CREACIÓN Y PRONÓSTICO DE INDICADORES DE DESEMPLEO EN BOGOTÁ DURANTE LA PANDEMIA DEL COVID-19 A PARTIR DE GOOGLE TRENDS

Autor:
LUIS ALBERTO GUERRERO PEÑARANDA

Tutor:
DIEGO ANDRÉS BUESAQUILLO SALAZAR

UNIVERSIDAD COLEGIO MAYOR DE CUNDINAMARCA

FACULTAD DE ADMINISTRACIÓN DE EMPRESAS Y ECONOMÍA

ECONOMÍA

BOGOTÁ

2022

Resumen:

En este documento se propone crear indicadores para la tasa de desempleo basados en Google Trends (GT). Para esto se seleccionan palabras clave relacionadas a la búsqueda de empleo, que tengan una relación fuerte y positiva con la tasa de desempleo bogotana. Seleccionadas las palabras se crean los Índices de Google Trends (IGT, IGT2, IGT3). Con los modelos SARIMA se escoge la mejor predicción por medio del MAPE y RMSE. De este resultado se concluye que el mejor indicador para el desempleo bogotano es IGT con los parámetros $(2,2,4)\ x\ (2,2,0,12)$. Por último se recalca que las predicciones son buenas en el corto plazo y que es posible saber el comportamiento (aumentos o disminuciones) de la tasa de desempleo por medio de los IGT con las palabras clave más relacionadas al desempleo, siempre teniendo en cuenta el contexto del periodo de tiempo estudiado

Palabras clave: Desempleo, Google Trends, SARIMA, Indicadores lideres

Clasificación JEL: C01, C53, E24, J64

1. Introducción

La presente investigación propone la creación de índices para el desempleo basado en Google Trends (GT) que puede ser usado en casos críticos como lo significó la pandemia de Covid-19.

Los antecedentes más destacados son Choi y Varían (2009) que usaron GT como pronóstico para el desempleo en Estados Unidos, ya que esta guardaba las tendencias de búsquedas en temas relacionados con el empleo. Para Colombia ya se ha hecho un estudio en la materia a partir de GT, allí Cardona y Rojas (2017) buscan predecir la tasa de desempleo en el corto plazo por el rezago que existe en la observación de los datos oficiales del mercado laboral.

El objetivo principal de la presente es evaluar la efectividad del Índice de Google Trends (IGT) para el desempleo como pronóstico durante la pandemia de Covid-19. De esta manera, se busca responder por qué los pronósticos de los indicadores de desempleo a partir de GT en Bogotá pueden ser una buena metodología a usar en el corto plazo. Se aclara que no se busca un dato preciso de desempleo, esta investigación pretende explicar qué puede pasar con el indicador económico (TD) en el periodo t, de acuerdo al comportamiento del IGT. Para esto se utiliza la tasa de desempleo (TD) calculada por el Departamento Administrativo Nacional de Estadísticas (DANE) para compararla con los resultados de los indicadores de GT (IGT, IGT2, IGT3), asimismo, se predicen y pronostican estos indicadores por medio de un modelo SARIMA.

Cabe destacar que la pandemia tuvo impacto significativo en varios factores relevantes de la economía como lo fueron el desempleo y la misma forma de estimarlo. Al frenar el contacto presencial humano varias empresas despidieron a sus empleados para lograr sostenerse en el mercado, generando un impacto mucho mayor que la crisis del 2009, según France 24 (2021). Es así que la TD trimestre móvil (mayo-julio) para Bogotá en el 2020, según el DANE, fue del 25,1%; también se evidencia un aumento significativo comparado con el mismo periodo del 2018 que fue de 10,6%. La recolección de datos para calcular el desempleo significó un problema en periodo pandémico, además se destaca el rezago que existe en su publicación oficial. El cálculo de la TD del DANE se hace a partir de la Gran Encuesta Integrada de Hogares (GEIH), su recolección de datos fue telefónica en zonas urbanas y presencial en zonas rurales; por esta razón es necesario pensar en un apoyo como lo son los índices y pronósticos, para saber qué puede ocurrir con el desempleo en el periodo inmediatamente presente por medio de un indicador.

Para Colombia se sabe que, según el DataReportal, en el 2020 se tenían 35 millones de usuarios de internet y el 86,27% utiliza como motor de búsqueda, según el Global Statcounter, a Google Chrome; utilizar un IGT servirá para ver sin rezago lo que puede estar sucediendo con la tasa de desempleo.

En la primera parte del documento se presenta una revisión literaria desde las investigaciones más importantes respecto al uso de Google Trends como ayuda para el pronóstico, hasta el caso colombiano hecho por Cardona y Rojas (2017). En la segunda parte encontrará la metodología, donde se habla de la construcción de indicadores para el desempleo por medio de Google Trends, así como la especificación del modelo SARIMA. En la tercera parte del documento se encontrarán los resultados del mejor modelo a usar y del mejor IGT, de acuerdo a los criterios de información AIC y BIC, así como el mejor predictor de acuerdo al MAPE y RMSE. Por último se encuentran las conclusiones.

Revisión literaria

En materia económica Google Trends ha servido para diversas investigaciones, en su mayoría relacionadas con el pronóstico inmediato (nowcasting). Aquí es relevante hacer una pequeña introducción de los pioneros en la materia en general y aquellos que han estudiado a partir del big data para nowcasting estudiando variables macroeconómicas relevantes; luego, la revisión se centrará únicamente en aquellos trabajos que incluyen a Google Trends como pronóstico para la tasa de desempleo.

Los pioneros en la materia fueron Ettredge, et al. (2005) ya que introdujeron el primer estudio basado en la posibilidad de examinar los datos guardados de las búsquedas internautas para encontrar estadísticas significativas en variables macroeconómicas, aunque no usan Google Trends y tampoco utilizan modelos de series temporales, podría decirse que introducen en su novedosa metodología, la relación existente entre los datos oficiales de la tasa mensual de desempleo de EEUU y los términos buscados relacionados con el empleo. En este trabajo encuentran una relación positiva entre el desempleo y las variables relacionadas con la búsqueda de empleo, donde hay una dominancia que explica las variables buscadas debido a las solicitudes de subsidio al desempleo.

Los estudios basados en big data y pronóstico inmediato, en los que se busca predecir otras variables de relevancia macroeconómica como lo son la tasa de inflación, resaltan Cavallo y Rigobon (2016)**,** Boettcher (2015), Griffioen et al. (2014)**;** las variables financieras que estudian Cerchiello y Giudici (2014)**,** Heston y Sinha (2014); el pronóstico y estudios relacionados con el PIB como los trabajos de Angelini, Bánbura y Runstler (2008), Angelini,

Camba-Mendez, Giannone y Reichlin (2011), Giannone, Reichlin y Simonelli (2009), Yiu y Chow (2010). Estas investigaciones las puede ver por separado para más comprensión o agrupadas en la revisión que hacen Kapetanios y Papailias (2018) para tener un panorama más amplio. Las investigaciones macroeconómicas con big data y Google Trends son bastante extensas, pero como se explicó desde un principio, aquí solo se revisarán aquellas con una relación directa al tema observado.

Los primeros en utilizar Google Trends fueron Ginsberg, et al. (2009), donde rastrean y predicen la propagación de la influenza antes que los organismos de control y prevención de enfermedades de Estados Unidos. Gracias a las frecuentes consultas relacionadas con síntomas de la enfermedad y las visitas médicas, lograron estimar el nivel de actividad de la influenza semanal en todas las regiones comprendidas del país.

Para los pronósticos del desempleo fueron Choi y Varian (2009b) los que utilizaron por primera vez Google Trends. Por medio de las peticiones de desempleo de los Estados Unidos, considerado como buen indicador del mercado de trabajo con rezagos por su publicación, utilizan modelos autorregresivos, ARIMA y un modelo AR(1) como base, con los que demuestran que Google Trends ayuda a predecir las solicitudes de beneficios por desempleo.

En contraste de este pronóstico del desempleo, en su primer trabajo relacionado con Google Trends, Choi y Varian (2009a) afirman que la herramienta no ayuda como tal a predecir el futuro sino más bien el presente inmediato, aquí se centran en diferentes variables como las ventas mensuales de Ford, luego toman ejemplos de ventas al por mayor, ventas de automóviles, predicción de viajes y visitas, donde por medio de diferentes modelos AR estacional simple y modelos de efectos fijos calculan series de predicciones por mes visto donde se calcula el error de predicción para tomar el Error Medio Absoluto (MAE, por sus siglas en inglés).

Es así que con el MAE se evidencia una mejora del 12% en predicción de las “Nuevas viviendas” y un 18% en las predicciones de “Vehículos por motor y piezas”. En concreto, estos dos últimos mencionados, son sintetizados en uno solo, donde se amplían ambos trabajos incluyendo modelos más sofisticados, allí Choi y Varian (2012) descubren que los modelos AR estacionales simples que incluyen variaciones destacadas de Google Trends tienen la tendencia a superar a los modelos que incluyen los predictores entre un 5% y un 20%.

Ahora bien, no todos los estudios se han hecho para Estados Unidos. El pronóstico de la tasa de desempleo de Alemania hecho por Askitas y Zimmermann (2009) tiene una característica relevante que los mismos autores resaltan, ya que ellos buscan demostrar que los datos pueden ser usados para predecir el comportamiento económico medido por estadísticas

tradicionales. También suman la relevancia de Google Trends como pronóstico en épocas de crisis económicas donde la información oficial con cálculos y metodologías tradicionales suele ser lenta. Este plus recae en su rapidez para obtener resultados que en últimas se traducen en argumentos para una adecuada toma de decisiones acertadas en el campo económico. En este estudio se sugiere un método innovador en el uso de datos en la actividad de internet, debido a que consideran que la econometría clásica no se ha aprovechado de forma adecuada, con éste método logran demostrar las fuertes correlaciones que existen entre la búsqueda de palabras clave relacionadas con el empleo y las tasas de desempleo usando los datos mensuales de Alemania.

Pasando un poco más al continente asiático, se tiene un estudio por parte del Banco Central de Turquía, en el que Chadwick y Gönül (2012) buscan investigar si los datos de las consultas de las búsquedas de Google pueden mejorar el desempeño de la predicción inmediata de la tasa mensual de desempleo no agrícola para Turquía entre el 2005 y 2012. Ellos, al igual que Choi y Varian (2009b), hacen uso de google insights for search para recopilación. Sumado a esto establecen que utilizar pocas variables que representen la información que se contiene en los datos de google es mucho más beneficioso metodológicamente ya que aumenta los grados de libertad, utilizan los modelos de regresión lineal y el procedimiento Bayesian Model Averaging, donde se selecciona un modelo de referencia y 45 modelos de predicción inmediata a través de BMA (Bayesian Model Averaging) y diagnósticos residuales. Tomando esto en cuenta, se concluye que los indicadores de Google Trends funcionan mejor en las predicciones inmediatas de la tasa de desempleo.

En un estudio de países europeos mucho más amplio, se destacan Barreira, et al. (2013), porque innovan el ejercicio llevándolo a una ampliación de estudio transnacional-interlingüístico que abarca economías bastante diferentes con niveles de ingreso distintos, al igual que unas estructura de desempleo opuestas que en últimas dependen de las realidades de cada país. Los países de estudio son Portugal, España, Francia e Italia, donde no solo se busca predecir la tasa de desempleo, sino también la venta de automóviles. En su metodología se toman los datos ajustados sobre las tasas mensuales de cada país que se recopilaron del almacén de estadísticas del Banco Central Europeo, se tomaron los datos mensuales de enero de 2005 a agosto de 2013; para la estimación se utilizaron los modelos autorregresivos ARIMA-X-12, modelo AR(p), con ellos se concluye que al agregar los datos provenientes de Google Trends, por lo general, lleva a mejoras en las predicciones en Portugal, Italia y Francia, y los datos ayudan a explicar la varianza de las estimaciones obtenidas en las ventas de automóviles en

Portugal más que en España. No obstante, estos resultados no apoyan la hipótesis de los autores en relación de que los datos puedan mejorar las predicciones de ventas de automóviles.

Para las economías latinoamericanas, se resalta el trabajo hecho en México, donde se utilizan las búsquedas de google relacionadas al empleo para su pronóstico, también se debate una discusión relacionada con el nowcasting y big data sobre los datos de recolección generados por internet al servicio de la predicción del desempleo. Aquí Campos y López-Araiza (2020) además de tener una discusión sobre un mejor predictor, introducen el machine learning en su metodología, utilizando el método de análisis de regresión LASSO (Least Absolute Shrinkage and Selection Operator) y los bosques aleatorios (Random Forest). También usan un modelo autorregresivo AR. Éste estudio se diferencia de los demás porque establece una mejor metodología, puesto que concluyen que en el método LASSO tiene mayores beneficios predictivos que un modelo AR. Al igual que en los trabajos anteriores, obtienen que el índice de Google Trends atrapa información útil que mejora las predicciones de la tasa de desempleo en México.

