar que dichas variables sean publicadas con retraso y se pueda anticipar la situación económica de España en el momento deseado.

1.3. Objetivos

El objetivo principal del TFG es la evaluación de distintos métodos para combinar series proporcionadas por la herramienta Google Trends para predecir en tiempo real el índice de sentimiento económico, el desempleo y la inflación en España. De ello deriva también, estudiar la relación entre diversos términos de búsqueda en Google y dichos indicadores.

De este modo, podemos listar los siguientes objetivos:

- Revisar la literatura existente sobre el uso de series de Google Trends para anticipar indicadores económicos y posteriormente, comprobar su eficacia para dicho fin.
- Conocer las bases teóricas de los indicadores económicos seleccionados, así como de los métodos para seleccionar y combinar series temporales.
- Construcción y evaluación de métodos que sean capaces de obtener una buena predicción de indicadores económicos tomando como datos de entrada las series de Google Trends.
- Justificar mediante la combinación de dichas series temporales la relación entre Google Trends y los indicadores económicos en España.

1.4. Estructura del TFG

Este Trabajo Final de Grado (TFG) se estructura en cinco capítulos fundamentales, cuya secuencia y contenido básico se describe a continuación:

En primer término, se incluye un capítulo introductorio, donde se plasman las intenciones y los objetivos del trabajo.

En segundo lugar, se contextualiza el tema mediante la detallada definición de conceptos como nowcasting, Google Trends, ESI, desempleo e inflación, así como la explicación de los modelos que se desarrollarán. Adicionalmente, se exponen las investigaciones previas relacionadas, abarcando de esta manera el marco teórico que sustenta el trabajo.

En tercer lugar, se describe la metodología aplicada para la ejecución del proyecto, se detalla la muestra utilizada y se explican los métodos de evaluación/validación de los modelos propuestos.

En cuarto lugar, se exponen los resultados, dónde se analizan los resultados obtenidos del estudio, así como su análisis e interpretación.

Finalmente, en el capítulo de conclusiones, se relacionan los objetivos detallados en el capítulo de introducción con las conclusiones obtenidas, proporcionando así una justificación clara de su cumplimiento. Asimismo, se presentan propuestas de mejora y se discuten las limitaciones que surgieron durante el estudio. Además, se identifica un posible legado que podría ser beneficioso para los interesados y, se ofrecen recomendaciones de posibles direcciones para futuros trabajos en esta línea de trabajo.

2. MARCO CONTEXTUAL

2.1. Modelos de series temporales

2.1.1. Series temporales

Una serie temporal (o simplemente una serie) es una secuencia de N observaciones (datos) ordenadas y equidistantes cronológicamente sobre una característica (serie univariante o escalar) o sobre varias características (serie multivariante o vectorial) de una unidad observable en diferentes momentos (Mauricio, 2007).

Los datos manejados en este TFG son series temporales, pues tanto los datos que proporciona Google Trends, como los indicadores económicos seleccionados constan de observaciones ordenadas y equidistantes en el tiempo. En el caso concreto que nos ocupa, la frecuencia de las series será mensual.

2.1.2. Modelos de *nowcasting*

La técnica del *nowcasting* consiste en llevar a cabo una serie de acciones para "predecir" el presente, el pasado reciente y el futuro cercano de una serie temporal. También puede definirse como la práctica de observar, mediante un modelo, las corrientes de información que se publican en tiempo real (Bańbura, Giannone, Modugno, & Reichlin, 2013). En el ámbito de la economía, se han utilizado modelos de *nowcasting* para predecir, entre otras cosas, el desempleo, el Producto Interior Bruto y la inflación. (Ettredge, Gerdes, & Karuga, 2005); (Camacho & Pérez-Quiros, 2010); (Bánbura & Modugno, 2010).

La importancia de los modelos de *nowcasting* en la economía radica en la necesidad de contar con herramientas que permitan evaluar la situación económica actual en tiempo real debido al retraso en la publicación de los indicadores tradicionales.

La predicción del nivel de inflación mediante modelos de *nowcasting* es un tema de investigación poco estudiado en comparación con otros indicadores económicos. Aun así, son relevantes las investigaciones realizadas por (Monteforte & Moretti, 2009); (Bánbura & Modugno, 2010) con respecto al Espacio Económico Europea.

En un estudio pionero, (Ettredge, Gerdes, & Karuga, 2005) propusieron un modelo de *nowcasting* que utilizaba el historial de búsqueda de varios buscadores para pronosticar la cantidad de

personas que acababan de perder su trabajo. Desde entonces, muchos autores han utilizado la información de Google Trends para desarrollar modelos de *nowcasting* y predicción en diversos ámbitos.

2.1.3. Estacionalidad

Las series temporales suelen descomponerse en cuatro componentes: nivel, tendencia, estacionalidad y ruido. La estacionalidad se refiere al patrón cíclico a corto plazo de una serie temporal, que se puede repetir varias veces dentro del período de tiempo analizado (Doménech, 2021).

Posiblemente, la razón más importante para efectuar el ajuste estacional de una serie sea la que propuso Persons: cada uno de los componentes es causado por fenómenos distintos. Esta misma idea elabora Granger, en particular en lo que toca a la estacionalidad, encontrando que hay al menos cuatro posibles causas de las fluctuaciones estacionales, que según este autor no tienen por qué ser completamente ajenas o distintas entre sí (Guerrero, 1990).

Se dice que una serie está ajustada por estacionalidad cuando se ha eliminado la componente estacional de la misma. Es importante tener en cuenta que, cuando se hacen modelos de *nowcasting*, no se deben mezclar series con estacionalidad y ajustadas por estacionalidad. En el caso concreto de este TFG, los datos de ESI proporcionados por la Comisión Europea ya están desestacionalizados. Sin embargo, los de Google Trends no lo están. Por ello, se procederá a desestacionalizar las series temporales que no lo estén.

El ajuste por estacionalidad se puede hacer fácilmente en R utilizando la función *decompose*, ya que el atributo *seasonal* contiene la componente estacional. Existen otros métodos más avanzados para realizar este ajuste, pero quedan fuera del alcance de este TFG.

2.2. Google Trends

2.2.1. Introducción

Google Trends (GT) es una herramienta que proporciona series temporales mostrando la popularidad (volumen) de la búsqueda de determinadas palabras clave, en un área de conocimiento definida, en el buscador de Google. Dichas series temporales tienen su origen en el año 2004 y su frecuencia puede variar de diaria a mensual, pudiéndose además limitar el área geográfica para la que deseen obtenerse las mismas.

Google Trends se presentó por primera vez el 11 de mayo de 2006. Luego, Google lanzó Google Insights for Search el 5 de agosto de 2008, un servicio avanzado y más detallado que proporcionaba datos sobre las tendencias de búsqueda a los usuarios. El 27 de septiembre de 2012, Google combinó Google Insights for Search con Google Trends (Jun, Choi, & Yoo, 2018).

2.2.2. Interfaz

Se puede acceder a la herramienta mediante interfaz de usuario o con una API.

A través de la interfaz de usuario, en la primera página se muestran los términos más buscados en los últimos 7 días.

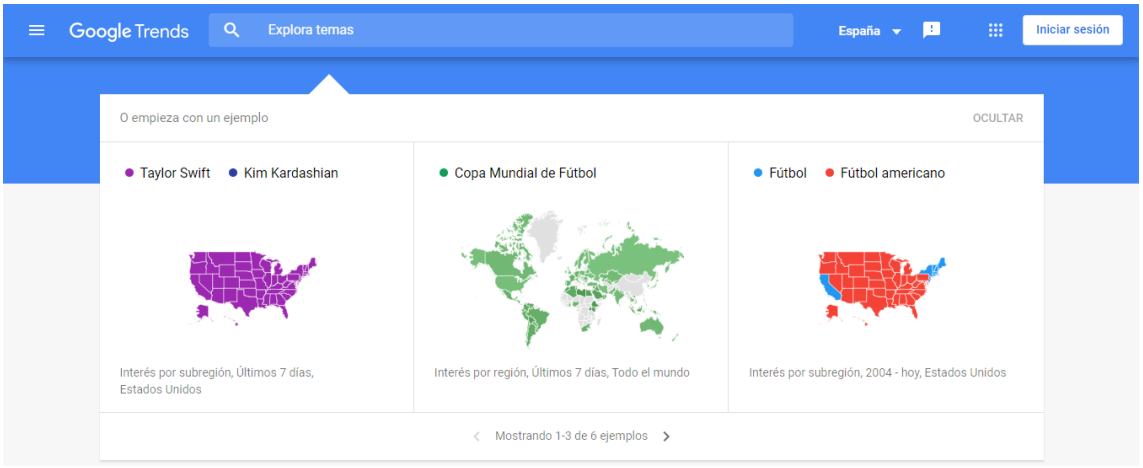

Figura 1. Interfaz Google Trends. Fuente: Google Trends

Se puede buscar una palabra concreta de la cual se desea obtener la popularidad en la barra 'explorar temas'. Dicho término se puede filtrar por:

- Lugar: Puedes elegir la región o país del cual quieres obtener la información.
- Tiempo: Puedes elegir cualquier fecha o intervalo.
- Categoría: Puedes revisar la categoría en los que despunta tu término.
- Plataforma: En función de tu objetivo puedes diferenciar resultados globales, de búsquedas web, imágenes, noticias, Google Shopping o YouTube.

En el ejemplo, se emplea la palabra 'PIB' y se filtra por variables; España, últimos 12 meses, todas las categorías y búsqueda de noticias.

Primero, GT muestra un gráfico de líneas de la evolución temporal de la búsqueda del término según el tiempo escogido. El eje X es el tiempo y el eje Y es una escala de 0 a 100 que muestra cómo de popular es la búsqueda.

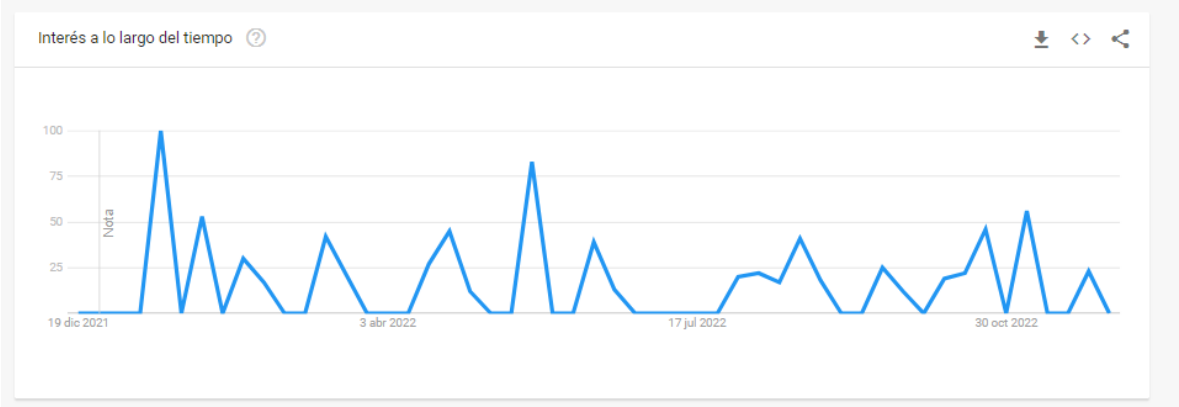

Figura 2. Evolución de las búsquedas del término PIB. Fuente: Google Trends

Más abajo GT se tiene el interés por subregión, temas relacionados con la búsqueda y consultas relacionadas. con ella.

Figura 3. Interés por subregión del término de búsqueda PIB. Fuente: Google Trends

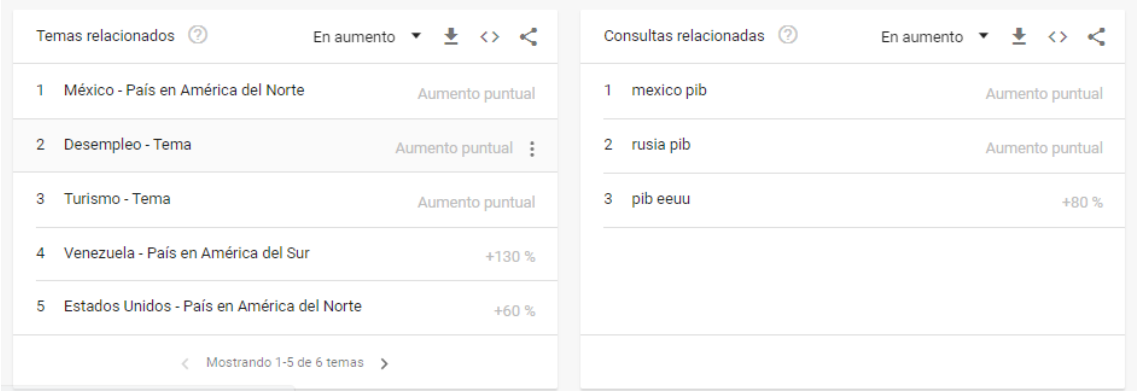

Figura 4. Términos y consultas relacionados con el término de búsqueda PIB. Fuente: Google Trends

2.2.3. Search Volume Index

Google Trends proporciona un índice del volumen de consultas que los usuarios introducen en Google en una zona geográfica determinada, denominado Search Volume Index (SVI).

Este se calcula mediante un submuestreo aleatorio en la población y se basa en la cuota de consultas: el volumen total de consultas para el término de búsqueda en cuestión dentro de una región geográfica concreta dividido por el número total de consultas en esa región durante el periodo de tiempo examinado.

La cuota de consulta máxima en el periodo de tiempo se normaliza a 100, y la cuota de consulta en la fecha inicial examinada se normaliza a cero. Las consultas son *"broad matched"* en el sentido de que consultas como [automóviles usados] se cuentan en el cálculo del índice de consulta para [automóvil]. (Varian & Choi, 2012).

2.3. Índice de sentimiento económico

2.3.1. Definición

El indicador de sentimiento económico (ESI) es un indicador compuesto elaborado por la Dirección General de Asuntos Económicos y Financieros (DG ECFIN) de la Comisión Europea. Su objetivo es seguir el crecimiento del Producto Interior Bruto a nivel de los estados miembros, la Unión Europea y la zona del euro (Europea U. , 2023).

2.3.2. Cuestionarios de los distintos sectores

El ESI se calcula a partir de cuestionarios realizados a distintos sectores de la economía. Los cuestionarios para el cálculo del ESI son:

1. Encuesta de coyuntura industrial: se pregunta a empresas industriales sobre la cartera de pedidos, el nivel de existencias, la producción, el grado de incertidumbre, la tendencia de los precios de venta, empleo, competitividad e inversión tanto a nivel actual como previsión a tres meses.

2. Estudio de Servicios: se pregunta a empresas de servicios sobre la condición global de la organización, el empleo total y los precios de venta en los últimos tres meses y los tres siguientes. Después, se realizan preguntas sobre la demanda y los factores que limitan el crecimiento de la empresa (de manera trimestral). Y, por último, hay una serie de preguntas acerca de la evolución de la inversión.

3. Análisis de la percepción de los consumidores: se recaba información de la población sobre la economía en su unidad familiar durante el último año y sobre la previsión para el próximo. También, se pregunta opinión acerca de los precios que pagan los consumidores por bienes y servicios y la tasa de desempleo en el país.

4. Estudio del comercio minorista: se pregunta a empresas minoristas cerca de la actividad del negocio, el volumen y los precios de venta, el empleo total y el volumen de stock en los últimos 3 meses y de cómo cree la empresa que dichas variables evolucionarán en los próximos 3 meses.

5. Encuesta de coyuntura del sector de la construcción: se pregunta acerca de las competencias, la producción, cartera de pedidos, empleo, grado de incertidumbre y el periodo de trabajo asegurado actualmente. Además, se requieren las expectativas de los próximos meses sobre los precios, el empleo, la producción y la cartera de pedidos.

(Europea C. , 2023)

2.3.3. Cálculo e interpretación

El ESI es una medida compuesta que toma en cuenta las respuestas de empresas y consumidores en cinco sectores distintos, obtenidas a través de las Encuestas de empresas y consumidores de la Unión Europea (UE). Se utiliza un promedio ponderado de los saldos de respuestas para cada sector, asignando pesos del 40 % a la industria, el 30 % a los servicios, el 20 % a los consumidores, el 5 % al comercio minorista y otro 5 % a la construcción.

Los saldos se construyen como la diferencia entre los porcentajes de encuestados que dan respuestas positivas y negativas. Los agregados de la UE y de la zona del euro se calculan sobre la base de los resultados nacionales y se ajustan estacionalmente.

ESI se escala a una media a largo plazo de 100 y una desviación estándar de 10. Por lo tanto, los valores superiores a 100 indican un sentimiento económico superior al promedio y viceversa (Europea U. , 2023).

2.4. Desempleo

2.4.1. Definición, medición y construcción

El indicador desempleo, también conocido como paro o desocupación, técnicamente se define como el conjunto de personas de determinades de edades que, han tomado medidas específicas para buscar empleo por su cuenta o han realizado gestiones para establecerse en un trabajo, pero no lo han logrado. Es decir, una situación en la que hay una parte de la población en edad de trabajar que no se encuentra empleada en ningún trabajo, pese a que lo demandan.

Para ello, se considera que el desempleado es todo aquel que, teniendo entre 16 y 67 años, está dispuesto a trabajar, es decir, en búsqueda de empleo, pero no lo encuentra (Morales, 2022).

La principal variable que utilizan los economistas para medir el desempleo es la tasa de desempleo. Se calcula de la siguiente forma:

$$Tasa\ de\ desempleo = \frac{Nº\ personas\ desempleadas}{Población\ activa} \times 100$$

Para saber el número exacto de personas desempleadas y las personas que conforman la población activa en España, el Instituto Nacional de Estadística (INE) realiza la Encuesta de Población Activa (EPA).

