Documentos técnicos de hábitat
No. 10 – julio 2023

Uso de redes neuronales en el pronóstico del PIB del subsector edificador en la ciudad de Bogotá

BOGOTÁ EL MEJOR HOGAR
ALCALDÍA MAYOR DE BOGOTÁ D.C.
SECRETARÍA DEL HÁBITAT
BOGOTÁ

Subsecretaría de Planeación y Política
Jaime Andrés Florez
Subsecretario

Subdirección de Información Sectorial
María Paula Salcedo Porras
Subdirectora

Equipo técnico - Subdirección de Información Sectorial
Daniela Sedano Saenz
Cristian Torres[1]

[1] maria.salcedo@habitatbogota.gov.co , cristian.torres@habitatbogota.gov.co, daniela.sedano@habitatbogota.gov.co.

Contenido

1. Objetivo

Mejorar el método de estimación de pronósticos sobre el Producto Interno Bruto-PIB del subsector edificador en la ciudad de Bogotá mediante la inclusión de la metodología de redes neuronales (McCulloch y Pitts, 1943), con lo cual se busca robustecer la herramienta desde dos frentes, incrementando el inventario de modelos usados y como fuente de perspectivas de corto plazo sobre los resultados de crecimiento o caída anual del valor agregado de la subrama, aprovechando la oferta de indicadores de actividad que posee el gremio edificador.

Objetivos específicos

1. Evaluar los enfoques utilizados para pronosticar series de tiempo con base en redes neuronales.

2. Evaluar el aporte de los modelos escogidos en el paso uno sobre el actual enfoque de estimación.

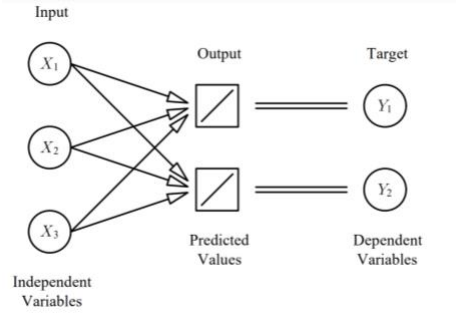
3. Generar un modelo de clasificación que permita evaluar a corto plazo las probabilidades de caída o crecimiento anual del valor agregado de edificaciones, dado los resultados de los indicadores de actividad al corte del requerimiento del cálculo.

2. Introducción

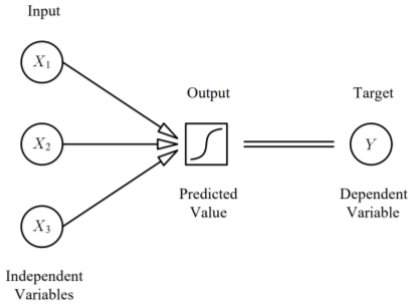
Una Red Neuronal Artificial (RNA) es un modelo matemático que tiene la intención de emular el funcionamiento biológico de las neuronas, imitando el proceso de aprendizaje y generalización, facilitando la automatización de reglas de aplicación y decisión. Adicionalmente, (Cybenko 1989; Funahashi 1989; Hornik, Stinchcombe and White 1989) demuestran que las RNA son una aproximación funcional universal de todas las especiaciones continuas con cualquier nivel de precisión requerido ***Ilustración 1***. De lo anterior se tienen dos importantes consecuencias: el proceso de estimación no requiere predecir la forma del modelo y su eficacia depende de la calidad y volumen de la base de entrenamiento. Las RNA en estadística pueden ser clasificadas como una familia flexible de modelos de regresión, discriminación y reducción de dimensiones o una clase de sistemas dinámicos no lineales compuestos por la interconexión de elementos básicos denominados *"neuronas"*, las cuales buscan recrear de manera básica el proceso de sinapsis[2], recreación que adquiere altísimos niveles de complejidad gracias a los procesos de adaptabilidad, autoorganización y aprendizaje que logran, estas características junto a la presencia de altos volúmenes de datos le han permitido abrirse campo en la práctica.

