



**VAES**  
Viceministerio de Análisis  
Económico y Social

## Proyectando el desempleo en República Dominicana con datos de *Google Trends*

Texto de discusión N.º 47

**Autora<sup>1</sup>:**

Reyna Gomera

Noviembre 2024  
Santo Domingo, D.N.

---

<sup>1</sup> El contenido de este documento es de exclusiva responsabilidad de la autora, por tanto, no refleja la posición del Ministerio de Economía, Planificación y Desarrollo.



Ministro:  
**Pável Isa Contreras**

**Viceministerio de Análisis Económico y Social  
(VAES)**

Viceministro:  
**Alexis Cruz Rodríguez**

**Dirección de Análisis Macroeconómico  
(DAM)**

Directora:  
**Yaurimar Terán Hernández**

Coordinadora de Proyecciones Económicas  
**Reyna Gomera**

Apoyo editorial  
**Dirección de Comunicación del  
Ministerio de Economía, Planificación y Desarrollo**



# Proyectando el desempleo en República Dominicana con datos de *Google Trends*<sup>2</sup>

## Resumen

Dada la relevancia de la tasa de desempleo y la aparición de nuevas fuentes de información ante el mayor acceso y uso de la Internet, diversas investigaciones han evaluado si la inclusión de información en tiempo cuasi real, como los índices de búsqueda de *Google Trends*, mejora la capacidad predictiva de diferentes estrategias de modelaje. En línea con estas investigaciones, el presente documento evalúa si para el caso de República Dominicana se observaría un mejor desempeño de los modelos de series de tiempo tradicionales (ARIMAX) permitiendo contar con pronósticos de corto plazo para el desempleo más robustos. La evidencia apunta a una mejora de las predicciones con la inclusión de búsquedas en Google de términos como “trabajo”, “liquidación” o portales de oferta laboral como “aldaba” y “LinkedIn”.

**Palabras clave:** desempleo, Google Trends, proyecciones, términos de búsqueda.

---

<sup>2</sup> Se agradecen los valiosos aportes de Mario Ávila. Así como los comentarios de miembros del Viceministerio de Análisis Económico y Social, en especial, de Néstor Viñals, Lisselotte Gálvez e Ilvin Mendoza.

## 1. Introducción

La evolución del desempleo es una medida clave del desempeño del mercado laboral y de la salud económica de un país. Históricamente, períodos caracterizados por altos niveles de desempleo suelen asociarse con episodios de baja actividad económica y viceversa, constituyendo entonces el desempleo y el crecimiento económico dos caras de una misma moneda. Por lo anterior, contar con información oportuna sobre la trayectoria presente y futura de este indicador, es de vital importancia para la toma de decisiones tanto en el nivel público como privado; en especial, cuando las publicaciones oficiales ocurren con cierto rezago debido a cuestiones logísticas<sup>3</sup>. Esta necesidad ha propiciado el uso de fuentes de datos alternativas con la finalidad de garantizar la disponibilidad de estadísticas de alta frecuencia.

En este contexto, y bajo el supuesto de que las necesidades y preferencias de los individuos serían reveladas mediante una simple búsqueda en la red ([Ettredge et al., 2005](#)), es decir, un individuo desempleado buscaría términos relacionados a ofertas laborales; una gran cantidad de trabajos empíricos ha incorporado datos de la Internet para evaluar cómo mejoran los pronósticos de corto plazo del desempleo ([Choi y Varian 2009b, 2012; D'Amuri y Marcucci 2009; Askitas y Zimmermann 2009; Tuhkuri 2015; Burop y Montes 2019](#)).

La evidencia empírica da cuenta de una mayor capacidad predictiva de los modelos ante la inclusión de índices de búsqueda, incluso en países en donde la proporción de personas sin empleo es alta ([Vicente et al. 2015; Simionescu y Cifuentes-Faura 2022](#)) y el nivel de informalidad laboral es elevado ([Trespalacios 2021; Chang y Del Río 2013](#)).

Hasta donde se tiene conocimiento, no existe una investigación que documente la mejora en los pronósticos de desempleo al utilizar datos de búsquedas en la red para República Dominicana, una economía con altos niveles de informalidad. Para contextualizar, en este país, la tasa de desempleo promedió 6.34 % durante el periodo 2004-2023, inferior a la media de la región de América Latina y El Caribe (7.52 %). No obstante, un gran número de

---

<sup>3</sup> En el caso de República Dominicana, los datos del mercado laboral son divulgados 90 días después del cierre del trimestre correspondiente.

personas ocupadas se encuentran en la informalidad, representando un 56.53 % de los ocupados totales en 2023.