La finalidad de la EPA es obtener datos de la población en relación con el mercado de trabajo: ocupados, activos, parados e inactivos.

La población activa o activos son aquellas personas de 16 o más años que se encuentran empleadas o están disponibles y en condiciones de serlo. Se subdividen en ocupados (se encuentran empleados) y parados (desempleados) (INE, 2023).

2.4.2. Aplicación

La tasa de desempleo es una de las principales variables macroeconómicas que permiten analizar el mercado laboral.

En un sentido más amplio de la palabra, el desempleo puede considerarse un indicador clave del estado económico nacional, puesto que altos niveles de desempleo pueden afectar negativamente a la generación de riqueza, la competitividad salarial y, por lo tanto, la calidad de vida de la población (Morales, 2022).

2.5. Inflación

2.5.1. Índice de precios de consumo

El Índice de Precios de Consumo (IPC) es una medida estadística de la evolución de los precios de los bienes y servicios que consume la población residente en viviendas familiares en España.

El conjunto de bienes y servicios, que conforman la cesta de la compra, se obtiene básicamente del consumo de las familias y la importancia de cada uno de ellos en el cálculo del IPC está determinada por dicho consumo (INE, 2023).

La fórmula empleada para calcular los índices del IPC es la fórmula de Laspeyres encadenado**.** El precio actual de los productos se multiplica por la cantidad actual y se divide por los precios y cantidades de un año base, por ejemplo, el año anterior (Bermejo, 2020).

2.5.2. Definición, medición y cálculo

La inflación se define como un incremento generalizado de los precios de bienes y servicios durante un determinando periodo que, a su vez crea una disminución del poder adquisitivo de las personas reduciendo su capacidad de compra y ahorro.

Existen cuatro tipos de inflación según el porcentaje y, están marcadas según la magnitud del aumento.

- Inflación moderada: Hace referencia a una subida de precio de forma lenta. Sin superar el 10% anual.
- Inflación galopante: Sucede cuando los precios aumentan las tasas de dos o tres dígitos, en un plazo medio de un año.
- Hiperinflación: Inflación anormal. El índice de precios aumenta en un 50% de forma mensual. Es decir, casi un 13.000% de forma anual.
- Deflación: Es la inflación negativa. En este caso los precios de los productos tienen una caída en vez de aumentar. Hay una disminución en los precios.

De forma más coloquial, el dinero se vuelve menos valioso (Economista, 2006).

Existen dos indicadores:

- Tasa de inflación: Es un cociente que toma como referencia el IPC un periodo y en el periodo anterior. La cual se calcula como:

$$Tasa\ inflación = \frac{(IPC_{final} - IPC_{inicial})}{IPC_{inicial}} \times 100$$

(De La Hoz, Uzcátegui, Borges, & Velazco, 2008)

- Deflactor del PIB: El cociente entre las variaciones del PIB a precios corrientes y a precios constantes indicaría la variación del deflactor del PIB entre dos períodos. Es decir, está midiendo la variación de los precios de todos los bienes y servicios generados por la economía, independientemente de su destino económico (consumo intermedio, consumo final, inversión o exportación) (Cristóbal, 2007).

2.5.3. Aplicación

Se considera uno de los factores más relevantes en el estudio de la macroeconomía y política monetaria de los bancos centrarles. Si la inflación está controlada, los recursos económicos se gestionan y asignan mejor, logrando mayor bienestar y crecimiento.

Por otra parte, (Redondo, 1993) dice que la inflación es el deterioro del poder adquisitivo de un signo monetario empleado como patrón de medida en el intercambio de bienes o servicios, incrementando el valor monetario de los mismos con lo cual disminuye el poder de compra de esa moneda (De La Hoz, Uzcátegui, Borges, & Velazco, 2008).

2.6. Métodos de selección de variables y reducción de la dimensionalidad

2.6.1. Introducción

Suelen existir múltiples series de GT correlacionadas con los indicadores económicos que se van a estudiar. Así, por ejemplo, la tasa de desempleo estará relacionada con la popularidad de las búsquedas tanto de "paro" como de "desempleo". Ante la disyuntiva de qué hacer con estas series correlacionadas entre sí, este apartado revisa distintos métodos de selección de variables y reducción de dimensionalidad.

2.6.2. Stepwise regression

2.6.2.1. Definición

La regresión *Stepwise* es un método para desarrollar un modelo estadístico que implica la adición o eliminación de variables predictoras mediante las pruebas F o T. Las pruebas sobre los coeficientes estimados se emplean para hacer la selección de las predictoras que serán incluidas o excluidas (Wang & L.Jain, 2003).

2.6.2.2. Ventajas y desventajas

Algunas de las ventajas de realizar regresión por pasos son:

- Tiene la capacidad de explorar un gran número de variables predictoras, permitiendo una búsqueda exhaustiva de las mejores variables para incluir en el modelo final.
- Comparado con otros métodos de selección, es especialmente rápido y eficiente en términos de tiempo de cómputo.
- La secuencia en la que se agregan o eliminan las variables proporciona información útil sobre la importancia de cada variable en la predicción de la variable objetivo.

Algunas de las desventajas son:

- La regresión por pasos puede enfrentar el desafío de tener demasiadas variables predictoras y muy pocos datos para estimar coeficientes significativos.
- Si dos variables predictoras están altamente correlacionadas, solo una puede ser incluida en el modelo, lo que puede ser un problema.
- Los valores de R-cuadrado a menudo son excesivamente altos, lo que puede no ser indicativo de una buena calidad del modelo.
- Las pruebas F y chi-cuadrado que se utilizan para evaluar las variables en el modelo no siempre tienen una distribución adecuada.
- Los valores pronosticados y los intervalos de confianza pueden ser demasiado estrechos
- Los valores P pueden tener un significado incorrecto y los coeficientes de regresión pueden estar sesgados.
- La colinealidad, es decir, la alta correlación entre las variables predictoras, puede dar problemas.

  (Wang & L.Jain, 2003)

2.6.2.3. Tipos

La regresión *Stepwise* implica la inclusión o exclusión de variables independientes en un modelo de regresión según su significancia estadística. Esto se puede llevar a cabo probando una a una las variables independientes y agregándolas si son significativas o incluyendo todas las posibles variables en el modelo y eliminando las no significativas. También se puede utilizar una combinación de ambos enfoques, dando lugar a tres enfoques distintos:

1. **La selección hacia adelante** comienza sin variables en el modelo, prueba cada variable a medida que se agrega al modelo y luego conserva las que se consideran estadísticamente más significativas, repitiendo el proceso hasta que los resultados sean óptimos.
2. **La eliminación hacia atrás** comienza con un conjunto de variables independientes, eliminando una a la vez y luego probando para ver si la variable eliminada es estadísticamente significativa.
3. **La eliminación bidireccional** es una combinación de los dos primeros métodos que prueban qué variables deben incluirse o excluirse**.**

(Hayes, 2022)

2.6.3. Análisis de componentes principales

2.6.3.1. Definición

El análisis de componentes principales (PCA) es un método estadístico que reduce la complejidad de espacios muestrales con muchas dimensiones, manteniendo su información. Se aplica a una muestra con *n* individuos y *p* variables (X_1, X_2, …, X_p), buscando un número menor de factores subyacentes ($z<p$) que expliquen aproximadamente lo mismo que las *p* variables originales. Los nuevos valores reducidos, conocidos como componentes principales, reemplazan los valores originales para caracterizar a cada individuo.

El análisis de componentes principales pertenece a la familia de técnicas conocida como unsupervised learning. En este caso, la variable respuesta Y no se tiene en cuenta ya que el objetivo no es predecir Y sino extraer información empleando los predictores, por ejemplo, para identificar subgrupos (Rodrigo J. A., 2017).

2.6.3.2. Cálculo de las componentes principales

Cada componente principal (Z_i) se obtiene por combinación lineal de las variables originales. La primera componente principal de un grupo de variables (X_1, X_2, …, X_p) es la combinación lineal normalizada de dichas variables que tiene mayor varianza:

$$Z_1 = \emptyset_{11}X_1 + \emptyset_{21}X_2 + \cdots + \emptyset_{P1}X_P$$

Que la combinación lineal sea normalizada implica que:

$$\sum_{j=1}^{p} \emptyset_{j1}^2 = 1$$

Los términos $\emptyset_{11}$, …, $\emptyset_{1p}$ reciben en el nombre de *loadings* y son los que definen a la componente $\emptyset_{11}$ es el *loading* de la variable X_1 de la primera componente principal. Los *loadings* pueden interpretarse como el peso/importancia que tiene cada variable en cada componente y, por lo tanto, ayudan a conocer qué tipo de información recoge cada una de las componentes.

Dado un set de datos X con n observaciones y p variables, se centralizan las variables para conseguir que todas tengan media cero y se resuelve un problema de optimización para encontrar el valor de los *loadings* con los que se maximiza la varianza. Una forma de resolver esta optimización es mediante el cálculo de *eigenvector-eigenvalue* de la matriz de covarianzas.

Una vez calculada la primera componente (Z_1) se calcula la segunda (Z_2) repitiendo el mismo proceso, pero añadiendo la condición de que la combinación lineal no pude estar correlacionada con la primera componente. Esto equivale a decir que Z_1 y Z_2 tienen que ser perpendiculares. El proceso se repite de forma iterativa hasta calcular todas las posibles componentes (min(n-1, p)) o hasta que se decida detener el proceso. El orden de importancia de las componentes viene dado por la magnitud del *eigenvalue* asociado a cada *eigenvector* (Rodrigo J. A., 2017).

2.6.4. LASSO

2.6.4.1. Regresión lineal

El modelo de regresión lineal (Legendre, Gauss, Galton y Pearson) considera que, dado un conjunto de observaciones $\{y_i, x_{i1}, \dots, x_{np}\}_{i=1}^{n}$ la media μ de la variable a predecir *y* tiene una relación lineal con la o las regresoras $X_1 \dots X_p$ tal que:

$$\mu_y=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p$$

La línea de regresión poblacional es el resultado de la ecuación anterior que muestra la relación entre las variables predictoras y la media de la variable a predecir.

Una descripción adicional que es comúnmente hallada en textos de estadística es:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i$$

Donde, se hace referencia al valor que toma *y* para una observación *i* concreta. Los resultados de una observación específica nunca serán iguales al promedio, por eso se añade el término de error ϵ.

Asimismo, se considera la misma interpretación para ambos casos:

- β_0: es la ordenada en el origen, hace referencia al valor promedio *y* cuando las variables explicativas son cero.
- β_j: es el efecto promedio que tiene sobre *y* el incremento en una unidad de la predictora x_j, siendo constantes el resto de las variables. También conocidos como coeficientes parciales de regresión.
- ϵ: es el residuo o error, la diferencia entre el valor real y el predicho por el modelo. Recoge la influencia sobre *y* de las variables que no se incluyen en el modelo.

Mayoritariamente, los valores β_0 y βj poblacionales no se conocen, por ello se obtienen sus estimaciones a partir de una muestra. El ajuste del modelo implica la estimación de los coeficientes de regresión que mejor se ajusten a los datos, maximizando la probabilidad de que el modelo origine los valores observados. El método más utilizado para el ajuste es el de mínimos cuadrados ordinarios (OLS).

En un modelo de regresión, los coeficientes de cada predictor indican su impacto en la variable respuesta. Sin embargo, las unidades de medida utilizadas para cada predictor pueden afectar la magnitud de los coeficientes, lo que dificulta la comparación de su importancia relativa. Por esto, se recomienda estandarizar todas las variables predictoras antes de ajustar el modelo. De esta manera, los coeficientes se pueden comparar directamente, y un coeficiente cercano a cero indica que el predictor tiene una influencia menor en la variable respuesta en comparación con otros predictores (Rodrigo J. A., 2020).

2.6.4.2. Regularización en regresión lineal

Las estrategias de regularización consisten en la incorporación de términos de penalización en el ajuste por mínimos cuadrados ordinarios (OLS) con el fin de evitar el sobreajuste del modelo, reducir la varianza, disminuir el impacto de la correlación entre los predictores y minimizar la influencia de aquellos predictores que son menos relevantes en el modelo. Como resultado de la aplicación de la regularización, se logran modelos con un mejor poder predictivo y capacidad de generalización (Rodrigo J. A., 2020).

2.6.4.3. Método LASSO

La regularización Lasso penaliza la suma del valor absoluto de los coeficientes de regresión $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$. A esta penalización se le conoce como *l1* y tiene el efecto de forzar a que los coeficientes de los predictores tiendan a cero. Dado que un predictor con coeficiente de regresión cero no influye en el modelo, *lasso* consigue excluir los predictores menos relevantes. El grado de penalización está controlado por el hiperparámetro λ. Cuando λ=0, el resultado es equivalente al de un modelo lineal por mínimos cuadrados ordinarios. A medida que λ aumenta, mayor es la penalización y más predictores quedan excluidos.

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j| = suma\ residuos\ cuadrados + \lambda \sum_{j=1}^{p}|\beta_j|$$

2.7. Crítica y propuesta al marco contextual

2.7.1. Trabajos previos relacionados

Internet forma parte de nuestro día a día desde hace ya unos años. En él se almacena toda clase de deseos, preocupaciones, pensamientos, etc.

La actividad en línea genera diariamente una gran cantidad de información, que deja un rastro digital en su camino. Esto constituye una ventaja para pronosticar indicadores económicos, ya que se puede obtener un análisis de estos en tiempo real y esto proporciona mayor capacidad de

maniobra. Una de las herramientas más populares que permite acceder a tal información es Google Trends.

Numerosos estudios han profundizado sobre la utilización de los datos extraídos de Google Trends para el pronóstico de variables sociales y económicas. Algunos de ellos, como (Varian & Choi, 2012) tienen un estudio en el que el objetivo es prever a corto plazo los valores de indicadores económicos tales como la tasa de desempleo, la confianza del consumidor y las ventas de automóviles.

(Guzmán, 2011) ha examinado los datos de Google como predictor de la inflación. (Vosen & Schmidt, 2011), por ejemplo, muestran que los datos de búsqueda de Google superan a los indicadores basados en encuestas en la previsión del consumo privado de Estados Unidos.
Basándose en esto, (Woo & Owen, 2019) documentan además la utilidad de incorporar datos de búsqueda en línea en modelos de previsión del consumo privado de Estados Unidos.

Más aún, algunos estudios han explorado la utilidad de la herramienta para la predicción de desempleo, incluyendo (Smith, 2016) para el Reino Unido, (González-Fernández & González-Velasco, 2018) para España, y (Maas, 2019) para los Estados Unidos.
(Narita & Yin, 2018) muestran que los datos de GSV se correlacionan significativamente con variables macroeconómicas como el PIB real, la inflación y los flujos de capital en los países en desarrollo de bajos ingresos.

Y otros, como (Eichenauer, Indergand, & Martínez, 2021) construyeron un índice de sentimiento económico diario (DESI). Se pretendía que este indicador fuese robusto, por ello se basaron en tres supuestos:

1. La información obtenida mensualmente proporciona la tendencia de la actividad de búsqueda con mayor precisión a largo plazo.
2. El análisis de las búsquedas realizadas durante algunas semanas resulta más adecuado con los datos semanales.
3. Para analizar el comportamiento a corto plazo de las búsquedas durante varios días se consideran más útiles los datos diarios.

Con el objetivo de lograr una mayor precisión en el análisis a corto plazo, se optó por desagregar temporalmente los datos semanales mediante el uso del marco estadístico propuesto por (Chow & Lin, 1971). Este método es ampliamente utilizado en estadística y permite descomponer la serie de menor a mayor frecuencia, utilizando como indicador la serie diaria.

2.7.2. Propuesta

El presente marco contextual genera una oportunidad de investigación que se busca completar o, en su defecto, obtener información adicional que contribuya a una mejor comprensión de la utilidad de Google Trends en el ámbito económico y determinar los límites de su alcance.

Los trabajos revisados anteriormente han encontrado una relación positiva entre las series de GT y los indicadores económicos. La selección de estas series se ha hecho de forma diversa en los distintos estudios. Algunos han utilizado componentes principales, otros han seleccionado series de forma intuitiva, sin dar mayor explicación sobre ello. La correlación entre las series de GT y los indicadores subyacentes pueden cambiar en el tiempo. Por ejemplo, hoy en día puede haber preocupación por la inflación y esto se refleja en el sentimiento económico. Sin embargo, hacía más de 15 años que la inflación no era una preocupación en España. Por este motivo, resulta conveniente estudiar métodos para automáticamente seleccionar palabras clave relacionadas con el indicador económico subyacente.

Este TFG pretende aportar en esta línea con la evaluación de tres métodos de selección y combinación de series temporales para la predicción a corto plazo de tres indicadores económicos.

3. METODOLOGÍA

3.1. Muestra

Se ha determinado que, con el objetivo de mantener la coherencia de Google Trends, se tomarán los datos correspondientes a los tres indicadores desde el 01/01/2007 hasta el 01/02/2023. La frecuencia de los datos con la que se trabajará es mensual. Esta decisión se detalla en el presente capítulo.

3.1.1. Índice de Sentimiento Económico

En una primera instancia, se han obtenido los datos correspondientes a ESI directamente desde la página web de la Comisión Europea utilizando RStudio. Para ello, se emplea el enlace que contiene el archivo comprimido con los datos, el cual es descargado mediante las funciones download.file() y unzip(), permitiendo su importación directa a la terminal de R. Se presenta la serie original en la siguiente figura.