Como se ha visto en las revisiones anteriores todas las investigaciones se basan en los índices de Google Trends, lo que diferencian a Chang y Del Río (2013) de los demás es la creación de su propio Índice de Google de Desempleo (IGD). Para construir éste índice se utilizó la fuente de información de aquella población que representaba a la población en búsqueda de empleo. En el estudio peruano se busca observar si Google Trends refleja el comportamiento de las variables macroeconómicas, donde toman como ejemplo el Índice de Empleo de Lima para Empresas de 100 y más Trabajadores (IE100). Ellos utilizan modelos ARMA y ARDL, a los que se les aplican indicadores como el error cuadrático medio (RECM) y el error absoluto medio (EAM). Los resultados obtenidos son prometedores ya que el IGD permite hacer nowcasting y un periodo hacia adelante como máximo, por lo que concluyen que el Índice de Google de Desempleo es un buen predictor en el corto plazo de IE100.

Por último se tiene el único estudio hecho en Colombia, basado en Google Trends y enfocado en el pronóstico de la tasa de desempleo. Allí Cardona y Rojas (2017) buscan estimar la predicción a corto plazo debido a los rezagos que se tienen en la publicación de los datos oficiales del DANE. Ellos estiman modelos de regresión lineal simple, modelos autorregresivos integrados de media móvil (ARIMA) y la versión ampliada por variables exógenas (ARIMAX). El resultado de este ejercicio es que los índices de Google Trends siguen acompañando el comportamiento de la tasa de desempleo, reflejando sus variaciones de corto plazo y sus patrones estacionales.

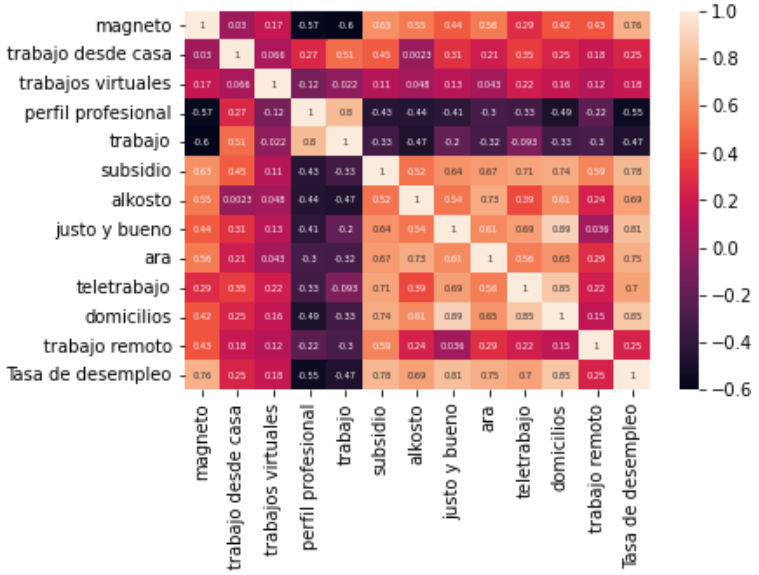
Teniendo en cuenta cada una de éstas metodologías, en este trabajo se crean tres indicadores de GT para la tasa de desempleo en Bogotá, por tanto, se establecerán varias estimaciones y se evaluarán para establecer cuál de todas presenta un mejor desempeño.

2. Metodología[1]

La metodología está fundamentada en procedimientos cuantitativos apoyados por el modelo SARIMA (Seasonal Autoregressive Integrated Moving Average), utilizando los datos de Google Trends (GT) y la Gran Encuesta Integrada de Hogares (GEIH), ambos en el periodo comprendido de 2017-2021 mensualmente. Las variables de análisis serán los indicadores IGT, IGT2, IGT3 y la tasa de desempleo (TD) calculada por el DANE. Se usará la tasa de desempleo trimestral móvil para Bogotá empalmada al año 2018.[2]

Antes de establecer cada uno de los modelos se debe calcular el Índice de Google Trends (IGT). En GT se buscan las palabras directamente relacionadas con la búsqueda de empleo, bolsas de trabajo y formas de empleo, como lo hacen Choi y Varian (2009b), por ejemplo: “El empleo”, “Computrabajo”, “Trabajos sin experiencia”, ”trabajos Colombia”, ”hoja de vida”, “Linkedln”, entre otros. Estas series dependen del periodo a tomar, para este caso se toman los promedios mensuales ya que su frecuencia es semanal.

Mapa de calor: correlaciones palabras clave GT y tasa de desempleo



[1] La metodología, fue desarrollada en su mayoría con el lenguaje Python.
[2] Revise el cuadro 3 de los anexos

Las palabras clave a tomar en cuenta para la construcción del IGT son aquellas que comparadas con la tasa de desempleo tienen una correlación superior o cercano a 0.7 y otros términos que tengan que ver con el periodo pandémico.

El mapa de calor contiene algunas de las correlaciones de las palabras clave con la tasa de desempleo bogotana, entre más claro sea, se tendrá una relación fuerte y positiva, por el contrario, entre más oscura sea, la relación será negativa y fuerte. Revise el cuadro 1 de los anexos para ver todas las correlaciones.

Se crearon tres indicadores de desempleo con datos de GT. Para el Índice de Google Trends (IGT), las palabras clave que lo conforman son "magneto", "subsidio", "alkosto", "justo y bueno", "ara", "teletrabajo" y "domicilios", ya que estos tiene una correlación aproximada o superior a 0.7; Cárdenas y Rojas (2017) también toman correlaciones relativamente altas, pero hacen el estudio por palabra clave y no formando un índice que es como lo hacen Chang y Del Rio (2013) y también como se hará en este trabajo.

Para el segundo indicador, IGT2, no se fue tan exigente con correlaciones altas; las palabras clave fueron "rappi", "teletrabajo", "subsidio", "servicio de domicilio", "justo y bueno", "trabajo desde casa", "domicilios". Por el contrario, se fue menos flexible para el tercer indicador, IGT3, y se tomaron varias correlaciones negativas con el fin de contrastar la importancia de escoger siempre las mejores palabras clave en el periodo determinado, este indicador es mucho más amplio y está conformado por "empleos", "CompuTrabajo", "magneto", "perfil profesional", "subsidio", "prestamo", "transmilenio", "rappi", "outsourcing", "teletrabajo", "teletrabajo sin experiencia", "servicio de domicilio", "justo y bueno" y "domicilio".

La explicación de que las palabras clave tengan una relación positiva y fuerte con la tasa de desempleo, es que las búsquedas efectivamente se relacionan para las personas que buscaron trabajo durante este periodo, ya sea en bolsas de trabajo, almacenes grandes y tipos de trabajo. Magneto es una plataforma de ofertas laborales; para la palabra subsidio, en pandemia se solicitaron muchos subsidios por la pérdida de empleos y el aumento de la pobreza; en lo que respecta a almacenes grandes como alkosto, justo y bueno y ara, por lo menos en los dos últimos se abrieron ofertas laborales, además se tiene en cuenta que para hacer parte de estas tiendas no se requiere un título profesional, por lo que su cobertura es mucho más grande; la palabra teletrabajo se explica por sí sola, al estar en cuarentena muchos buscaron esta alternativa, e incluso la mayoría de trabajos tomaron esta modalidad.

De este modo, se utiliza la metodología de Chang y Del Río (2013) de indexación de las series ponderando por la inversa de la desviación estándar, que consiste en calcular el promedio

(1) y la desviación estándar (2) para la palabra clave; luego se obtienen los ponderadores alpha que es la inversa de la desviación estándar dividido sobre la sumatoria de las inversas de la desviación estándar (3); por último, se crea los Índices de Google Trends (IGT, IGT2, IGT3) luego de sumar el producto de las palabra clave en el periodo t con sus respectivos ponderadores alpha (4)[3].

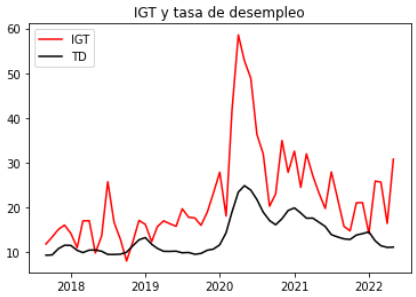
$$\bar{x}_i = \sum_{t=1}^{T} x_{ti} \qquad (1)$$

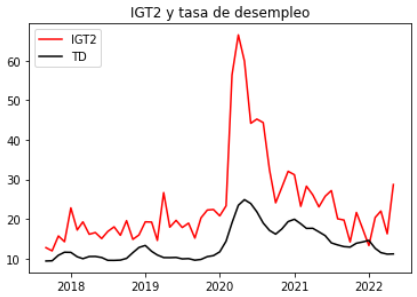
$$\sigma_i = \sqrt{\frac{\sum_{t=1}^{T}(x_{ti} - \bar{x}_i)^2}{T-1}} \qquad \forall\ t = 1\ a\ T \qquad (2)$$

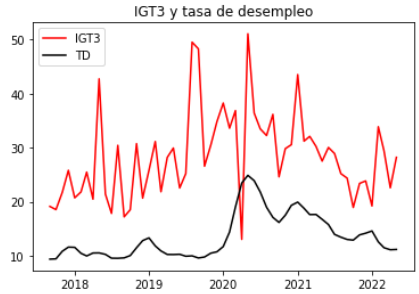
$$\alpha_i = \frac{\frac{1}{\sigma_i}}{\sum_{i=1}^{I}\frac{1}{\sigma_i}} \qquad (3)$$

$$IGT_t = \sum_{i=1}^{I} x_{ti}\alpha_i \qquad (4)$$

Gráficos Indicadores de Google Trends (IGT, IGT2, IGT3) y tasa de desempleo







[3] Véase un ejemplo para la creación de IGT en el periodo de septiembre 2017 en el cuadro 2 de los anexos.

Una vez construido los Índices de Google Trends (IGT, IGT2, IGT3), obsérvese la relación que hay entre ellos y la tasa de desempleo (TD) para Bogotá. Se evidencia que la línea roja, que representa los indicadores, sigue bastante bien el comportamiento del desempleo (para el caso de IGT e IGT2), también debe tener en cuenta, que a diferencia de la TD (línea negra), esta no tiene rezago y su dato se puede saber con anticipación, es decir, que se puede saber qué está pasando con el indicador de desempleo basado en GT en un momento t y no en un momento t+1, como sucede con la tasa de desempleo. Ahora bien se aclara que no se busca un dato preciso de desempleo, ya que de esto se encarga el DANE, lo que se busca es saber qué puede pasar con el indicador económico, si puede aumentar o disminuir en el periodo t, de acuerdo al comportamiento del IGT.

Gráficamente, se evidencia y recalca la importancia de que las palabras clave tengan una relación positiva y fuerte con la tasa de desempleo. El comportamiento de IGT3 no sigue el comportamiento de la TD, por lo que se considera estéril e ineficaz desde un principio.

A continuación se aplican los tests de Dickey-Fuller Aumentado (ADF) para saber si los índices tienen o no raíz unitaria, es decir, si se tiene o no tendencia estocástica en las series temporales.

Test ADF:

```
Prueba Dickey-Fuller Ampliada para IGT
ADF Statistic: -2.753878
p-value: 0.065170
Critical Values:
  1%: -3.553
  5%: -2.915
  10%: -2.595
La hipótesis nula no se rechaza, los datos tienen raíz unitaria y no
son estacionarios

Prueba Dickey-Fuller Ampliada para IGT2
ADF Statistic: -2.444950
p-value: 0.129449
Critical Values:
  1%: -3.553
  5%: -2.915
  10%: -2.595
La hipótesis nula no se rechaza, los datos tienen raíz unitaria y no
son estacionarios

Prueba Dickey-Fuller Ampliada para IGT3
```

```
    ADF Statistic: -6.287167
    p-value: 0.000000
    Critical Values:
     1%: -3.553
     5%: -2.915
     10%: -2.595
    La hipótesis nula se rechaza, los datos no tienen raíz unitaria y
son estacionarios
```

Tanto IGT como IGT2, resultan tener raíz unitaria, por tanto, no son estacionarios. Nótese que el p-valor de IGT e IGT2 es mayor al 5%. En contraste IGT3 no tiene raíz unitaria. Para que los datos sean estacionarios y eliminar la raíz unitaria, se toman la primera y segunda diferencia con el fin de utilizar el modelo SARIMA. Se aclara que se seguirá trabajando con IGT3, a este no se le aplican las diferencias ya que no es necesario.[4]

Especificación del modelo:

El modelo SARIMA (Seasonal Autoregressive Integrated Moving Average), es una extensión del ARIMA que tiene en cuenta la estacionalidad, se expresa como $(p,d,q)\ x\ (P,D,Q,s)$. Por una parte (p,d,q) representa los parámetros para el modelo ARIMA, por otra (P,D,Q,s), contiene los componentes estacionales autorregresivos de integración y promedio móvil, asimismo, el periodo de estacionalidad es s.

