Sección de Matemáticas
Universidad de La Laguna

Alejandro Carvajal Martínez

Influencia del COVID-19 en las predicciones del número de parados en Canarias usando Google Trends

Influence of COVID-19 on the predictions of the number of unemployed in the Canary Islands using Google Trends

Trabajo Fin de Grado
Grado en Matemáticas
La Laguna, Julio de 2021

Dirigido por
Enrique Francisco González Dávila
Elisa María Jorge González

Enrique Francisco González Dávila
Departamento de Matemáticas, Estadística e Investigación Operativa
Universidad de La Laguna
38200 La Laguna, Tenerife

Elisa María Jorge González
Departamento de Matemáticas, Estadística e Investigación Operativa
Universidad de La Laguna
38200 La Laguna, Tenerife

Agradecimientos

Quiero agradecer a mi familia el apoyo en todo momento. En especial, a mis padres, madrina y abuelos. También agradecer a mis amigos por estar siempre ahí. A Gustavo por ayudarme y explicarme cualquier duda que tuviese en estos 5 años.

Por último agradecer a mi tutor, Enrique González Dávila, por coger esta línea de trabajo conmigo y ayudarme siempre que lo he necesitado.

Alejandro Carvajal Martínez
La Laguna, 5 de julio de 2021

Resumen · Abstract

Resumen

La predicción del número de parados en tiempo de crisis y acontecimientos esporádicos de gran impacto plantea grandes retos para cualquier tipo de modelo de prediccción. La utilización de herramientas que introduzcan información en tiempo real son cada vez más demandadas. Las búsquedas en la red por aquellas personas que demandan un empleo, desean cambiarlo o simplemente ven peligrar su puesto de trabajo, son un termómetro de la evolución de la serie de número de parados de una determinada región. En este trabajo se plantea un modelo de predicción del número de parados para Canarias haciendo uso de las búsquedas en internet obtenidas a través de Google Trends. La influencia de la pandemia de la COVID-19 en los modelos permitirá evaluar la efectividad de estos para corregir series econométricas en caso de grandes catástrofes.

Palabras clave: *Paro– Google Trends –ARIMA –Gretl .*

Abstract

Predicting the number of unemployed in times of crisis and sporadic high-impact events poses great challenges for any type of forecasting model. The use of tools that can input information in real time is becoming more and more in demand. Web searches by people who are looking for a job, want to change their job or simply see their job in danger, are a thermometer of the evolution of the number of unemployed people in a given region. In this paper we propose a prediction model of the number of unemployed for the Canary Islands using Internet searches obtained through Google Trends. The influence of the COVID-19 pandemic on these models will make it possible to evaluate their effectiveness in correcting econometric series in the event of major catastrophes.

Keywords: *Unemployment – Google Trends – ARIMA – Gretl .*

Contenido

Introducción

Durante los últimos años, el desempleo ha sido uno de los principales problemas sociales en nuestra comunidad, creando inestabilidad en el ámbito del mercado laboral y grandes desigualdades. La última Encuesta de Población Activa (EPA), correspondiente al primer trimestre de este 2021, estima la tasa de paro en las Islas en un 25,42 %, sin embargo, la tasa de desempleo llega al pico del 34,2 % al incluir en el cálculo a las personas afectadas por ERTE. Esto significa que 34 de cada cien canarios acceden al mercado laboral en búsqueda de oferta, propiamente aquellas personas que se encuentran desempleadas, como aquellas otras que tienen la incertidumbre de su continuidad en la empresa y deciden buscar alternativas [1].

Estos datos de desempleo estructural se han agravado debido a la crisis del COVID-19 en la que nos hemos visto inmersos, pero la comunidad Canaria, ya desde antes de la pandemia tenía un desempleo estructural elevado, una de las más altas de España. En 2019, antes de que el coronavirus dinamitara la actividad turística, el Archipiélago ya se situaba en el puesto número once de toda Europa con mayor tasa de paro.

Conocer con antelación la evolución de las cifras de paro tiene suma importancia para poder establecer políticas de ayuda e reinserción. Son muchas las familias que, por desgracia, se encuentran en una situación de precariedad. Unas políticas que se apoyen en resultados de estudios sociales podría lograr un menor número de parados.

Cada vez más las personas requieren información inmediata, esto hace que sus necesidades queden reflejadas de forma casi automática en las redes. A destacar en este área, el efecto directo que realizan las redes sociales ofreciendo a los usuarios nuevas formas de buscar trabajo que nada tienen que ver con los métodos tradicionales. Las redes sociales se han convertido en una parte íntegra de nuestra vida y hoy en día es raro ver a una persona que no tenga una cuenta de Facebook, Twitter, Instagram o LinkedIn.

El presente trabajo se centra en la influencia que tiene sobre el número de parados en Canarias las búsquedas en tiempo real que se producen en Google.

En particular, se utilizará la plataforma Google Trends para obtener el índice de palabras relacionadas con la búsqueda de empleo que realizan las personas en dicha región en un periodo de tiempo seleccionado. Esta información será fusionada con modelos de series temporales estándar, como son los autorregresivos integrados de medias móviles (ARIMA). Adicionalmente, será utilizado el hecho de la pandemia de la COVID-19, y los confinamientos asociados a ella, para evaluar la influencia de la introducción de dichas variables en los diferentes modelos de predicción utilizados.

El trabajo consta de dos partes. Una primera parte teórica donde se introduce conceptos generales que ayudarán a entender el problema. Como son las series temporales, modelos disponibles y sistemas de búsqueda de información en internet. Y una segunda parte práctica, en la que se realiza un estudio particularizado sobre el número de parados en Canarias. Se seleccionarán las mejores palabras claves utilizadas en búsqueda de internet relacionadas con esta variable objetivo, se evaluarán diferentes modelos y se realizará un estudio de validación. Para llevar a cabo estos procedimientos se utilizará el software econométrico Gretl, totalmente libre y disponible en la dirección web: http://gretl.sourceforge.net.

En ocasiones, las imprecisiones de nuestras variables pueden hacer que las condiciones iniciales se vean afectadas y presenten un resultado aleatorio a largo plazo, haciendo que los datos se vean afectados por diversos acontecimientos esporádicos (pandemias, demografía, dinámica de la población). Una posibilidad radical que impediría la predicción exacta distinguiendo un comportamiento genuinamente aleatorio. Se estudiará como se comporta nuestra variable si está afectada por estos acontecimientos.

1

Series temporales

Una serie temporal es el resultado de observar los valores de una variable a lo largo del tiempo en intervalos regulares (cada día, cada mes, cada año, etc.) sobre una característica (serie univariante) o sobre varias características (serie multivariante) de una unidad observable en diferentes momentos (Mauricio, 2007 [2]).

La representación matemática frecuente de una serie temporal univariante es:

$$y_t : t = 1, ..., N \tag{1.1}$$

donde y_t es la t-ésima observación ($1 \leq t \leq N$) de la serie y N es el número de observaciones que consta la serie completa (el tamaño o la longitud de la serie).

Las N observaciones $y_1, y_2, ..., y_N$ se pueden obtener en un vector columna $y \equiv [y_1, y_2, ..., y_N]'$ de orden $N \times 1$.

Uno de los principales objetivos al analizar una serie temporal es intentar predecir el comportamiento para adelantarnos a su evolución, teniendo en cuenta que las condiciones se mantendrán en el tiempo. Este análisis se encargará de las series en las que sabiendo los valores pasados, no se pueda pronosticar claramente un comportamiento futuro de la variable, lo que se denomina como serie temporal aleatoria.

De acuerdo con Mauricio (2007) [2] la predicción que se obtenga sobre el valor que va a tener la variable en el futuro no se podrá considerar totalmente exacta, pero sí que será próxima a la realidad, ya que estudiará la regularidad que hay en los datos a la hora de estimar el modelo de comportamiento de dicha variable.

1.1. Componentes de una serie temporal

Desde el punto de vista clásico de análisis de series temporales, cualquier serie es el sumatorio de cuatro componentes, los cuales son: tendencia (T), variaciones estacionales (VE), variaciones cíclicas (V) y por último, variaciones

accidentales (R). No siempre que se analiza una serie temporal y se procede a su descomposición, existen cada uno de los componentes (Pilar, 2018 [3]).

Según Caparrós (2011) [4] podemos definir las cuatro componentes como:

- **Tendencia (T)**
  Muestra el movimiento de la variable a largo plazo, debido a cambios tecnológicos, demográficos o institucionales. Es necesario la disposición de un alto número de observaciones durante muchos años para corroborar si hay un patrón de los datos, de manera que aumenten, disminuyan o se mantengan constantes.

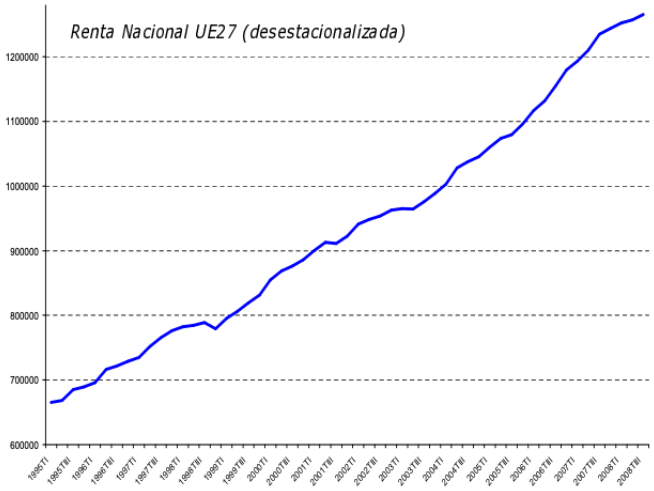

Figura 1.1: Tendencia de una serie temporal

Observando la figura 1.1 se puede concluir que la renta nacional de la UE ha seguido en este periodo una evolución continuamente creciente, es decir, la serie presenta una tendencia creciente.

- **Variaciones estacionales (VE)**
  Las variaciones estacionales son oscilaciones que se presentan con una duración de tiempo anuales o inferiores a un año. Suelen ser repetitivas mostrando el efecto de encontrarnos en una determinada época (climatología, festividades,...).
  El número de parados produce un comportamiento estacional de la serie en periodos concretos como Navidad o Semana Santa produciendo un aumento de las contrataciones. La disminución del desempleo en estos casos sigue un patrón estacional.

Figura 1.2: Variaciones estacionales de una serie temporal

En la figura 1.2 se observa como la venta de helados en 2014 se disparó en los meses correspondientes al verano.

- **Variaciones cíclicas (C)**
  Recoge los movimientos originados por el ciclo económico con periodos de duración desconocidos y superiores a un año. Para que una serie pueda reconocer la componente cíclica es necesario tener una serie larga y con un número completo de ciclos.

- **Variaciones accidentales o residuales (R)**
  Las variaciones accidentales recoge aquellos factores asociados al muy corto plazo y que quedan fuera del control del analista. Dentro de este componente, también denominada residual, se encuentran aquellos factores eventuales, esporádicos e imprevisibles, pero fácilmente reconocibles como una catástrofe natural.

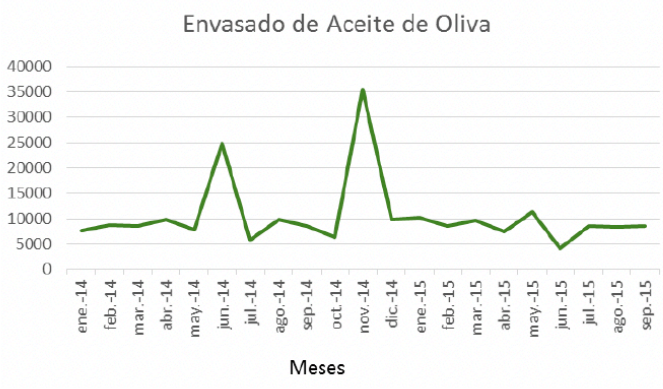

Figura 1.3: Variaciones accidentales de una serie temporal

La forma en que estos cuatro componentes se analicen dará como resultado la serie temporal. Hay muchas formas de interacción entre ellos, pero generalmente se eligen los siguientes tres: esquema aditivo, esquema multiplicativo y esquema multiplicativo mixto.

- *Esquema Aditivo:*

$$Y_t = T_t + VE_t + C_t + R_t \tag{1.2}$$

La componente se expresa en la misma unidad que las observaciones. La variación residual, en este modelo, es independiente de las demás componentes, es decir, la magnitud de dichos residuos no depende del valor que tome cualquier otra componente de la serie.

- *Esquema Multiplicativo:*

$$Y_t = (T_t)(VE_t)(C_t)(R_t) \tag{1.3}$$

En este modelo la tendencia se denota en la misma unidad que las observaciones. No se cumple la hipótesis de independencia del esquema aditivo. Otro tipo de modelo multiplicativo que si la cumple es el llamado esquema multiplicativo mixto.

- *Esquema Multiplicativo Mixto:*

$$Y_t = (T_t)(VE_t)(C_t) + (R_t) \tag{1.4}$$

La elección de uno de estos métodos de interacción es solo el comienzo de un tedioso camino para la obtención de un modelo que consiga desarrollar adecuadamente la evolución de la serie a estudiar. Aunque para obtener este modelo, primero se debe determinar la tendencia determinista de la serie, a través de la cual se conoce la existencia de reglas matemáticas que permitirá obtener resultados futuros precisos. Debido a su naturaleza inestable, los únicos componentes que no pueden mostrarnos la trayectoria regular de la construcción del modelo son los componentes accidentales.

1.2. Análisis de una serie temporal

Hay tres tipos o métodos para el análisis de series de tiempo: análisis clásico, análisis causal y la metodología Box-Jenkins (Quesada, 2015 [5]):

- *Método clásico:* En este método se pretende descomponer la serie temporal en los distintos componentes previamente desarrollados: tendencia, variaciones cíclicas, variaciones estacionales y variaciones residuales, de manera que se

pueda esbozar en el futuro y conseguir predecir la evolución de la variable. Dentro de esta metodología clásica existen variantes incorporadas recientemente como son los modelos de espacio de estados, los cuales usan el desglose en las cuatro componentes pero introducen relaciones entre ellas a través de modelos (Casals et al., 2016 [6]).