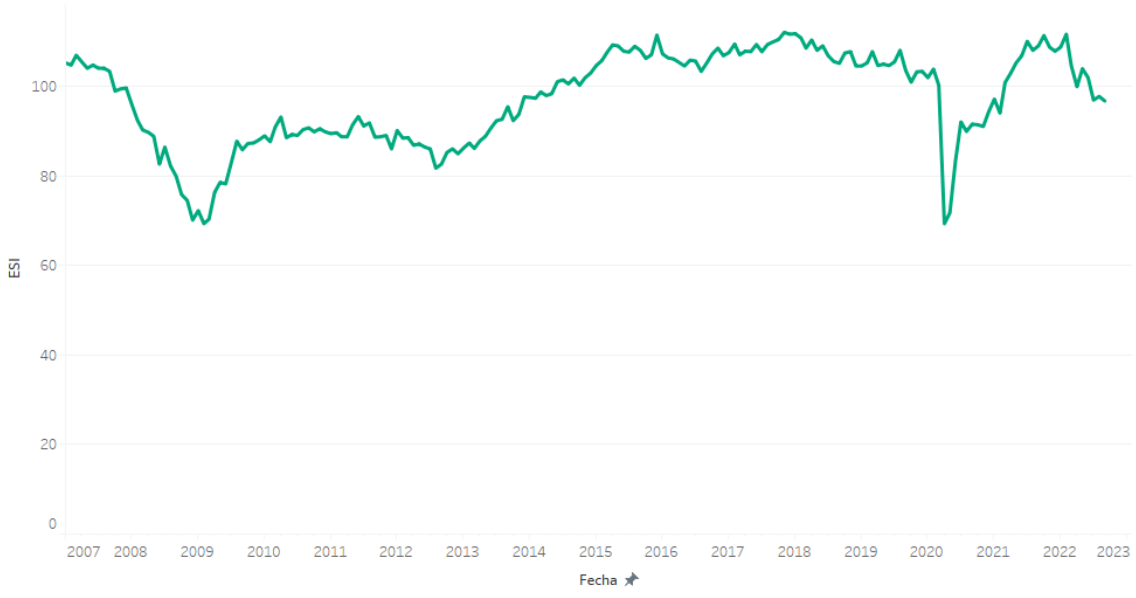

Figura 5. Serie original ESI. Fuente: Elaboración propia

3.1.2. Desempleo

Con el fin de obtener la serie de desempleo, se ha tomado como referencia los datos del Servicio de Empleo Público Estatal (SEPE) correspondientes al paro registrado. Para ello, se han descargado los datos en formato '.xlsx' desde el año 2001 hasta la actualidad, aunque posteriormente solo se utiliza los correspondientes al periodo comprendido entre 2007 y 2023, tal y como se ha mencionado en el apartado 3.1.

Los datos iniciales descargados presentan el siguiente formato:

Servicio Público de Empleo Estatal

Evolución del paro registrado según sectores. Últimos 10 años.

		TOTAL	AGRICULTURA	INDUSTRIA	CONSTRUCCIÓN	SERVICIOS	SIN EMPLEO ANT.
2014	ENERO	4.814.435	208.174	512.531	649.211	3.071.282	373.237
	FEBRERO	4.812.486	216.083	507.583	643.061	3.067.530	378.229
	MARZO	4.795.866	230.937	502.018	629.169	3.046.322	387.420
	ABRIL	4.684.301	224.699	493.736	617.966	2.961.616	386.284
	MAYO	4.572.385	215.807	479.471	593.772	2.896.348	386.987
	JUNIO	4.449.701	220.465	463.961	574.631	2.812.743	377.901
	JULIO	4.419.860	220.889	454.163	559.917	2.800.225	384.666
	AGOSTO	4.427.930	213.995	459.943	560.079	2.815.386	378.527
	SEPTIEMBRE	4.447.650	199.139	453.223	548.465	2.856.994	389.829
	OCTUBRE	4.526.804	223.745	456.266	539.490	2.918.218	389.085
	NOVIEMBRE	4.512.116	215.165	450.962	530.425	2.927.158	388.406
	DICIEMBRE	4.447.711	212.526	453.397	543.114	2.861.883	376.791
2015	ENERO	4.525.691	228.384	452.644	535.257	2.938.627	370.779
	FEBRERO	4.512.153	228.851	446.109	525.166	2.938.404	373.623
	MARZO	4.451.939	224.790	439.216	516.319	2.889.380	382.234
	ABRIL	4.333.016	209.571	427.661	496.870	2.816.496	382.418
	MAYO	4.215.031	195.429	414.787	479.350	2.747.670	377.795
	JUNIO	4.120.304	202.456	400.648	467.644	2.685.783	363.773
	JULIO	4.046.276	200.131	389.367	457.133	2.641.480	358.165
	AGOSTO	4.067.955	194.167	395.169	461.776	2.664.356	352.487
	SEPTIEMBRE	4.094.042	181.720	391.140	451.874	2.707.511	361.797
	OCTUBRE	4.176.369	203.315	394.046	448.039	2.768.583	362.386
	NOVIEMBRE	4.149.298	196.162	388.735	437.821	2.767.128	359.452

Figura 6. Formato inicial de los datos de desempleo. Fuente: Elaboración propia

Ya que el archivo en cuestión contiene información redundante y, los datos no se encuentran en el formato adecuado para su procesamiento en R, se efectúa una sencilla transformación de estos: se sustituye el mes por la fecha completa de interés y se toma la columna 'total' como la variable que indica el desempleo. Como resultado, se obtiene una columna que muestra la fecha junto con el correspondiente indicador.

	A	B
1	Fecha	Desempleo
2	01/01/2007	2.082.508
3	01/02/2007	2.075.275
4	01/03/2007	2.059.451
5	01/04/2007	2.023.124
6	01/05/2007	1.973.231
7	01/06/2007	1.965.869
8	01/07/2007	1.970.338
9	01/08/2007	2.028.296
10	01/09/2007	2.017.363
11	01/10/2007	2.048.577
12	01/11/2007	2.094.473
13	01/12/2007	2.129.547
14	01/01/2008	2.261.925
15	01/02/2008	2.315.331
16	01/03/2008	2.300.975
17	01/04/2008	2.338.517
18	01/05/2008	2.353.575

Figura 7. Formato final de los datos de desempleo. Fuente: Elaboración propia

La serie original de desempleo presenta la siguiente forma.

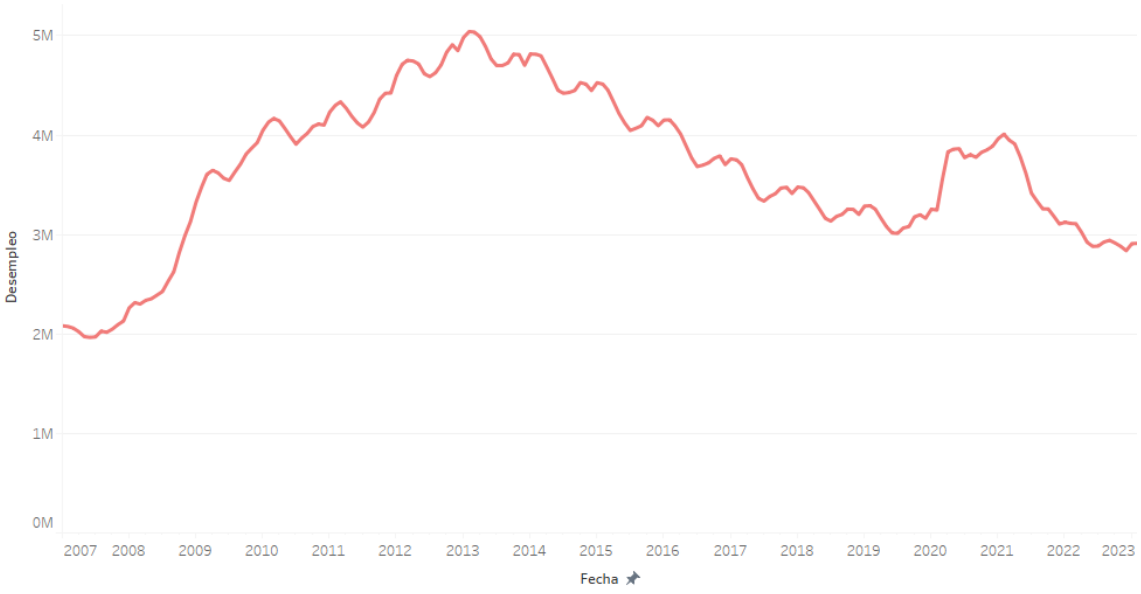

Figura 8. Serie original desempleo. Fuente: Elaboración propia

3.1.3. Inflación

En el caso de la serie de inflación, se toma como referencia los datos estatales de IPC proporcionados por el Instituto Nacional de Estadística (INE), también en extensión '.xlsx' y se realiza un preprocesado y tratamiento de datos similar al aplicado con paro registrado.

Primeramente, se tienen los datos de la siguiente forma:

	A	B	C	D	E	F	G	H
1	**Resultados nacionales**							
2	**Índices nacionales**							
3								
4	**Índices nacionales: general y de grupos ECOICOP**							
5	**Unidades: Índice**							
6								
7		**Índice**						
8		**2023M03**	**2023M02**	**2023M01**	**2022M12**	**2022M11**	**2022M10**	**2022M09**
9	**Índice general**		110,703	109,668	109,899	109,734	109,866	109,498
10								
11								
12	**Notas:**							
13								
14	**Fuente:**							
15	Instituto Nacional de Estadística							
16								
17								
18								

Figura 9. Formato inicial de los datos de IPC. Fuente: Elaboración propia

Con el objetivo de lograr una mayor consistencia y coherencia en el análisis, se busca obtener el mismo formato que se utiliza en paro registrado:

	A	B
1	FECHA	IPC
2	01/01/2007	81,129
3	01/02/2007	81,184
4	01/03/2007	81,800
5	01/04/2007	82,930
6	01/05/2007	83,158
7	01/06/2007	83,311
8	01/07/2007	82,704
9	01/08/2007	82,818
10	01/09/2007	83,090
11	01/10/2007	84,167
12	01/11/2007	84,770
13	01/12/2007	85,125
14	01/01/2008	84,598
15	01/02/2008	84,730
16	01/03/2008	85,482
17	01/04/2008	86,402
18	01/05/2008	86,985

Figura 10. Formato final de los datos de IPC. Fuente: Elaboración propia

La figura siguiente muestra la serie original del IPC.

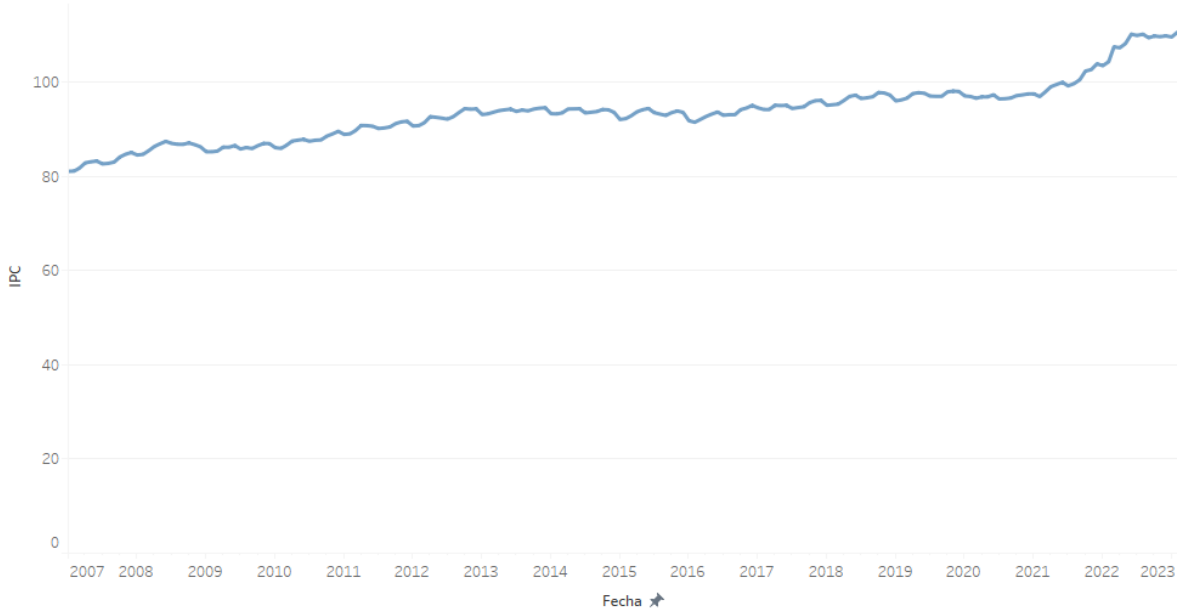

Figura 11. Serie original IPC. Fuente: Elaboración propia

3.2. Términos de búsqueda en Google Trends

Como se menciona en las secciones anteriores, el enfoque consiste en entrenar los modelos utilizando series temporales obtenidas de Google Trends. Por tanto, debe escogerse a priori una serie de términos de búsqueda relacionados con los indicadores a predecir.

En un principio, los términos se han seleccionado en base a los cuestionarios mediante los que se construye ESI. Sin embargo, puesto que la mera elección de estos términos no es objeto de estudio de este trabajo y, se puede ver que tienen correlación con los otros dos indicadores también, se decide utilizar los mismos términos de búsqueda para los tres indicadores económicos. Estos son Trabajo a tiempo parcial, Emprender, Cobrar paro, Paro, Desempleo, Crisis, Recesión económica, Recesión, Inflación, Materia Prima, Construcción, Quiebra, Bancarrota, Plan de ahorro, Ahorrar, Comprar, Amazon, Idealista, LinkedIn, InfoJobs, Hipoteca, Rebajas.

Cabe destacar que, los datos de GT están disponibles desde 2004, pero a partir de 2007 Google mejoró la calidad de su base de datos. De esta manera, los datos comienzan en enero de 2007 por dos razones comentadas anteriormente por (Eichenauer, Indergand, & Martínez, 2021). En un primer momento, es relevante mencionar que durante este periodo se presentó una notable innovación tecnológica con el lanzamiento del primer iPhone, lo cual generó un significativo

aumento en el uso de internet móvil. Además, la muestra de datos abarca un periodo de 20 meses previos al colapso de Lehman Brothers, considerado como el inicio de la recesión económica global de 2008/2009. De esta manera, se dispone de una amplia ventana temporal que abarca dos de las más importantes recesiones económicas, permitiendo una evaluación más amplia de los indicadores.

3.3. Métodos de evaluación

3.3.1. On sample

El primer método que se utiliza para evaluar los modelos es la evaluación 'On sample'. Esta consiste en evaluar el rendimiento del modelo utilizando los mismos datos que se utilizaron para entrenarlo. Es por ello que se realizan las predicciones y evaluación de todo el conjunto de datos que tenemos con las métricas que se explican en el presente capítulo.

3.3.2. Ventana deslizante

El segundo método utilizado es la evaluación con ventana deslizante. Este método divide el conjunto de datos en 'ventanas' de un tamaño específico y, con este subconjunto de datos entrena el modelo y posteriormente lo evalúa para los h periodos siguientes. Por tanto, el tamaño de ventana representa el número de observaciones consecutivas, la cual dependerá del tamaño de la muestra. El horizonte de pronóstico es el número de periodos que se pronostican a partir de la última observación de entrenamiento y, que depende de la periodicidad de los datos. Y, también se debe elegir cuantos periodos desplazamos la ventana entre periodo y periodo.

En este caso, se prueban distintos tamaños de ventana en combinación con distintos horizontes con el fin de ver que combinación proporciona mejores resultados. Concretamente, se prueban horizontes temporales de 1 a 12, que equivaldría a predecir cuál es el sentimiento económico, desempleo e IPC del próximo mes, hasta dentro de un año. Asimismo, se prueban ventanas desde 5 hasta 10 años.

Las validaciones se harán con las mismas métricas que en la evaluación 'on sample', a excepción del RSQUARE.

3.3.3. Métricas de evaluación

En ambos métodos de evaluación, 'on sample' y ventana deslizante, se validan los modelos con las medidas de error siguientes:

- Error Medio Absoluto (MAE)

  Es el promedio de la diferencia absoluta entre el valor observado y los valores predichos. Cuanto menor sea el nivel de MAE más ajustada estará la predicción del modelo.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- Error Porcentual Medio Absoluto (MAPE)

  Se obtiene al dividir la suma de los errores absolutos de pronóstico por la cantidad total de pronósticos realizados y expresar el resultado como un porcentaje. Al igual que con el MAE, cuanto menor sea el MAPE, más preciso será el modelo.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|} \times 100$$

- Error Cuadrático Medio (RMSE)

Representa a la raíz cuadrada de la distancia cuadrada promedio entre el valor real y el valor pronosticado. Se interpreta igual que el MAE y el MAPE, a menor RMSE, se tiene un mejor modelo.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- R al cuadrado (R-SQUARE)

R^2 mide lo bien que un modelo de regresión se ajusta a los datos reales. En otras palabras, se trata de una medida de la precisión general del modelo. R al cuadrado es también conocido como el coeficiente de determinación. Un elevado R cuadrado indica que el modelo es bueno para realizar predicciones de la variable en cuestión.

$$RSQUARE = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

3.4. Ajuste y configuración de las técnicas de selección de variables

Puesto que se ha explicado en el apartado 2.6 la carga teórica que compone los métodos de selección de variables y reducción de dimensionalidad que se van a utilizar, en este apartado se detallan los criterios y parámetros que se utilizan en dichos métodos.

Tanto en la evaluación "on sample" como en la evaluación mediante ventana deslizante, se siguen los mismos pasos, con la única diferencia de que en la ventana deslizante se realiza la selección de variables y el entrenamiento del modelo en cada ventana.

En el análisis de selección de variables utilizando el enfoque Stepwise, se empleará el método *forward*, dado que es la opción predeterminada de la función *step* en R. Además, se evaluará la adición secuencial de variables al modelo en base al criterio de información de Akaike (AIC).