En ese sentido, el propósito de esta investigación es evaluar si en el contexto nacional, se observaría una mejora de la capacidad predictiva de modelos convencionales de series de tiempo (modelo autorregresivo integrado de media móvil -ARIMA-) del desempleo ante la inclusión de índices de búsquedas de *Google Trends* relacionados con el término empleo y portales de búsqueda de ofertas laborales. Los resultados obtenidos confirman en principio la hipótesis planteada, aunque la ganancia es inferior relativo a lo documentado en países desarrollados. Esta dinámica resulta esperable, dada las características del mercado laboral dominicano y el avance relativamente reciente del acceso y uso del internet en el país.

La investigación se estructura de la siguiente forma. En la sección [2](#) se presentan estudios previos que incorporan información de *Google Trends* para la mejora de pronósticos, en particular, de indicadores del mercado laboral como la tasa de desempleo. La descripción de los datos está contenida en la sección [3](#). La sección [4](#) contiene los modelos de series de tiempo utilizados. En la sección [5](#) se contrastan los pronósticos arrojados por los modelos que utilizan información de *Google Trends* y el modelo de referencia. Finalmente, se presentan las conclusiones de la investigación.

## 2. Revisión de la literatura

En una gran cantidad de trabajos empíricos, se ha analizado cómo el uso de datos con disponibilidad en tiempo quasi real mejora las predicciones de corto plazo -pasado reciente, presente, y el futuro muy próximo- de indicadores convencionales. Esta interrogante ha sido abordada en diferentes ámbitos, desde el estudio de la propagación de enfermedades virales (ver [Johnson et al. 2004; Eysenbach, 2006; Ginsberg et al. 2009](#)), hasta el análisis del comportamiento de los consumidores ([Choi y Varian, 2009a, 2009b, 2012; Goel et al. 2010](#); entre otros).

Los estudios parten de la premisa de que los individuos con miras a maximizar su beneficio optan por tomar decisiones informadas, y dada la amplitud de la Internet, esta necesidad de más información suele traducirse en la mayoría de los casos en una búsqueda en la web. En esencia, la búsqueda de ciertos

términos revelaría información sobre las necesidades y preferencias de los individuos ([Ettredge et al., 2005](#)). Por ejemplo, en el caso de las personas desempleadas resultaría esperable que tiendan a buscar en la red información sobre potenciales ofertas laborales. En ese sentido, basándose en este supuesto, en el ámbito macroeconómico se han usado diversas medidas de popularidad sobre términos de búsquedas, con la finalidad de mejorar pronósticos del consumo, la inversión y, en línea con esta investigación, del desempleo.

Relativo a los indicadores del mercado laboral, varios han sido los autores que han documentado mejoras en los pronósticos, en particular, del desempleo; utilizando información procedente de *Google Trends*. Por ejemplo, para los Estados Unidos, [Choi y Varian \(2009b, 2012\)](#) y [Tuhkuri \(2015\)](#) demuestran que, al incluir índices de búsqueda, las proyecciones de desempleo mejoran marginalmente relativo a los resultados de modelos convencionales. Por su parte, [D'Amuri y Marcucci \(2009\)](#) obtienen un mejor desempeño con una disminución del error cuadrático medio de 19.8 %. En el caso de [Burop y Montes \(2020\)](#), concluyen que la combinación de índices de búsqueda, preferiblemente no lineal, podría mejorar de forma sustancial el pronóstico sobre el crecimiento del empleo; contrario al caso en donde solo se utilizan los índices individuales.

[Askitas y Zimmermann \(2009\)](#) obtienen mejoras para el pronóstico del desempleo en Alemania. Los autores hacen uso de cuatro grupos de términos vinculados a “*unemployment agencies*”, “*unemployment rate*”, “*personnel consultant*” y “*job search energies*”, este último resulta de una combinación de términos vinculados a través del operador booleano “OR”.

[McLaren y Shanbhogue \(2011\)](#), para el caso de Reino Unido, demuestran que los pronósticos del desempleo y del precio de las viviendas mejoran al incluir estos índices de búsqueda, incluso aun cuando los datos de *Google Trends* no reflejan el volumen de búsqueda y el cálculo se basa en una submuestra, resultando en índices distintos según la ventana temporal seleccionada. Este resultado descansa en que la dinámica de las series resulta más relevante para el pronóstico, que el volumen actual de búsquedas por sí mismo.