- *Método causal:* Se usan ecuaciones para explicar la evolución futura de las variables, por las cuales se podrá asociar con otras que afecten más directamente. Por consiguiente, se sabrá el valor futuro de la variable que se está estudiando.

- *Metodología Box-Jenkins:* El modelo analítico diseñado por Box-Jenkins (1970) intenta expresar la evolución de la variable a partir de sus propios valores pasados. En definitiva, si se consigue entender el patrón de comportamiento que desarrolla la variable que se está estudiando, se podrá predecir su comportamiento futuro. Este último será el análisis que se usa en este trabajo.

1.2.1. Clasificación descriptiva de series temporales

Para realizar la clasificación descriptiva de una serie temporal tenemos dos opciones: series de tiempo estacionarias o series de tiempo no estacionarias.

- **Series estacionarias**
  Una serie temporal es estacionaria si la media y la variabilidad se mantienen constantes a lo largo del tiempo, es decir, no muestra tendencia en su trayectoria.

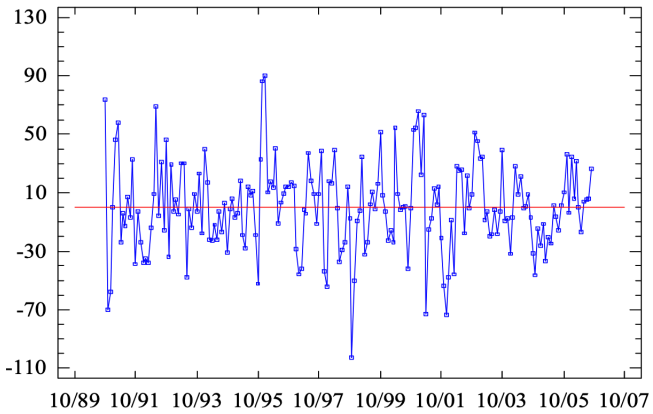

Figura 1.4: Serie estacional

- **Series no estacionarias**
  Una serie es no estacionaria si la media y/o la variabilidad cambian a lo largo del tiempo. Desde un punto de vista gráfico, en las series no estacionarias pueden mostrar una tendencia, es decir, la media crece o decrece a lo largo del tiempo. Además, pueden presentar efectos estacionales.
  En muchas ocasiones, este tipo de serie una vez se realiza una transformación o ajuste por otras variables, propicia una serie estacionaria.

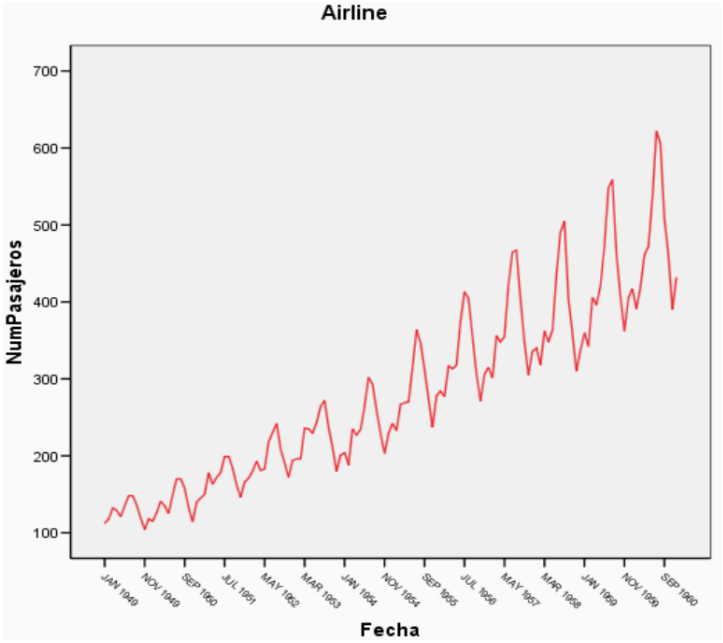

Figura 1.5: Serie no estacional

Esta introducción en las características y comportamientos de las series temporales servirá de gran ayuda para entender posteriormente el desarrollo del trabajo.

2

Metodología

2.1. Google Trends

Google Trends es una herramienta de acceso libre y gratuito de Google. Permite medir la popularidad de búsqueda de una palabra o frases y conocer el nivel de utilización de un determinado término. Proporciona valores relativos de uso en una escala de 0 a 100, donde 100 representa el nivel más alto sobre las búsquedas que se realizan respecto a un término o palabra clave. (ENyD, 2015 [7]).

Esta plataforma permite realizar búsqueda de términos (palabras, grupo de palabras o frases) desde 2004 hasta la actualidad en diferentes regiones geográficas y periodos de tiempo establecidos (diarios, semanales, mensuales,...). Adicionalmente, pueden ser introducidos 5 términos al mismo tiempo, relativizándolos con respecto a uno de ellos. La información puede ser descargada o representada gráficamente, ayudando a la búsqueda de patrones y cambios a lo largo del tiempo así como en diferentes zonas geográficas.

Los usuarios tienen la opción de especificar la palabra cuyo interés quieren conocer o bien pueden buscar por país / región, hora (por ejemplo, búsquedas realizadas en las últimas 10 horas), categoría (canciones, arte, deporte, etc.) y tipo de búsqueda (noticias, búsqueda de YouTube o imágenes) de dicho término.

Algunas razones para usar Google Trends según Bernal (2020) [8] son:

- Obtener diferentes ideas de contenido ya que esta plataforma proporciona los términos más buscados por las personas. Es de gran utilidad en políticas de marketing para las empresas.
- Encontrar tendencias. La búsqueda de las palabras más influyentes en un determinado sector puede proporcionar un aumento de las ventas sobre un negocio.
- Mejorar las predicciones sobre un tema.

En este trabajo se usará para mejorar las predicciones del número de parados.

2.1.1. Aplicaciones de Google Trends

Según Redondo (2013) [9] el primer estudio en utilizar Google Trends fue realizado por Ginsberg (2009) [10], en el campo de la medicina. Descubrieron que gracias a las búsquedas de 46 términos asociados a la gripe se consiguieron predecir brotes de gripe con dos semanas de antelación a los informes del Centro de Control y Prevención de Enfermedades.

Choi y Varian (2012) [11] pudieron predecir las ventas de inmuebles y automóviles a través de los datos procedentes de Google Trends. Para los casos de flujo turístico partieron de la premisa que Google era usado para planear viajes y, por lo tanto, un incremento de búsquedas de un determinado lugar supondría un aumento del flujo turístico.

Uno de los ámbitos en los que más ha intervenido Google Trends ha sido en el ámbito económico. McLaren y Shanbhogue (2011) [12] analizaron el mercado de viviendas en Reino Unido. Guzmán (2011) [13] demostró que se podía predecir el nivel de inflación. Para ello, usó los datos conseguidos en 36 encuestas y concluyó que los datos obtenidos de Google Trends tenían un menor error de predicción que todos los indicadores previamente validados.

También ha sido un tema recurrente el uso de Google Trends para predecir el nivel de la tasa de paro, especialmente donde la mayoría de la población de una región tiene acceso a internet. Entre otros, Askitas y Zimmermann (2009) [14] para Alemania y D'Amuri y Marcucci (2009) [15] para Estados Unidos, aplicaron este tipo de herramientas obteniendo resultados favorables.

2.2. Modelos propuestos para las predicciones

En esta sección se hablará sobre la metodología ARIMA, la cual fue desarrollada a lo largo de la década de los 70 por George E.P. Box, fundador del departamento de Estadística de la Universidad de Wisconsin y, por Gwilym M. Jenkins, profesor de Ingeniería de Sistemas de la Universidad de Lancaster, creadores de la metodología más utilizada para el análisis de series temporales.

La metodología Box-Jenkins para los Modelos Autorregresivos Integrados de Medias Móviles, cuyas siglas en inglés dan nombre a los modelos ARIMA, se pueden diferenciar en cuatro fases: identificación, estimación, examen de diagnóstico y pronóstico. Es fundamental que las observaciones duren en el tiempo. Gracias a estos modelos, se conseguirá una predicción futura del comportamiento de la variable a estudiar gracias a los valores pasados de dicha variable.

A continuación, se analizarán las diferentes fases comentadas anteriormente para identificar el mejor modelo que va a seguir la serie (Quesada, 2015 [5]):

- **Identificación**
  El objetivo de la primera fase es determinar qué modelo es el más adecuado para representar la serie. Para realizar este proceso previamente se necesita que la serie temporal sea estacionaria. En caso de que no sea así, habría que diferenciarla tantas veces como hiciese falta hasta lograrlo. Una vez conseguido, se procederá a evaluar la función de autocorrelación (FAC) y la función de autocorrelación parcial (FACP) muestrales, cuyo gráfico varía en función de la clase de modelo que se tenga y de su grado de parametrización.
  En nuestro caso, la identificación se hará con uno de los programas informáticos más utilizados, el 'Análisis TRAMO'. Sus principales aplicaciones son la previsión, el ajuste estacional, la estimación de la tendencia, la interpolación y la estimación de los efectos de calendario. Con este análisis se obtendrán los valores de los parámetros que definen el modelo ARIMA. Se ha usado en numerosos estudios de predicciones dentro del sector turístico y en este trabajo se hará lo propio usándolo para predecir el número de parados en Canarias.

- **Estimación**
  Una vez definido el modelo ARIMA en la fase anterior, se determinarán los parámetros $(p, d$ y $q)$ de los términos autorregresivos y de promedios móviles que se han incorporado en el modelo.

- **Examen de diagnóstico**
  Fijado el modelo ARIMA, es necesario proceder a la validación del modelo, realizando exámenes de diagnósticos sobre los residuales del ajuste, asegurándonos de que el modelo que se ha obtenido se ajuste a los datos de tal forma que se aproximen a la realidad.

- **Pronóstico**
  Cuando se haya seleccionado el modelo idóneo que se va a utilizar en nuestro estudio, será posible emplearlo para realizar un pronóstico, es decir, predecir el comportamiento futuro de la variable estudiada.

2.2.1. Definiciones y nociones básicas de los modelos ARIMA

A continuación se introducen algunas de las definiciones de concepto básico para comprender la terminología usada en series temporales y su posterior estudio (Mauricio, 2007 [2]):

Definición 2.1. ***(Proceso estocástico)*** *Un proceso estocástico es una secuencia de variables aleatorias, ordenadas y equidistantes cronológicamente, referidas a una (proceso univariante) o a varias (proceso multivariante) características que dependerá del tiempo y del fenómeno probabilístico.*

$..., Y_{-1}, Y_0, Y_1, Y_2, ...(Y_t : t = 0, \pm1, \pm2, ...); Y_t$ es una variable aleatoria univariante referida a la unidad observable considerada en el momento t.

$..., Y_{-1}, Y_0, Y_1, Y_2, ...(Y_t : t = 0, \pm1, \pm2, ...); Y_t$, donde $Y_t \equiv [Y_{t1}, Y_{t2}, ..., Y_{tM}]'$ con $(M \geq 2)$ es una variable aleatoria multivariante referida a la unidad observable considerada en el momento t.

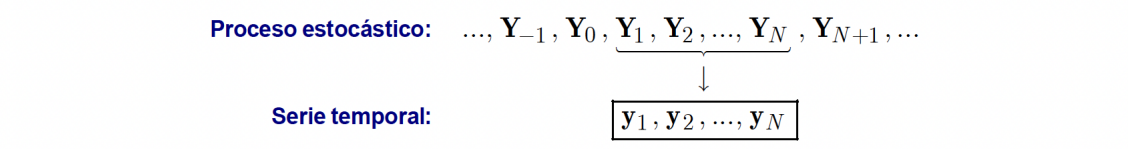

Figura 2.1: Proceso estocástico

Definición 2.2. ***(Proceso estacionario)*** *Un proceso estocástico (Y_t) es estacionario cuando las propiedades estadísticas de cualquier secuencia finita $Y_{t1}, Y_{t2}, ..., Y_{tn}$ con $(n \geq 1)$ de componentes de (Y_t) son semejantes a las de la secuencia $Y_{t1+h}, Y_{t2+h}, ..., Y_{tn+h}$ para cualquier número entero $h = 0, \pm1, \pm2, ...$*

Definición 2.3. ***(Proceso no estacionario)*** *Un proceso estocástico (Y_t) es no estacionario cuando las propiedades estadísticas de al menos una secuencia finita $Y_{t1}, Y_{t2}, ..., Y_{tn}$ con $(n \geq 1)$ de componentes de (Y_t), son diferentes de las de la secuencia $Y_{t1+h}, Y_{t2+h}, ..., Y_{tn+h}$ para al menos un número entero $h > 0$.*

Un proceso puede ser no estacionario en la media, en la varianza o en otras características de la distribución de las variables. En el caso de que la serie no sea estable en el tiempo diremos que es no estacionaria en la media. Cuando la varianza es perturbada por el tiempo, diremos que es no estacionaria en la varianza. Los procesos no estacionarios más importantes son los procesos integrados, los cuales veremos más adelante.

Definición 2.4. ***(Ruido blanco)*** *Podemos definir el ruido blanco como una serie consecutiva en el tiempo de variables aleatorias cuya media es cero, su varianza es constante y covarianza nula.*

$$\begin{aligned} 1. E[z_t] &= 0, t = 1, 2, ... \\ 2. Var(z_t) &= \sigma^2, t = 1, 2, ... \\ 3. Cov(z_t, z_{t-k}) &= 0, k = \pm1, \pm2, ... \end{aligned} \tag{2.1}$$

Puesto que su media y su varianza son constantes a lo largo del tiempo, se podrá decir que son series de tiempo estacionarias.

Definición 2.5. ***(Paseo aleatorio)*** *Un proceso estocástico univariante no estacionario (Y_t) es un paseo aleatorio cuando*

$$Y_t = \mu + Y_{t-1} + A_t, \forall t = 0, \pm 1, \pm 2, ..., \tag{2.2}$$

donde μ es un parámetro (que en muchas ocasiones vale cero) y $(A_t) \sim IID(0, \sigma_A^2)$.