Por otra parte, en PCA se seguirá la siguiente metodología:

- Escoger aquellos términos de búsqueda cuya correlación con el indicador sea mayor que un umbral predefinido, en este caso 0,5. En caso de no encontrarse términos que cumplan este criterio, se seleccionan las tres variables con la mayor correlación, independientemente de su valor.
- Se determina el número óptimo de componentes utilizando el "criterio del codo". El objetivo es identificar el punto en el que agregar más componentes principales no proporciona una mejora significativa en la varianza explicada. Inicialmente, se establece un umbral de varianza explicada de 0.7.
- Se realiza una regresión lineal utilizando las componentes principales seleccionadas y el indicador económico como variable dependiente.

Por último, en el método Lasso se realiza una previa validación cruzada con 5 ‘folds’ y se utilizará para entrenar el modelo el valor de lambda obtenido mediante el criterio del *1se*. Este criterio se ha mostrado como una opción efectiva para seleccionar un lambda que ofrezca un equilibrio entre ajuste y regularización en el modelo.

4. RESULTADOS

4.1. Introducción

En este apartado se realiza una exposición de los resultados obtenidos con cada modelo, para cada indicador económico. Por claridad del documento, para la evaluación con ventana deslizante se han incluido aquí las gráficas correspondientes a MAPE. Las métricas MAE y RMSE se han calculado e incluido sus gráficos en el Anexo.

4.2. Estimación para ESI

4.2.1. Primer método de evaluación: on sample

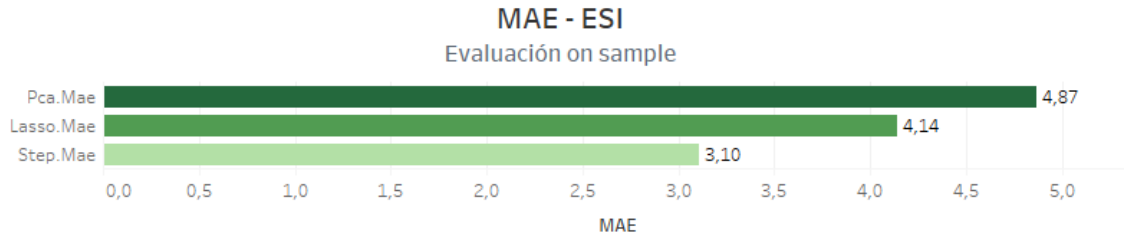

Figura 12. Resultados MAE evaluación 'on sample' para ESI. Fuente: Elaboración propia

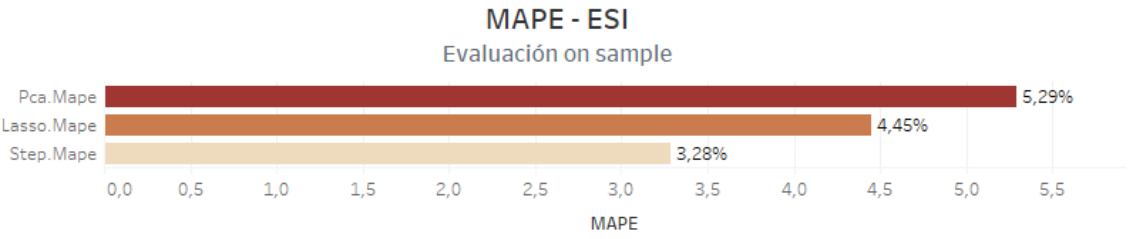

Figura 13. Resultados MAPE evaluación 'on sample' para ESI. Fuente: Elaboración propia

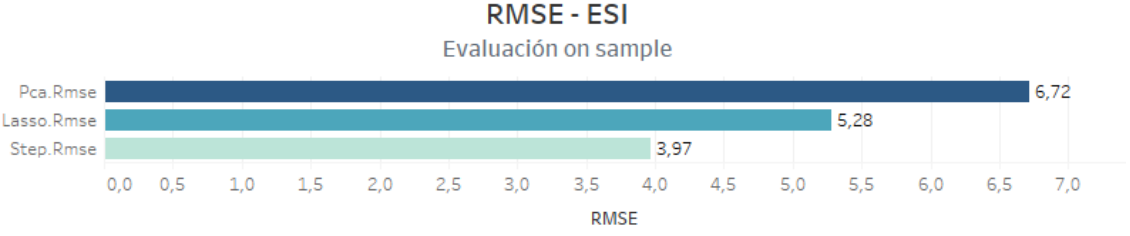

Figura 14. Resultados RMSE evaluación 'on sample' para ESI. Fuente: Elaboración propia

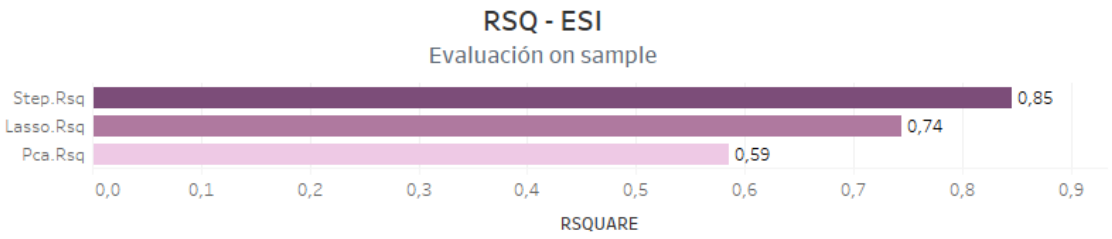

Figura 15. Resultados RSQUARE evaluación 'on sample' para ESI. Fuente: Elaboración propia

En la evaluación 'on sample' para ESI, se observa que Stepwise es el modelo que mejores resultados proporciona, puesto que el error obtenido es menor respecto a Lasso y PCA. Sus métricas son:

- Un MAE de 3,10 significa que, en promedio, las predicciones del modelo difieren en 3,10 unidades de la variable objetivo de los valores reales observados.
- Un valor MAPE de 3,28 % indica que el error promedio porcentual en las predicciones es del 3,28% en relación con los valores reales.
- Un valor RMSE de 3,96 significa que, en promedio, los errores pronosticados difieren de los valores observados en 3,96 unidades en la misma escala de medida que los datos originales.
- RSQUARE de 0,84. Es decir, el modelo es capaz de explicar el 84% de la variación en la variable dependiente a través de las variables independientes incluidas en el modelo.

4.2.1. Segundo método de evaluación: ventana deslizante

4.2.1.1 Modelo 1: Stepwise

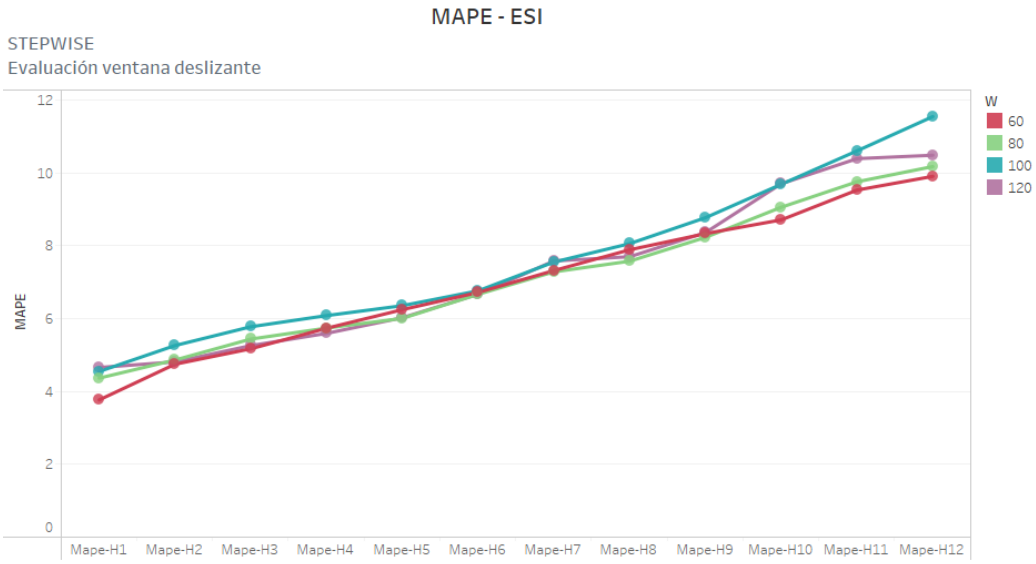

Figura 16. Resultados MAPE- Stepwise evaluación ventana deslizante para ESI. Fuente: Elaboración propia

Con el modelo Stepwise para ESI, generalmente se observa que:

- A medida que aumenta el horizonte, el error de pronóstico aumenta también.
- La tendencia es creciente, para ventanas más grandes, el error es mayor. Esto sugiere que el modelo es capaz de hacer predicciones precisas para el ESI en general.
- Con un horizonte de 6 periodos, los resultados son los mismos independientemente del tamaño de ventana.
- El resultado óptimo se alcanza con un tamaño de ventana de 60 y un horizonte de 1, obteniendo un MAPE de 3,9% aproximadamente.

4.2.1.2. Modelo 2: PCA + Regresión lineal

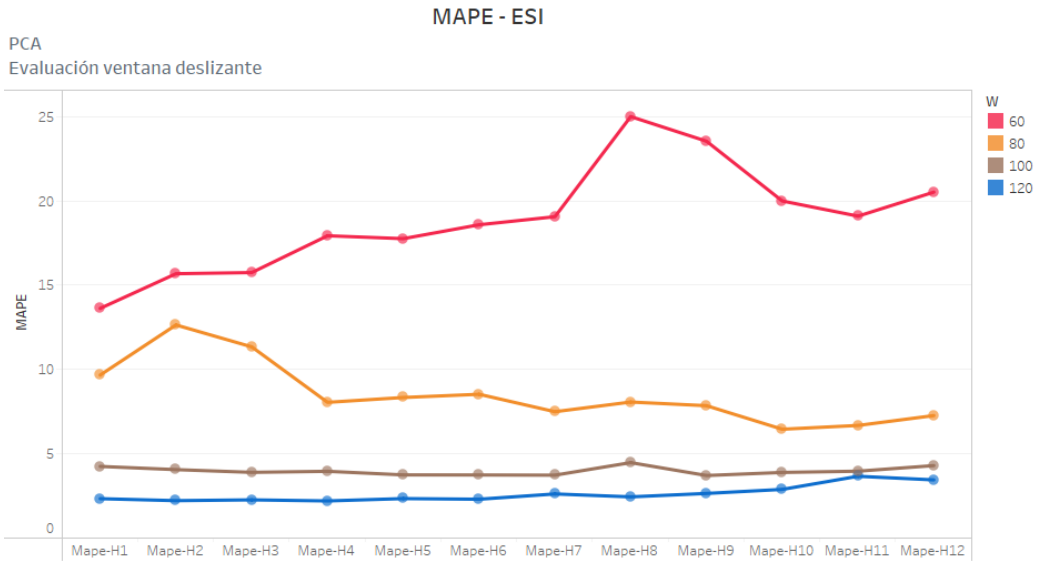

Figura 17. Resultados MAPE – PCA evaluación ventana deslizante para ESI. Fuente: Elaboración propia

Con el modelo PCA + regresión lineal para ESI, se observa:

- En general, los valores del MAPE disminuyen a medida que el tamaño de la ventana aumenta. Esto indica que el modelo tiende a mejorar sus predicciones cuando se utiliza más información histórica.
- Para un tamaño de ventana de 100 y 120 el error no oscila tanto como con ventanas de tamaño 60 y 80.
- Con este modelo el horizonte que se escoja no es relevante para ventanas de tamaño 100 y 120.
- En general, los valores óptimos se obtienen con una ventana de tamaño 120, con un MAPE de aproximadamente 2%.

4.2.1.3. Modelo 3: Lasso

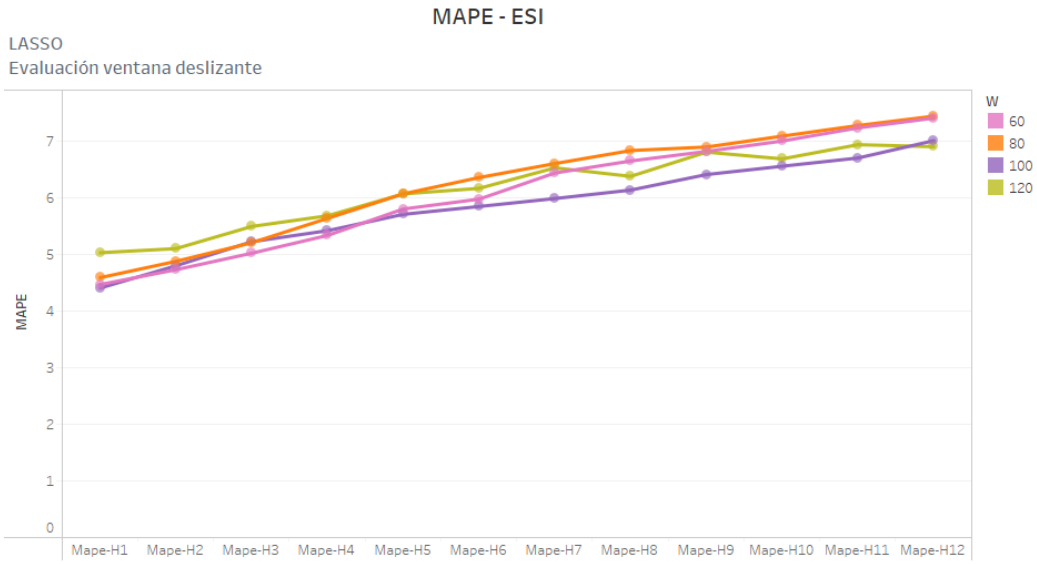

Figura 18. Resultados MAPE – Lasso evaluación ventana deslizante para ESI. Fuente: Elaboración propia

Para regularización Lasso con ESI, se obtiene una validación similar a Stepwise:

- Generalmente, a medida que aumentamos la ventana aumenta el error.
- Lo mismo ocurre con los horizontes, para horizontes mayores, se tiene un error mayor.
- No hay mucha variación en el error en cuanto a los distintos tamaños de ventana conforme se cambia de horizonte.
- El mejor resultado se obtiene con una ventana de tamaño 60 o 100 y con horizonte 1, con un MAPE de 4,4% aproximadamente.

4.3. Estimación para Desempleo

4.3.1. Primer método de evaluación: on sample

Figura 19. Resultados MAE evaluación 'on sample' para desempleo. Fuente: Elaboración propia

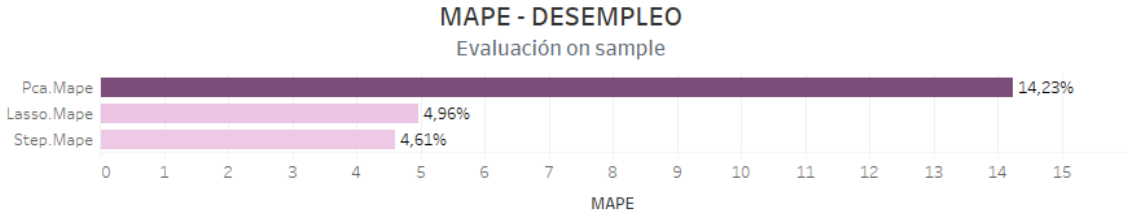

Figura 20. Resultados MAPE evaluación 'on sample' para desempleo. Fuente: Elaboración propia

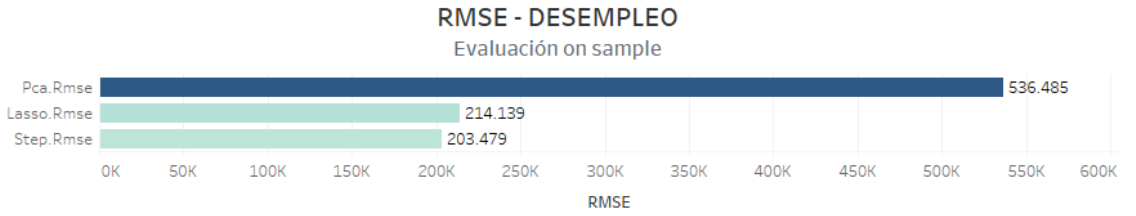

Figura 21. Resultados RMSE evaluación 'on sample' para desempleo. Fuente: Elaboración propia

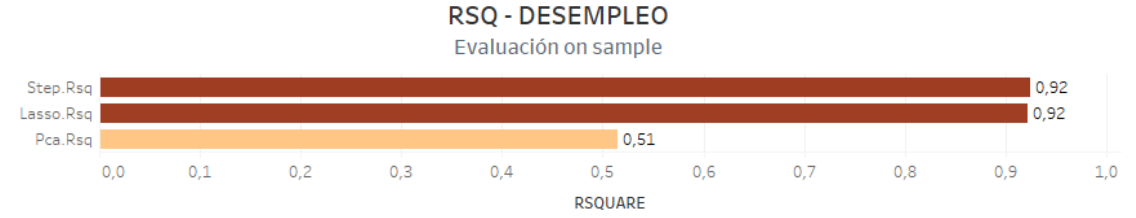

Figura 22. Resultados RSQUARE evaluación 'on sample' para desempleo. Fuente: Elaboración propia

En la evaluación on sample para Desempleo, de nuevo se observa que Stepwise es el modelo que mejores resultados proporciona, puesto que se obtiene un error menor respecto a los otros modelos. Sin embargo, la diferencia entre Lasso y Stepwise en este caso no es significativa.

A la hora de interpretar los resultados para este indicador con MAE y RMSE, hay que tener en cuenta la magnitud de los datos reales, cuya media es de 3 millones.

- Un valor de MAE de 162.327 significa que, en promedio, las predicciones del modelo difieren en 162.327 unidades de la variable objetivo de los valores reales observados.
- MAPE de 4,61 % indica que el error promedio porcentual en las predicciones es del 4,61% en relación con los valores reales.
- RMSE de 203.470 significa que, en promedio, los errores pronosticados difieren de los valores observados en 3,96 unidades en la misma escala de medida que los datos originales.
- RSQ de 0,92. Es decir, el modelo es capaz de explicar el 92% de la variación en la variable dependiente a través de las variables independientes incluidas en el modelo.