Esta estrategia también ha sido probada en economías con un alto nivel de desempleo. [Vicente et al. \(2015\)](#), para España, utilizan modelos ARIMA e

incluyen como variables explicativas el índice de búsqueda de “ofertas laborales” y un indicador de confianza empresarial relacionado con las perspectivas de empleo, arribando a la conclusión de que las previsiones de desempleo mejoran, incluso en un contexto de crisis económica. [Simionescu y Cifuentes-Faura \(2022\)](#) presentan el análisis para Portugal y España, con la novedad de predicciones a nivel regional para el caso español. Los autores demuestran que se obtiene una mejoría de las predicciones de las tasas de desempleo en ambos países, y en especial, en las regiones españolas. A resultados similares exclusivamente para el caso español, arriban [Mulero y García-Hiernaux \(2022\)](#) al desagregar por edad, género y brecha digital.

Para países de la región de América Latina y El Caribe, [Campos y López-Araiza \(2020\)](#) utilizan la combinación de “empleo” y “bolsa de trabajo” haciendo uso del operador booleano “OR” y, obtienen mejores predicciones para el caso del desempleo mexicano. [Trespalacios \(2021\)](#) arriba a conclusiones similares para Colombia, dividiendo en dos grupos los índices de búsqueda, un primero referente a palabras relacionadas con el término empleo y, un segundo con plataformas de empleo.

Para Perú, [Chang y Del Río \(2013\)](#) evalúan cómo mejora la capacidad predictiva de modelos ARIMA y de corrección de errores (MCE) al incorporar información de *Google Trends*. En sus investigaciones utilizan el índice de empleo de Lima para empresas de 100 y más trabajadores (IE100) en lugar de la tasa de desempleo, debido al alto nivel de informalidad de la economía. Pese a las características del mercado laboral peruano, los autores concluyen que la capacidad predictiva de los modelos mejora significativamente.

Hasta donde se tiene conocimiento, no existe una investigación que documente la mejora en los pronósticos de desempleo en la República Dominicana al utilizar datos de las búsquedas en Google. Los estudios más próximos analizan la mejoraría en las proyecciones de consumo privado, inversión, actividad económica y turismo incluyendo a un modelo Bayesiano Estructural de Series de Tiempo datos de búsqueda procedentes de Google ([Santana 2017a, 2017b](#)). En ese sentido, esta investigación intenta evaluar si datos procedentes de la Internet, específicamente de *Google Trends*, mejoran la capacidad predictiva de modelos de desempleo para el caso nacional.

### 3. Data

En esta sección se presentan los datos utilizados. En primer lugar, se describe la serie de desempleo para República Dominicana. Posteriormente, se hace referencia a los índices de búsqueda de Google Trends seleccionados para el pronóstico del desempleo.

#### 3.1 Tasa de desempleo

Los datos de desempleo son divulgados por el Banco Central de la República Dominicana (BCRD) aproximadamente 90 días después del cierre del trimestre correspondiente. La serie utilizada corresponde al desempleo abierto, que hace referencia al número de personas que declararon buscar activamente empleo como porcentaje de la fuerza laboral. Este indicador se prefiere en lugar del desempleo ampliado, debido a que este último considera los individuos que no buscaron, pero que estaban disponibles para trabajar; y que, por tanto, no forman parte de la población objetivo de esta investigación.

La muestra comprende el periodo desde primer trimestre de 2004 hasta el primer trimestre de 2024, para un total de 81 observaciones. Entre 2004 y 2016, el cálculo de la tasa de desempleo se realiza en base a datos de la Encuesta Nacional de Fuerza de Trabajo (ENFT), instrumento que recopilaba información con una frecuencia semestral. Por tal razón, los datos fueron interpolados utilizando una extensión del algoritmo de [Chow y Lin \(1971\)](#) en R. Dada la revisión efectuada entre 2014 y 2016, para el periodo posterior el cálculo se realiza en base a datos procedentes de la Encuesta Nacional Continua de Fuerza de Trabajo (ENCFT), levantada con una frecuencia trimestral; realizándose un empalme estadístico de ambas series.

#### 3.2 Índice de búsquedas de *Google Trends*

Los datos procedentes de Google indican la frecuencia con la cual los usuarios efectúan la búsqueda de un término, relativo al total de búsquedas que se realizan en la plataforma. Este índice de popularidad es suministrado por *Google Trends* en base a una muestra de datos en tiempo real, desde 2004 hasta 72 horas antes de la búsqueda.