Formalmente, el modelo SARIMA se puede expresar de la siguiente manera:

$$\phi(L)\Phi(L^s)(1-L)^d(1-L^s)^D y_t = \theta(L)\odot(L^s)\varepsilon_t \quad (1)$$

Donde:

$$\Phi(L)y_t = \left(1-\Phi_1 L^1-\Phi_2 L^2-\cdots-\Phi_p L^P\right)y_t \quad (2)$$

$$\odot(L)\varepsilon_t = \left(1-\odot_1 L^1-\odot_2 L^2-\cdots-\odot_q L^Q\right)\varepsilon_t \quad (3)$$

Por tanto se tienen P términos autorregresivos estacionales (coeficientes $\Phi_1,\dots,\Phi_P$) , Q términos de promedio móvil estacional (coeficientes $\odot_1,\dots,\odot_Q$), mientras que D es la diferenciación estacional basada en los *s* periodos estacionales.

Los mejores modelos se obtiene usando dos criterios de información:

- Criterio de Información Akaike (AIC por sus siglas en inglés): este criterio se basa en la inferencia frecuentista, formalmente se tiene que (k) es el número de parámetros del modelo y (L) considera la verosimilitud del modelo, ambos parámetros se involucran en la ecuación:

[4] Véase los anexos para ver las diferencias de cada indicador, así como sus autocorrelaciones para cada caso.

$$AIC = 2k - 2\ ln\ ln\ (L) \quad (4)$$

- Criterio de Información Bayesiano (BIC por sus siglas en inglés): este criterio está basado en la probabilidad bayesiana, está centrado en la suma de los cuadrados de los residuos, encontrando los rezagos que minimizan el modelo. También se define como el número de parámetros (k) por el logaritmo natural de la muestra (n), menos dos veces el logaritmo natural del valor máximo de la función de verosimilitud ($\hat{L}$), formalmente:

$$BIC = kln(n) - 2\ ln\ ln\left(\hat{L}\right) \quad (5)$$

Para ambos criterios se escoge el de menor valor y se comparan los resultados de los dos modelos seleccionados, uno por AIC y el otro por BIC. Cabe resaltar que el BIC es mucho más robusto ya que este penaliza la sobre-parametrización, sin embargo, para este caso pueden haber desventajas con el BIC, ya que la muestra es pequeña, por tanto es posible que el criterio bayesiano escoja un modelo simple.

De acuerdo a las diferencias para IGT y su periodo de estacionalidad (s), tanto para AIC como BIC, se encontraron 80 posibles modelos (160 en total), de los cuales se escogen el primer parámetro de cada criterio para modelar.

En IGT2, ambos criterios de información (AIC y BIC) coinciden en los mismos parámetros para su minimización, se encontraron 16 posibles modelos. Por otra parte, en IGT3 se tuvieron 81 posibles modelos por minimización de cada criterio (162 en total). Véase todos los parámetros de acuerdo al criterio en los anexos.

Adicionalmente, cada uno de los modelos fue sometido a dos indicadores de desempeño que compara el rendimiento de precisión de pronóstico: Error porcentual Absoluto Medio (MAPE, por sus siglas en inglés) y Raíz del Error Cuadrático Medio (RMSE, por sus siglas en inglés).

3. Resultados

3.1 Parámetros según AIC y BIC:

En IGT el modelo con mejor AIC (266.102), se encontró que los parámetros de $(p, d, q)\ x\ (P, D, Q, s)$. resultaron ser $(2, 2, 4)\ x\ (2, 2, 0, 12)$, la presente tiene en cuenta los

doce periodos del año, así como las dos diferencias a las que se somete el IGT para que no tenga raíz unitaria y sea estacional. Por otra parte, en el modelo con mejor BIC (277.872), los parámetros para un mejor modelo fueron $(0, 2, 4)\ x\ (2, 2, 0, 12)$.

En IGT2 el AIC y BIC concordaron con los parámetros que fueron $(0, 2, 2)\ x\ (2, 2, 0, 12)$.

Los parámetros para IGT3 se utilizaron para la modelación, es importante recordar, que al ser un indicador compuesto por palabras clave con relaciones negativas con la tasa de desempleo, su análisis y utilidad es estéril. Véase en los anexos todos los resultados de AIC y BIC para los tres indicadores creados.

3.2 Validación del modelo:

El modelo según AIC con los parámetros $(2, 2, 4)\ x\ (2, 2, 0, 12)$ de IGT, obtuvo un MAPE de 0.35%, de acuerdo a Chen et al. (2009) ante un MAPE más bajo, el pronóstico será mucho mejor. Asimismo, para la Raíz del Error Cuadrático Medio (RMSE), también se busca el menor valor; para el modelo anteriormente dicho se obtuvo un RMSE de 6.52. Por otra parte, el modelo con mejor BIC con los parámetros $(0, 2, 4)\ x\ (2, 2, 0, 12)$, registró un MAPE de 0.90% y un RMSE de 10.98.

En el caso de IGT2 AIC y BIC concordaron con un solo conjunto de parámetros para el mejor modelo y se obtuvo un MAPE de 6.06% y un RMSE de 18.83

En síntesis y para mayor comprensión:

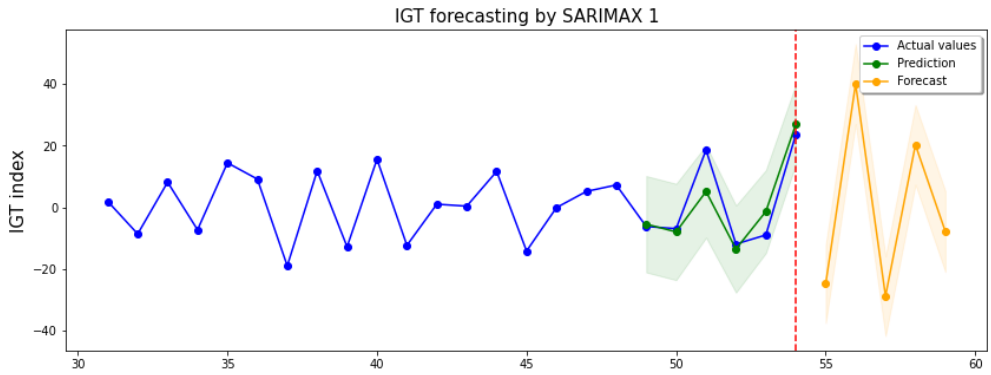
Cuadro de resultados

		SARIMA	MAPE	RMSE
IGT	**AIC**	$(2, 2, 4)\ x\ (2, 2, 0, 12)$	0.35%	6.52
	BIC	$(0, 2, 4)\ x\ (2, 2, 0, 12)$	0.90%	10.98
IGT2	**AIC Y BIC**	$(0, 2, 2)\ x\ (2, 2, 0, 12)$	6.06%	18.83

De acuerdo a estos resultados, el mejor modelo es para el IGT escogido por minimización de AIC. Véase los diagnósticos, predicciones y pronósticos en los anexos.

A continuación se presenta la predicción y pronóstico de IGT.

Gráfico de la predicción y pronóstico del mejor modelo



Como puede observarse en el gráfico anterior, la predicción sigue bastante bien el comportamiento de los valores actuales, es decir, los valores de IGT hasta su último periodo (mayo-2022). Por otra parte el forecast indica que para el pronóstico de un periodo adelante, se espera una disminución del IGT, a dos periodos adelante un aumento. Aquí se destaca que al ser una estimación instantánea, solo se pueden tener en cuenta los pronósticos de corto plazo, ya sea un periodo adelante o dos.

4. Conclusiones

La presente investigación evidencia que con datos de Google Trends (GT) se puede crear un indicador líder para la tasa de desempleo. Para este caso, los que muestran un mejor comportamiento son IGT e IGT2. Se evidencia que los indicadores deben construirse con palabras clave que tengan una relación fuerte y positiva con la tasa de desempleo (TD) para que siga un comportamiento similar, en lo que respecta a sus aumentos y disminuciones a lo largo del tiempo.

En lo que respecta a la predicción tomada con los modelos SARIMA, se recalca que estos son buenos, especialmente el de IGT por AIC con los parámetros $(2,2,4)\ x\ (2,2,0,12)$, pero únicamente en el corto plazo. De aquí se declara que las palabras clave no siempre van a tener una relación fuerte y positiva a lo largo del tiempo debido a que se tienen contextos diferentes.

Adicionalmente se resalta que en el pronóstico de corto plazo IGT puede seguir el comportamiento de la tasa de desempleo (no su valor), así como también en el periodo presente y sin rezago. Sin embargo, con respecto a lo anterior se reconocen limitaciones, ya que la tasa de desempleo tomada es un dato trimestre móvil y no mensual, a pesar de esto, los indicadores siguen su comportamiento, ya que hay doce datos trimestre móvil para Bogotá y otras ciudades

principales. No obstante para futuras investigaciones se recomienda comparar los datos mensuales, ya sea nacional o para ciudades principales y áreas metropolitanas, que son datos mensuales de acuerdo a la GEIH del DANE, evidentemente sujeta a supuestos, ya que el acceso a internet y uso de GT es limitado, por lo menos a nivel nacional.

Referencias bibliográficas

Angelini, E., Ba´nbura, M., Runstler, G. (2008). "Estimating and Forecasting the Euro Area Monthly National Accounts From a Dynamic Factor Model", ECB Working Paper Series, 953. Recuperado de https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp953.pdf

Angelini, E., Camba-Mendez, G., Giannone, D., Reichlin, L. (2011). "ShortTerm Forecasts of Euro Area GDP Growth", The Econometrics Journal, 14(1), C25-C44. Recuperado de https://www.jstor.org/stable/23127215

Askitas, N., Zimmermann, K. (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly*, 107-120. Recuperado de https://ftp.iza.org/dp4201.pdf

Barreira, N., Godinho, P., y Melo, P. (2013). Nowcasting unemployment rate and new car sales in south-western Europe with Google Trends. *Netnomics*, 129–165. doi:https://doi.org/10.1007/s11066-013-9082-8

Boettcher, I. (2015). "Automatic Data Collection on the Internet (Web Scraping)", New Techniques and Technologies for Statistics, Eurostat Conference, 9-13 March 2015. Recuperado de https://ec.europa.eu/eurostat/cros/system/files/NTTS2015%20proceedings.pdf_en

Campos, R., López-Araiza, S. (2020). GRANDES DATOS, GOOGLE Y DESEMPLEO. *Estudios Económicos*, *35*(1), 125-151. Recuperado de https://www.jstor.org/stable/26863997

Cardona, L., Rojas, J. (2017). Pronósticos para la tasa de desempleo en Colombia. *Archivos de economía*. Recuperado de https://colaboracion.dnp.gov.co/CDT/Estudios%20Econmicos/468.pdf

Cavallo, A., Rigobon, R. (2016). "The Billion Prices Project: Using Online Prices for Measurement and Research". The Journal of Economic Perspectives, 30(2), 151-178. Recuperado de https://www.aeaweb.org/articles?id=10.1257/jep.30.2.151

Cerchiello, P., Giudici, P. (2014). "How to Measure the Quality of Financial Tweets". Working Paper, ECB Workshop on using big data for forecasting and statistics, 07-08/04/2014, Frankfurt.

Chadwick, M., Gönül, Ş. (2012). Nowcasting Unemployment Rate in Turkey: Let's Ask Google. *Central Bank of the Republic of Turkey*. Recuperado de https://www.tcmb.gov.tr/wps/wcm/connect/635829fc-6c5f-4249-ab78-96fdc3a62cc8/WP1218.pdf?MOD=AJPERESyCACHEID=ROOTWORKSPACE-635829fc-6c5f-4249-ab78-96fdc3a62cc8-m3fw5-6

Chamberlin, G. (2010). Googling the present. *Economic and Labour Market Review*, 59-95. doi:https://doi.org/10.1057/elmr.2010.166

Chang, J., Del Río, A. (2013). Google Trends: Predicción del nivel de empleo agregado en Perú usando datos en tiempo real, 2005-2011. *Banco Central de Reserva del Perú*. Recuperado de https://www.bcrp.gob.pe/docs/Publicaciones/Documentos-de-Trabajo/2013/documento-de-trabajo-15-2013.pdf

Chen, L., Fan, S. y Lee, W-J. "Short-Term Load Forecasting Using Comprehensive Combination Based on Multimeteorological Information," in IEEE Transactions on Industry Applications, vol. 45, no. 4, pp. 1460-1466, July-aug. 2009, doi: 10.1109/TIA.2009.2023571.