4.3.2. Segundo método de evaluación: ventana deslizante

4.3.1.1. Modelo 1: Stepwise

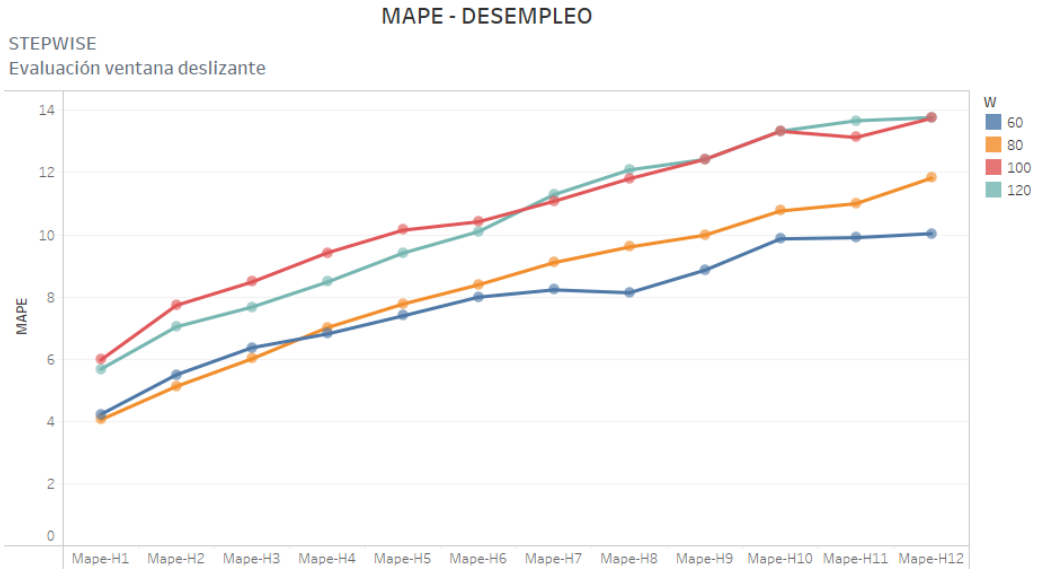

Figura 23. Resultados MAPE – Stepwise evaluación ventana deslizante para desempleo. Fuente: Elaboración propia

Con el modelo Stepwise para Desempleo, generalmente se observa que:

- A medida que aumenta el horizonte, el error de predicción también aumenta.
- Con tamaños de ventana mayores, un error mayor. A excepción de los primeros horizontes, donde un tamaño de ventana de 80 proporciona un error menor a 60 y un tamaño de 120 proporciona un error menor a 100.
- A partir del horizonte 7, es indiferente tomar un tamaño de ventana de 100 o de 120.
- El resultado óptimo se encuentra con un tamaño de ventana de 80 y un horizonte de 1, obteniendo un MAPE de 4,1% aproximadamente.

4.3.1.2. Modelo 2: PCA + Regresión lineal

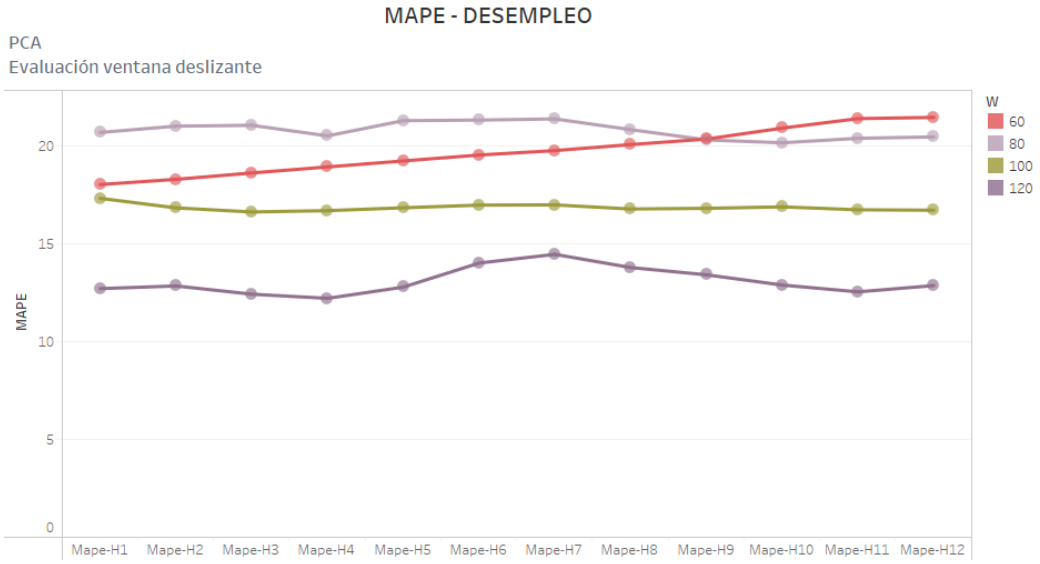

Figura 24. Resultados MAPE – PCA evaluación ventana deslizante para desempleo. Fuente: Elaboración propia

Con el modelo PCA + regresión lineal para Desempleo, se obtiene que:

- Por lo común, los valores del MAPE disminuyen a medida que el tamaño de la ventana aumenta. Es decir, el modelo mejora cuando utilizamos más información para entrenarlo.
- El horizonte no influye mucho en el error de pronóstico.
- El mejor resultado se consigue con una ventana de 120 y horizonte 4, con un MAPE aproximadamente de 12%. Algo más alto que en los otros modelos.

4.3.1.3. Modelo 3: Lasso

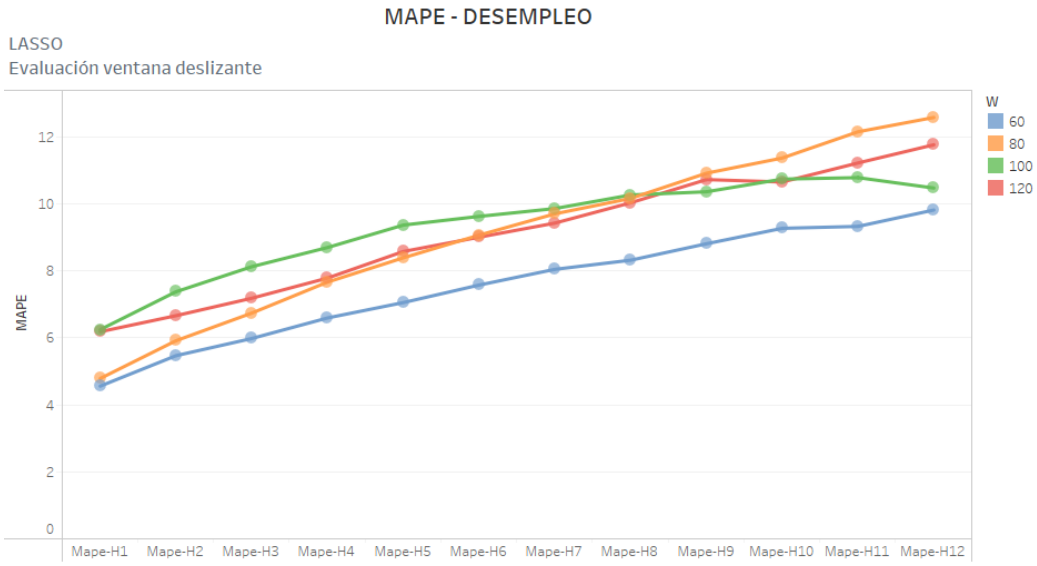

Figura 25. Resultados MAPE – Lasso evaluación ventana deslizante para desmepleo. Fuente: Elaboración propia

Para regularización Lasso con Desempleo, se observa que:

- Para ventanas de tamaño 60 el error es menor en comparación a los otros tamaños de ventana, teniendo el valor óptimo con horizonte 1 y aproximadamente un MAPE de 4,2%.
- El criterio a mayor ventana mayor error no reside, puesto que una ventana de 80 en horizontes de 8 a 12 proporciona un error peor respecto a las ventanas de tamaño 100 y 120.
- A mayor horizonte, error mayor. Excepto para la ventana de tamaño 100. Esto significa que el modelo no predice muy bien valores a largo plazo.

4.4. Estimación para IPC

4.4.1. Primer método de evaluación: on sample

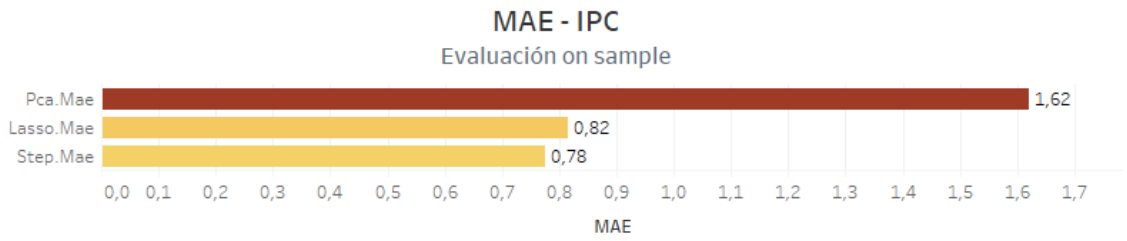

Figura 26. Resultados MAE evaluación 'on sample' para IPC. Fuente: Elaboración propia

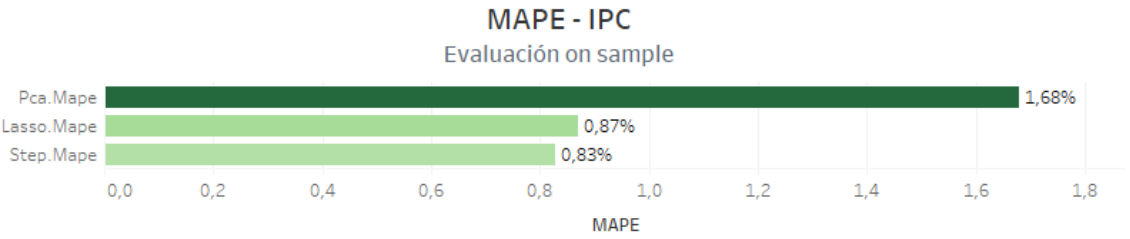

Figura 27. Resultados MAPE evaluación 'on sample' para IPC. Fuente: Elaboración propia

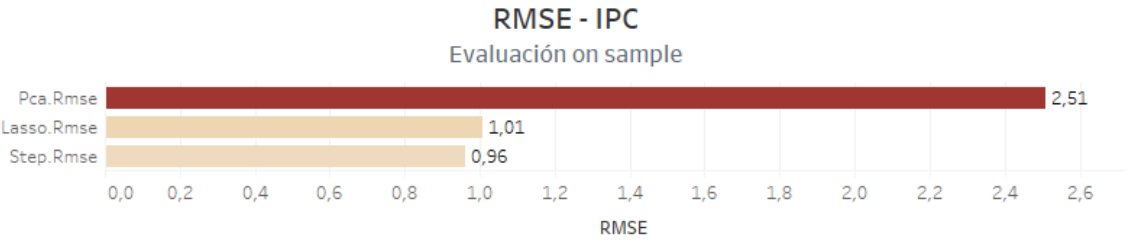

Figura 28. Resultados RMSE evaluación 'on sample' para IPC. Fuente: Elaboración propia

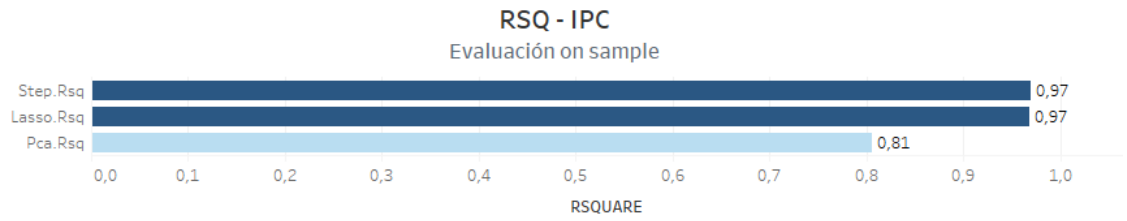

Figura 29. Resultados RSQUARE evaluación 'on sample' para IPC. Fuente: Elaboración propia

En la evaluación on sample para IPC, Stepwise también es el modelo que tiene un error más bajo. Además, este indicador es el que mejores resultados proporciona en todos los modelos, cuyas métricas son:

- Un valor de MAE de 0,78, lo que significa que, en promedio, los valores predichos difieren en menos de una unidad de los valores reales.
- Un MAPE de 0,83 % indica que el error promedio porcentual en los valores predichos es del 0,83% respecto a los valores reales.
- RMSE de 0,96 significa que, en promedio, los valores predichos difieren de los valores observados en 0,96 unidades.
- RSQ de 0,97 tanto en Lasso como en Stepwise. Es decir, estos modelos son capaces de explicar el 97% de la variabilidad de los datos.

4.4.2. Segundo método de evaluación: ventana deslizante

4.4.2.1. Modelo 1: Stepwise

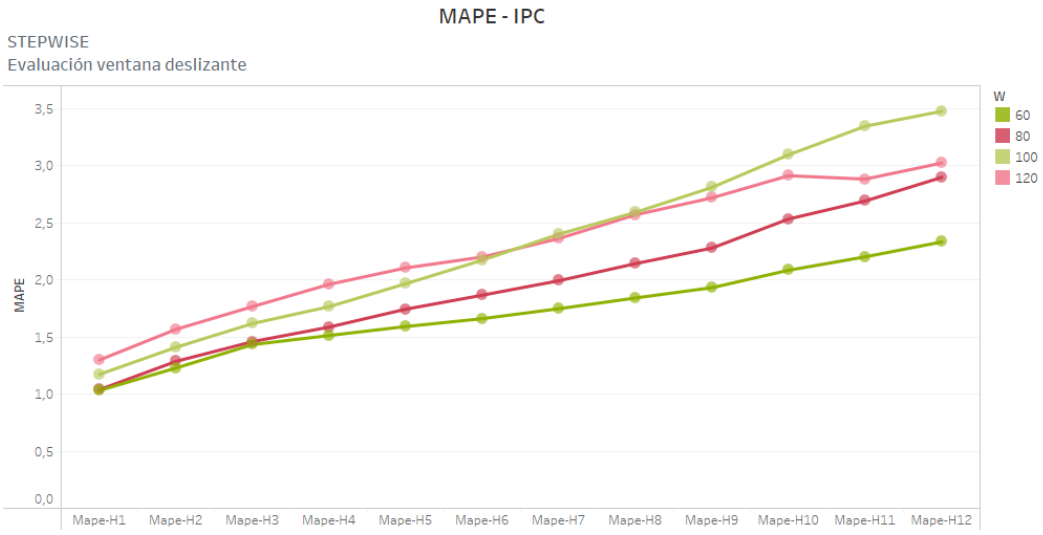

Figura 30. Resultados MAPE - Stepwise evaluación ventana deslizante para IPC. Fuente: Elaboración propia

Con el modelo Stepwise para IPC, de manera general se observa que:

- A medida que aumenta el horizonte, el error de pronóstico también aumenta.
- Con tamaños de ventana mayores, un error mayor. Excepto para la ventana de tamaño 100 que supera a todos los demás tamaños.

- Conforme se aumenta el horizonte, la variación del error es mayor en los distintos tamaños de ventana.
- El resultado óptimo se alcanza con un tamaño de ventana de 60 u 80 y un horizonte de 1, obteniendo un MAPE de 1% aproximadamente.

4.4.2.2. Modelo 2: PCA + Regresión lineal

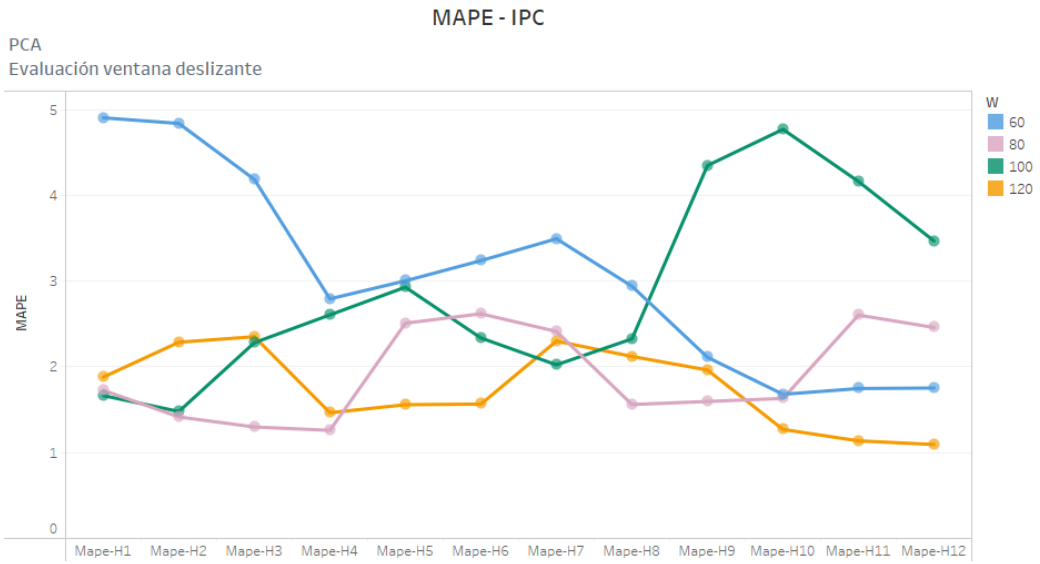

Figura 31. Resultados MAPE - PCA evaluación ventana deslizante para IPC. Fuente: Elaboración propia

Con el modelo PCA + regresión lineal para IPC, se tienen unos resultados bastante variados:

- Para una ventana de tamaño 60, si aumentamos el horizonte, el error disminuye.
- Para el resto de ventanas, el valor del error dependerá del horizonte que se tome, no se tiene ninguna tendencia.
- Aun así, los resultados óptimos se proporcionan con una ventana de 120 y horizonte 12, con un MAPE de 1% aproximadamente. A pesar de ser el indicador que menor error tiene de pronóstico, este caso no se había observado en indicadores anteriores.