Definición 2.6. ***(Residuo)*** *Los residuos de un modelo son las diferencias entre los valores de las variables y la correspondiente predicción usando la función de regresión. Matemáticamente, la definición de residuo para la i-ésima observación se puede escribir como:*

$$e_i = y_i - f(x_i; \widehat{\beta}), \tag{2.3}$$

donde y_i es el valor observado en el i-ésimo tiempo y $f(x_i; \widehat{\beta})$ es la predicción que se realiza con el modelo fijado para las variables explicativas en el i-ésimo tiempo y los parámetros beta estimados.

Los residuos pueden ser nulos, positivos o negativos. Si el residuo es nulo, no existe error en la estimación, ya que los valores observados son iguales a los valores estimados. Si es positivo entonces el valor observado es mayor que el estimado y, por el contrario, si el residuo es negativo entonces el valor observado será menor que el valor estimado.

Actualmente hay diversas maneras para estimar el rendimiento y evaluar como de bien se comporta el modelo ajustado a una serie temporal, algunas de ellas son: la raíz del error cuadrático medio (RMSE), el error absoluto medio (MAE) y el error porcentual absoluto medio (MAPE). De acuerdo con Mauricio (2007) [2] se puede definir los errores como:

Definición 2.7. ***(Raíz del error cuadrático medio)*** *Sea $\widehat{y}_1, \widehat{y}_2, ..., \widehat{y}_t$ una secuencia de valores predichos. La raíz del error cuadrático medio (RMSE, del inglés Root Mean Squared Error) asociada con dicha secuencia es:*

$$RMSE \equiv \sqrt{\frac{\sum_{t=1}^{T} (\widehat{y}_t - y_t)}{T}} \tag{2.4}$$

Representa la raíz cuadrada de la distancia cuadrada promedio entre el valor real y el valor pronosticado. Indica cómo de cerca están los datos observados de los valores predichos del modelo. Cuanto más se aproxime RMSE a cero mejor será la predicción. Es el criterio más importante para comprobar si estamos ante una buena predicción (González, 2020 [16]).

Definición 2.8. ***(Error absoluto medio)*** *Sea $\widehat{y}_1, \widehat{y}_2, ..., \widehat{y}_n$ una secuencia de valores predichos. El error absoluto medio (MAE, del inglés Mean Absolute Error) asociado con dicha secuencia es:*

$$MAE \equiv \frac{\sum_{i=1}^{n} |y_i - \widehat{y}_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}, \tag{2.5}$$

es decir, el promedio de la diferencia absoluta entre el valor observado y los valores predichos.

Definición 2.9. ***(Error porcentual absoluto medio)*** *Mide el tamaño del error absoluto 2.8 en téminos porcentuales. El error porcentual absoluto medio (MAPE, del inglés Mean Absolute Percentage Error) es:*

$$MAPE \equiv \frac{100}{n} \sum_{i=1}^{n} |\frac{y_i - \widehat{y}_i}{y_i}|, \tag{2.6}$$

Se empezará la siguiente sección hablando sobre los primeros modelos que forman el ARIMA. Estos modelos son la fusión de los modelos autorregresivos (AR), los procesos integrados (I) y los modelos de medias móviles (MA).

2.2.2. Los Modelos Autorregresivos (AR)

En este apartado se comentarán los modelos autorregresivos, los cuales son modelos de procesos estacionarios que se utilizan para representar la dependencia de los valores de una serie temporal con su pasado. Tal y como indica Peña (2010) [18], son los modelos más simples, ya que representan la dependencia entre dos variables aleatorias usando regresión lineal.

Hay que recordar que el modelo que explica el comportamiento de una variable y_t como función lineal de otra x_t es el modelo de regresión simple, este se puede representar mediante la siguiente ecuación:

$$y_t = c + bx_t + a_t$$

donde a es una variable aleatoria normal, con media nula y varianza constante y c y b son constantes a determinar. Aplicando la dependencia en cierta serie temporal, z_t, y tomando $y_t = z_t$ y $x_t = z_{t-1}$, conseguiremos el modelo autorregresivo de orden uno, también llamado AR(1). Se dirá que una serie temporal z_t tiene un modelo AR(1) si ha sido originada por:

$$z_t = c + \phi z_{t-1} + a_t \tag{2.7}$$

donde a_t es un proceso de ruido blanco con varianza σ^2 y c y $-1 < \phi < 1$ son constantes a determinar.

Puede ocurrir que la dependencia lineal del valor actual de la serie temporal, z_t, dependa no sólo de z_{t-1}, sino también de los p retardos anteriores,

$z_{t-2}, ..., z_{t-p}$. Debido a esto, se conseguirá un modelo autorregresivo de orden p. Diremos que una serie z_t estacionaria sigue un modelo autorregresivo de orden p si:

$$\tilde{z_t} = \phi_1 \tilde{z}_{t-1} + ... + \phi_p \tilde{z}_{t-p} + a_t \tag{2.8}$$

donde a_t un proceso de ruido blanco y $\tilde{z_t} = z_t - \mu$, siendo μ la media del proceso estacionario z_t. La ecuación de un modelo AR(p) es:

$$(1\text{-}\ \phi_1 B - .... - \phi_p B^p)\tilde{z_t} = a_t, \tag{2.9}$$

y denominando $\phi_p(B) = 1 - \phi_1 B - ...\phi_p B^p$ al polinomio de grado p donde el primer término es la unidad, tenemos:

$$\phi_p(B)\tilde{z_t} = a_t \tag{2.10}$$

que es la expresión general de un modelo autorregresivo.

Estos modelos poseen "memoria larga", debido a que el valor actual tiene correlación con todos los valores anteriores, siendo los coeficientes decrecientes. Los modelos autorregresivos no pueden representar series en los que el valor actual de la serie tenga correlación con unos pocos valores que le preceden, de forma que la función de autocorrelación (FAC) tenga sólo unas pocas autocorrelaciones diferentes de cero. Estos procesos los veremos en el próximo apartado, denominándose modelos de media móvil.

2.2.3. Los Modelos de Medias Móviles (MA)

Los modelos de medias móviles o procesos MA (del inglés *Moving Average*) son una aproximación de series temporales univariantes. Este modelo detalla que la variable de salida depende del valor actual y de varios de los valores que le preceden. Una de las grandes diferencias que guardan con los modelos anteriores es que los modelos de medias móviles son siempre estacionarios.

Posteriormente, se combinará estos modelos con los autorregresivos para formar los procesos ARMA. Estos modelos estocásticos estacionarios son de mucha utilidad para representar una gran cantidad de series temporales.

Se define el modelo de media móvil de primer orden, MA(1), basándose en la idea de que la innovación de ayer permanece hoy parcialmente. Se representa mediante la ecuación:

$$\tilde{z_t} = a_t - \theta a_{t-1} \tag{2.11}$$

donde a_t un proceso de ruido blanco con varianza σ^2 y $\tilde{z_t} = z_t - \mu$ siendo μ la media del proceso.

Generalizando, se puede escribir modelos en los cuales su valor actual no dependa únicamente de la última innovación, sino de las q últimas innovaciones. Se obtiene el modelo MA(q), cuya representación es:

$$\tilde{z_t} = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - ... - \theta_q a_{t-q}$$

Introduciendo la notación de operadores:

$$\tilde{z_t} = (1 - \theta_1 B - \theta_1 B^2 - ... - \theta_q B^q) a_t \tag{2.12}$$

escribiéndose de forma más simple como:

$$\tilde{z_t} = \theta_q(B) a_t \tag{2.13}$$

Cabe recordar que los modelos autorregresivos se caracterizaban por tener una "memoria larga", en cambio, los modelos de medias móviles tienen la propiedad de tener una "memoria muy corta", ya que el valor actual sólo está correlado con un pequeño número de valores que le preceden en la serie.

2.2.4. Los Modelos Autorregresivos de Medias Móviles (ARMA)

Los modelos autorregresivos de medias móviles (en inglés *AutoRegressive Moving Average*) es la fusión de los modelos vistos anteriormente. Este modelo se denomina ARMA (p, q), donde p es el orden de la parte autorregresiva y q es el orden de media móvil.

De acuerdo con M. Alonso (2007) [17], los procesos ARMA permiten representar procesos donde los primeros q valores son cualesquiera, mientras que los siguientes decrecen.

El proceso más simple es el ARMA (1,1) y se escribe de la siguiente manera:

$$\tilde{z_t} = \phi_1 \tilde{z}_{t-1} + a_t - \theta_1 a_{t-1},$$

o, con notación de operadores:

$$(1 - \phi_1 B)\tilde{z_t} = (1 - \theta_1 B) a_t, \tag{2.14}$$

donde $|\theta_1| < 1$ para que sea invertible y $|\phi_1| < 1$ para que el proceso sea estacionario. Generalizando, el proceso ARMA(p, q) será:

$$(1\text{-}\phi_1 B - ... - \phi_p B^p)\tilde{z_t} = (1 - \theta_1 B - ... - \theta_q B^q) a_t \tag{2.15}$$

o, en notación compacta,

$$\phi_p(B)\tilde{z_t} = \theta_q(B) a_t$$

Uno de los motivos que explica por qué los procesos ARMA son frecuentes en la práctica, es que en el caso de tener procesos que sean suma de otros, si alguno de ellos tiene estructura de un modelo autorregresivo, resultará ser un proceso ARMA.

2.2.5. Procesos integrados(I)

Comenzaremos a estudiar los procesos no estacionarios. Los procesos no estacionarios más relevantes son los procesos integrados debido a que cuando son diferenciados se consiguen procesos estacionarios. La propiedades más importantes que tienen estos procesos es la manera en que llega a desaparecer la dependencia con respecto al tiempo. Los modelos ARMA son procesos estacionarios en los que van disminuyendo las autocorrelaciones y se acercan a cero a los pocos retardos. Tal y como menciona Peña (2010) [18], en los procesos integrados estas autocorrelaciones decrecen con el tiempo y son capaces de descubrir coeficientes de autocorrelación diferentes de cero hasta retardos más altos.

Siendo z_t una serie no estacionaria, en el caso de llamar $w_t = \nabla z_t$ a una nueva serie, podremos observar que los valores se mueven alrededor de una media constante y se asemeja a una serie estacionaria. Se concluye que si z_t se convierte en una serie estacionaria a través de una diferencia, este proceso se denominara proceso integrado de primer orden, siendo por tanto la cantidad de diferencias que se necesiten para lograr un proceso estacionario. Generalizando, se dirá que un proceso es integrado de orden $h \geq 0$, y se representará por $I(h)$, cuando se consiga un proceso estacionario al diferenciarlo h veces.

2.2.6. Los Modelos Autorregresivos Integrados de Media Móvil (ARIMA)

En esta sección se hablará sobre los modelos en los que se basa nuestro estudio. Previamente se ha comentado como se forman los procesos autorregresivos, de medias móviles e integrados, los cuales son los que forman este modelo (Peña, 2010 [18]). Uniendo estos tres procesos obtenemos:

$$(1 - \phi_1 B - ... - \phi_p B^p)(1 - B)^d z_t = c + (1 - \theta_1 B - ... - \theta_q B^q)a_t$$

que serán denotados como procesos ARIMA (p, d, q). En esta notación p es el orden de la parte autorregresiva estacionaria, d es el número de raíces unitarias y q es el orden de la parte media móvil. Utilizando el operador diferencia, $\nabla = 1 - B$, el proceso anterior puede escribirse:

$$\phi_p(B)\nabla^d z_t = c + \theta_q(B)a_t. \tag{2.16}$$

El nombre ARIMA proviene de las iniciales del inglés *Autoregresive Integrated Moving Average*, donde 'integrado' de orden d hace referencia a que el proceso z_t puede ser obtenido como suma del proceso estacionario $\omega_t = \nabla^d z_t$. En efecto, resulta que:

$$z_t = (1 - B)^{-1}\omega_t = \sum_{j=-\infty}^{t} \omega_t.$$

Cuando se introducen diferencias estacionales estos modelos pueden ser reescritos de forma más reducida como los modelos ARIMA estacionales que serán descritos en la siguiente sección. Estos se pueden escribir de forma simplificada como el modelo ARIMA $(p, d, q) \times (P, D, Q)$.

2.2.7. Modelos ARIMA estacionales

Como se ha visto anteriormente se puede convertir series no estacionarias en estacionarias haciendo diferencias regulares. También se ha explicado que a través de diferencias estacionales se puede eliminar la estacionalidad.

Según M. Alonso (2007) [17], se puede convertir una serie temporal con estacionalidad en estacionaria por medio de la transformación:

$$w_t = \nabla_s^D \nabla^d z_t$$

donde d el número de diferencias regulares y D es el número de diferencias estacionales. Para los casos que exista dependendencia estacional se podrá usar el modelo ARIMA para series temporales estacionarias incluyendo aparte de la dependencia regular, la dependencia estacional, la cual está vinculada a observaciones alejados por s periodos (con datos mensuales se usará $s = 12$). Si se modela por separado la dependencia estacional y la regular se obtiene el modelo ARIMA estacional multiplicativo, cuyo resultado es:

$$\Phi_P(B^s)\phi_p(B)\nabla_s^D\nabla^d z_t = \theta_q(B)\Theta_Q(B^s)a_t, \tag{2.17}$$

donde $\phi_p = (1-\phi_1 B - ...\phi_p B^P)$ es el operador AR regular de orden p, $\Phi_P(B^s) = (1 - \Phi_1 B^s - ...\Phi_P B^s P)$ es el operador AR estacional de orden P, $\nabla^d = (1-B)^d$ las diferencias regulares y $\nabla_s^D = (1 - B^s)$ representa las diferencias estacionales.

$$\Theta_Q(B^s) = (1 - \Theta_1 B^s - ... - \Theta_Q B^{sQ})$$

es el operador media móvil estacional de orden Q,

$$\theta_q(B) = (1 - \theta_1 B - ...\theta_q B^q)$$

es el operador media móvil regular de orden q y a_t es un proceso de ruido blanco.

Esta clase de modelos, introducidos por Box y Jenkins representan muchas series estacionales que encontramos en la práctica y se escriben de forma simplificada como el modelo ARIMA $(p, d, q) \times (P, D, Q)_s$.