Choi, H., Varian, H. (2009a). *Predicting the Present with Google Trends.* USA: Google Inc. Recuperado de https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1659302

Choi, H., Varian, H. (2009b). Predicting Initial Claims for Unemployment Benefits. Recuperado de https://static.googleusercontent.com/media/research.google.com/es//archive/papers/initialclaimsUS.pdf

Choi, H., Varian, H. (2012). Predicting the Present with Google Trends. *THE ECONOMIC RECORD, 88*, 2–9. Recuperado de https://doi.org/10.1111/j.1475-4932.2012.00809.x

Ettredge, M., Gerdes, J., y Karuga, G. (2005). Using Web-based Search Data to Predict Macroeconomic Statistics. *COMMUNICATIONS OF THE ACM*, 87-92. Recuperado de https://www.researchgate.net/publication/200110929_Using_Web-based_search_data_to_predict_macroeconomic_statistics

Giannone, D., Reichlin, L., Simonelli (2009). "Nowcasting Euro Area Economic Activity in Real-Time: The Role of Confidence Indicators", National Institute Economic Review, 210, 90-97. Recuperado de https://www.jstor.org/stable/23881019

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., y Brilliant, L. (2009). Detecting influenza epidemics using search engine. *Nature, 457*, 1012-1015. doi:https://doi.org/10.1038/nature07634

Griffioen, R., de Haan, J., Willenborg, L. (2014). "Collecting Clothing Data from the Internet", Statistics Netherlands Technical Report. Recuperado de https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2014/UNECE-ILO_2014_Griffioen_deHaan_Willenborg.pdf

Heston, S. L., Sinha, N. R. (2014). "News versus Sentiment: Comparing Textual Processing Approaches for Predicting Stock Returns", Working Paper. Recuperado de https://finpko.ku.edu/myssi/FIN938/Heston%20&%20Sinha_News%20vs%20Sentiment_WP_2014.pdf

Varian, H. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives, 28*(2), 3–28. doi:10.1257/jep.28.2.3

Yiu, M. S., Chow, K. K. (2010). "Nowcasting Chinese GDP: Information Content of Economic and Financial Data", China Economic Journal, 3(3), 223-240. Recuperado de https://www.researchgate.net/publication/228272415_Nowcasting_Chinese_GDP_Information_Content_of_Economic_and_Financial_Data

Anexos

Cuadros:

Cuadro 1: *Correlaciones palabras clave Google Trends con tasa de desempleo*

	Tasa de desempleo		
empleos	-0,40918	buscar empleo	-0,25764
CompuTrabajo	-0,42458	subsidio	0,78123
trabajos	-0,46620	prestamo	-0,53294
linkedin	0,21578	gota a gota	-0,21672
trabajos sin experiencia	-0,01312	desempleado	0,41423
el empleo	-0,49583	exito	0,40854
busco trabajo	-0,44914	falabella	0,41844
ofertas de trabajo	-0,42777	transmilenio	-0,65645
magneto	0,75863	rappi	0,53565
trabajo desde casa	0,24936	alkosto	0,69322
elempleo	-0,51717	arturo calle	0,39152
Jooble	0,06629	jeronimo martins	-0,34327
Trabajos	-0,46620	yumbo	0,58545
trabajos virtuales	0,17782	olimpica	0,44227
portal de empleo	-0,31467	justo y bueno	0,80808
sena empleo	-0,52845	ara	0,74505
formato hoja de vida	-0,48847	servicio de aseo	-0,09389
hoja de vida	-0,45139	servicio de empleo	-0,32772
vacantes	-0,00165	empleo compensar	-0,24231
cv	-0,20408	empleo colsubsidio	0,20694
minerva	-0,26322	empleo call center	-0,14447
formato hoja de vida	-0,48847	outsourcing	-0,52701
perfil laboral	0,25289	teletrabajo	0,69706
perfil profesional	-0,55100	teletrabajo sin experiencia	0,42820
curriculum	-0,40889	servicio de domicilio	0,52041
elempleo.com	-0,45250	domicilios	0,84864
buscar trabajo	-0,19131	domicilio	0,87223
servicio de empleo	-0,34366	trabajo remoto	0,24702
trabajo	-0,46998	**Tasa de desempleo**	**1,00000**

Cuadro 2: *Ejemplo de indexación para IGT septiembre 2017*

Mes	magneto	teletrabajo	subsidio	alkosto	justo y bueno	ara	domicilio
sep-17	10,50	14,00	11,75	8,50	9,75	36,50	20,00
oct-17	12,00	16,00	12,10	8,80	8,60	29,20	20,80
Promedio (1)	11,25	15,00	11,93	8,65	9,18	32,85	20,40
Desviación estándar (2)	1,06	1,41	0,25	0,21	0,81	5,16	0,57
1/desviación (3)	0,94	0,71	4,04	4,71	1,23	0,19	1,77
Sumatoria de la inversa de las desviaciones (4)	13,60						
Ponderaciones (3)/(4)	0,07	0,05	0,30	0,35	0,09	0,01	0,13
IGT							
(sep-17)*Ponderaciones (3)/(4) : (A)	0,73	0,73	3,49	2,95	0,88	0,52	2,60
IGT sep-17 indexado= Sumatoria (A)	**11,90**						

Cuadro 3: *Empalme tasa de desempleo a base 2018*

Año	Trimestre	Base 2005 (a)	Base 2018 (b)	Empalme
2020	May - Jul*	25,1	[(Jun - Ago* de (b))*25.1]/24.1	24,9
	Jun - Ago*	24,1	[(Jul - Sep* de (b))*24.1]/22.0	23,9
	Jul - Sep*	22,0	[(Ago - Oct de (b))*22.0]/19.1	21,8
	Ago - Oct	19,1	[(Sep - Nov de (b))*19.1]/17.2	19,0
	Sep - Nov	17,2	[(Oct - Dic de (b))*17.2]/16.3	17,1
	Oct - Dic	16,3	[(Nov 20 - Ene 21 de (b))*16.3]/17.6	16,2
	Nov 20 - Ene 21	17,6	[(Dic 20 - Feb 21 de (b))*17,6]/19.5	17,5
	Dic 20 - Feb 21	19,5	[(Ene - Mar de (b))*19.5]/20.1	19,4
2021	Ene - Mar	20,1	19,9	
	Feb - Abr	19,2	18,8	
	Mar - May	18,1	17,6	
	Abr - Jun	18,1	17,7	
	May - Jul	17,0	16,8	
	Jun - Ago	16,1	15,8	
	Jul - Sep	14,4	14,0	
	Ago - Oct	13,2	13,5	
	Sep - Nov	12,3	13,0	
	Oct - Dic	11,3	12,9	
	Nov 21 - Ene 22	13,9	13,9	
	Dic 21 - Feb 22		14,2	
2022	Ene - Mar		14,6	
	Feb - Abr		12,6	
	Mar - May		11,5	
	Abr - Jun		11,1	
	May - Jul		11,2	
Fuente: DANE cálculos propios				

Cuadro 4: *Resultados de elección mejor AIC para IGT*

index	(p,q)x(P,Q)	AIC
0	2,4,2,0	266,10
1	4,4,2,0	267,23
2	0,4,2,0	268,30
3	2,4,4,0	270,09
4	2,4,2,2	270,12
5	0,4,0,2	270,27
6	4,4,4,0	270,28
7	4,4,2,2	270,62
8	0,4,0,0	271,60
9	2,4,0,2	271,65
10	4,4,0,2	271,94
11	0,4,2,2	272,29
12	0,4,4,0	272,31
13	2,4,0,0	273,17
14	2,4,0,4	273,23
15	0,2,2,0	273,71
16	2,4,4,2	274,10
17	4,4,0,4	274,15
18	2,4,2,4	274,32
19	4,4,2,4	274,79
20	0,2,0,2	274,93
21	4,4,4,2	274,99
22	0,4,0,4	275,06
23	0,2,0,0	276,20
24	4,2,2,0	276,27
25	0,4,4,2	276,52
26	2,2,2,0	277,00
27	0,2,0,4	277,75
28	0,4,2,4	278,05
29	2,4,4,4	278,13
30	4,4,4,4	278,15
31	4,0,2,0	278,53
32	4,2,0,2	278,59
33	2,2,0,2	279,11
34	4,2,0,0	279,84
35	0,2,4,0	279,88
36	4,2,4,0	280,12
37	4,2,2,2	280,28
38	0,2,2,2	280,33
39	2,2,0,0	280,62
40	2,2,4,0	280,85
41	2,2,2,2	281,11
42	4,2,0,4	281,14
43	2,2,0,4	281,93
44	4,4,0,0	282,25
45	4,0,0,2	282,70
46	0,2,2,4	283,30
47	4,0,4,0	284,12
48	4,0,2,2	284,20
49	4,2,2,4	284,23
50	4,2,4,2	284,26
51	0,2,4,2	284,49
52	2,2,4,2	284,86
53	2,2,2,4	284,86
54	4,0,0,4	285,22
55	4,0,0,0	285,47
56	4,0,2,4	287,30
57	4,0,4,2	287,65
58	4,2,4,4	288,30
59	0,2,4,4	288,46
60	2,2,4,4	288,92
61	2,0,2,0	289,37
62	2,0,0,2	290,65
63	4,0,4,4	291,23
64	2,0,0,0	291,45
65	2,0,4,0	292,86
66	2,0,2,2	293,14
67	2,0,0,4	293,69
68	2,0,4,2	296,86
69	2,0,2,4	296,91
70	2,0,4,4	300,89
71	0,0,2,0	328,42
72	0,0,0,0	328,63
73	0,0,0,2	328,78
74	0,0,4,0	332,40
75	0,0,2,2	332,41
76	0,0,0,4	332,41
77	0,0,4,2	336,40
78	0,0,2,4	336,40
79	0,0,4,4	340,40

Cuadro 5: *Resultados de elección mejor BIC para IGT*

index	(p,q)x(P,Q)	BIC
0	0,4,2,0	277,87
1	2,4,2,0	278,41
2	0,4,0,0	278,43
3	0,4,0,2	279,84
4	0,2,0,0	280,31
5	0,2,2,0	280,55
6	0,2,0,2	281,77
7	4,4,2,0	282,27
8	2,4,0,0	282,75
9	2,4,0,2	283,96
10	0,4,2,2	284,60
11	0,4,4,0	284,62
12	2,4,4,0	285,13
13	2,4,2,2	285,16
14	2,2,2,0	286,57
15	4,4,0,2	286,98
16	0,2,0,4	287,32
17	0,4,0,4	287,36
18	2,2,0,0	287,46
19	4,4,4,0	288,05
20	4,0,2,0	288,10
21	2,4,0,4	288,27
22	4,4,2,2	288,40
23	4,2,2,0	288,58
24	2,2,0,2	288,68
25	4,2,0,0	289,41
26	0,2,4,0	289,45
27	0,2,2,2	289,90
28	4,2,0,2	290,90
29	0,4,4,2	291,56
30	2,4,4,2	291,88
31	4,4,0,4	291,93
32	2,4,2,4	292,10
33	4,0,0,2	292,27
34	4,0,0,0	292,31
35	0,4,2,4	293,10
36	2,2,4,0	293,15
37	2,2,2,2	293,42
38	2,2,0,4	294,24
39	4,4,0,0	294,56
40	4,2,4,0	295,16
41	4,4,2,4	295,30
42	4,2,2,2	295,32
43	4,4,4,2	295,50
44	2,0,0,0	295,56
45	0,2,2,4	295,60
46	4,2,0,4	296,18
47	2,0,2,0	296,21
48	4,0,4,0	296,42
49	4,0,2,2	296,51
50	0,2,4,2	296,80
51	2,0,0,2	297,49
52	4,0,0,4	297,53
53	2,4,4,4	298,64
54	2,2,4,2	299,90
55	2,2,2,4	299,90
56	4,4,4,4	301,39
57	4,2,2,4	302,01
58	4,2,4,2	302,04
59	4,0,2,4	302,34
60	2,0,4,0	302,43
61	4,0,4,2	302,69
62	2,0,2,2	302,71
63	2,0,0,4	303,26
64	0,2,4,4	303,50
65	2,2,4,4	306,69
66	4,2,4,4	308,81
67	4,0,4,4	309,00
68	2,0,4,2	309,16
69	2,0,2,4	309,22
70	2,0,4,4	315,93
71	0,0,0,0	330,00
72	0,0,2,0	332,52
73	0,0,0,2	332,89
74	0,0,4,0	339,23
75	0,0,2,2	339,25
76	0,0,0,4	339,25
77	0,0,4,2	345,97
78	0,0,2,4	345,97
79	0,0,4,4	352,70