4.4.2.3. Modelo 3: Lasso

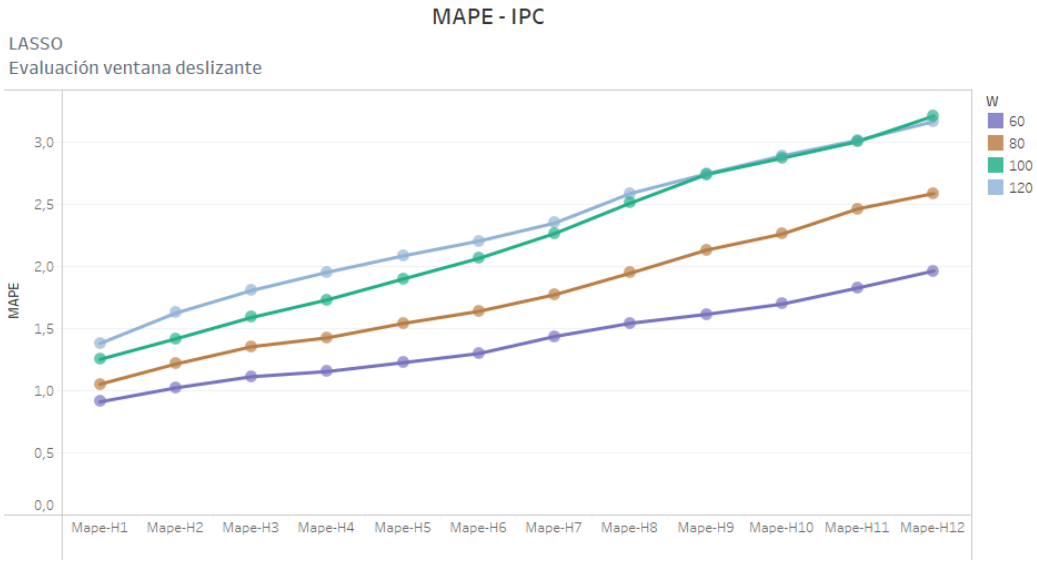

Figura 32. Resultados MAPE - Lasso evaluación ventana deslizante para IPC. Fuente: Elaboración propia

Para regularización Lasso con IPC, se observa que:

- Cuanto menor es la ventana, menor es el error de pronóstico.
- Si se aumenta el horizonte, el error aumenta también.
- A partir del horizonte 9, las ventanas 100 y 120 no difieren en sus resultados.
- El 'mejor' error se obtiene con tamaño de ventana 6 y horizonte 1, con un MAPE de 0,9% aproximadamente.

5. CONCLUSIONES

5.1. Síntesis del TFG

El Trabajo Fin de Grado en cuestión y toda la literatura revisada, evidencia la adecuación del uso de Google Trends para pronosticar indicadores económicos en tiempo real en España. Lo cual coincide con investigaciones previas en este campo.

Asimismo, se han estudiado, construido y evaluado varios métodos de combinación de series temporales con los que se obtiene un buen pronóstico de los indicadores, tomando como datos de entrada GT. En los tres indicadores estudiados, la métrica MAPE oscila entre 1% y 14% para Stepwise, entre 3 y 25% para PCA y entre 4,5 y 13% para Lasso en la evaluación con ventana deslizante. Por ello, se afirma que se encuentran modelos con un error aceptable en el pronóstico de los indicadores. Generalmente, el error de pronóstico más bajo se obtiene con Stepwise o Lasso, ventanas pequeñas y horizonte 1. En PCA, a mayor nivel de ventana y horizonte, mejores resultados se obtienen, pero el error a veces deja de ser aceptable.

Esto tiene sentido, ya que los modelos se vuelven menos precisos a medida que se utilizan datos históricos. De igual manera, el error de pronóstico es mayor para horizontes más largos, lo que indica que es más difícil predecir el sentimiento empresarial, el nivel de desempleo y el nivel de inflación a largo plazo. Sin embargo, en modelos como PCA que se necesita mucha información para obtener un modelo preciso, el aumento de la ventana puede aumentar la calidad del modelo.

Debido a todas las desventajas que presenta Stepwise comentadas en el capítulo 2 y la poca variación de resultados que hay entre este y Lasso, se puede afirmar que este último es el mejor modelo del estudio para combinar series de GT y predecir los indicadores.

Por todo lo expuesto, se justifica la relación entre las series temporales que ofrece Google Trends y los indicadores económicos en nuestro país, siendo una fuente de datos que proporciona una oportunidad para medir la evolución de la economía española en su conjunto.

5.2. Análisis del marco legal y ético

Este estudio emplea datos provenientes de la Unión Europea, el INE y el SEPE. Cabe destacar que dichos datos son de carácter público y, por tanto, no deberían generar controversias en lo que respecta a la protección de datos. Asimismo, no existen problemas de interoperabilidad ni de propiedad intelectual.

Por otra parte, es importante mencionar que ninguno de los modelos empleados permite la discriminación por razones de raza, sexo, creencias, ideologías o cualquier otro atributo protegido. Esto se debe a que ninguna de las variables utilizadas en los modelos contempla dichas categorías.

5.3. Relación del trabajo con los estudios cursados

Este trabajo combina habilidades aprendidas en la asignatura de 'Economía digital' con técnicas avanzadas de análisis de datos para explorar las relaciones entre las búsquedas en Google y las variables económicas y sociales. Además, se explora cómo estas relaciones pueden utilizarse para realizar predicciones útiles y tomar decisiones informadas en la economía y sociedad.

Por otro lado, se utilizan modelos estadísticos como PCA, aprendido en las asignaturas ‘Modelos descriptivos y predictivos I’ y ‘Modelos descriptivos y predictivos II’ y otros como Regresión lineal y Stepwise, los cuales se enseñan en ‘Métodos estadísticos para la toma de decisiones I’ y ‘Métodos estadísticos para la toma de decisiones II’. Sin embargo, la técnica de regularización Lasso ha sido aprendida en el trascurso de realización del trabajo.

Para las técnicas de visualización y *storytelling*, se han utilizado conocimientos adquiridos en la asignatura ‘Visualización de datos’, demostrando un buen dominio de la herramienta Tableau, la cual ha sido utilizada también en las prácticas de empresa.

Además, para llevar a cabo un proyecto de tal magnitud se ha tomado como referencia la estrategia, planificación y gestión de los proyectos realizados en las asignaturas ‘Proyecto I’, ‘Proyecto II’ y ‘Proyecto III’.

También se ha utilizado el lenguaje de programación R, adquirido en varias de las asignaturas nombradas previamente.

Por último, se ha hecho hincapié en muchas de las competencias transversales
obtenidas a lo largo de los 4 años de carrera, como pueden ser 'Aplicación y pensamiento práctico', 'Planificación y gestión del tiempo', 'Innovación y creatividad', 'Análisis y resolución de problemas', 'Diseño y proyecto' y 'Aprendizaje permanente'.

5.4. Legado

Se pone a disposición de aquellos interesados en reproducir o profundizar en el estudio realizado, toda la información necesaria en el siguiente enlace: https://github.com/mcabrod/TFG.git. Para ello, se ha creado un repositorio público en GitHub que contiene todo el código en R utilizado, los datos correspondientes a los tres indicadores analizados, las series de Google Trends descargadas, los términos de búsqueda empleados y los archivos de Tableau que permiten visualizar los gráficos presentados en la memoria del trabajo.

5.5. Limitaciones del trabajo

Las limitaciones que presenta este trabajo, son las mismas que se han experimentado en investigaciones previas. La primera de ellas es inherente al propio uso de datos de GT: no se conoce la manera en la que Google realiza el submuestreo aleatorio para calcular el índice, por lo que no es posible cuantificar el error de muestreo.

Por otro lado, como se ha comentado en el apartado 2.1.3. existen métodos de ajuste por estacionalidad que quedan fuera del alcance de este trabajo, como puede ser X-13ARIMA-SEATS y un mejor tratamiento de la estacionalidad de los datos de GT probablemente mejore la calidad de la predicción.

5.6. Trabajo futuro

Este TFG se sitúa en el ámbito de la economía digital y se enfoca en la aplicación de técnicas de análisis de series temporales utilizando datos obtenidos de la plataforma Google Trends. Esta herramienta ha sido ampliamente utilizada en investigaciones previas para la predicción y comprensión de variables económicas y sociales relevantes. El objetivo principal de este trabajo

es contribuir al avance de esta línea de trabajo y mejorar la comprensión de los patrones de comportamiento de los consumidores y su relación con las variables económicas.

De este modo, se podría seguir el estudio analizando probando otros modelos de predicción, o una combinación de los ya realizados, como PCA + Lasso. También, se podría cambiar los términos de búsqueda y ver si los resultados varían mucho de los proporcionados o incluso seguir variando los propios parámetros de las funciones, probar más tamaños de ventana y distintos horizontes. Y, por supuesto, aplicar el estudio a más indicadores económicos como el Producto Interior Bruto, prima de riesgo, etc. Todos aquellos indicadores que puedan servir de utilidad tanto a nivel público como a nivel empresarial.

6. REFERENCIAS

Referencias

Angel, R. (1993). Curso práctico de contabilidad general superior. Tomo II. Tercera edición. Venezuela.

Banbura, M., & Modugno, M. (2010). Maximum likelihood estimation of factor models on data sets with arbitrary pattern of missing data. *European Central Bank.* Working Paper 1189.

Bańbura, M., Giannone, D., Modugno, M., & Reichlin, L. (2013). Now-casting and the real-time data flow. *Handbook of Economic Forecasting. Vol 2, 195-237.*

Bermejo, I. (2020). Qué es y cómo se calcula el IPC. *Larazon.es*.

Camacho, M., & Perez-Quiros, G. (2010). Introducing the euro-sting: Short-term indicator of euro area growth. *Journal of Applied Econometrics*, 25(4), 663-694.

Chow, G. C., & Lin, A. L. (1971). Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. *The review of Economics and Statistics*, 372-375.

Cristóbal, A. C. (2007). El índice deflactor del producto interior bruto. *Revista Indice*, 8-9.

De la Hoz, B., Uzcátegui, S., Borges, J., & Velazco, A. (2008). La inflación como factor distorsionante de la información financiera. *Revista Venezolana de Gerencia*, 13(44), 556-572.

Doménech, J. (2021). *Economía digital. p-38.*

Economista, E. (2006). *elEconomista.es*.

Eichenauer, V. Z., Indergand, R., Martínez, I. Z., & Sax, C. (2022). Obtaining consistent time series from Google Trends. *Economic Inquiry*, 60(2), 694-705.

Ettredge, M., Gerdes, J., & Karuga, G. (2005). Using web-based search data to predict macroeconomic statistics. *Communications of the ACM*, 48(11), 87-92.

Europea, C. (2023). *Comision Europea*. Recuperado el enero de 2023

Europea, U. (2023). *Eurostat*. Recuperado el enero de 2023

González-Fernández, M., & González-Velasco, C. (2018). Can Google econometrics predict unemployment? Evidence from Spain. *Economics letters* , 42-45.

Guerrero, V. M. (1990). Desestacionalización de series de tiempo económicas: introducción a la metodología. *Comercio Exterior*, 40(11), 1035-1046.

Guzman, G. (2011). Internet search behavior as an economic forecasting tool: The case of inflation expectations. *Journal of economic and social measurement*, 36(3), 119-167.

Hayes, A. (2022). Stepwise Regression: Definition, Uses, Example, and Limitations. *Investopedia.*

INE. (2023). *Instituto Nacional de Estadística*. Recuperado el enero de 2023

Jun, S. P., Yoo, H. S., & Choi, S. (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. *Technological forecasting and social change*, 130, 69-87.

Maas, B. (2020). Short-term forecasting of the US unemployment rate. *Journal of Forecasting*, 39(3), 394-411.

Mauricio, J. A. (2007). Universidad Complutense de Madrid. *Introducción al análisis de series temporales*. 10.

Monteforte, L., & Moretti, G. (2013). Real-time forecasts of inflation: The role of financial variables. *Journal of Forecasting*, 32(1), 51-61.

Morales, F. C. (2022). *Rankia.*

Narita, M. F., & Yin, R. (2018). In search of information: use of google trends' data to narrow information gaps for low-income developing countries. *International Monetary Fund.*

Rodrigo, J. A. (2017). Cienciadedatos.net. *Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE*.

Rodrigo, J. A. (2020). Cienciadedatos.net. *Regularización Ridge, Lasso y Elastic Net con Python*.

Smith, P. (2016). Google's MIDAS touch: Predicting UK unemployment with internet search data. *Journal of Forecasting*, 35(3), 263-284.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic record*, 88, 2-9.

Vosen, S., & Schmidt, T. (2011). Forecasting private consumption: survey-based indicators vs. Google trends. *Journal of forecasting*, 30(6), 565-578.

Wang, G. C., & Jain, C. L. (2003). Regression analysis: modeling & forecasting. *Institute of Business Forec.*

Woo, J., & Owen, A. L. (2019). Forecasting private consumption with Google Trends data. *Journal of Forecasting*, *38*(2), 81-91.

ANEXO

Objetivos de desarrollo sostenible (ODS)

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenibles	Alto	Medio	Bajo	No Procede
ODS 1. Fin de la pobreza.			×	
ODS 2. Hambre cero.			×	
ODS 3. Salud y bienestar.		×		
ODS 4. Educación de calidad.		×		
ODS 5. Igualdad de género.				×
ODS 6. Agua limpia y saneamiento.				×
ODS 7. Energía asequible y no contaminante.				×
ODS 8. Trabajo decente y crecimiento económico.	×			
ODS 9. Industria, innovación e infraestructuras.	×			
ODS 10. Reducción de las desigualdades.				×
ODS 11. Ciudades y comunidades sostenibles.		×		
ODS 12. Producción y consumo responsables.	×			
ODS 13. Acción por el clima.			×	
ODS 14. Vida submarina.				×
ODS 15. Vida de ecosistemas terrestres.				×
ODS 16. Paz, justicia e instituciones sólidas.				×
ODS 17. Alianzas para lograr objetivos.			×	

Tabla 1: Relación del trabajo con los ODS. Fuente: ETSINF

Reflexión sobre la relación del TFG con los ODS y con el/los ODS más relacionados.

En 2015, la Organización de las Naciones Unidas aprobó la Agenda 2030 sobre el Desarrollo Sostenible, una oportunidad para que los países y sus sociedades emprendan un nuevo camino con el que mejorar la vida de todos, sin dejar a nadie atrás. La Agenda define un total de 17 Objetivos de Desarrollo Sostenible (ODS) de aplicación universal para impulsar el crecimiento económico, el compromiso con las necesidades sociales y la protección del medio ambiente.

Este Trabajo de Fin de Grado tiene el potencial de contribuir a la consecución de varios de los Objetivos de Desarrollo Sostenible.

Por ejemplo, en relación al ODS **Trabajo decente y crecimiento económico,** los modelos propuestos pueden ayudar a predecir el comportamiento de la economía y tomar medidas para mantener un crecimiento sostenible o reducir el impacto de una posible crisis económica.

En cuanto **al ODS Producción y consumo responsable**, los modelos pueden ajustar la producción de bienes y servicios de una forma más responsable al predecir el sentimiento económico de los consumidores, el desempleo y la inflación. Lo cual está relacionado con el ODS Acción por el clima al reducir la contaminación.

Por último, relacionado con **el ODS Industria, innovación e infraestructuras**, los modelos pueden ofrecer la oportunidad al país donde se prevea crecimiento económico de invertir en infraestructuras como carreteras y hospitales, así como en educación de calidad durante el crecimiento económico.

En conclusión, al utilizar Google Trends para analizar y predecir la situación socioeconómica, estos modelos pueden ser extrapolado a diferentes campos como salud, bienestar, educación y sostenibilidad.