El cálculo del índice se realiza como en la expresión (1). Cada observación sobre la búsqueda de un término  $q$ , realizada en la zona geográfica  $i$  en un rango temporal  $t$  ( $n_{q,i,t}$ ) es dividida por el total de búsquedas de esa zona y el rango

de tiempo ( $N_{i,t}$ ) obteniéndose su popularidad relativa. Posteriormente, el resultado es escalado en un rango de 0 a 100, donde cero indica baja popularidad del término y no la ausencia de búsquedas; y 100 refiere una alta relevancia.

$$IGT_{q,i,t} = \frac{\left( \frac{n_{q,i,t}}{N_{i,t}} \right)}{\max_t \left( \frac{n_{q,i,t}}{N_{i,t}} \right)} \times 100 \quad (1)$$

Previamente *Google Trends* realiza una limpieza de los datos. Se eliminan las búsquedas que han sido realizadas por pocas personas, y aquellas que han sido efectuadas en el ordenador de forma repetida en un corto tiempo. Asimismo, son filtradas las búsquedas que contienen apostrofes u otros caracteres especiales.

Debido a que el objetivo de esta investigación es probar si el pronóstico de desempleo mejora con la inclusión de índices de *Google Trends*, en una primera etapa se efectuó la identificación y selección de términos relevantes. El proceso inició con la división de los términos de búsqueda en dos grupos. En un primer grupo se evaluaron términos estándares de búsquedas de trabajo (p. ej. “empleo”, “empleo rd”, “trabajo”, “vacantes”, “oferta laboral”). En una segunda agrupación se contemplaron los portales de empleo disponibles para República Dominicana como p. ej. “Aldaba”, “Tu Empleo RD”, “Chiripas”, “LinkedIn”, “Concursa Map”, entre otros.

En un primer proceso de filtrado, se eliminaron las series con baja o nula popularidad, y se seleccionaron las que tenían mayor correlación con el nivel de desempleo. En total se evaluaron 30 series<sup>4</sup>, de las cuales ocho fueron preseleccionadas: “empleo”, “trabajo”, “vacante”, “liquidación”, “preaviso”, “prestación”, “LinkedIn” y “Aldaba”. Finalmente, utilizando el algoritmo *Forward Stepwise Selection*, que consiste en estimar varios modelos con los potenciales regresores, y seleccionar el modelo que resulte con en el menor criterio de información de Aikaike (AIC); en total se seleccionaron los siguientes cinco índices: “empleo”, “trabajo”, “liquidación”, “LinkedIn” y “Aldaba”.

---

<sup>4</sup> Ver tabla A1 en anexos para más detalle sobre las variables consideradas.

#### 4. Metodología

El modelo *benchmark* para el pronóstico del desempleo abierto se realiza siguiendo la estrategia planteada por [Box y Jenkins \(1976\)](#). En la etapa de identificación se utilizan las herramientas tradicionales como la visualización de la serie, el correlograma y las pruebas de raíces unitarias. Formalmente, se evalúa la presencia de raíces unitarias utilizando las pruebas de Dickey-Fuller aumentada ([Dickey y Fuller, 1979](#)) y KPSS ([Kwiatkowski et al., 1992](#)).

Luego de efectuar las transformaciones pertinentes a la serie de desempleo ( $y_t$ )<sup>5</sup>, se procede a identificar el número de rezagos óptimos utilizando el correlograma. Como resultado, la forma del modelo base es un ARIMA (1,1,0). No obstante, dado los efectos de la pandemia en el mercado laboral dominicano, y en consecuencia la presencia de valores atípicos influyentes en el año 2020, los residuos no siguen una distribución normal; por lo cual se requirió la inclusión de un impulso:  $\zeta_{02/20} = 1$  cuando  $t = 2020Q2$  y cero (0) en caso contrario. En definitiva, la forma del modelo ARIMAX (1,1,0) es la siguiente:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \gamma_1 \zeta_{02/20} \quad (2)$$

Se realizan otras especificaciones con la finalidad de evaluar si se obtiene una mejora en el pronóstico del desempleo, al incluir al modelo ARIMAX (1,1,0) los índices de *Google Trends*. En una primera especificación (MOD1) de este tipo, se considera como variable explicativa exógena el índice de búsqueda que se refiere al término trabajo ( $X_t^{trabajo}$ ). La expresión que define el modelo es la siguiente:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \gamma_1 \zeta_{02/20} + \theta_1 X_t^{trabajo} \quad (3)$$

Otra especificación considera el índice de búsqueda “Aldaba”, único portal de empleo entre los evaluados con una alta popularidad y vinculación fuerte con la serie de desempleo abierto. Siendo el segundo modelo alternativo (MOD2) de la siguiente forma:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \gamma_1 \zeta_{02/20} + \theta_2 X_t^{aldaba} \quad (4)$$