Cuadro 6: *Resultados de elección mejor AIC y BIC para IGT2*

index	(p,q)x(P,Q)	AIC
0	(0, 2, 2, 0)	279,54
1	(0, 2, 0, 2)	279,57
2	(0, 2, 2, 2)	283,54
3	(0, 2, 0, 0)	288,95
4	(2, 2, 0, 2)	294,11
5	(2, 2, 2, 0)	294,27
6	(2, 2, 2, 2)	298,11
7	(2, 2, 0, 0)	299,97
8	(2, 0, 0, 2)	304,40
9	(2, 0, 2, 0)	304,40
10	(2, 0, 2, 2)	308,40
11	(2, 0, 0, 0)	311,32
12	(0, 0, 2, 0)	327,16
13	(0, 0, 0, 2)	327,68
14	(0, 0, 2, 2)	331,15
15	(0, 0, 0, 0)	340,36

index	(p,q)x(P,Q)	BIC
0	(0, 2, 2, 0)	286,38
1	(0, 2, 0, 2)	286,41
2	(0, 2, 0, 0)	293,05
3	(0, 2, 2, 2)	293,11
4	(2, 2, 0, 2)	303,68
5	(2, 2, 2, 0)	303,84
6	(2, 2, 0, 0)	306,81
7	(2, 2, 2, 2)	310,42
8	(2, 0, 0, 2)	311,24
9	(2, 0, 2, 0)	311,24
10	(2, 0, 0, 0)	315,42
11	(2, 0, 2, 2)	317,97
12	(0, 0, 2, 0)	331,26
13	(0, 0, 0, 2)	331,79
14	(0, 0, 2, 2)	337,99
15	(0, 0, 0, 0)	341,73

Cuadro 7: *Resultados de elección mejor AIC para IGT3*

index	(p,q)x(P,Q)	AIC
0	0,1,0,1	326,63
1	0,1,2,0	328,43
2	0,1,0,2	328,44
3	0,1,1,1	328,46
4	0,2,0,1	328,57
5	1,1,0,1	328,58
6	1,2,0,1	329,57
7	0,1,2,1	329,96
8	0,2,2,0	330,21
9	0,1,1,2	330,23
10	1,1,2,0	330,27
11	2,1,0,1	330,27
12	0,2,0,2	330,39
13	1,1,0,2	330,40
14	0,2,1,1	330,41
15	1,1,1,1	330,43
16	0,1,1,0	330,56
17	1,2,2,0	331,00
18	1,2,0,2	331,11
19	1,2,1,1	331,21
20	2,2,0,1	331,41
21	2,1,2,0	331,82
22	0,2,2,1	331,86
23	1,1,2,1	331,88
24	0,1,2,2	331,94
25	1,2,2,1	331,96
26	2,1,0,2	331,98
27	2,1,1,1	332,02
28	0,2,1,2	332,16
29	1,1,1,2	332,17
30	0,2,1,0	332,56
31	1,1,1,0	332,56
32	1,2,1,2	332,60
33	2,2,0,2	332,79
34	2,2,1,1	332,96
35	2,2,2,0	333,33
36	2,1,2,1	333,43
37	2,2,2,1	333,62
38	1,2,2,2	333,71
39	2,1,1,2	333,73
40	2,0,0,1	333,75
41	0,2,2,2	333,85
42	1,1,2,2	333,88
43	2,2,1,2	334,21
44	1,2,1,0	334,28
45	2,1,1,0	334,40
46	2,0,2,0	334,56
47	2,2,2,2	335,04
48	2,0,0,2	335,30
49	2,0,1,1	335,38
50	2,1,2,2	335,43
51	2,2,1,0	336,21
52	2,0,2,1	336,46
53	1,0,0,1	336,54
54	2,0,1,2	337,03
55	2,0,1,0	337,97
56	1,0,0,2	338,17
57	1,0,1,1	338,21
58	2,0,2,2	338,42
59	1,0,2,0	338,50
60	1,0,1,0	339,90
61	1,0,2,1	339,91
62	1,0,1,2	339,95
63	1,0,2,2	341,80
64	0,1,0,0	341,89
65	0,2,0,0	343,78
66	1,1,0,0	343,78
67	0,0,0,1	345,11
68	0,0,0,2	345,27
69	0,0,1,1	345,49
70	1,2,0,0	345,68
71	2,1,0,0	345,78
72	0,0,2,0	346,68
73	0,0,1,2	346,96
74	0,0,2,1	347,24
75	2,2,0,0	347,67
76	0,0,1,0	347,98
77	0,0,2,2	348,89
78	2,0,0,0	350,33
79	1,0,0,0	352,74
80	0,0,0,0	364,06

Cuadro 8: *Resultados de elección mejor BIC para IGT3*

index	(p,q)x(P,Q)	BIC
0	1,2,0,1	348
1	2,2,0,1	349
2	1,2,0,2	351
3	1,2,1,1	352
4	2,2,2,0	352
5	1,2,2,0	352
6	2,2,0,2	352
7	1,2,1,0	352
8	2,2,1,1	352
9	0,2,0,1	352
10	2,2,1,0	354
11	0,2,0,2	354
12	0,2,1,1	354
13	1,2,2,1	355
14	1,2,1,2	355
15	2,2,2,1	355
16	0,2,1,0	355
17	0,2,2,0	355
18	2,2,1,2	355
19	0,2,1,2	357
20	0,2,2,1	358
21	1,2,2,2	359
22	2,2,2,2	359
23	0,2,2,2	362
24	1,2,0,0	362
25	2,2,0,0	363
26	2,1,0,1	365
27	2,1,2,0	367
28	2,1,0,2	368
29	2,1,1,1	368
30	2,1,1,0	369
31	0,2,0,0	369
32	2,1,2,1	370
33	2,1,1,2	371
34	1,1,0,1	371
35	1,1,1,0	373
36	2,1,2,2	374
37	1,1,0,2	374
38	1,1,1,1	374
39	1,1,2,0	375
40	2,1,0,0	377
41	1,1,2,1	378
42	1,1,1,2	378
43	1,1,2,2	381
44	1,1,0,0	384
45	2,0,0,1	386
46	2,0,2,0	387
47	0,1,0,2	387
48	2,0,2,1	388
49	0,1,1,1	388
50	0,1,0,1	388
51	2,0,0,2	389
52	0,1,1,0	389
53	2,0,1,1	389
54	0,1,2,0	390
55	2,0,2,2	391
56	0,1,1,2	391
57	2,0,1,2	392
58	0,1,2,1	392
59	2,0,1,0	392
60	0,1,2,2	396
61	2,0,0,0	401
62	1,0,0,1	402
63	1,0,1,0	402
64	1,0,0,2	403
65	1,0,1,1	403
66	1,0,2,0	404
67	0,1,0,0	405
68	1,0,1,2	407
69	1,0,2,1	407
70	1,0,2,2	410
71	1,0,0,0	416
72	0,0,0,2	425
73	0,0,1,1	427
74	0,0,1,0	429
75	0,0,1,2	429
76	0,0,2,0	429
77	0,0,0,1	430
78	0,0,2,1	430
79	0,0,2,2	433
80	0,0,0,0	449

Resultados modelos

Modelo SARIMA mejor AIC para IGT:

```
      SARIMAX Results
==========================================================================
Dep. Variable: IGT                       No. Observations: 55
Model: SARIMAX(2, 2, 4)x(2, 2, 0, 12)    Log Likelihood   -124.051
                                         AIC               266.102
                                         BIC               278.407
Sample:0                                 HQIC              269.956
- 55
Covariance Type: opg
==============================================================================
                 coef    std err          z      P>|z|    [0.025 0.975]
------------------------------------------------------------------------------
ar.L1         -1.7581    0.096    -18.285      0.000    -1.947  -1.570
ar.L2         -0.9860    0.103     -9.579      0.000    -1.188  -0.784
ma.L1         -0.4715    3.758     -0.125      0.900    -7.838   6.895
ma.L2         -1.3533    1.831     -0.739      0.460    -4.941   2.235
ma.L3          0.3005    3.503      0.086      0.932    -6.565   7.166
ma.L4          0.6072    2.349      0.259      0.796    -3.997   5.211
ar.S.L12      -0.7192    0.240     -2.995      0.003    -1.190  -0.249
ar.S.L24      -0.9396    0.127     -7.400      0.000    -1.188  -0.691
sigma2        39.0184    132.921    0.294      0.769   -221.501  299.538
==============================================================================
Ljung-Box (L1) (Q):     0.29         Jarque-Bera (JB): 0.07
Prob(Q)                 0.59         Prob(JB):         0.96
Heteroskedasticity (H): 1.96         Skew:             0.04
Prob(H) (two-sided):    0.30         Kurtosis:         2.77
===================================================================================

MAPE  SARIMAX 1 : 0.35 %
RMSE SARIMAX 1 : 6.52
```

Modelo SARIMA mejor BIC para IGT:

```
SARIMAX Results
==========================================================================
Dep. Variable: IGT                       No. Observations:   55
Model: SARIMAX(0, 2, 4)x(2, 2, 0, 12)    Log Likelihood     -127.151
                                         AIC                 268.301
                                         BIC                 277.872
Sample: 0                                HQIC                271.299
- 55
Covariance Type: opg
==============================================================================
                 coef    std err          z      P>|z|    [0.025 0.975]
------------------------------------------------------------------------------
ma.L1         -2.9236   3561.039     -0.001    0.999  -6982.432 6976.585
ma.L2          2.8313   6861.790      0.000    1.000  -1.34e+04 1.35e+04
ma.L3         -0.8626   3236.097     -0.000    1.000  -6343.497 6341.772
ma.L4         -0.0451    160.395     -0.000    1.000  -314.414  314.324
ar.S.L12      -0.6639      0.354     -1.875    0.061  -1.358    0.030
ar.S.L24      -0.7757      0.252     -3.078    0.002  -1.270    -0.282
sigma2        91.5793   3.26e+05      0.000    1.000  -6.4e+05  6.4e+05
==============================================================================
Ljung-Box (L1) (Q):     0.11         Jarque-Bera (JB): 0.86
Prob(Q):                0.74         Prob(JB)          0.65
Heteroskedasticity (H): 2.00         Skew:             0.40
Prob(H) (two-sided):    0.29         Kurtosis:         2.72
==============================================================================

MAPE  SARIMAX 2 : 0.90 %
RMSE SARIMAX 2 : 10.98
```

Modelo SARIMA mejor AIC y BIC para IGT2:

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                               IGT2   No. Observations:                   55
Model:             SARIMAX(0, 2, 2)x(2, 2, [], 12)   Log Likelihood                -134.770
Date:                             Tue, 27 Sep 2022   AIC                             279.539
Time:                                     00:58:25   BIC                             286.376
Sample:                                          0   HQIC                            281.681
                                               - 55
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -1.9487     28.330     -0.069      0.945     -57.474      53.576
ma.L2          0.9994     29.057      0.034      0.973     -55.951      57.950
ar.S.L12      -0.8849      0.276     -3.208      0.001      -1.426      -0.344
ar.S.L24      -0.3748      0.370     -1.012      0.311      -1.100       0.351
sigma2       365.8971   1.06e+04      0.034      0.973   -2.05e+04    2.12e+04
===================================================================================
Ljung-Box (L1) (Q):                   3.24   Jarque-Bera (JB):                 1.40
Prob(Q):                              0.07   Prob(JB):                         0.50
Heteroskedasticity (H):               0.53   Skew:                            -0.53
Prob(H) (two-sided):                  0.33   Kurtosis:                         2.78
===================================================================================