Gráficos MAE y RMSE

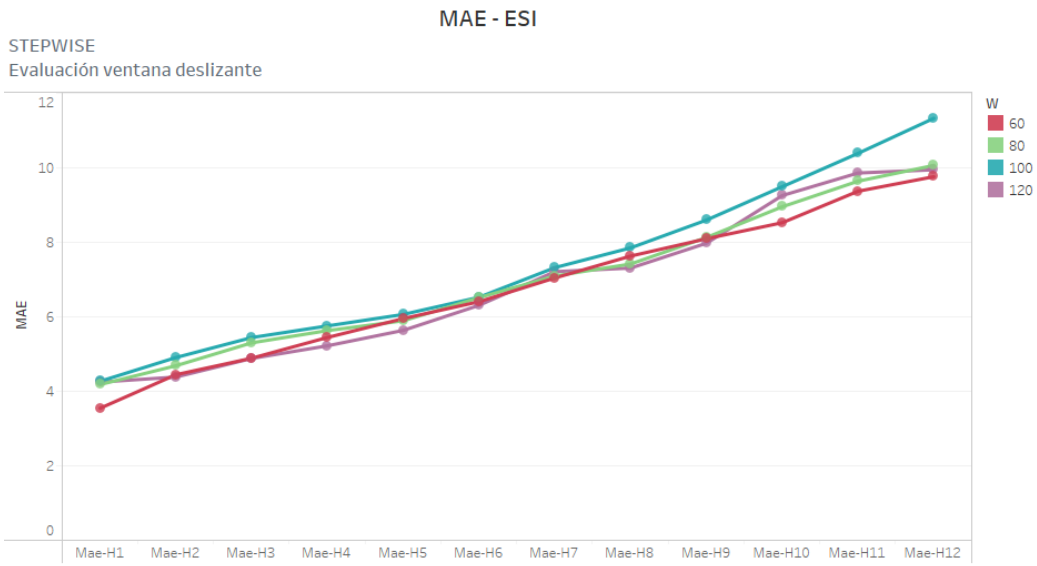

Figura 33. Resultados MAE - Stepwise evaluación ventana deslizante para ESI. Fuente: Elaboración propia

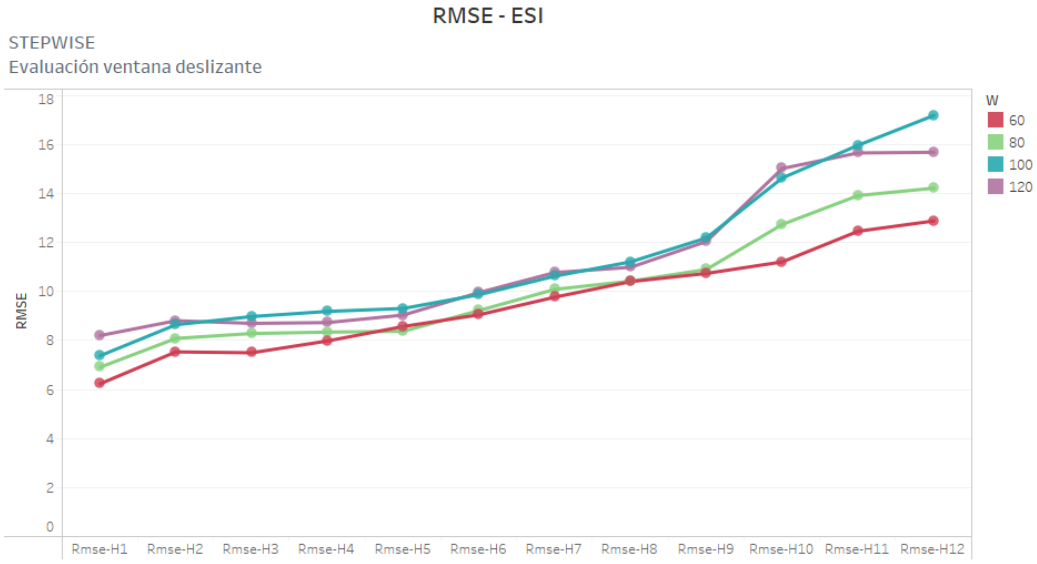

Figura 34. Resultados RMSE - Stepwise evaluación ventana deslizante para ESI. Fuente: Elaboración propia

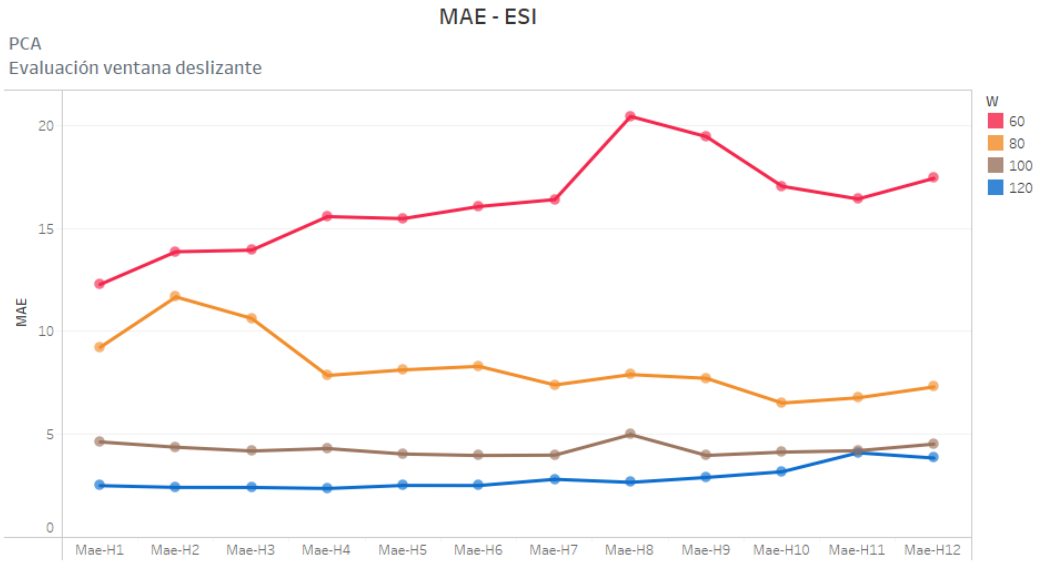

Figura 35. Resultados MAE - PCA evaluación ventana deslizante para ESI. Fuente: Elaboración propia

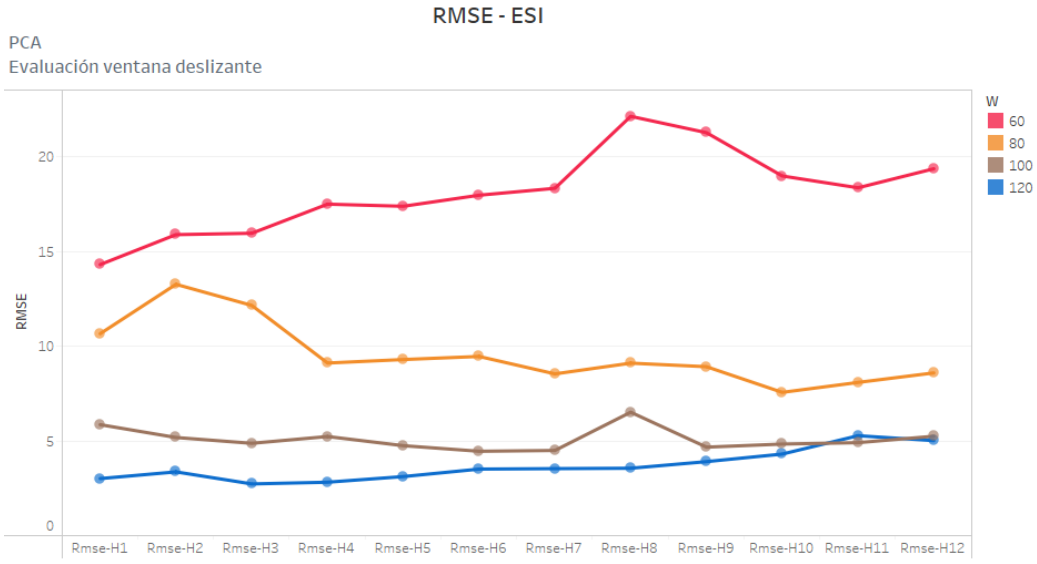

Figura 36. Resultados RMSE - PCA evaluación ventana deslizante para ESI. Fuente: Elaboración propia

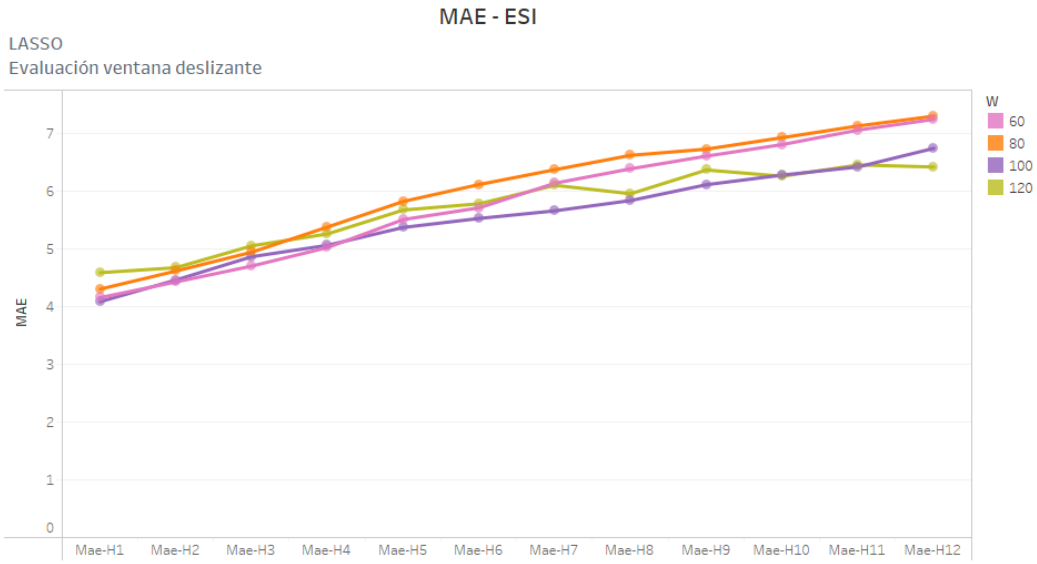

Figura 37. Resultados MAE - Lasso evaluación ventana deslizante para ESI. Fuente: Elaboración propia

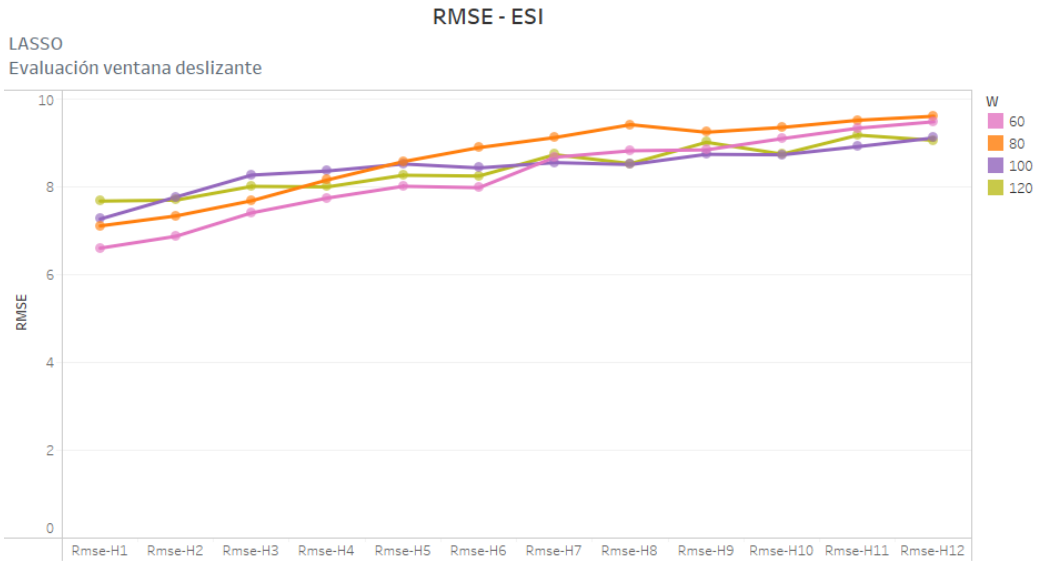

Figura 38. Resultados RMSE - Lasso evaluación ventana deslizante para ESI. Fuente: Elaboración propia

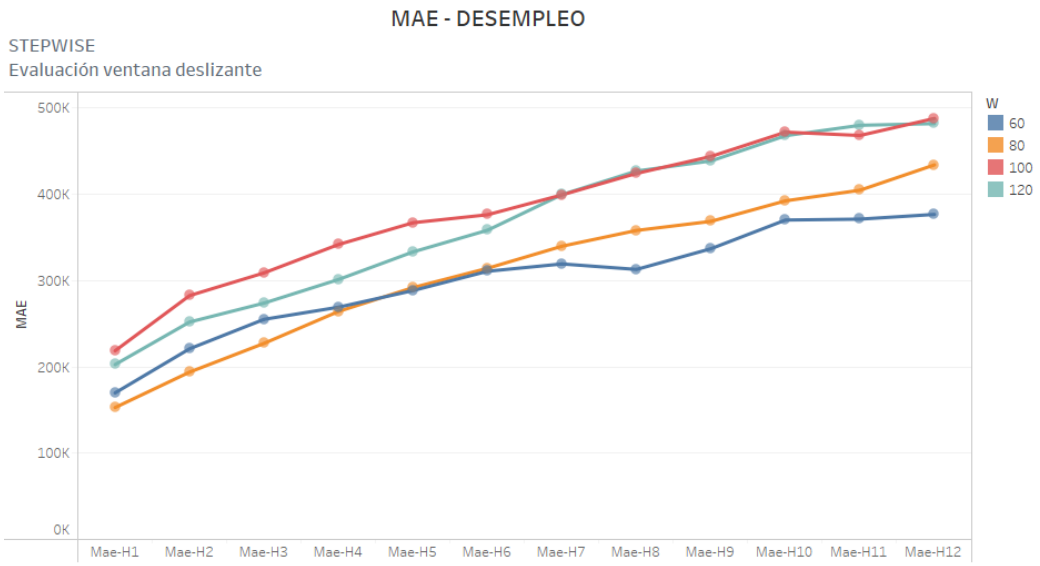

Figura 39. Resultados MAE - Stepwise evaluación ventana deslizante para desempleo. Fuente: Elaboración propia

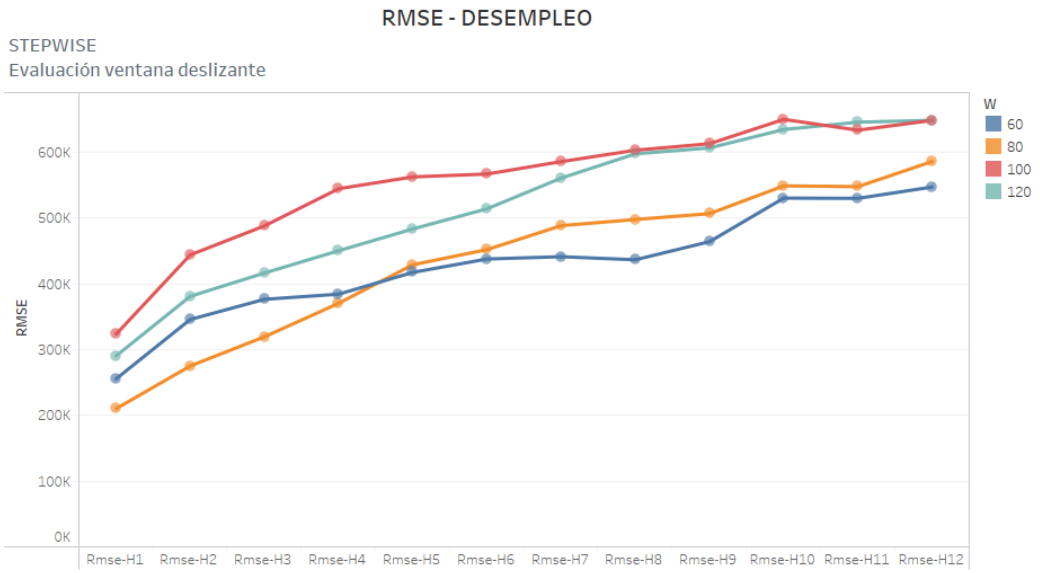

Figura 40. Resultados RMSE - Stepwise evaluación ventana deslizante para desempleo. Fuente: Elaboración propia

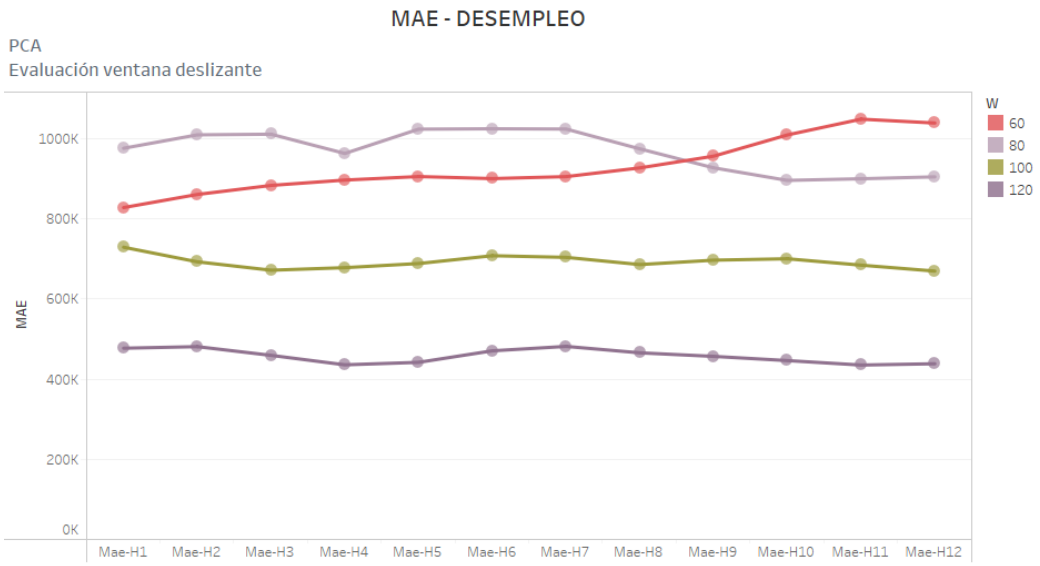

Figura 41. Resultados MAE - PCA evaluación ventana deslizante para desempleo. Fuente: Elaboración propia