Una tercera especificación (MOD3) incluye la combinación de dos índices: el primero “liquidación” ( $X_t^{liquidacion}$ ), para aproximar la búsqueda de información

---

<sup>5</sup> Se requirió una primera diferenciación del logaritmo de la serie de desempleo abierto.

sobre el pago de las prestaciones sociales e indemnizaciones en caso de ser requerido, y un segundo término ( $X_t^{trabajo}$ ) que refiere a la búsqueda de empleo incluido en la primera especificación. De este modo, la expresión es la siguiente:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \gamma_1 \zeta_{02/20} + \theta_1 X_t^{trabajo} + \theta_2 X_t^{liquidacion} \quad (5)$$

Por último, utilizando Análisis de Componentes Principales (PCA, por sus siglas en inglés) una técnica de reducción de dimensionalidad desarrollada por [Hotelling \(1933\)](#), la data inicial de *Google Trends* (5 -dimensiones) se reduce a un subespacio de menor tamaño que captura la mayor varianza posible de los datos, de forma que las componentes son ortogonales, por lo que contribuye a solucionar el potencial problema de multicolinealidad. El objetivo de esta especificación (MOD4) es analizar cómo se comporta el pronóstico cuando se incluye la mayor cantidad de información, contenida en los indicadores adelantados, y no solo el resultado de un término de búsqueda que pudiese estar influenciado por diversas razones distintas a la necesidad de encontrar un empleo. Además, utilizar varios índices podría contribuir a mitigar que los resultados estén influenciados por la selección de un buen término ([Tuhkuri, 2015](#)).

Se calculan dos componentes que en total explican el 85.76 % de la varianza de los datos de Google. Las búsquedas de “empleo”, “liquidación” y “LinkedIn” se encuentran positivamente correlacionadas con el primer componente. En cambio, la búsqueda de “trabajo” y “Aldaba” correlaciona de forma negativa. Con respecto al segundo componente, todos los signos se mantienen. Sin embargo, el índice “aldaba” es el de mayor contribución. El modelo con PCA está dado por la siguiente expresión:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \gamma_1 \zeta_{02/20} + \theta_3 X_t^{PC1} + \theta_4 X_t^{PC2} \quad (6)$$

La relevancia de cada uno de los modelos alternativos presentados relativo al modelo *benchmark* es evaluada utilizando los criterios de información de [Akaike \(1974\)](#) y [Schwarz \(1978\)](#). Otro aspecto valorado es el desempeño fuera de muestra, empleando medidas de precisión de proyección como la raíz del error cuadrático medio (RMSE, por sus siglas en inglés) y el error porcentual absoluto medio (MAPE).

## 5. Resultados

Luego de especificar la forma funcional del modelo *benchmark* y los cuatro modelos con el índice de *Google Trends* se procede a analizar el desempeño dentro, y fuera de muestra en base a las métricas especificadas en la sección previa.

La estimación del mejor modelo base (MODO) se presentan en la tabla 1. La inclusión del impulso resulta ser oportuna, con un AIC y BIC inferior a modelos que no contemplan dicha intervención. La variable  $\zeta_{02/20}$  incluida para controlar por los efectos de la pandemia del COVID-19, indica que, durante el segundo trimestre de 2020, la tasa de desempleo abierto se redujo en aproximadamente la mitad de lo observado en el trimestre previo. Este movimiento se explica principalmente por la reducción del número de individuos que buscaron activamente un empleo, como consecuencia de las restricciones de movilidad.

Al comparar dentro de la muestra los resultados del MODO con las cuatro alternativas que incluyen los índices procedentes de Google Trends, se observa una reducción del criterio Akaike para los modelos MOD1 y MOD3 (ver tabla 1). Entre estas dos alternativas, el MOD1 resulta ser el modelo más adecuado dada la significancia de los estimadores y los criterios de información. En el caso de los restantes, MOD2 y MOD4, no se observa una mejoría significativa dentro de muestra respecto al modelo base.

**Tabla 1.**

Estimación de los resultados de los modelos ARIMAX para el desempleo de República Dominicana

	MODO	MOD1	MOD2	MOD3	MOD4
$y_{t-1}$	-0.3704*** (0.1137)	-0.3577*** (0.1158)	-0.3667*** (0.1151)	-0.3596*** (0.11172)	-0.3684*** (0.1165)
$X_t^{trabajo}$		-0.2886** (0.1384)		-0.2857** (0.1415)	
$X_t^{aldaba}$			0.0202 (0.0276)		
$X_t^{liquidacion}$				0.0015 (0.0148)	
$X_t^{PC1}$					0.0541** (0.0275)
$X_t^{PC2}$					0.0023 (0.0205)
$\zeta_{02/20}$	-0.7055*** (0.0784)	-0.6433*** (0.0814)	-0.7002*** (0.0784)	-0.6440*** (0.0816)	-0.6588*** (0.0797)

	MODO	MOD1	MOD2	MOD3	MOD4
Desviación estándar de las innovaciones	0.0094	0.0088	0.0093	0.0088	0.0088
Criterio Akaike (AIC)	-116.6739	-118.8766	-115.2071	-116.8871	-116.6257
Criterio Schwarz (BIC)	-110.0598	-110.0578	-106.3883	-105.8637	-105.6022

Nota: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Fuente: Elaboración de la autora.