En nuestro caso, trabajaremos con $s = 12$ ya que trataremos con datos mensuales, el modelo quedaría de la siguiente manera:

$$(p, d, q) \times (P, D, Q)_{12} \tag{2.18}$$

El modelo de las series temporales que son anuales pueden escribirse como:

$$(1\text{-}\phi_1 B_{12} - ... - \phi_P B^{12} P)(1 - B^{12})z_t = (1 - \Theta_1 B^{12} - ... - \Theta_Q B^{12} Q)\alpha_t \quad (2.19)$$

donde $t = 1, ..., T$, y como estamos relacionando meses de diferentes años el modelo ARIMA se expresa en B^{12}. Al ser el mismo modelo para todos los meses, se puede aplicar a la serie inicial z_t y conseguir así la serie de residuos α_t.

En definitiva, el modelo ARIMA estacional multiplicativo se centra en que la dependencia estacional (2.19) es igual para todos los periodos.

En el software que vamos a utilizar para nuestro estudio, existe un conjunto de métodos estadísticos para el ajuste estacional y otros análisis descriptivos que suelen estar englobados dentro de la herramienta X-13-ARIMA-SEATS. Este método ha sido puesto en práctica en lugares tan influyentes como la Oficina del Censo de EE.UU, la Oficina de Estadística de Australia y las oficinas de estadística de muchos otros países.

El método que se utiliza para el ajuste estacional se centra en el algoritmo X-11. El objetivo es evaluar los cuatro componentes que tiene una serie temporal para posteriormente eliminar la componente estacional estableciendo una serie temporal estacionalmente ajustada.

Es posible incorporar a estos modelos variables auxiliares o exógenas, denontándolos como modelos ARIMAX.

2.3. Gretl

Gretl es un software desarrollado para la estimación de modelos econométricos y el análisis estadístico. Para analizar series temporales, el programa ofrece diversas opciones: se muestran los correlogramas de las funciones de autocorrelación y autocorrelación parcial sobre la variable que hayamos escogido, permitiendo seleccionar el número de retardos (Pérez, 2016 [19]).

Algunas de las propiedades que más destacan es el uso de distintos métodos de series temporales tales como modelos ARIMA o modelos GARCH, VAR y VECM. También sobresale la gran variedad de poder usar diferentes estimadores (máxima verosimilitud, mínimos cuadrados, métodos de sistema y de ecuación única , GMM). Tiene una sencilla interfaz que está disponible en diversos idiomas y posee múltiples facilidades para el intercambio de datos con Octave, Python y R.

Como hemos comentado anteriormente, para la identificación se usará 'Análisis TRAMO', el cual no viene incorporado inicialmente en Gretl pero se puede descargar de manera gratuita en su página web oficial. En ocasiones, la transformación logarítmica de la serie temporal puede producir una estabilización de la variabilidad y con ello conseguir que las condiciones de aplicación de los diferentes modelos de series temporales introducidos puedan verificarse. Procederemos a continuación con las aplicaciones y resultados.

3

Aplicaciones y resultados

La implementación de las herramientas sobre series temporales introducidas en los capítulos anteriores serán ejemplificadas sobre datos reales. En este capítulo se desarrolla el estudio para comprobar si las búsquedas en tiempo real influyen en mejorar las predicciones del número de parados. El número de parados se sitúa en marzo de 2021 en 280.650 personas, el 25,42 % del total de la población activa. Esto es debido a que vivimos en una comunidad en la que dependemos en gran medida del sector turístico, el cual se ha visto mermado debido a la gran relevancia que ha tenido para este sector la pandemia del COVID-19. La importancia que ha tenido esta pandemia ha producido un aumento exponencial en el número de parados.

Han sido numerosos los investigadores que a lo largo de los años se han dado cuenta que la gran cantidad de información que procede de las búsquedas en tiempo real podría resultar beneficioso para mejorar los estudios de fenómenos sociales (Choi y Varian, 2012 [11]; Guzmán, 2011 [13]). Esto es debido a que la información de las búsquedas en internet de las personas revelan sus necesidades y puede contribuír a la predicción de anormalidades, tal como sucede en nuestro caso con el COVID-19. Respecto al fenómeno estudiado en este trabajo, numerosos autores sugieren el uso de datos de Internet para pronosticar el desempleo (D'Amuri y Marcucci, 2009 [15] ; McLaren y Shanbhogue, 2011 [12]), y los resultados muestran que, de hecho, pueden considerarse útiles en la estimación.

El propósito de este trabajo es determinar si el uso de la información de Google Trends junto con datos de la serie temporal de la población desempleada puede mejorar la precisión y la previsión del número de parados y permitir ofrecer una "inmediata"previsión para su uso. De forma esquemática el proceso aplicado sigue los siguientes pasos:

- Se hace un estudio sobre cuáles serán las variables que entrarán en el modelo.
- Con estas variables y con los datos del número de parados se dispondrá a realizar el análisis TRAMO-SEATS con el propósito de obtener el modelo ARIMA para nuestra serie temporal.

- Una vez obtenido el modelo que se ajuste a nuestra serie, se averiguará si las variables que se han elegido tienen la suficiente relevancia para mejorar las predicciones.
- Se realizará el mismo proceso cogiendo los residuos del número de parados y se analizará otro modelo para estos. Se comparará si con este nuevo modelo se consigue un menor error que para el caso anterior.
- Finalmente, se hará un pronóstico del número de parados hasta diciembre de 2021 seleccionando el modelo que mejor se adapte a nuestra variable, es decir, el que tenga menor error.

3.1. Número de parados

Los datos del número de parados se han obtenido de la página oficial del Instituto Canario de Estadística (ISTAC), los cuales serán modelizados en forma logarítmica con el fin de reducir la sensibilidad de las estimaciones a observaciones extremas.

Una vez se han visto todos los conceptos básicos y desarrollado los modelos ARIMA, se dispondrá a realizar la parte práctica, donde se hará un análisis sobre los cambios de cifras en el desempleo que se ha experimentado en Canarias de marzo de 2006 hasta marzo de 2021. Posteriormente, se hará una predicción sobre como se comportará nuestra variable.

Figura 3.1: Representación del desempleo en Canarias

En la figura 3.1 se puede localizar de manera clara el cambio de tendencia que se produce en Canarias a principios del año 2008. Se puede observar como aumenta exponencialmente desde esta fecha hasta tocar un máximo a mediados de 2013. A continuación, se tiene una tendencia bajista hasta el mes de marzo del año 2020 en el que los datos vuelven a subir debido a otra gran crisis en la que todavía nos vemos sumergidos, la crisis del COVID-19.

El análisis se va a llevar a cabo en los periodos comprendidos entre marzo de 2006 y marzo de 2021, incluyendo así las dos grandes crisis de los últimos tiempos.

Los datos oficiales del número de parados son los siguientes:

Año\Mes	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre
2006			136536	135885	133443	132328	130291	127768	125720	126004	125702	122153
2007	126736	129444	130648	133091	135268	135557	132139	132797	133912	137259	138813	139081
2008	148560	154109	158842	166241	171511	175824	176101	176339	180805	190469	199889	202993
2009	215976	227706	238344	246500	249661	248325	246264	245966	248858	248108	248786	248783
2010	253930	257939	261534	266042	265916	265973	262661	260794	259751	260273	256693	254620
2011	258311	261176	261202	256773	254341	251963	253629	252478	251990	258264	266213	265569
2012	273983	278898	283699	288752	293228	290664	289785	288666	288813	287820	289032	284915
2013	289517	291474	291672	295824	296362	293054	290375	286081	284072	284309	283378	274053
2014	276034	275551	277498	276786	274412	270059	268422	266668	266637	265791	265385	260682
2015	259743	258687	256851	254016	251941	248223	243906	242649	242313	247162	248639	247529
2016	245984	243632	242888	242063	240067	236876	232069	230185	230901	232874	234774	229233
2017	231774	229900	230779	225702	222749	220079	217895	217045	221006	221794	219698	216087
2018	215701	215082	213768	213141	212411	208594	206747	207017	205430	208101	209975	207015
2019	209419	209466	209235	207618	206041	202683	204662	205173	204529	210131	210893	208249
2020	211164	207837	227634	254981	261074	261714	257649	257406	254280	262487	268319	269437
2021	279230	283477	280650									

Tabla 3.1: Número de parados

El análisis de modelos ARIMA se desarrollará usando Gretl, el cual nos valdrá para realizar todas las predicciones que necesitemos.

Recapitulando, se analizará la dirección del desempleo en Canarias añadiendo variables auxiliares obtenidas de Google Trends a lo largo del periodo comprendido entre 2006 y 2021. El modelo se establecerá hasta el mes previo que se produjese el COVID-19 (febrero de 2020) y se comprobará así cuan bueno es realizando las predicciones hasta los datos que tenemos (marzo de 2021). La segunda opción será hallar los residuos del número de parados y realizar el mismo estudio que en el caso anterior. Por último, se comprobará que opción se comporta mejor con nuestros datos y se hará un pronóstico sobre como será la evolución del desempleo hasta diciembre de 2021.

3.2. Detección de palabras clave

La relevancia y el potencial de Google Trends reside en la gran cantidad de información que contiene Google, siendo el motor de búsqueda más grande

del mundo. Otro de los aspectos a remarcar es la facilidad con la que se puede acceder a esta información.

Algunas de las consecuencias de extraer información de grandes cantidades de datos son los múltiples errores que se pueden cometer y la inapropiada elección de una determinada palabra para entrar al modelo. Estos errores pueden darse debido a que nos podamos encontrar con que la información obtenida de Google Trends tenga correlaciones falsas. El motivo de que dos o más fenómenos estén correlados no significa que haya una conexión de causa-efecto entre estos. Esta correlación podría ser aleatoria o que tuviesen otra variable en común. Para reducir este riesgo, se realizará un listado de posibles palabras más influyentes que podría buscar una persona cuando se queda en paro. En este proceso se han incluido 30 palabras (trabajo, erte, infoempleo, infojobs, linkedin, paro, sepe, inem, ...) de las cuales, introduciéndolas en Google Trends y realizando el coeficiente de correlación entre las palabras con la variable en estudio (número de parados), se han obtenido cinco palabras que usaremos para mejorar nuestro modelo, que son:

- Paro
- Sepe
- Infojobs
- Seguridad Social
- Empleo

Aquí podemos ver la tendencia de estas palabras en Google:

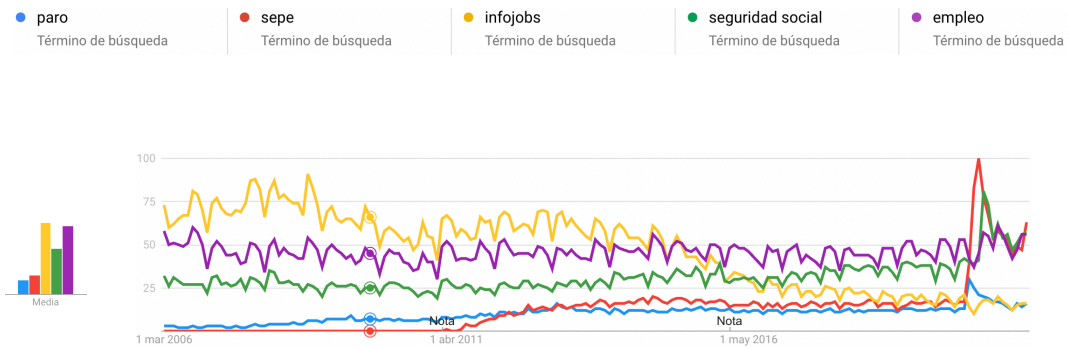

Figura 3.2: Tendencia de las palabras escogidas

Descargaremos los datos de interés sobre cada una de las palabras desde Google Trends y estos serán introducidos en la base de datos con el número de parados que hay en cada mes. A continuación, la serie de palabras claves, así como con uno y dos retardos, será correlacionada con el número de parados para averiguar si esa palabra tiene efecto en el mismo mes en que la persona se queda en paro o si tiene un mayor efecto a uno o dos meses vista.

Aplicando el coeficiente de correlación de estos meses (el propio mes y los dos con retardo) en cada una de las cinco palabras escogidas, se elige aquellos coeficientes que estuviesen más cercanos a 1, ya que obtendremos así en que mes tiene mayor correlación dicha palabra con el número de parados.

Los coeficientes de correlación de cada palabra en el mes sin retardo (0), con un mes de retardo (1) y con dos meses de retardo (2) son:

Variables\Retardos	0	1	2
Paro	0,48361205	**0,5037634**	0,50365964
Sepe	0,26465733	**0,26553353**	0,25990584
Seguridad social	0,25930669	0,25747929	**0,26786434**
Infojobs	0,1883868	**0,19250135**	0,18927434
Empleo	0,1403921	**0,10889822**	0,10139553

Tabla 3.2: Retardos de las variables auxiliares

En la tabla 3.2 se puede observar que en todas las palabras se ha escogido el valor que más próximo esté de uno, salvo para el caso de la palabra 'Empleo'. Se ha considerado que lo más lógico cuando una persona se quede en paro en un determinado mes tenga mayor repercusión como mínimo en el mes siguiente, ya que antes de que no acabe el mes no sabríamos si las búsquedas han sido mayores o menores a los meses anteriores.

El resultado de escoger dichos retardos nos da la siguiente gráfica:

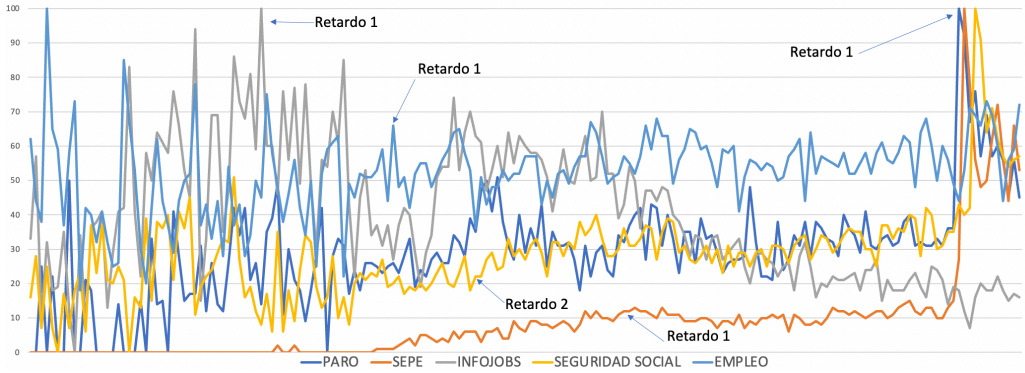

Figura 3.3: Selección de las variables auxiliares y sus retardos

Una vez conseguida la basa de datos final, será introducida en Gretl para estudiar si estas búsquedas pueden ayudar a mejorar las predicciones del número de parados.