MAPE  SARIMAX 1 : 6.06 %
RMSE SARIMAX 1 : 18.83
```

Modelo SARIMA mejor AIC para IGT3:

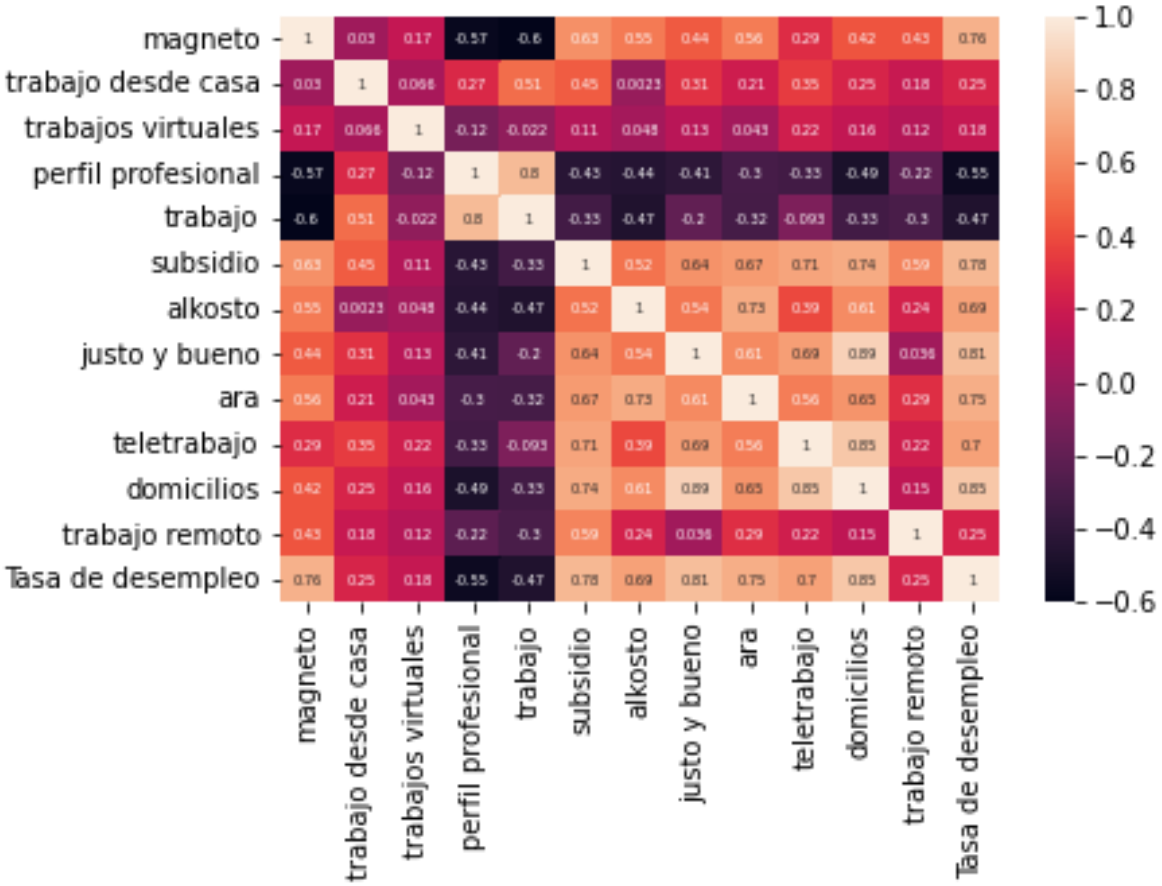
```
                               SARIMAX Results
==========================================================================================
Dep. Variable:                               IGT3   No. Observations:                   57
Model:             SARIMAX(0, 1, 1)x(0, 1, 1, 12)   Log Likelihood                -160.315
Date:                             Tue, 27 Sep 2022   AIC                             326.629
Time:                                     01:50:09   BIC                             331.982
Sample:                                 09-01-2017   HQIC                            328.614
                                      - 05-01-2022
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.7616      0.128     -5.943      0.000      -1.013      -0.510
ma.S.L12      -0.9993    204.822     -0.005      0.996    -402.443     400.444
sigma2        55.1596   1.13e+04      0.005      0.996   -2.21e+04    2.22e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.23   Jarque-Bera (JB):                 0.56
Prob(Q):                              0.63   Prob(JB):                         0.76
Heteroskedasticity (H):               0.89   Skew:                            -0.21
Prob(H) (two-sided):                  0.83   Kurtosis:                         3.36
===================================================================================


MAPE3  SARIMAX 1 : 0.23 %
RMSE3 SARIMAX 1 : 6.77
```

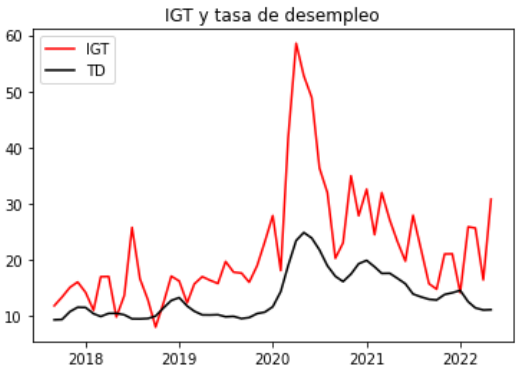
Modelo SARIMA mejor AIC para IGT3:

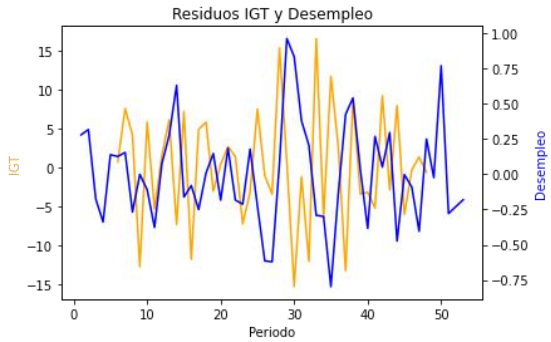
```
                     SARIMAX Results
==========================================================================================
Dep. Variable:                               IGT3   No. Observations:                   55
Model:             SARIMAX(1, 1, 2)x(0, 1, [1], 12)   Log Likelihood                -164.792
Date:                              Tue, 27 Sep 2022   AIC                             339.584
Time:                                      02:29:06   BIC                             348.272
Sample:                                           0   HQIC                            342.768
                                                - 55
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4227      0.157     -2.693      0.007      -0.730      -0.115
ma.L1         -1.9680     90.138     -0.022      0.983    -178.635     174.699
ma.L2          0.9998     91.585      0.011      0.991    -178.504     180.504
ma.S.L12      -0.9979     84.500     -0.012      0.991    -166.616     164.620
sigma2        77.5797   1.05e+04      0.007      0.994   -2.05e+04    2.06e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.92   Jarque-Bera (JB):                 2.83
Prob(Q):                              0.34   Prob(JB):                         0.24
Heteroskedasticity (H):               0.87   Skew:                            -0.58
Prob(H) (two-sided):                  0.80   Kurtosis:                         2.47
===================================================================================