Para evaluar cada modelo, en términos de su capacidad predictiva, se procedió a dividir la muestra en dos partes, 80 % para la estimación de los parámetros y 20 % para pruebas fuera de muestra. Debido a que la muestra de prueba resulta pequeña, para garantizar la robustez de los resultados se realizó la prueba de [Diebold y Mariano \(1995\)](#), que contrasta la precisión de dos métodos distintos de proyección<sup>6</sup>. La tabla 2 contiene las dos métricas principales, RMSE y MAPE, utilizadas para evaluar el desempeño fuera de muestra; así como el resultado de la prueba Diebold y Mariano (DM).

**Tabla 2.**

#### Medidas de evaluación de capacidad predictiva fuera de muestra

	MODO	MOD1	MOD2	MOD3	MOD4
RMSE	0.2859	0.2677	0.2843	0.2678	0.2749
MAPE	0.1488	0.1391	0.1479	0.1391	0.1484
Prueba DM (valor p)		0.0336	0.0348	0.0335	0.6093

Fuente: Elaboración de la autora.

De acuerdo con los indicadores de evaluación, la mejoría en el pronóstico de la tasa de desempleo abierto oscila entre 0.25 % y 6.54 %. Los modelos MOD1 y MOD3 presentan la mayor disminución del RMSE y MAPE, con una reducción de más del 6 % respecto al modelo base. En cuanto al MOD4, que incluye todos los términos de búsqueda seleccionados, la disminución del RMSE fue de 3.86 % mientras que el MAPE solo se redujo en 0.25 %.

De acuerdo con los resultados de la prueba DM, existe evidencia estadística de que solo tres de los modelos (MOD1, MOD2 y MOD3) que incluyen datos de *Google Trends* poseen una capacidad predictiva superior al modelo base.

Los resultados evidencian que con la inclusión de indicadores de *Google Trends*, el pronóstico de corto plazo del desempleo en República Dominicana mejora. Los hallazgos se alinean con las evidencias de investigaciones similares, aunque la ganancia es inferior relativo a países desarrollados. Esta dinámica

<sup>6</sup> Las hipótesis para una prueba de dos colas:  $H_0: E(L(e_1)) = E(L(e_2))$ ,  $H_1: E(L(e_1)) \neq E(L(e_2))$ .

resulta esperable debido, en primer lugar, a las características del mercado laboral dominicano, cuyos niveles de informalidad alcanzan en promedio un 57 %, y a que la proporción de individuos que usaban internet a finales de la primera década del siglo XXI no superaba el 50 % de la población, siendo hasta 2022 que más del 80 % de los dominicanos tenían acceso a la red.

## 6. Conclusión

El desempleo es un indicador relevante sobre el estado del mercado laboral y la actividad económica, por lo cual tener información en tiempo casi real, puede contribuir al proceso de toma de decisiones de los hacedores de política y el sector privado. En ese sentido, la presente investigación se nutre de datos procedentes de *Google Trends* sobre la popularidad de varios términos vinculados a la búsqueda de empleo, con la finalidad de obtener predicciones en tiempo quasi real y evaluar si mejora la capacidad predictiva de los modelos.

Al comparar el desempeño dentro y fuera de muestra de los modelos que incluyen los índices de *Google* con el de referencia, se obtiene evidencia de que la información contribuye a mejorar el pronóstico de desempleo. Los modelos que incluyen los índices referentes a la búsqueda de términos como “trabajo” o “liquidación” presentan un RMSE y MAPE significativamente menor al del modelo base. En el caso del modelo que incluye una combinación lineal de todos los términos contemplados, la mejora evidenciada es menor relativo a los modelos que incluyen solo un término, pero no es desdeñable. La prueba de Diebold y Mariano (1995) confirma los resultados, a excepción del modelo con PCA.