3.3. Obtención del modelo ARIMA

En esta sección se empezará la primera fase para la obtención del modelo ARIMA. El objetivo principal es determinar los valores de p, d y q, es decir, conseguir el mejor modelo que se va a adaptar a la serie.

En primer lugar, se aplicará el logaritmo al número de parados para mejorar la estabilidad del modelo. Posteriormente, se verá si la media del proceso es o no constante para saber si se trata de una serie estacionaria.

La figura 3.4 muestra la función de autocorrelación (FAC) y la función de autocorrelación parcial (FACP), donde se puede ver que la función de autocorrelación va disminuyendo poco a poco y de forma progresiva, y la función de autocorrelación parcial indica que existe una correlación significativa en la fase 1 seguido de correlaciones que no son significativas.

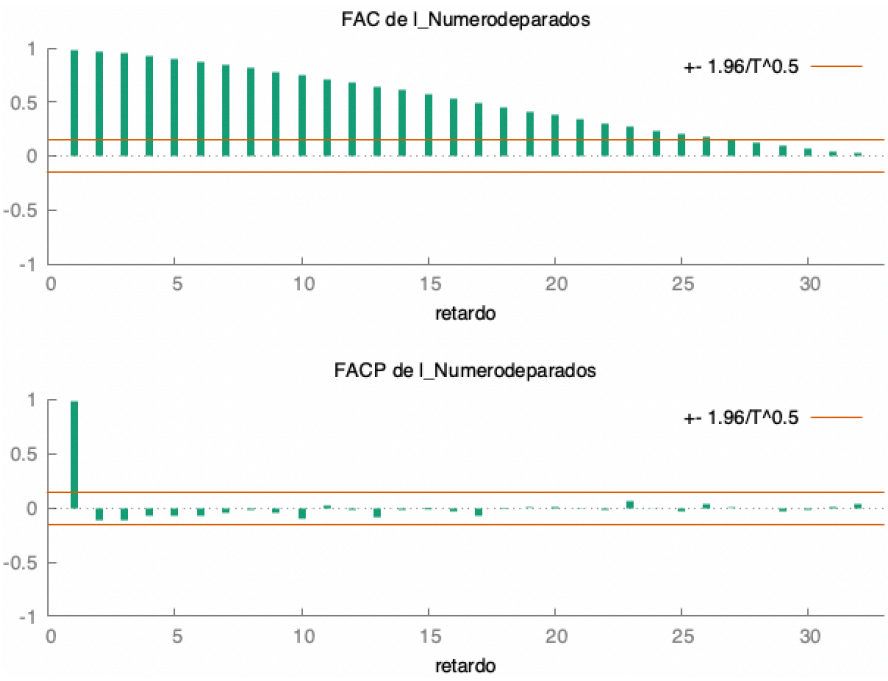

Figura 3.4: FAC y FACP

Para comprobar que estamos ante una serie estacionaria haremos los contrates de hipótesis, que nos ayudarán a verificar la estacionalidad en media. Para ello, empleamos el contraste aumentado de Dickey-Fuller.

$$\begin{cases} H_0 : \text{Existe una raíz unitaria (la serie no es estacionaria en media)} \\ H_1 : \text{No existe raíz unitaria (la serie es estacionaria en media).} \end{cases}$$

El resultado es:

```
Contraste aumentado de Dickey-Fuller para l_Numerodeparados
contrastar hacia abajo desde 13 retardos, con el criterio AIC
tamaño muestral 154
la hipótesis nula de raíz unitaria es: [a = 1]

  contraste con constante
  incluyendo 13 retardos de (1-L)l_Numerodeparados
  modelo: (1-L)y = b0 + (a-1)*y(-1) + ... + e
  valor estimado de (a - 1): -0.0194268
  estadístico de contraste: tau_c(1) = -3.53455
  valor p asintótico 0.00717
  Coef. de autocorrelación de primer orden de e: -0.021
  diferencias retardadas: F(13, 139) = 11.959 [0.0000]

  con constante y tendencia
  incluyendo 13 retardos de (1-L)l_Numerodeparados
  modelo: (1-L)y = b0 + b1*t + (a-1)*y(-1) + ... + e
  valor estimado de (a - 1): -0.018542
  estadístico de contraste: tau_ct(1) = -3.36746
  valor p asintótico 0.05581
  Coef. de autocorrelación de primer orden de e: -0.024
  diferencias retardadas: F(13, 138) = 8.426 [0.0000]
```

Figura 3.5: Contraste de Dickey-Fuller

Observando el resultado del contraste de Dickey-Fuller (Figura 3.5), se obtiene que el p-valor $< 0,01 < 0.05$, por lo tanto se rechaza H_0, lo que implica que la serie es constante en media.

Una vez se ha comprobado que la serie es constante en media, quedará verificar que la serie sea constante en varianza para corroborar que estamos ante una serie estacional. Para confirmar que es constante en varianza se hará el gráfico rango-media. El test de hipótesis y sus respectivos resultados son:

$$\begin{cases} \mathrm{H}_0 : \text{El rango no varía en función de la media (la varianza es constante)} \\ \mathrm{H}_1 : \text{El rango varía en función de la media (la varianza no es constante)} \end{cases}$$

```
Estadísticos de rango-media para l_Numerodeparados
Utilizando 14 submuestras de tamaño 12

                          Rango           media
2006:03 - 2007:02         0.111314         11.7696
2007:03 - 2008:02         0.165153         11.8310
2008:03 - 2009:02         0.360145         12.1328
2009:03 - 2010:02         0.0790081        12.4229
2010:03 - 2011:02         0.0438820        12.4727
2011:03 - 2012:02         0.101564         12.4696
2012:03 - 2013:02         0.0330366        12.5737
2013:03 - 2014:02         0.0782602        12.5630
2014:03 - 2015:02         0.0701949        12.4969
2015:03 - 2016:02         0.0582659        12.4200
2016:03 - 2017:02         0.0578615        12.3649
2017:03 - 2018:02         0.0704412        12.3026
2018:03 - 2019:02         0.0397860        12.2512
2019:03 - 2020:02         0.0409919        12.2421

Pendiente de 'rango' con respecto a 'media' = -0.151654
El valor p para H0: [Pendiente = 0 es 0.107472]
```

Figura 3.6: Estadísticos de rango-media para la variable 'Número de parados'

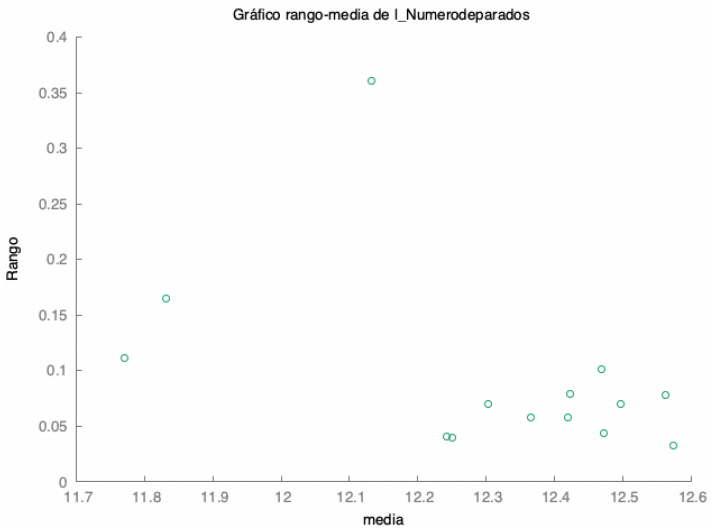

Figura 3.7: Rango-media del 'Número de parados'

El p-valor es superior a 0.05, por tanto aceptamos H_0 y se demuestra que la varianza es también constante.

Cuando sabemos que la serie es estacionaria en media y en varianza, se comprobará cuál es el modelo fijado para los datos del número de parados y las variables auxiliares (paro, sepe, infojobs, seguridad social y empleo).

Se establece el rango que se quiera que tenga el modelo, la primera opción que se hará será coger el rango de la muestra hasta febrero de 2020, antes de que empezase en España la pandemia del COVID-19. Se comprobará así la optimalidad con la que se puede ajustar el modelo.

El modelo fijado es el siguiente:

```
MODEL FITTED

     NONSEASONAL     P= 0     D= 2     Q= 1
        SEASONAL    BP= 0    BD= 1    BQ= 1
     PERIODICITY    MQ= 12

         MEAN     =     0.00000

         SE       = *******



                ARIMA PARAMETERS

          TH    =   -0.5731
          SE    =    *****
          BTH   =   -0.6871
          SE    =    *****


RESIDUALS
   X  10.0D-1
```

Figura 3.8: Modelo de la variable 'Número de parados' fijado hasta 02/20

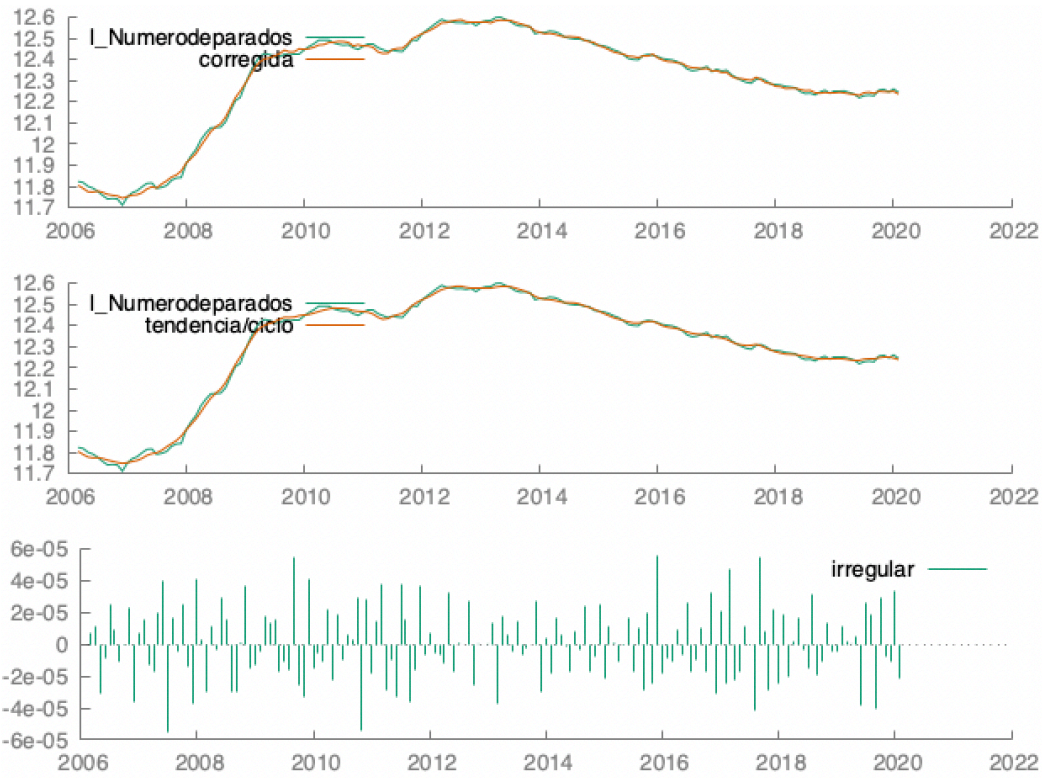

Figura 3.9: Optimalidad del modelo hasta 02/20

A partir de la detección del mejor modelo (Figura 3.8), Gretl nos muestra que estamos ante un ARIMA $(0,2,1) \times (0,1,1)_{12}$, donde (0,1,1) son las variaciones estacionales o parte cíclica de la serie. En la figura 3.9, se puede observar que la línea naranja es el modelo ajustado al número de parados, línea verde, la cual casi no se puede dislumbrar ya que el modelo se ajusta con gran exactitud.

3.4. Validación del modelo y predicciones

Después de que se tenga el modelo fijado se dispondrá a comprobar si las variables auxiliares tienen la suficiente relevancia para corregir el modelo fijado ante un brusco cambio de tendencia como es el caso a partir de marzo de 2020.

```
Modelo 2: ARMAX, usando las observaciones 2007:05-2020:02 (T = 154)
Estimado usando X-13-ARIMA (MV exacta)
Variable dependiente: (1-L)^2(1-Ls) l_Numerodeparados

                  coeficiente   Desv. típica     z      valor p
  ------------------------------------------------------------
  const           -8.16308e-05  0.000152488    -0.5353  0.5924
  theta_1         -0.568457     0.0653613      -8.697   3.40e-18 ***
  Theta_1         -0.682621     0.0605146     -11.28    1.64e-29 ***
  PARO             1.01966e-05  5.65240e-05     0.1804  0.8568
  SEPE            -0.000302501  0.000476484    -0.6349  0.5255
  INFOJOBS        -3.27597e-05  4.33916e-05    -0.7550  0.4503
  SEGURIDADSOCIAL  8.79282e-05  8.44306e-05     1.041   0.2977
  EMPLEO           2.57905e-05  5.20214e-05     0.4958  0.6201

Media de la vble. dep. -0.000256   D.T. de la vble. dep.    0.015427
Media de innovaciones  -0.000126   D.T. innovaciones        0.010939
R-cuadrado              0.996823   R-cuadrado corregido     0.996694
Log-verosimilitud       472.8995   Criterio de Akaike      -927.7990
Criterio de Schwarz    -900.4664   Crit. de Hannan-Quinn   -916.6966

                     Real Imaginaria     Módulo Frecuencia
  ------------------------------------------------------------
  MA
   Raíz  1           1.7591     0.0000     1.7591     0.0000
  MA (estacional)
   Raíz  1           1.4649     0.0000     1.4649     0.0000
  ------------------------------------------------------------
```

Figura 3.10: Modelo ARMAX sobre el número de parados fijado hasta 02/20 ($*0,1\,\%$ significativo,$**0,05\,\%$ significativo, $***0,01\,\%$ significativo)

Analizando este modelo (Figura 3.10), todas las variables auxiliares tienen un p-valor mayor que 0,05, por lo que se acepta la hipótesis nula H_0, y las variables son no significativas.