MAPE3  SARIMAX 2 : 0.72 %
RMSE3 SARIMAX 2 : 8.39
```

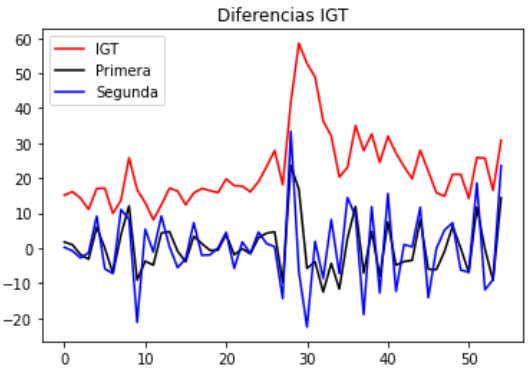
Gráficos:

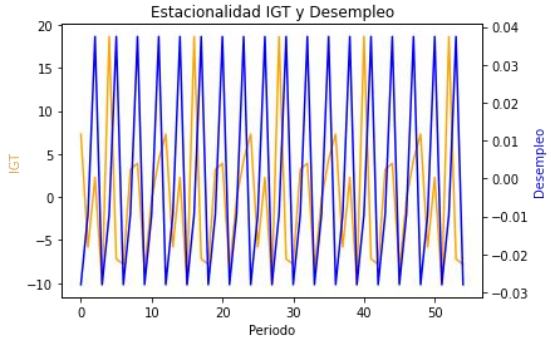
Gráfico 1: *Mapa de calor de correlaciones entre palabras clave GT y tasa de desempleo*

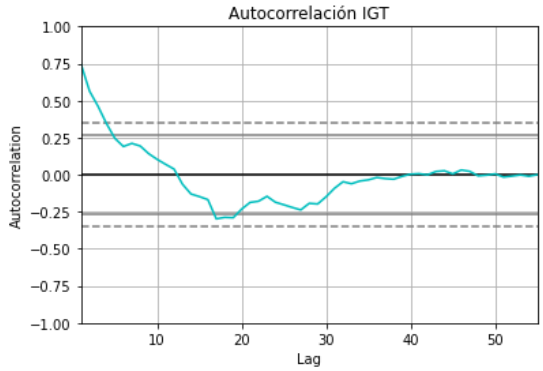


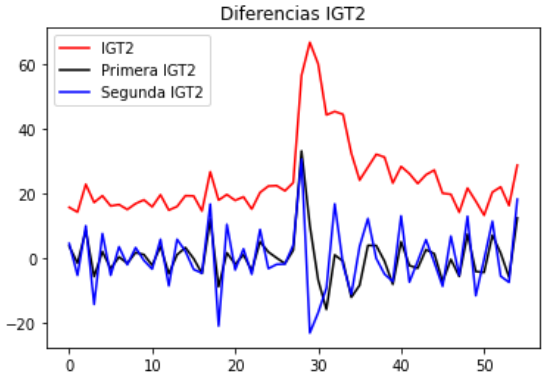


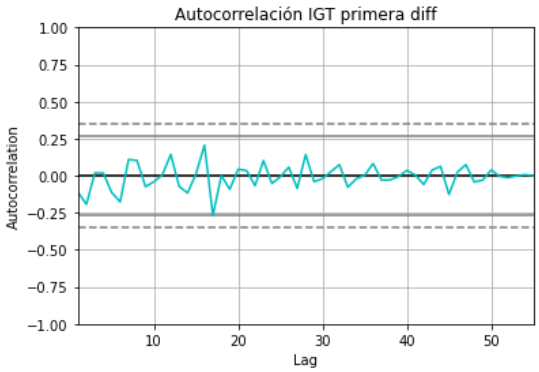


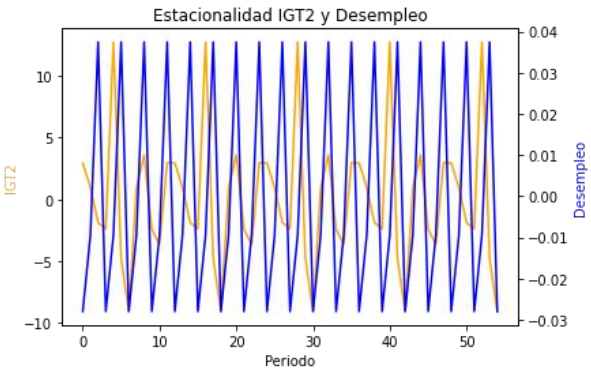


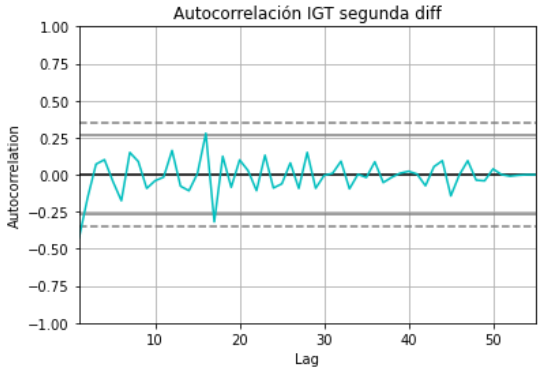


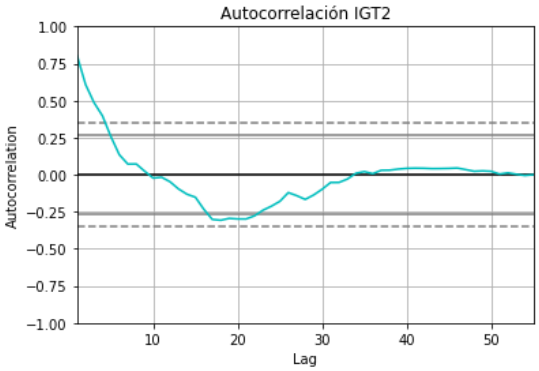

Autocorrelación IGT
Autocorrelation
Lag

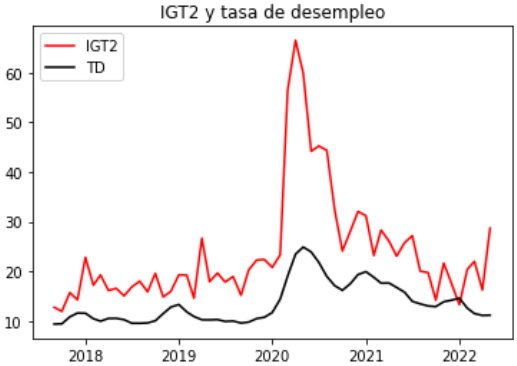

Diferencias IGT2
IGT2
Primera IGT2
Segunda IGT2

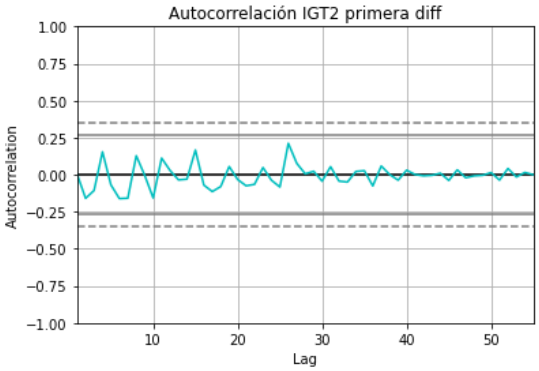

Autocorrelación IGT primera diff
Autocorrelation
Lag

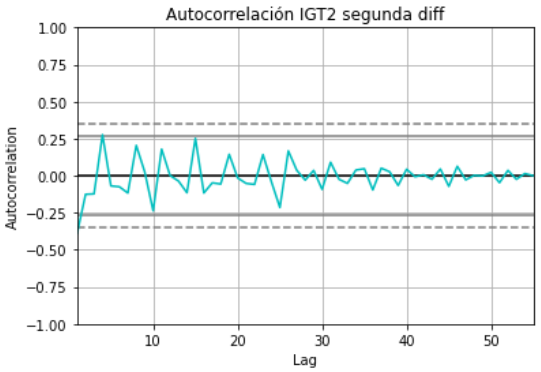

Estacionalidad IGT2 y Desempleo
IGT2
Desempleo
Periodo

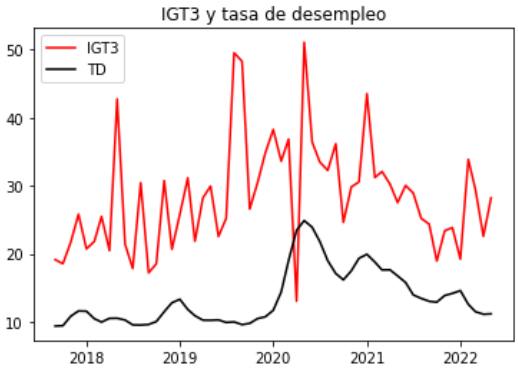

Autocorrelación IGT segunda diff
Autocorrelation
Lag


Autocorrelación IGT2
Autocorrelation
Lag


IGT2 y tasa de desempleo
IGT2
TD


Autocorrelación IGT2 primera diff
Autocorrelation
Lag


Autocorrelación IGT2 segunda diff
Autocorrelation
Lag


IGT3 y tasa de desempleo
IGT3
TD

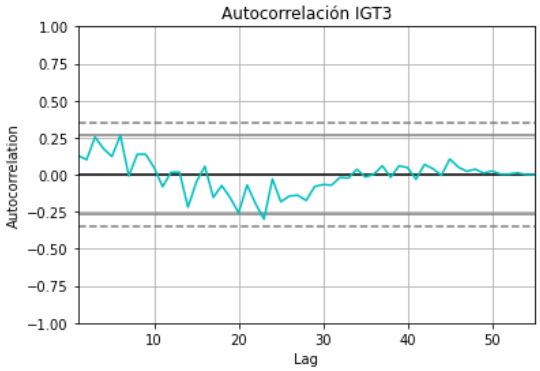

Autocorrelación IGT3
Autocorrelation
Lag

Gráfico 18: *Diagnósticos SARIMA de IGT para mejor AIC*

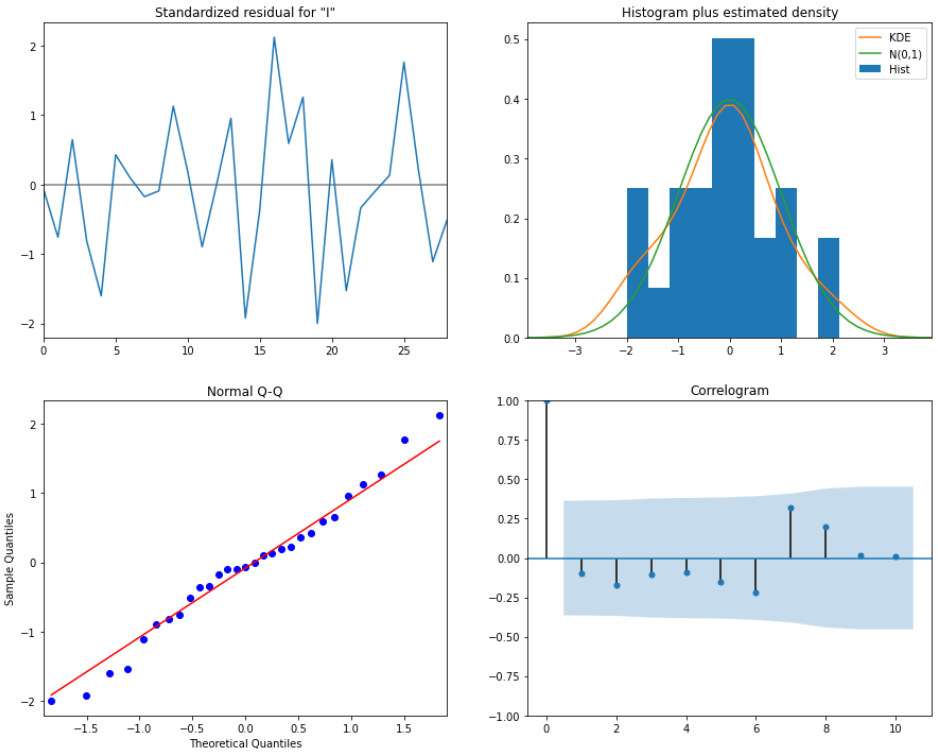