En conclusión, la información procedente de *Google Trends* contribuye a predecir de forma contemporánea la tasa de desempleo abierto para el caso de República Dominicana. En lo adelante, la agenda de investigación se enfocará en explorar nuevas metodologías de pronóstico como *Random Forest* o regresión *LASSO*, así como la incorporación de otras variables macroeconómicas de alta frecuencia. En paralelo, se exploraría cómo la inclusión de los índices de búsqueda de *Google Trends* mejoraría los pronósticos del empleo para varios grupos demográficos y macro regiones del país.

## 7. Bibliografía

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 716-723.
- Askitas, N., & Zimmermann, K. F. (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly* 55, 107-120.
- Borup, D., & Montes Schütte, E. C. (2020). In search of a job: Forecasting employment growth using Google Trends. *Journal of Business & Economic Statistics*, 186-200. Obtenido de <https://doi.org/10.1080/07350015.2020.1791133>
- Box, G. E., & Jenkins, G. M. (1976). Time Series Analysis: Forecasting and Control. *Holden-Day, S. Francisco*.
- Campos Vázquez, R. M., & López-Araiza, S. E. (2020). Grandes datos, google y desempleo. *Estudios Económicos*, 125-151.
- Chang Kcomt, J., & Del Río Lazo, 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*.
- Choi, H., & Varian, H. (2009a). Predicting the Present with Google Trends. *Google Inc.*
- Choi, H., & Varian, H. (2009b). Predicting Initial Claims for Unemployment Benefits. *Google Inc.*
- Choi, H., & Varian, H. (2012). Predicting the Present with Google Trends. *Google Inc.*
- Chow, G. C., & Lin, A.-I. (1971). Best Linear Unbiased Interpolation, Distribution, and Extrapolation of Time Series by Related Series. *The Review of Economics and Statistics, MIT Press*, 372-375.
- D'Amuri, F., & Marcucci, J. (2009). Google it! Forecasting the US unemployment rate with a Google job search index. *MPRA*.
- Dickey, D. A., & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association*, 427-431.
- Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 134-144.
- Ettredge, M., Gerdes, J., & Karuga, G. (2005). Using Web-based search data to predict macroeconomic statistics. *Communications of the Association for Computing Machinery*, 87:92.
- Eysenbach, G. (2006). The Open Access Advantage. *Journal of Medical Internet Research*, 1-6.
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 1012-1014.

- Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences*, 17486-17490.
- Google. (2024). *FAQ about Google Trends data*. Obtenido de Google Trends: [https://support.google.com/trends/answer/4365533?hl=en&ref\\_topic=6248052&sjid=1426693974411918838-NA](https://support.google.com/trends/answer/4365533?hl=en&ref_topic=6248052&sjid=1426693974411918838-NA)
- Hotelling, H. (1993). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 417-441.
- Johnson, H. A., Wagner, M. M., Hogan, W. R., Chapman, W., Olszewski, R. T., Dowling, J., & Barnas, G. (2004). Anlysis of Web Access Logs for Surveillance of Influenza. *IOS Press*, 1202-1206.
- Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 159-178.
- McLaren, N., & Shanbhogue, R. (2011). Using internet search data as economic indicators. *Bank of England Quarterly Bulletin*, 134-140.
- Mulero, R., & Garcia-Hiernaux, A. (2022). Forecasting unemployment with Google Trends: age, gender and digital divide. *Empirical Economics*, 587-605.
- Santana, L. (2017a). Nowcasting con Google Trends: Dinámica de la Actividad Económica Mensual, el Consumo Privado y la Inversión basada en datos de Google Trends y un Modelo Bayesiano Estructural de Series de Tiempo. *Banco Central de la República Dominicana*.
- Santana, L. (2017b). Nowcasting y Proyecciones de Corto y Mediano Plazo de la Dinámica del Sector Turístico de la República Dominicana: Un enfoque basado en Probabilidades de Google Trends y un Modelo Bayesiano Estructural de Series de Tiempo.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 461-464.
- Simionescu, M., & Cifuentes-Faura, J. (2022). Can unemployment forecasts based on Google Trends help government design better policies? An investigation based on Spain and Portugal. *Journal of Policy Modeling*, 1-21.
- Trespalacio Cárdenas, L. M. (2021). Modelo de nowcasting para pronostivas la tasa de desempleo de Colombia utilizando Google Trends.
- Tuhkuri, J. (2015). Big Data: Do Google Searches Predict Unemployment?
- Vicente, M. R., López-Menéndez, A. J., & Pérez, R. (2015). Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing? *Technological Forecasting and Social Change*, 132-139.