Se van a realizar las predicciones de este modelo desde marzo de 2020 hasta marzo de 2021. Seguidamente, se examinará si este modelo es capaz de predecir la gran subida que produjo la crisis del COVID-19.

Figura 3.11: Predicción de 03/2020 hasta 03/2021

Como podemos observar en la figura 3.11, las predicciones no son buenas. El modelo no consigue cambiar la tendencia con los datos previos y las variables auxiliares no son significativas.

Se hará el mismo proceso cambiando el rango de la muestra a meses en los que ya se podría visualizar los datos producidos por la pandemia. Una vez que se cogan algunos datos influenciados por esta crisis, se observará si el modelo puede corregir la tendencia y hacer una mejor predicción.

Estableciendo el rango hasta abril de 2020, teniendo así dos meses (marzo y abril) bajo los efectos de la pandemia. Los resultados son:

```
Modelo 4: ARMAX, usando las observaciones 2007:05-2020:04 (T = 156)
Estimado usando X-13-ARIMA (MV exacta)
Variable dependiente: (1-L)^2(1-Ls) l_Numerodeparados

                  coeficiente   Desv. típica     z     valor p
  ------------------------------------------------------------
  const            0.000155989  0.000247827    0.6294  0.5291
  theta_1         -0.449136     0.0818340     -5.488   4.06e-08 ***
  Theta_1         -0.689546     0.0771312     -8.940   3.89e-19 ***
  PARO             0.000111275  6.61277e-05    1.683   0.0924   *
  SEPE             0.000788057  0.000553860    1.423   0.1548
  INFOJOBS        -3.58077e-05  5.32791e-05   -0.6721  0.5015
  SEGURIDADSOCIAL  0.000162686  0.000102642    1.585   0.1130
  EMPLEO           1.36771e-05  6.42990e-05    0.2127  0.8316

Media de la vble. dep.  0.000628   D.T. de la vble. dep.   0.017764
Media de innovaciones  -0.000181   D.T. innovaciones       0.014314
R-cuadrado              0.994446   R-cuadrado corregido    0.994223
Log-verosimilitud       437.1129   Criterio de Akaike     -856.2258
Criterio de Schwarz    -828.7771   Crit. de Hannan-Quinn  -845.0773

                     Real Imaginaria     Módulo Frecuencia
  ------------------------------------------------------------
  MA
   Raíz  1           2.2265     0.0000     2.2265     0.0000
  MA (estacional)
   Raíz  1           1.4502     0.0000     1.4502     0.0000
```

Figura 3.12: Modelo ARMAX sobre el número de parados fijado hasta 04/20 ($*0,1\,\%$ significativo, $**0,05\,\%$ significativo, $***0,01\,\%$ significativo)

En este caso, la palabra 'Paro' es significativa al $0,1\,\%$ y veremos a continuación si las predicciones mejoran respecto a la anterior.

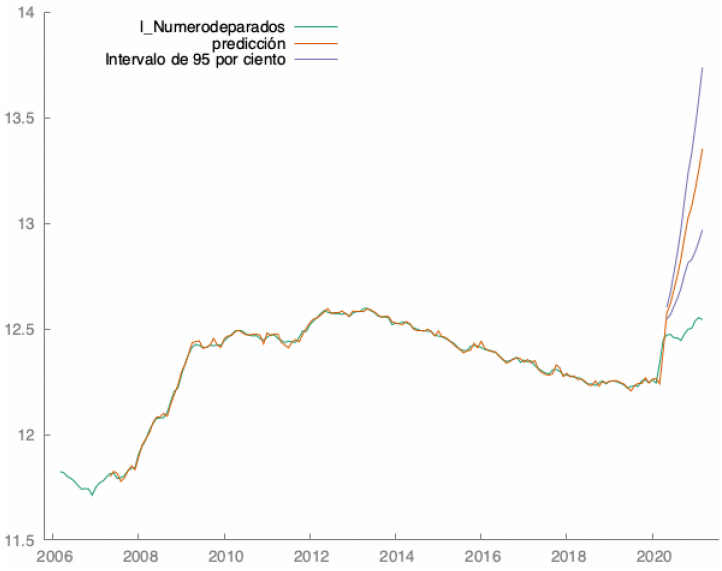

Figura 3.13: Predicción de 05/2020 hasta 03/2021

En la figura 3.13, podemos darnos cuenta de que ahora la predicción en vez de seguir una tendencia bajista, como en el caso anterior, presenta una tendencia más alcista que la propia crisis de la pandemia. Las causas que expliquen este comportamiento puede deberse al plan de amortiguación que el gobierno español hizo a través de los ERTES, en el que numerosas empresas se vieron salvadas a tener que dejar a más gente en paro.

Realizaremos una tabla con los diferentes errores que hemos definido anteriormente (2.7, 2.8 y 2.9). Se irá añadiendo un mes a la anterior predicción hasta llegar a marzo de 2021, con el fin de comprobar qué sucede según vamos teniendo más datos de la pandemia.

Predicciones	MAE	RMSE	MAPE	Palabras relevantes
03/2020 - 03/2021	0,29724	0,30756	2,3794	
04/2020 - 03/2021	0,084446	0,096498	0,67523	
05/2020 - 03/2021	0,44187	0,49396	3,5325	PARO *
06/2020 - 03/2021	0,44	0,48844	3,5169	PARO ** SEPE***
07/2020 - 03/2021	0,18078	0,19417	1,4449	PARO *** SEG.SOCIAL *
08/2020 - 03-2021	0,051603	0,056148	0,41245	PARO **
09/2020 - 03/2021	0,021142	0,025996	0,16912	PARO ***
10/2020 - 03/2021	0,056326	0,06262	0,44943	PARO **
11/2020 - 03/2021	0,027789	0,033014	0,22164	PARO **
12/2020 - 03/2021	0,021895	0,0251	0,17458	PARO ** SEPE*
01/2021 - 03/2021	0,020967	0,022653	0,16712	PARO **
02/2021 - 03/2021	0,032089	0,041695	0,25577	PARO **
03/2021-	0,051754	0,051754	0,41255	PARO **

Tabla 3.3: Errores de las predicciones del número de parados usando el modelo ARIMA $(0, 2, 1) \times (0, 1, 1)_{12}$

Observando la tabla 3.3, la raíz del error cuadrático medio (RMSE) va disminuyendo según se vayan teniendo más datos de los meses en los que ya sabemos que hay una pandemia, siendo de 0,3 en la predicción que realizamos de marzo de 2020 a marzo de 2021, mientras que si la hacemos en septiembre de ese mismo año es inferior a 0,03. También, el error porcentual absoluto medio (MAPE) disminuye de un 2,37 % en la primera predicción a un 0,17 % en caso de que fuese realizada en septiembre. Según se va variando el rango hacia meses a posteriori, el modelo se va ajustando y la variable 'Paro' pasa a ser significativa al 0,05 %. Las variables 'Sepe' y 'Seguridad social' entran de forma significativa en el modelo algunos meses.

A continuación se realizará una tabla entre el MAPE de la estimación y de la predicción con variables auxiliares y otra sin ellas. Posteriormente, se comprobará si las variables auxiliares influyen realmente y verifican la hipótesis de que las búsquedas en internet ayudan a mejorar las futuras predicciones sobre el número de parados.

Estimación		**Predicción**	
Desde mar-06 hasta		**De**	
feb-20	0,066998	mar-20 / mar-21	2,3794
mar-20	0,071833	abr-20 / mar-21	0,67523
abr-20	0,077229	may-20 / mar-21	3,5325
may-20	0,079128	jun-20 / mar-21	3,5169
jun-20	0,079768	jul-20 / mar-21	1,4449
jul-20	0,080877	ago-20 / mar-21	0,41245
ago-20	0,080955	sept-20 / mar-21	0,16912
sept-20	0,081745	oct-20 / mar-21	0,44943
oct-20	0,081936	nov-20 / mar-21	0,22164
nov-20	0,081766	dic-20 / mar-21	0,17458
dic-20	0,081891	ene-21 / mar-21	0,16712
ene-21	0,082087	feb-21 / mar-21	0,25577
feb-21	0,081829	mar-21	0,41255

Tabla 3.4: MAPE de la estimación y predicción con variables auxiliares del modelo $(0, 2, 1) \times (0, 1, 1)_{12}$

Estimación		Predicción	
Desde mar-06 hasta		De	
feb-20	0,06705	mar-20 / mar-21	2,308
mar-20	0,071652	abr-20 / mar-21	0,683
abr-20	0,078188	may-20 / mar-21	4,407
may-20	0,077309	jun-20 / mar-21	2,176
jun-20	0,07842	jul-20 / mar-21	1,460
jul-20	0,080039	ago-20 / mar-21	0,41946
ago-20	0,080645	sept-20 / mar-21	0,17047
sept-20	0,081713	oct-20 / mar-21	0,58543
oct-20	0,081942	nov-20 / mar-21	0,23665
nov-20	0,081892	dic-20 / mar-21	0,18357
dic-20	0,081718	ene-21 / mar-21	0,16963
ene-21	0,081781	feb-21 / mar-21	0,26022
feb-21	0,081284	mar-21	0,46036

Tabla 3.5: MAPE de la estimación y predicción sin variables auxiliares del modelo $(0, 2, 1) \times (0, 1, 1)_{12}$

Como podemos ver en la tabla 3.4, la estimación va aumentando poco a poco hasta estabilizarse en 0,081 % según se vayan incorporando más meses en la obtención de los paramétros del modelo.

En la tabla 3.5, podemos observar que las estimaciones son muy similares a las obtenidas en la estimación con variables auxiliares. En cuanto a las predicciones, observamos que el MAPE es considerablemente inferior en las predicciones que se realizan con variables auxiliares frente a las que se realizan sin ellas. El modelo con las variables auxiliares tarda 4 meses de pandemia en ajustarse adecuadamente, pero una vez se ajuste el error cometido es inferior.

3.5. Fijación del modelo ARIMA sobre los residuales del modelo de regresión

En esta sección se realizará la segunda opción para hacer las predicciones, se estudiarán los residuos del número de parados para la obtención de un nuevo modelo. Luego, se analizará como se comporta este modelo en nuestra variable 'Número de parados'.

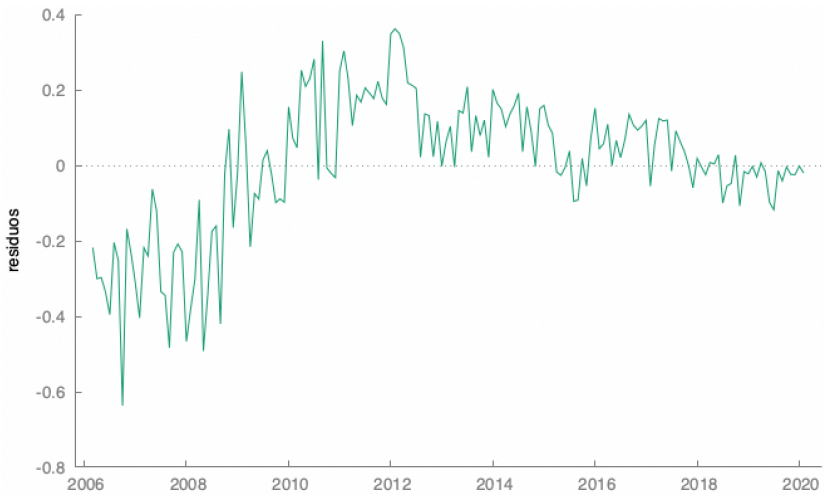

Figura 3.14: Residuos

El gráfico de los residuos (3.14) muestra que la media se considera constante ya que estos oscilan alrededor de cero. Por otro lado, se puede observar la aleatoriedad en la amplitud de las oscilaciones que recorren el comportamiento de la serie, por tanto se puede considerar que la varianza también es constante.

Se realizarán los mismos pasos que hemos hecho para el modelo anterior. El resultado del mejor modelo fijado para la variable 'residuos' es:

```
MODEL FITTED

     NONSEASONAL      P= 0      D= 1      Q= 1
     PERIODICITY    MQ= 12

         MEAN      =      0.00000

         SE        = *******


                ARIMA PARAMETERS

       TH     =    -0.8255
       SE     =     *****
```

Figura 3.15: Modelo de la variable 'residuos' fijado hasta 02/20

El modelo residual fijado es el ARIMA $(0,1,1)\times(0,0,0)_{12}$. Nos disponemos a analizar el nuevo modelo examinando que variables son signiticativas.

```
Modelo 16: ARMAX, usando las observaciones 2006:04-2020:02 (T = 167)
Estimado usando X-13-ARIMA (MV exacta)
Variable dependiente: (1-L) l_Numerodeparados

                  coeficiente   Desv. típica     z     valor p
  ------------------------------------------------------------
  const             0.00258408   0.00175382     1.473   0.1406
  theta_1           0.557493     0.0651234      8.561   1.12e-17 ***
  PARO             -3.64197e-05  6.31414e-05   -0.5768  0.5641
  SEPE             -0.00134622   0.000548450   -2.455   0.0141   **
  INFOJOBS         -4.77005e-05  4.82315e-05   -0.9890  0.3227
  SEGURIDADSOCIAL  -8.82479e-05  7.26086e-05   -1.215   0.2242
  EMPLEO            0.000209039  5.83797e-05    3.581   0.0003   ***

Media de la vble. dep.  0.002516   D.T. de la vble. dep.   0.017124
Media de innovaciones  -0.000016   D.T. innovaciones       0.014313
R-cuadrado              0.996578   R-cuadrado corregido    0.996472
Log-verosimilitud       472.0295   Criterio de Akaike     -928.0590
Criterio de Schwarz    -903.1150   Crit. de Hannan-Quinn  -917.9348

                     Real Imaginaria     Módulo Frecuencia
  ---------------------------------------------------------
  MA
   Raíz  1          -1.7937     0.0000     1.7937     0.5000
```

Figura 3.16: Modelo ARMAX sobre el número de parados para el modelo $(0,1,1) \times (0,0,0)_{12}$ fijado hasta 02/20 ($*0,1\,\%$ significativo,$**0,05\,\%$ significativo,$***0,01\,\%$ significativo)

En este caso, la variable 'Empleo' es significativa al $0,01\,\%$. La expresión del modelo estimado quedaría de la siguiente manera:

$$W_t = 0,00258408 + 0.557493\varepsilon_{t-1} - 3,64e^{-5}V_{PARO} - 0,001346V_{SEPE} - 4,77e^{-5}V_{INFOJOBS} - 8,824e^{-5}V_{SEGURIDADSOCIAL} + 0,000209V_{EMPLEO}$$

Se va a realizar las predicciones con el modelo ARIMAX obtenido desde marzo de 2020 hasta marzo de 2021.