Gráfico 19: *Diagnósticos SARIMA de IGT para mejor BIC*

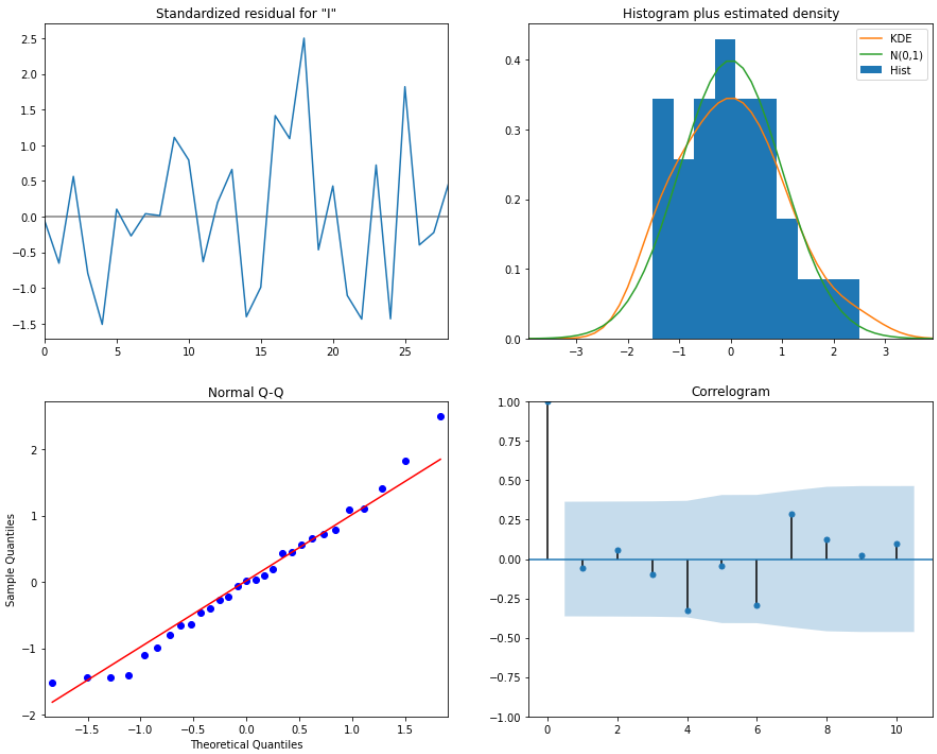


Gráfico 20: *Predicción SARIMA de IGT para mejor AIC*

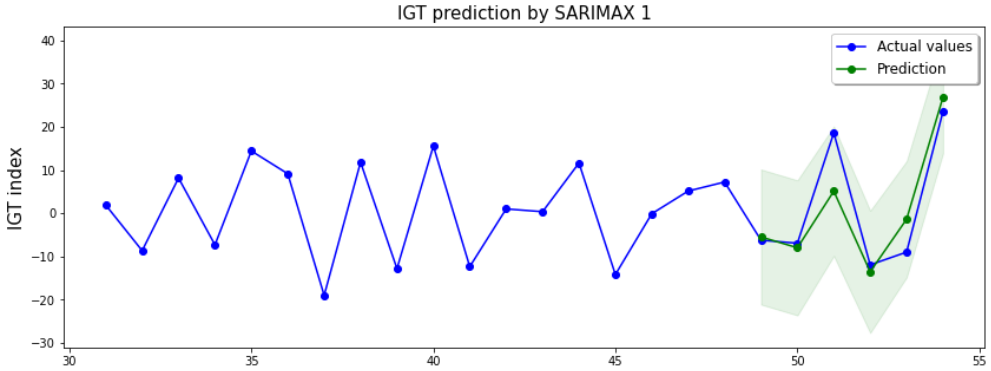


Gráfico 21: *Forecast SARIMA de IGT para mejor AIC*

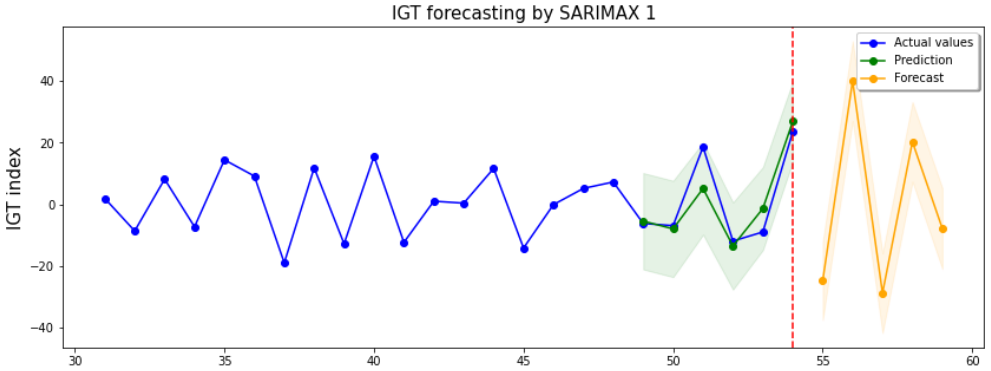


Gráfico 22: *Predicción SARIMA de IGT para mejor BIC*

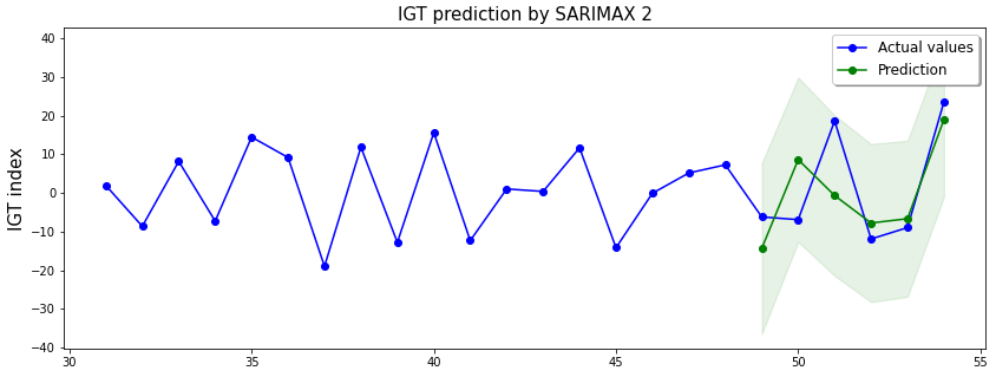


Gráfico 23: *Forecast SARIMA de IGT para mejor BIC*

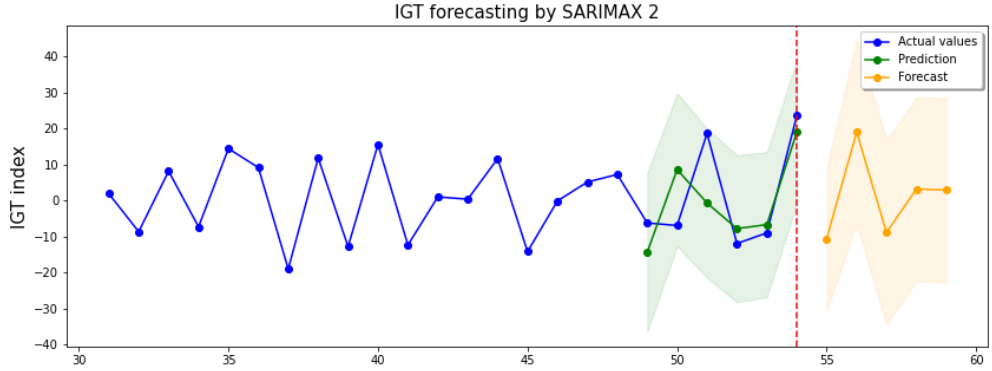


Gráfico 24: *Diagnósticos SARIMA de IGT2 para mejor AIC y BIC*

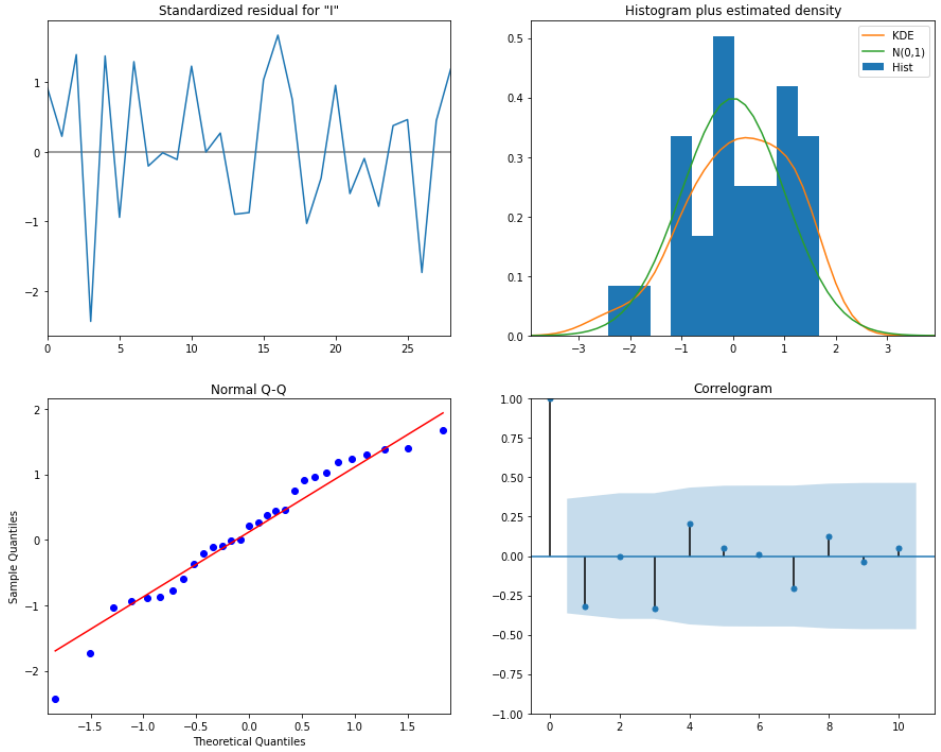


Gráfico 25: *Predicción SARIMA de IGT2 para mejor AIC y BIC*

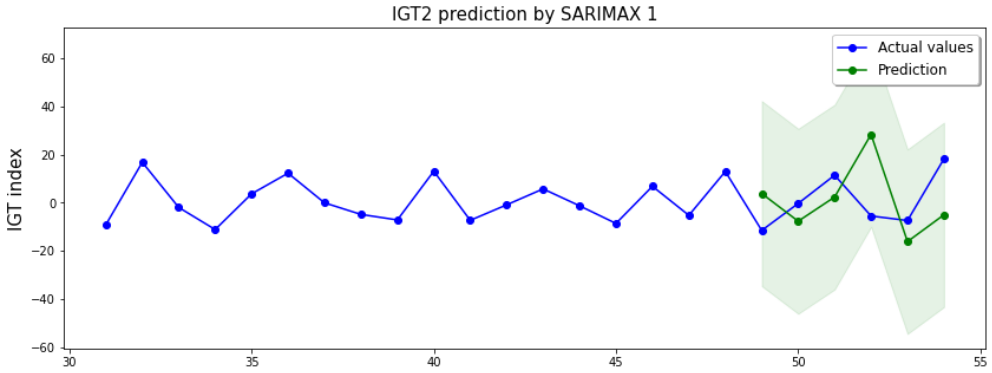


Gráfico 26: *Forecast SARIMA de IGT2 para mejor AIC y BIC*

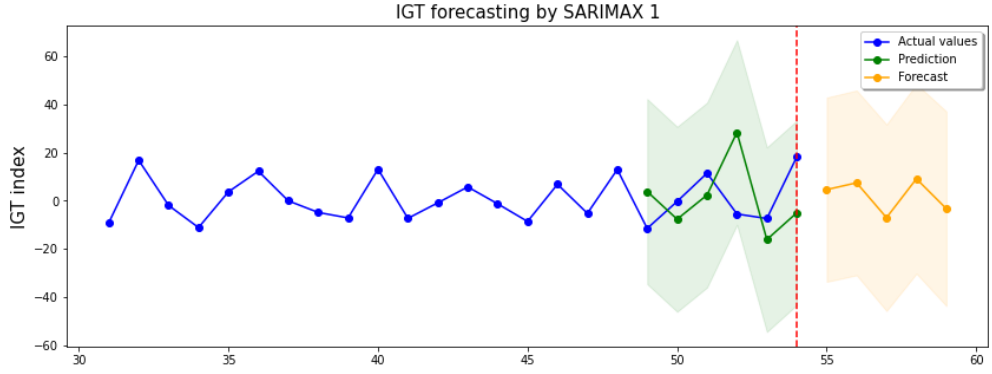


Gráfico 27: *Diagnósticos SARIMA de IGT3 para mejor AIC*

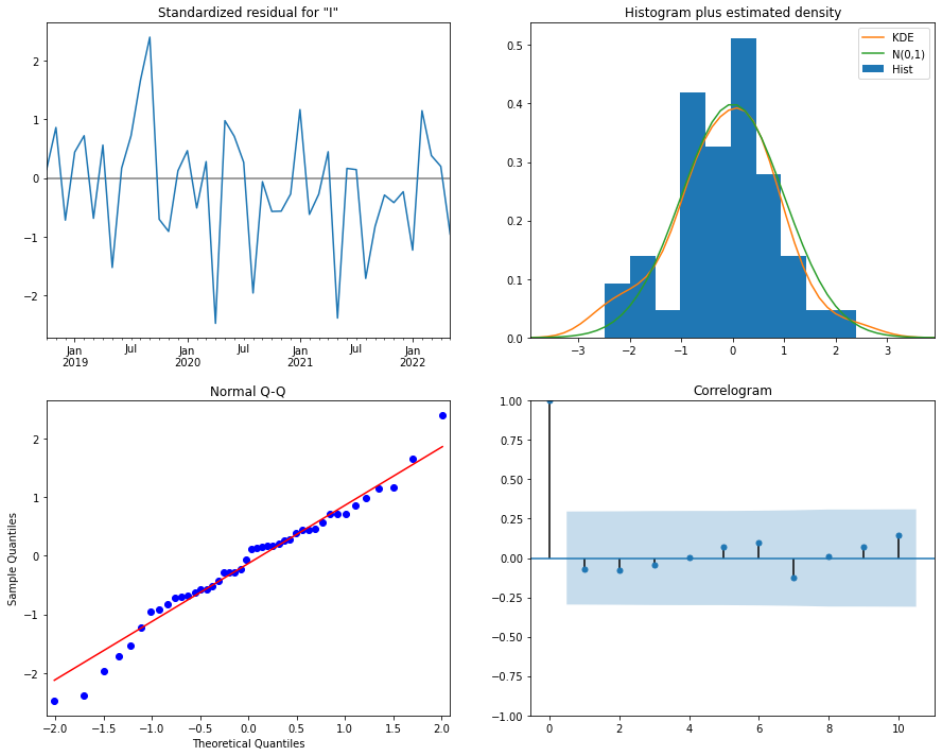


Gráfico 28: *Predicción SARIMA de IGT3 para mejor AIC*

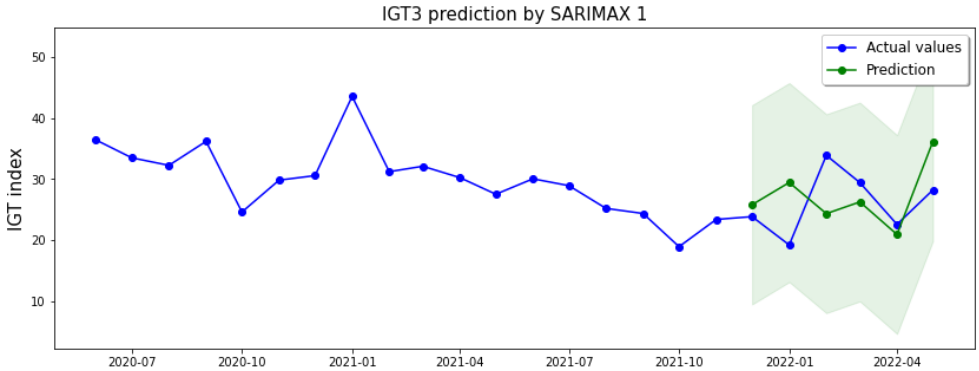


Gráfico 29: *Forecast SARIMA de IGT3 para mejor AIC*

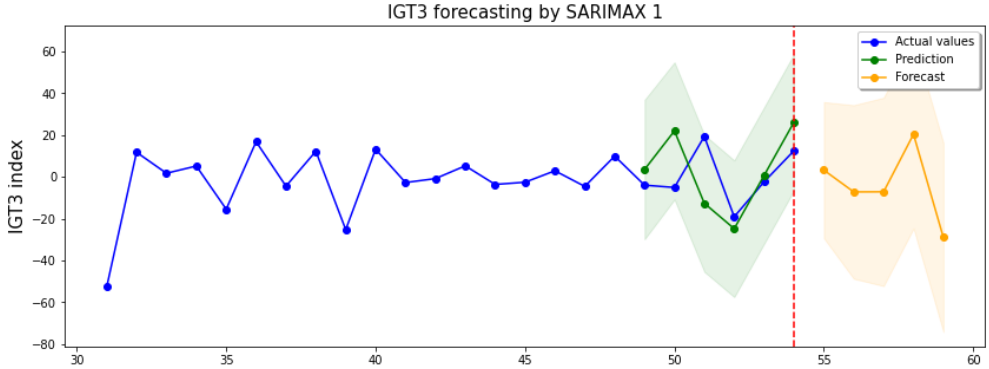


Gráfico 30: *Diagnósticos SARIMA de IGT3 para mejor BIC*

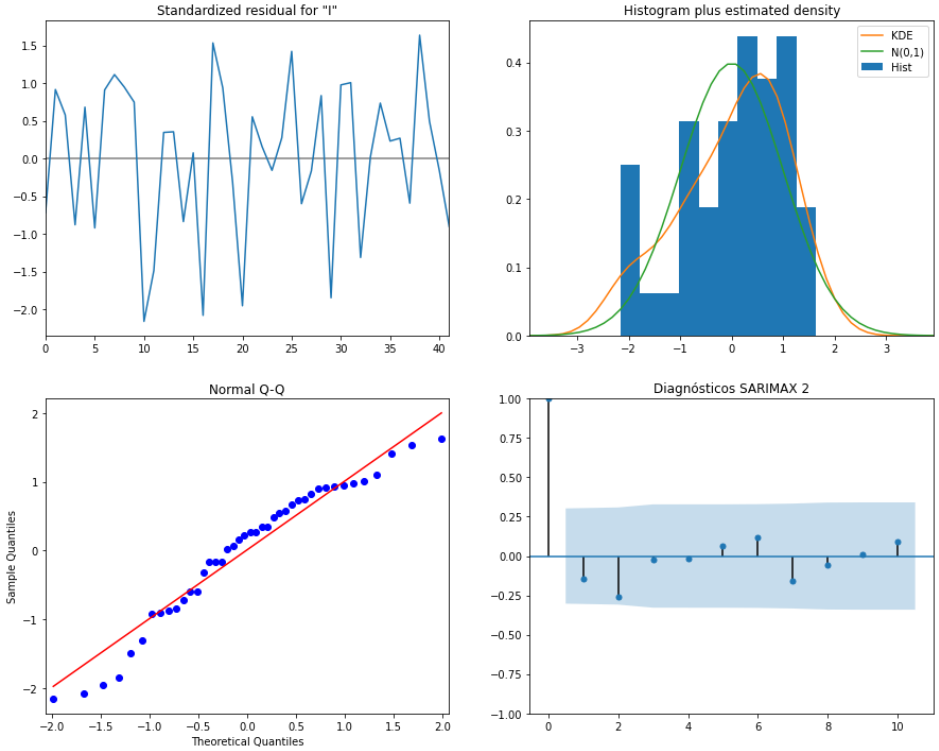


Gráfico 31: *Predicción SARIMA de IGT3 para mejor BIC*

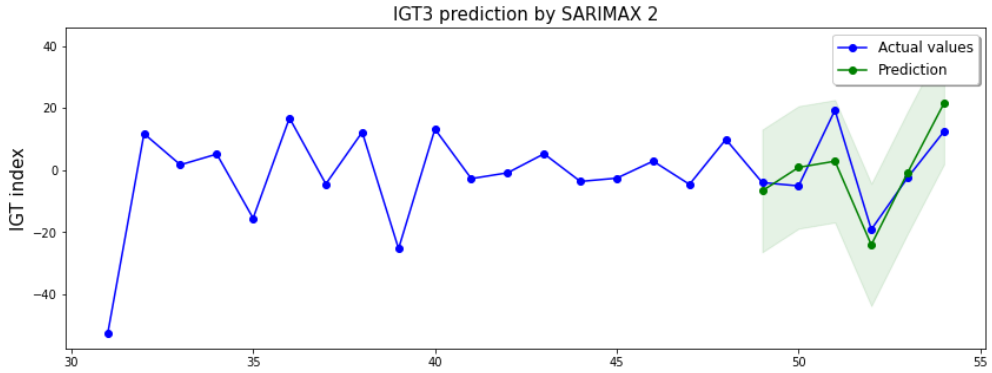


Gráfico 32: *Forecast SARIMA de IGT3 para mejor BIC*