## 8. Anexos

**Tabla A1**

Estadísticas descriptivas de la tasa de desempleo e índices de Google Trends, 2004-2024

	Obs.	Media	Mediana	Desv. Estándar	Min	Max
Tasa de desempleo (%)	81	6.35	6.33	1.27	3.20	8.73
Empleo	81	50.36	44.00	16.231	26.67	97.33
Trabajo	81	61.46	57.00	12.814	42.00	91.00
Vacante	81	38.53	34.00	33.296	0.00	93.67
Oferta laboral	81	2.56	0.00	8.7111	0.00	54.00
Desempleo	81	11.35	11.67	6.5406	0.00	33.33
Cesantía	81	11.42	13.67	7.332	0.00	33.33
Liquidación	81	35.11	37.67	22.879	0.00	72.33
Preaviso	81	31.56	39.00	23.378	0.00	67.00
Indemnización laboral	81	1.21	0.00	5.6254	0.00	33.33
Prestación	81	53.72	61.33	26.248	0.00	93.00
Despido	81	21.65	28.00	13.691	0.00	41.33
Renuncia	81	14.01	15.00	7.041	0.00	46.67
Ministerio de Trabajo	81	26.57	31.33	20.461	0.00	70.33
Teletrabajo	81	8.64	0.00	17.339	0.00	71.00
Trabajo remoto	81	7.75	0.00	19.773	0.00	90.00
Aldaba	81	50.82	49.33	24.719	0.00	90.33
LinkedIn	81	37.02	51.00	27.985	0.00	79.00
Tu Empleo RD	81	21.98	16.33	25.138	0.00	98.33
Corotos empleo	81	10.42	0.00	13.744	0.00	55.67
Chiripas	81	23.67	26.00	24.034	0.00	77.33
Jooble	81	14.86	0.00	23.685	0.00	93.33
RD Trabaja	81	0.74	0.00	6.7037	0.00	60.33
EmpleateYa	81	14.87	4.33	19.968	0.00	77.00
Computrabajo	81	5.67	2.67	13.013	0.00	87.67
Tecoloco	81	7.97	0.00	17.76	0.00	85.00
BuscoJobs	81	0.84	0.00	5.9396	0.00	51.00
Opcionempleo	81	3.47	0.00	8.7633	0.00	36.67
Concursa Map	81	8.81	0.00	17.859	0.00	76.67

Fuente: Elaborado con datos del BCRD y *Google Trends*.

**Tabla A2**

*P*-values de las pruebas ADF y KPSS del logaritmo de desempleo ( $\ln Y_t$ )

	Periodo muestral					
	2004:01-2024:01		2004:01-2020:04		2021:01-2024:01	
	ADF	KPSS	ADF	KPSS	ADF	KPSS
$\ln Y_t$	0.4968	0.1000	0.6667	0.1000	0.9457	0.0636
$(1-L)\ln Y_t$	0.0100	0.1000	0.0100	0.1000	0.0158	0.1000

Fuente: Elaboración de la autora.

**Tabla A3**

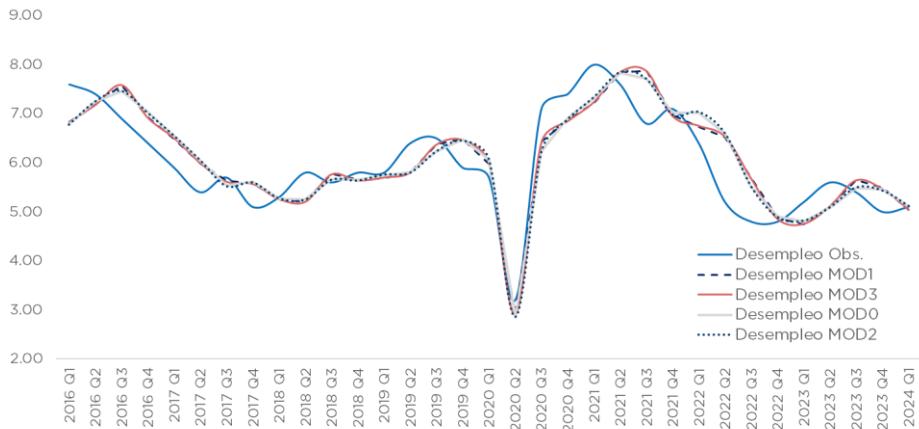
Resultado de la prueba de *Ljung-Box* a los residuales

	MODO	MOD1	MOD2	MOD3	MOD4
P-value	0.4018	0.5818	0.2640	0.4293	0.2169

Fuente: Elaboración de la autora.

**Gráfica A1**

Resultados del modelo de referencia y alternativas con datos de *Google Trends*, T1 2016 – T1 2024.



Fuente: Elaboración de la autora.