Figura 3.17: Predicción del número de parados utilizando el modelo $(0,1,1) \times (0,0,0)_{12}$

Como era de esperar y pasaba con el modelo obtenido anterior, la predicción a partir de marzo de 2020 no se puede considerar buena. Aún teniendo datos en tiempo real, resulta muy difícil realizar buenas predicciones. Si tuviésemos al menos un mes de la pandemia como estimación mejoraríamos considerablemente esas predicciones.

El siguiente paso será realizar las predicciones con el modelo fijado cambiando el rango de la muestra.

Predicciones	MAE	RMSE	MAPE	Palabras relevantes
03/2020 - 03/2021	0.28798	0.29523	2.305	EMPLEO*** SEPE **
04/2020 - 03/2021	0.13007	0.13456	1.04	EMPLEO***
05/2020 - 03/2021	0.023242	0.027657	0.18597	EMPLEO***
06/2020 - 03/2021	0.024324	0.029121	0.19476	EMPLEO***
07/2020 - 03/2021	0.026857	0.031164	0.21465	EMPLEO***
08/2020 - 03-2021	0.036864	0.044165	0.29428	EMPLEO***
09/2020 - 03/2021	0.040872	0.046009	0.32628	EMPLEO***
10/2020 - 03/2021	0.077116	0.07998	0.61556	EMPLEO***
11/2020 - 03/2021	0.017817	0.022944	0.14202	EMPLEO***
12/2020 - 03/2021	0.024522	0.028569	0.19544	EMPLEO***
01/2021 - 03/2021	0.028751	0.02984	0.22913	EMPLEO***
02/2021 - 03/2021	0.0096828	0.013456	0.077185	EMPLEO***
03/2021-	0.019534	0.019534	0.15571	EMPLEO***

Tabla 3.6: Errores de las predicciones de número de parados usando el modelo $(0,1,1) \times (0,0,0)_{12}$

En la tabla 3.6 se puede observar que en el caso de que entrasen los dos primeros meses a la estimación (marzo y abril), en los que ya se podía magnificar la influencia que iba a tener el COVID-19, el porcentaje de error absoluto (MAPE) se reduce en más de un 90 %, pasando de tener un 2,30 % de error a un 0,18 % a partir de mayo de 2020. La palabra 'Empleo' es significativa al $0,01$ % en todos los meses, entrando en una ocasión la variable 'Sepe'.

Como se hizo para el modelo anterior, se van a realizar dos tablas comparando los errores MAPE, tanto en la estimación como en la predicción, para el caso donde el modelo introduzca variables auxiliares y otro donde no tenga en cuenta estas variables.

Estimación		**Predicción**	
Desde mar-06 hasta		**De**	
feb-20	0,088042	mar-20 / mar-21	2,305
mar-20	0,093452	abr-20 / mar-21	1,04
abr-20	0,097355	may-20 / mar-21	0,18597
may-20	0,097289	jun-20 / mar-21	0,19476
jun-20	0,097106	jul-20 / mar-21	0,21465
jul-20	0,097051	ago-20 / mar-21	0,29428
ago-20	0,096541	sept-20 / mar-21	0,32628
sept-20	0,096598	oct-20 / mar-21	0,61556
oct-20	0,098429	nov-20 / mar-21	0,14202
nov-20	0,097939	dic-20 / mar-21	0,19544
dic-20	0,097482	ene-21 / mar-21	0,22913
ene-21	0,098335	feb-21 / mar-21	0,077185
feb-21	0,097807	mar-21	0,15571

Tabla 3.7: MAPE de la estimación y predicción con variables auxiliares del modelo $(0,1,1) \times (0,0,0)_{12}$

Estimación		Predicción	
Desde mar-06 hasta		De	
feb-20	0,091584	mar-20 / mar-21	2,8167
mar-20	0,097302	abr-20 / mar-21	1,789
abr-20	0,10058	may-20 / mar-21	0,39025
may-20	0,10033	jun-20 / mar-21	0,19056
jun-20	0,099919	jul-20 / mar-21	0,20417
jul-20	0,10006	ago-20 / mar-21	0,31898
ago-20	0,099754	sept-20 / mar-21	0,31442
sept-20	0,09978	oct-20 / mar-21	0,75942
oct-20	0,10113	nov-20 / mar-21	0,18323
nov-20	0,10067	dic-20 / mar-21	0,20565
dic-20	0,10012	ene-21 / mar-21	0,22913
ene-21	0,10136	feb-21 / mar-21	0,084374
feb-21	0,1009	mar-21	0,199783

Tabla 3.8: MAPE de la estimación y predicción sin variables auxiliares del modelo $(0,1,1) \times (0,0,0)_{12}$

En la tabla 3.7, la estimación va aumentando hasta estabilizare en 0,098 %. Mientras que en la tabla 3.8, todas las estimaciones poseen un error más elevado a las obtenidas en la estimación con variables auxiliares. En cuanto a las predicciones, se observa que el MAPE en la mayoría de los meses, es menor en las predicciones que se realizan con variables auxiliares frente a las que se realizan sin ellas. El modelo con las variables auxiliares tarda 2 meses de pandemia en ajustarse, pero una vez se ajuste el error cometido es inferior al 0,7 %. Se comprueba que existe una disminución del error en las predicciones al introducir la ayuda del uso de la información 'en línea'.

Finalmente, se puede contemplar lo bien que se comporta este modelo con los datos de los residuos del número de parados llegando a estar muchos meses por debajo del 0,03 de error absoluto medio. Realizando una comparación de los errores del modelo cogiendo los residuos (tabla 3.6), con los errores del anterior modelo (tabla 3.3), podemos asegurar que se comete un menor error cuando se cogen los residuales. En el caso de que se tuviese que predecir desde junio de 2020 a marzo de 2021, el MAPE cometido por el primer modelo ($(0,2,1) \times (0,1,1)_{12}$) es de un 3,51 %, mientras que para el segundo ($(0,1,1) \times (0,0,0)_{12}$) es solamente de un 0,19 %. Se demuestra, por tanto, que el mejor modelo para la predicción del número de parados es el ARIMA $(0,1,1) \times (0,0,0)_{12}$.

Figura 3.18: Predicción de 02/2021 al 03/2021

En la figura 3.18 queda reflejado el buen ajuste que existe entre el comportamiento real que ha tenido el desempleo en Canarias en los últimos años y el modelo. Esto nos hace pensar que el pronóstico que vamos obtener para el resto de 2021 con los valores que hemos analizado, puede ser cercano a la realidad.

3.6. Pronóstico

Una vez que se ha conseguido el mejor modelo para nuestra serie temporal, se podrá desarrollar un pronóstico sobre como se comportará el desempleo en Canarias durante un periodo establecido.

Se realizará un análisis de predicción con una duración de 9 meses, es decir, hasta diciembre de 2021 con un intervalo de confianza del 95 %. En este pronóstico no se realizará la transformación del logaritmo del número de parados para poder observar con mayor facilidad la cifra estimada de desempleados a finales de año.

Los resultados de las predicciones son los siguientes:

Observaciones	Predicción	Desv. Típica	Intervalo del 95%
2021 Abril	279948.50	3898.97	272306.65 - 287590.34
2021 Mayo	280740.27	6838.92	267336.23 - 294144.31
2021 Junio	281532.04	8850.97	264184.45 - 298879.63
2021 Julio	282323.81	10483.73	261776.08 - 302871.55
2021 Agosto	283115.58	11894.43	259802.93 - 306428.24
2021 Septiembre	283907.35	13154.71	258124.60 - 309690.11
2021 Octubre	284699.13	14304.37	256663.06 - 312735.19
2021 Noviembre	285490.90	15368.28	255369.62 - 315612.18
2021 Diciembre	286282.67	16363.16	254211.46 - 318353.87

Tabla 3.9: Predicciones del número de parados hasta 12/2021 utilizando el modelo $(0,1,1) \times (0,0,0)_{12}$

Figura 3.19: Pronóstico del 04/2021 hasta 12/2021

En la figura 3.19 se observa que la predicción sobre la conducta futura de nuestra variable tiene como resultado una pequeña disminución del número de parados en los primeros meses. Aunque se volverá a tomar la tendencia creciente, con lo que se espera que el desempleo en Canarias no mejore durante este año 2021.

Por último, hay que tener en cuenta que aunque supiésemos todos los coeficientes del modelo, es muy difícil que esto se dé exactamente en la realidad, ya que siempre que hablamos de un pronóstico hay múltiples factores externos

que no podremos tener en cuenta y, por tanto, existe un margen de error que no se podrá suprimir.

3.7. Resultados y discusión

En esta sección se compararán ambos modelos (con y sin variables auxiliares) y se verán los resultados que hemos obtenido en este trabajo.

En primer lugar, se comprobará que el mejor modelo que se adapta a nuestra variable 'Número de parados' es el modelo $(0,1,1) \times (0,0,0)_{12}$.

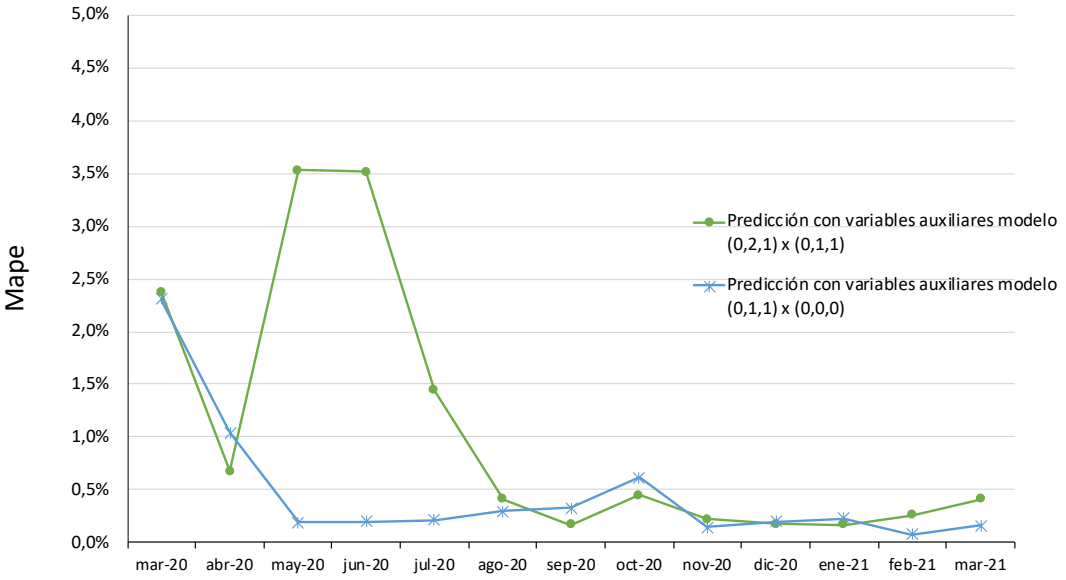

Figura 3.20: MAPE de las predicciones con el modelo $(0,2,1) \times (0,1,1)_{12}$ y el modelo $(0,1,1) \times (0,1,1)_{12}$ ambos incluyendo variables auxiliares

Analizando la figura 3.20, se puede observar como la predicción del modelo $(0,2,1) \times (0,1,1)_{12}$ tarda seis meses en corregir el efecto que ocasionó la pandemia del COVID-19. Mientras que la predicción del modelo $(0,1,1) \times (0,1,1)_{12}$ necesita solamente tres meses.

En segundo lugar, se relacionará el MAPE obtenido del modelo $(0,2,1) \times (0,1,1)_{12}$ con variables auxiliares, con el mismo modelo sin estas variables. El resultado nos muestra la siguiente gráfica:

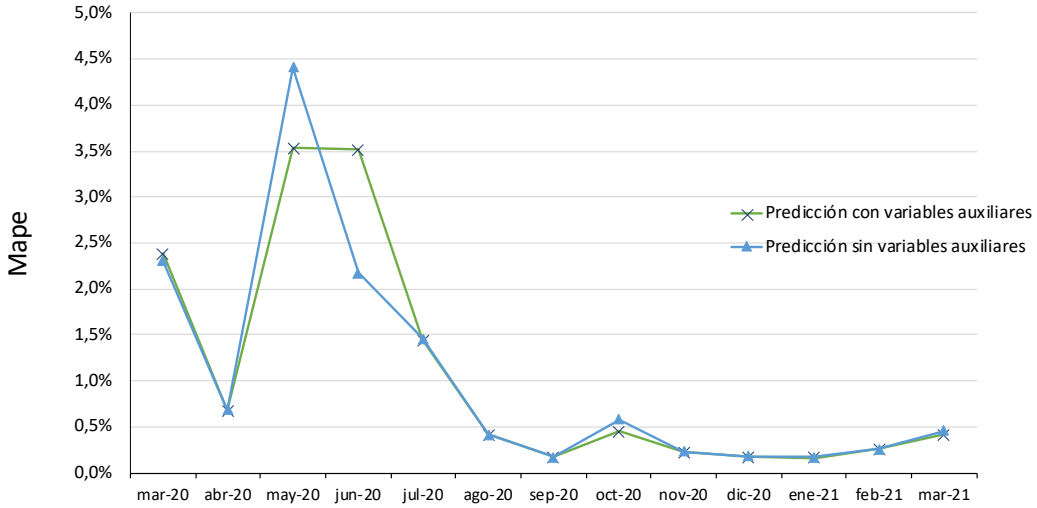

Figura 3.21: MAPE de las predicciones con el modelo $(0,2,1) \times (0,1,1)_{12}$ con variables auxiliares y sin variables auxiliares

En la figura 3.21, se observa como en la predicción con las variables auxiliares se obtiene un menor error en los tres primeros meses. En cambio, en el mes de junio se comete un peor ajuste que el modelo sin las variables auxiliares. Los motivos que expliquen este comportamiento puede deberse a que al principio de la pandemia no éramos conscientes de la situación que iba a ocurrir en los siguientes meses, lo que produjo una mayor incertidumbre y provocase un aumento del número de búsquedas que no correspondía con lo que estaba sucediendo. En los últimos meses de la predicción, el error es menor al 0,5 %, siendo muy similar en ambos casos.