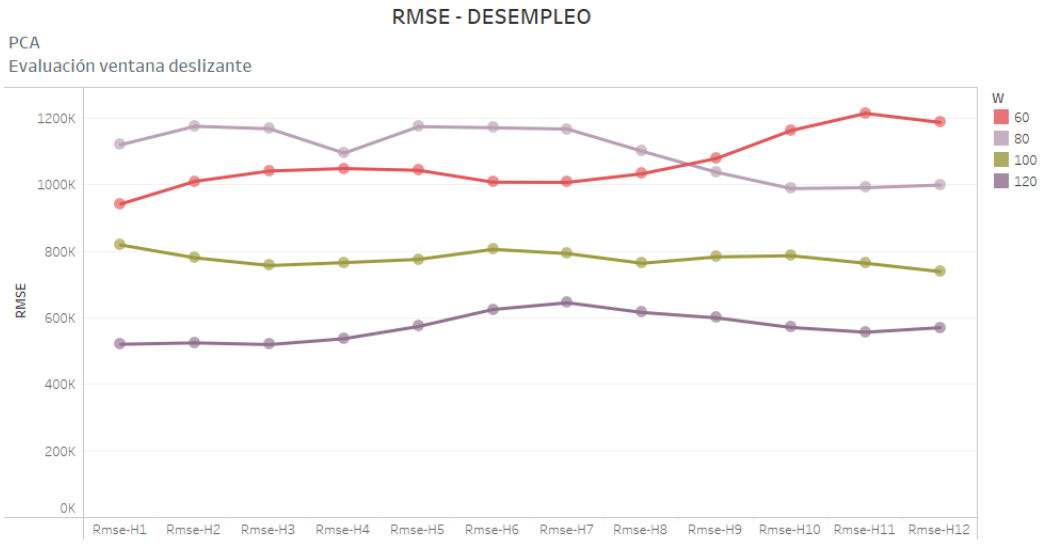

Figura 42. Resultados RMSE - PCA evaluación ventana deslizante para desempleo. Fuente: Elaboración propia

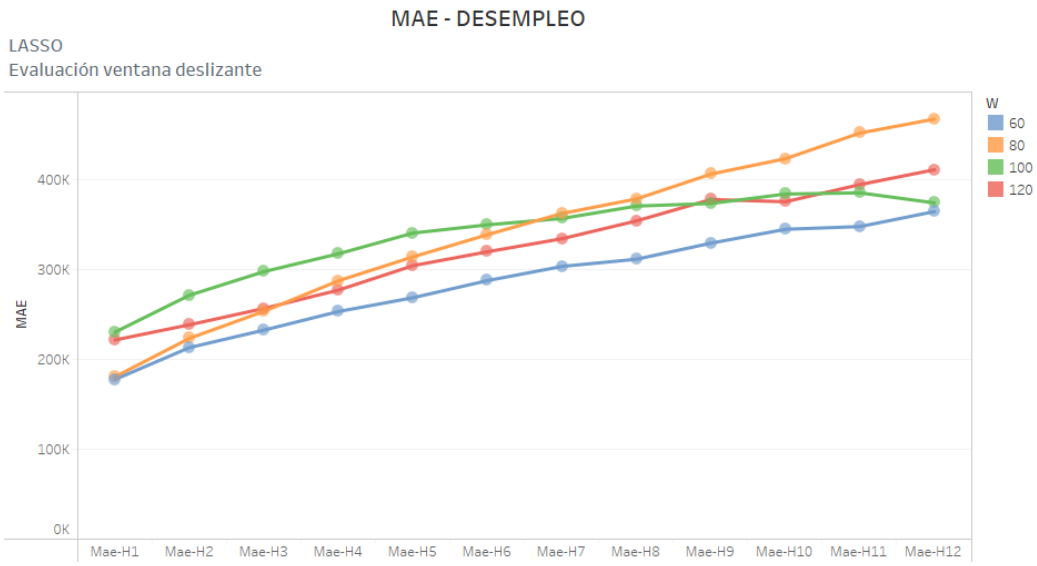

Figura 43. Resultados MAE - Lasso evaluación ventana deslizante para desempleo. Fuente: Elaboración propia

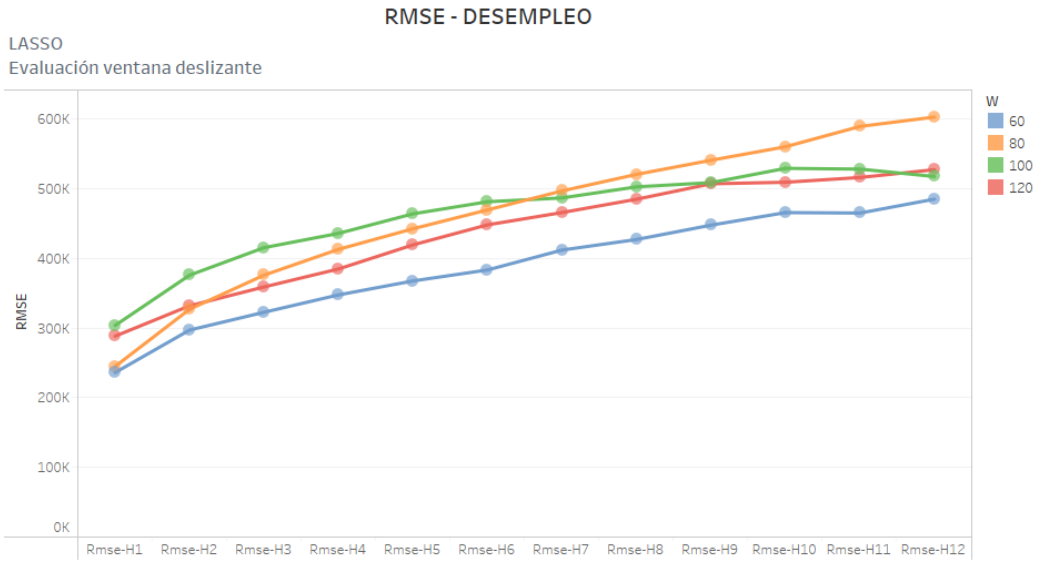

Figura 44. Resultados RMSE - Lasso evaluación ventana deslizante para desempleo. Fuente: Elaboración propia

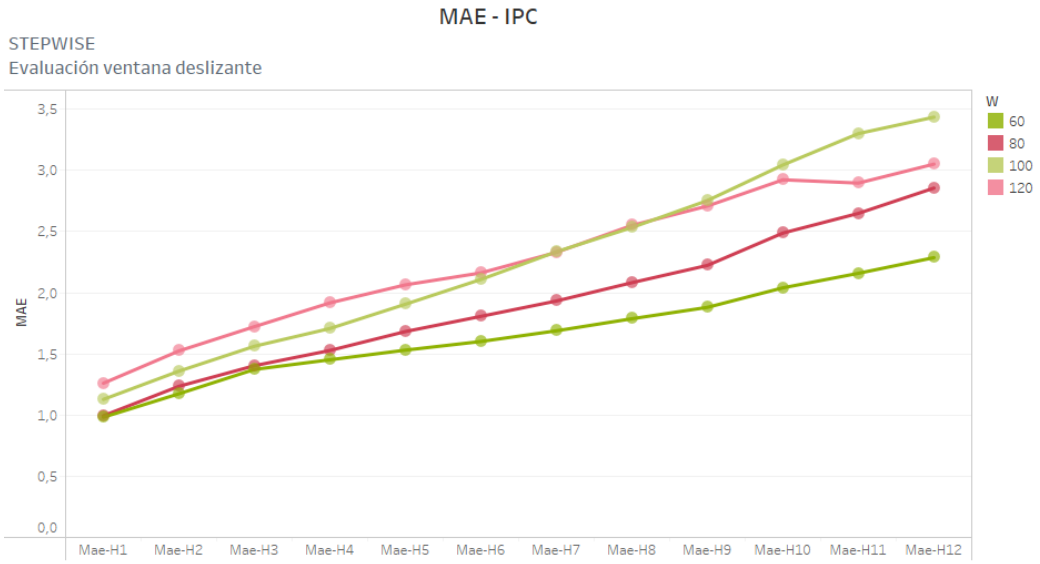

Figura 45. Resultados MAE - Stepwise evaluación ventana deslizante para IPC. Fuente: Elaboración propia

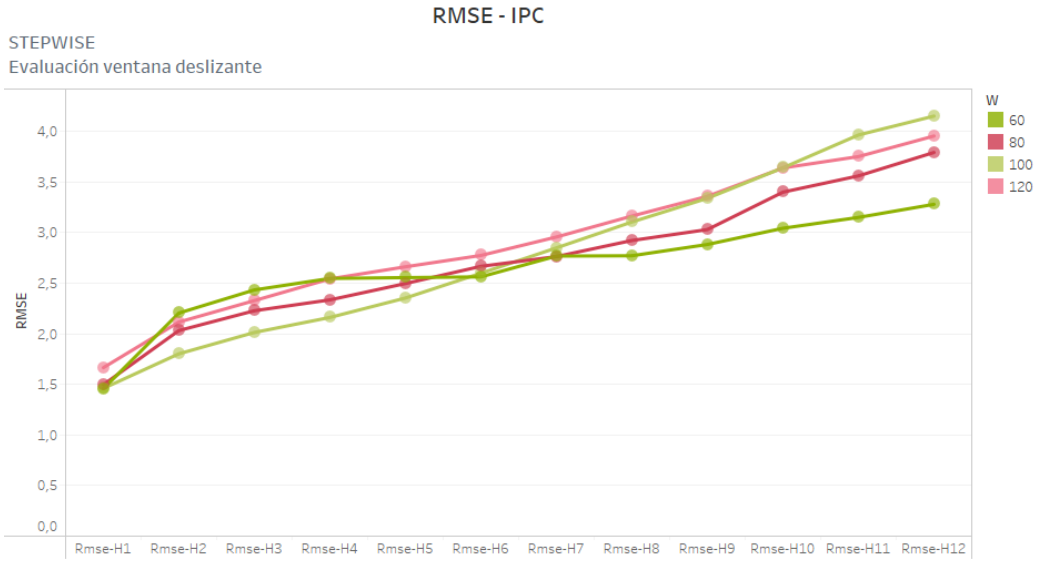

Figura 46. Resultados RMSE - Stepwise evaluación ventana deslizante para IPC. Fuente: Elaboración propia

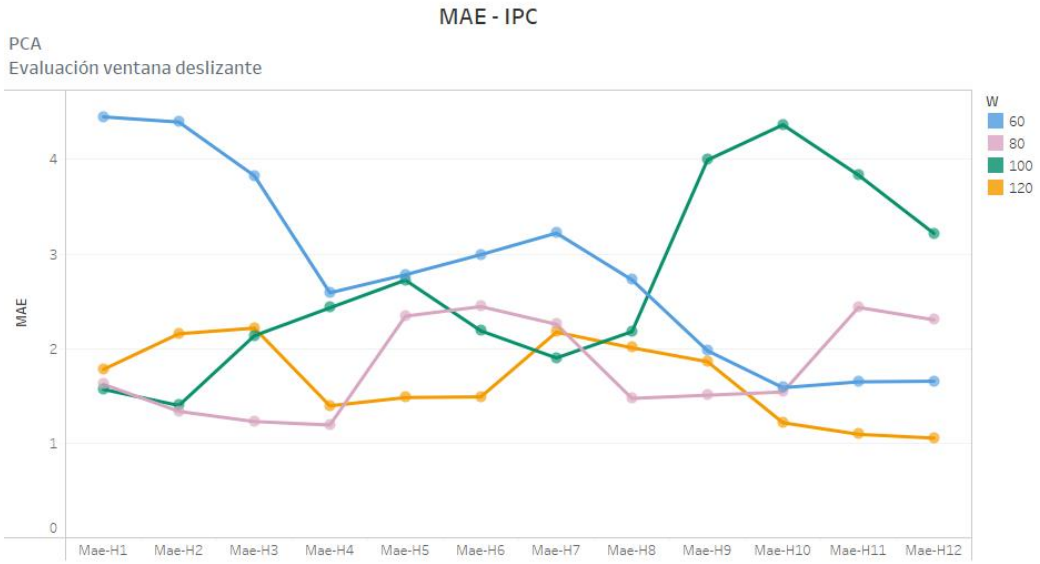

Figura 47. Resultados MAE - PCA evaluación ventana deslizante para IPC. Fuente: Elaboración propia

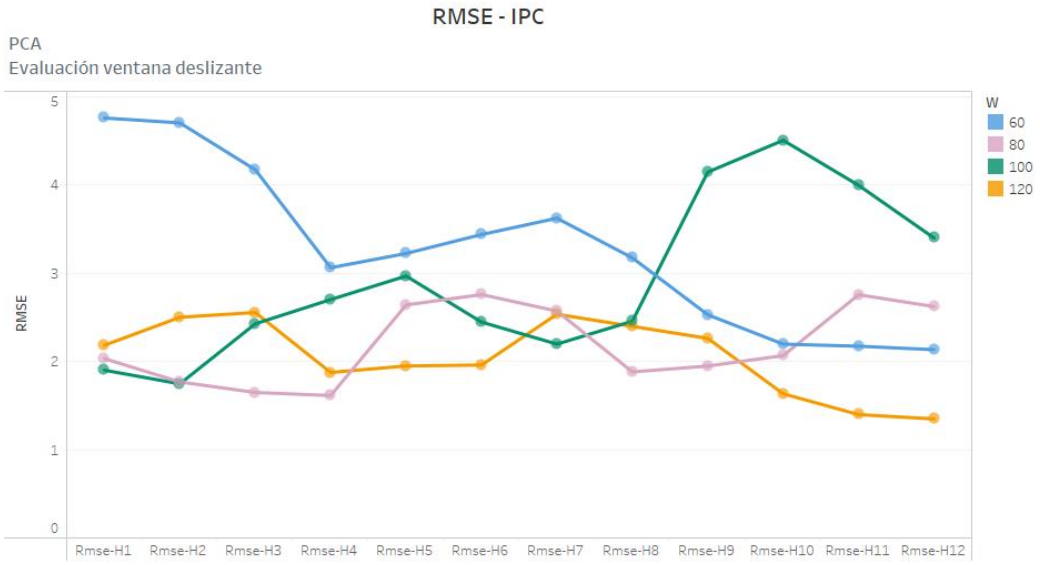

Figura 48. Resultados RMSE - PCA evaluación ventana deslizante para IPC. Fuente: Elaboración propia

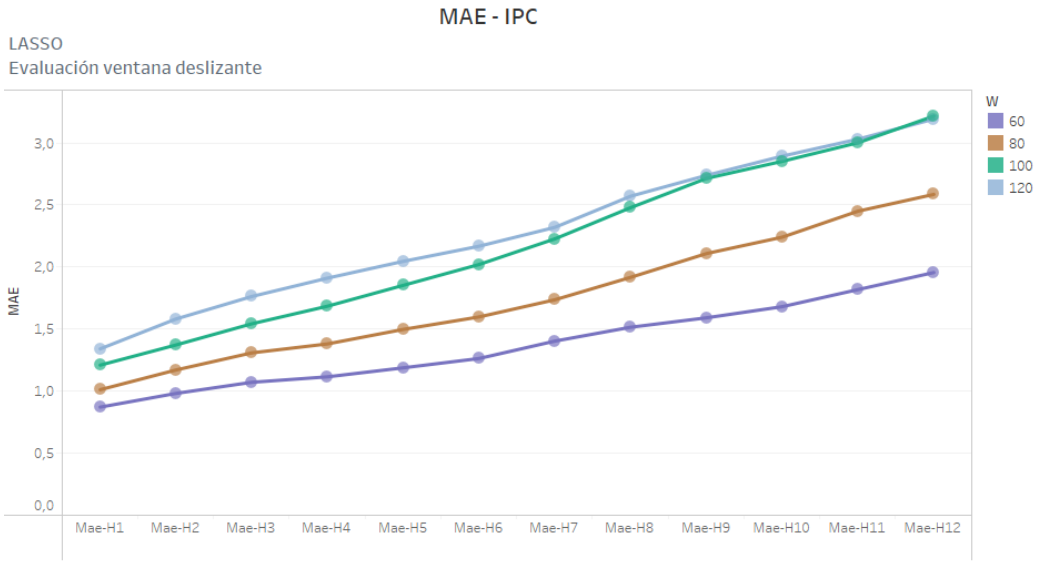

Figura 49. Resultados MAE - Lasso evaluación ventana deslizante para IPC. Fuente: Elaboración propia

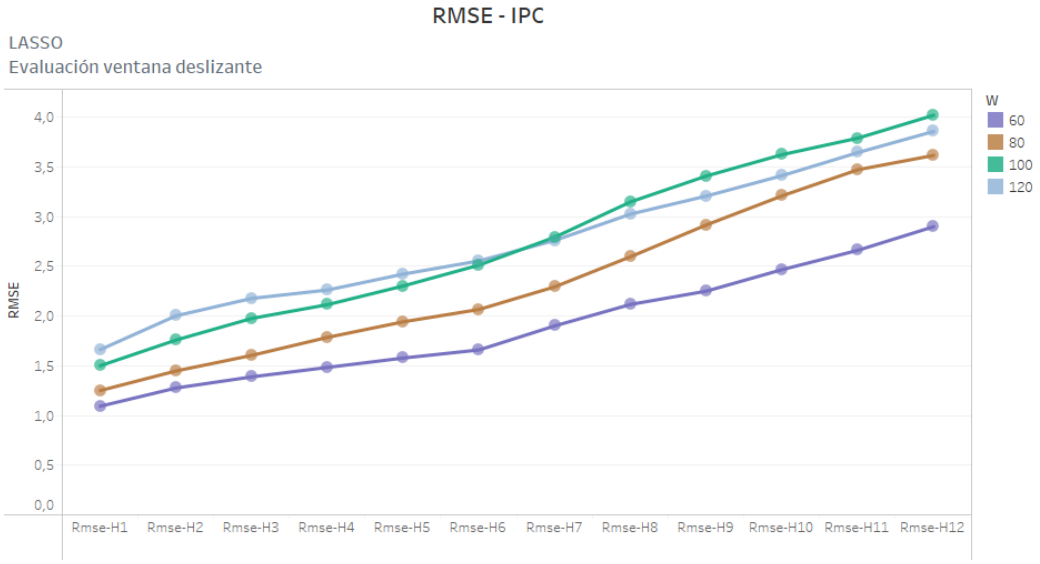

Figura 50. Resultados RMSE - Lasso evaluación ventana deslizante para IPC. Fuente: Elaboración propia