Hay que recordar que el estudio lo hemos hecho sobre el logaritmo de 'Numero de parados', por tanto, el error que se va a cometer si no hubiésemos aplicado esta transformación será mayor. Se han estudiado determinadas predicciones y se ha obtenido que en el caso de tener un error de un 0,67 %, sería equivalente a que el error sobre el número de parados (sin la transformación logarítmica) fuese de un 8,79 %.

En último lugar, se verá lo que ocurre en el segundo modelo $(0,1,1) \times (0,0,0)_{12}$ comparando los errores con variables auxiliares y sin ellas.

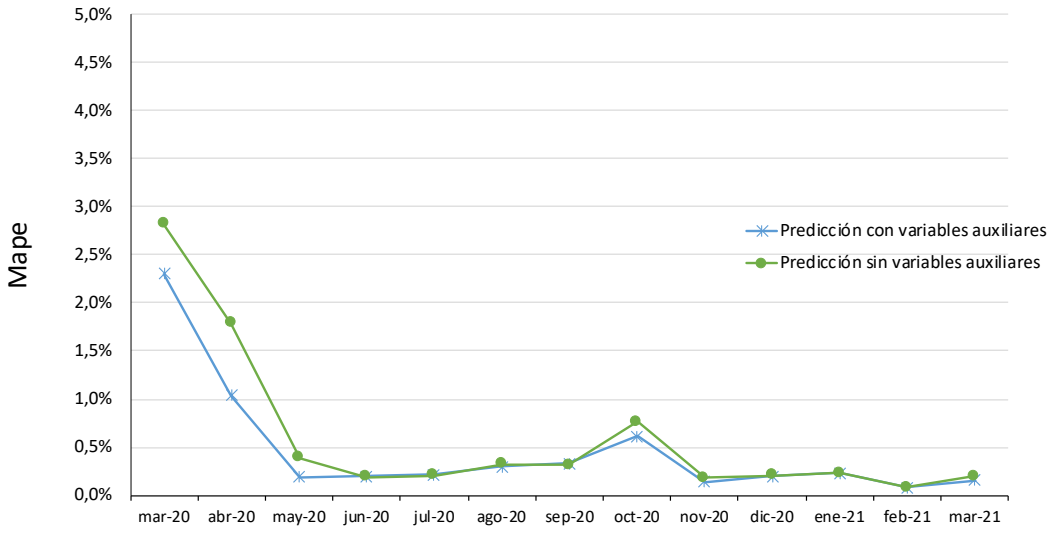

Figura 3.22: MAPE de las predicciones con el modelo $(0,1,1) \times (0,0,0)_{12}$ con variables auxiliares y sin variables auxiliares

En la figura 3.22 se contempla una disminución del error en todos los meses para el caso de las predicciones con variables auxiliares. En definitiva, en ambos modelos hemos obtenido que las predicciones con las variables auxiliares producen un menor error. Se confirma la hipótesis de nuestro trabajo sobre la mejora de las predicciones gracias a la introducción en los modelos de búsquedas en tiempo real.

4

Conclusiones

Para concluir el estudio realizado sobre la evolución del desempleo en Canarias a lo largo de los últimos años, voy a hacer una pequeña valoración sobre lo conseguido.

Una vez hecho el análisis de series temporales que se perseguía, hemos obtenido que el uso de las búsquedas de información a través de Google de la población canaria en tiempo real ayudan considerablemente a reducir el error y conseguir unas mejores predicciones sobre el número de parados. Los modelos tardan en ajustarse 2-3 meses debido a este cambio tan grande de tendencia que se produjo a causa de la pandemia, pero una vez ajustados el error que se comete al predecir con la información 'en línea' es inferior.

A lo largo del 2021 no se espera un cambio de tendencia en el número de parados, sino todo lo contrario, seguirá aumentando hasta casi tocar a final de año los 290.000 parados, esto supondría que la situación económica de Canarias no mejoraría hasta entrado el año 2022. Cabe recordar que la tasa de paro en Canarias es del $25,42\,\%$, siendo la segunda comunidad autónoma con más desempleo. Esto es debido a que el sector turístico representa aproximadamente el $30\,\%$ del PIB canario. Canarias ha sufrido especialmente el impacto de la pandemia por las limitaciones de movimientos. Ha supuesto un obstáculo casi infranqueable para el turismo, produciendo la disminución de la demanda de turistas procedentes de todas partes del mundo y, por consiguiente, que el número de parados en esta comunidad sea tan elevado.

Aún así, no debemos olvidar que estamos ante una predicción y el futuro es incierto, por lo que si todo transcurre tal y como nuestros resultados indican, es posible que se eleve el número de parados debido a que desaparecerán los ERTES y muchas empresas tendrán que volver a despedir. Con una visión a largo plazo y cuando todo vuelva a la normalidad 'prepandémica', el número de parados debería disminuír ya que volveremos a tener un gran flujo de turistas a nuestras islas. Aunque cualquier mínima variación accidental, como una nueva variante del COVID-19 o una demora en la producción de vacunas, podría llevarnos a una recuperación más tardía.

Finalmente, he de citar que la realización de este trabajo me ha ayudado a desarrollar nuevos conocimientos y me ha supuesto una importante visión para el futuro sobre cómo realizar un estudio social y, sin ninguna duda, me serán de gran utilidad para una posterior formación en Big Data.

Bibliografía

[1] Montero, M. Á. (2021) *Canarias sufre la mayor subida del paro del país y la menor del empleo.* El Día: La Opinión de Tenerife. [en línea]. [Fecha de consulta: 5-04-2021]. Disponible en: https://www.eldia.es/economia/2021/05/06/canarias-sufre-mayor-subida-paro-51396814.html.

[2] Mauricio, J.A. (2007). *Introducción al Análisis de Series Temporales.* 1º Edición. Disponible en: https://www.ucm.es/data/cont/docs/518-2013-11-11-JAM-IAST-Libro.pdf.

[3] Pilar, M. (2018). *Series Temporales.* Universidad Autónoma de Madrid. [en línea]. [Fecha de consulta: 7-04-2021]. Disponible en: https://www.studocu.com/es/document/universidad-autonoma-de-madrid/estadistica/apuntes/series-temporales/2267437/view.

[4] Caparrós, A. (2011). *Análisis clásico de series temporales.* [en línea]. [Fecha de consulta: 10-04-2021]. Disponible en: http://webpersonal.uma.es/~antonio/series.pdf.

[5] Quesada, A. (2015). *Análisis de la evolución temporal del desempleo en España.* Universidad de Jaén. [en línea]. [Fecha de consulta: 10-04-2021]. Disponible en: http://tauja.ujaen.es/bitstream/10953.1/4405/1/TFG-Quesada-Cazalla%2CAna.pdf.

[6] Casals, García-Hiernaux, Jerez, Sotoca, Trindade (2016). State-Space methods for time series Analysis. *State-Space methods for time series Analysis.*

[7] Escuela de Negocios y Dirección. (2014). *Qué es y para qué sirve Google Trends.* Universidad Europea Miguel de Cervantes. [en línea]. [Fecha de consulta: 13-04-2021]. Disponible en: https://www.escueladenegociosydireccion.com/revista/business/marketing-digital/google-trends/#:~:text=Google%20Trends%20es%20una%20herramienta,de%20tiempo%20determinado%2C%20permiti%C3%A9ndote%20identificar.

[8] Bernal, W. (2020). *Google Trends: qué es y cómo usar la herramienta en tu estrategia.* RD Station. [en línea]. [Fecha de consulta: 15-04-2021]. Disponi-

ble en: https://www.rdstation.com/es/blog/que-es-google-trends/.

[9] Redondo, J. (2013). *Uso de Google Trends para predecir el nivel y la estructura del desempleo en España.* Universidad Politécnica de Valencia. [en línea]. [Fecha de consulta: 16-04-2021]. Disponible en: https://riunet.upv.es/bitstream/handle/10251/31028/TFC_JORGE_REDONDO_CABALLERO.pdf?sequence=1&isAllowed=y.

[10] Gingsberg, J.; Mohebbi, M. H.; Patel, R. S. ; Brammer, L.; Smolinksi, M. S.; Brilliant, L. (2009) *Detecting influenza epidemics using search engine querey data.* nº457 pp. 1012-1014

[11] Choi, H., Varian, H. (2012). *Predicting the present with Google Trends.* [en línea]. [Fecha de consulta: 25-04-2021]. Disponible en: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-4932.2012.00809.x.

[12] Mclaren, N.; Shanboghe, R. (2011). *Using Internet Search Data as Economic Indicators.* Bank of England Quarterly Bulletin. 2011 Q2

[13] Guzman, G. (2011). *Internet search behaviour as an economic forecasting tool.* The Journal of Economic and Social Measurament.
nº 36 (1-3) pp. 337-386.

[14] Askitas, N.; Zimmerann, K.F. (2009). *Google Econometrics and Unemployment Forecasting.* nº 55(2) pp. 107-120.

[15] D'amuri, F.; Marcucci, J. (2009). *Google it! Forecasting the US Unemployment Rate with A Google Jobs Search Index.* Bank of England Quarterly Bulletin. nº31 2010.

[16] González, L. (2020). *Evaluando el error en los modelos de regresión* [en línea]. [Fecha de consulta: 18-04-2021]. Disponible en: https://aprendeia.com/evaluando-el-error-en-los-modelos-de-regresion/.

[17] M. Alonso, A. (2007). *Introducción al Análisis de Series Temporales.* Universidad Carlos III de Madrid. [en línea]. [Fecha de consulta: 12-04-2021]. Disponible en: http://halweb.uc3m.es/esp/Personal/personas/amalonso/esp/seriestemporales.pdf.

[18] Peña, D. (2010). *Análisis de Series Temporales.* Alianza Editorial, S.A., Madrid, 2010.

[19] J. Pérez, C. (2016). *Guía rápida de Gretl* Universidad Carlos III de Madrid. [en línea]. [Fecha de consulta: 20-04-2021]. Disponible en: http://www.eco.uc3m.es/docencia/econometria/Datos/Guia%20Rapida%20de%20Gretl.pdf.

Influence of COVID-19 on the predictions of the number of unemployed in the Canary Islands using Google Trends

Sección de Matemáticas
Universidad de La Laguna

Alejandro Carvajal Martínez
Facultad de Ciencias · Sección de Matemáticas
Universidad de La Laguna
alu0101045160@ull.edu.es

Abstract

PREDICTING NUMBER OF UNEMPLOYED in times of crisis and sporadic high-impact events poses great challenges for any type of forecasting model. The use of tools that can input information in real time is becoming more and more in demand. Web searches by people who are looking for a job, want to change their job or simply see their job in danger, are a thermometer of the evolution of the number of unemployed people in a given region. In this paper we propose a prediction model of the number of unemployed for the Canary Islands using Internet searches obtained through Google Trends.

1. Introduction

THE PRESENT WORK focuses on the influence that real-time searches on Google have on the number of unemployed in the Canary Islands. In particular, the Google Trends platform will be used to obtain the index of words related to the job search carried out by people in said region in a selected period of time.

2. Methodology

GOOGLE TRENDS is a free and freely available tool from Google. It allows to measure the search popularity of a word or phrases and to know the level of use of a certain term. In this work it will be used to improve the predictions of the unemployment number. The prediction of the number of unemployed will be done using ARIMA models. Thanks to these models, a future prediction of the behavior of the variable to be studied will be obtained thanks to the past values of this variable.

3. Applications and results

THE DATA on the number of unemployed have been obtained from the official website of the Canary Islands Institute of Statistics (ISTAC), which will be modeled in logarithmic form. The number of unemployed in the Canary Islands is:

Figure 1: *Representation of unemployment in the Canary Islands*

A list will be made of the possible most influential words that a person could search for when unemployed. Introducing these words in Google Trends and performing the correlation coefficient between the words with the variable under study (number of unemployed), we have obtained five words that we will use to improve our model, which are:

Figure 2: *Chosen word trend*

The values of p, d and q will be determined, that is, the best model that fits the series. The model obtained is an ARIMA $(0,2,1) \times (0,1,1)_{12}$. Subsequently, the predictions of this model from March 2020 to March 2021 will be carried out and it will be examined whether this model is able to predict the large rise that produced the COVID-19 crisis. The same will be done with the residuals and it will be tested which model is better. Once obtained that the best model for our time series is the residual (ARIMA $(0,1,1) \times (0,0,0)_{12}$), a forecast of how unemployment will behave in the Canary Islands can be elaborated. Finally, we will see what happens in the model $(0,1,1) \times (0,0,0,0)_{12}$ comparing the errors with and without auxiliary variables.

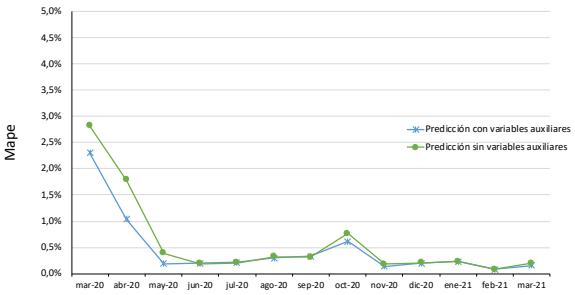

Figure 3: *MAPE of predictions with the model* $(0,1,1) \times (0,0,0)_{12}$ *with auxiliary variables and without auxiliary variables*

4. Conclusion

THE TIME SERIES ANALYSIS that was pursued has been carried out and it has been obtained that the use of information searches through Google of the Canary Islands population in real time helps considerably to reduce the error and get better predictions on the number of unemployed.

References

[1] Peña, D. (2010). *Análisis de Series Temporales.* Alianza Editorial, S.A., Madrid, 2010.

[2] Caparrós, A. (2011). *Análisis clásico de series temporales.* [en línea]. [Fecha de consulta: 10-04-2021]. Disponible en: url-http://webpersonal.uma.es/ antonio/series.pdf.