Ilustración 1 Equivalencia entre la RNA y los modelos estadísticos más comunes

Panel 1. Modelo de regresión lineal múltiple

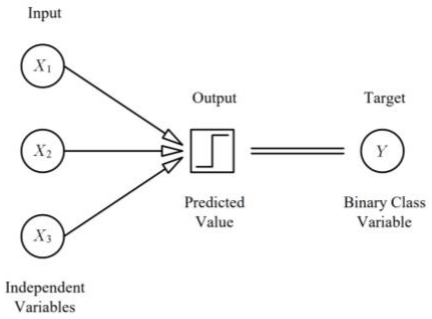


Panel 2. Modelo de regresión logística

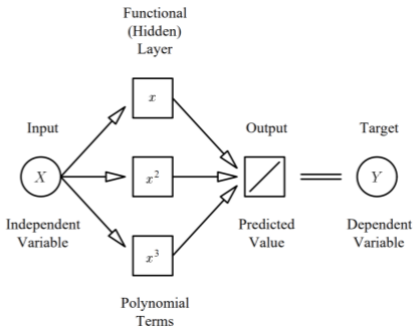


[2] La sinapsis es una aproximación especializada entre neuronas, ya sea entre dos neuronas de asociación, una neurona y una célula receptora, o entre una neurona y una célula efectora. En estos contactos se lleva a cabo la transmisión del impulso nervioso. Este se inicia con una descarga química que origina una corriente eléctrica en la membrana de la célula emisora (denominada presináptica); una vez que este impulso nervioso alcanza el extremo del axón, la conexión es la encargada de excitar o inhibir la acción de otra célula llamada célula receptora (denominada postsináptica).

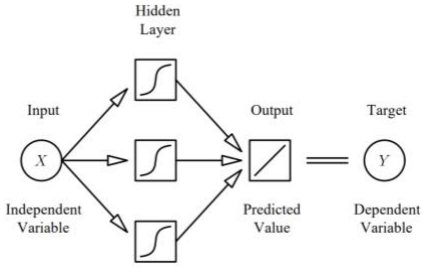
Panel 3. Función de discriminante lineal



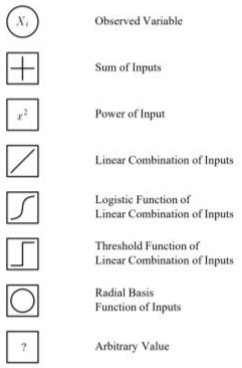
Panel 4. Modelo de regresión polinomial



Panel 5. Modelo de regresión no lineal simple



Panel 6. Convenciones



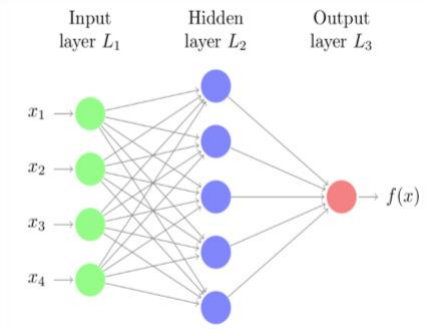
Fuente: (Gibson, 2017). Elaboración SDHT-SIS.

Este enfoque plantea analogía entre los estímulos recibidos por el cerebro y las variables independientes, donde el impacto de cada impulso es determinado por el proceso de entrenamiento, el cual consiste en enfrentar la capacidad de adaptación y autoorganización de la RNA a una base donde el dato a estimar ya tiene una realización, con lo cual es posible contrastar y calibrar mediante medidas basadas en el error. La ***Ilustración 2*** muestra la forma más común de representar una RNA, en esta la primera sección de la red neuronal (color verde) se conoce como capa de entrada y recibe los datos en bruto, es decir, el valor de los predictores. La parte intermedia (color azul), conocida como capa oculta, recibe los valores de la capa de entrada, ponderados por los pesos (flechas grises), finalmente se tiene la denominada capa de salida, la cual combina los valores que salen de la capa intermedia y genera la predicción. Dentro de la unidad fundamental de la RNA solo ocurren dos operaciones: la suma ponderada de sus entradas y la aplicación de una función de activación. En la primera parte, se multiplica cada valor de entrada $x_i$por su peso asociado w_i y se suman junto con el sesgo. A continuación, este valor se

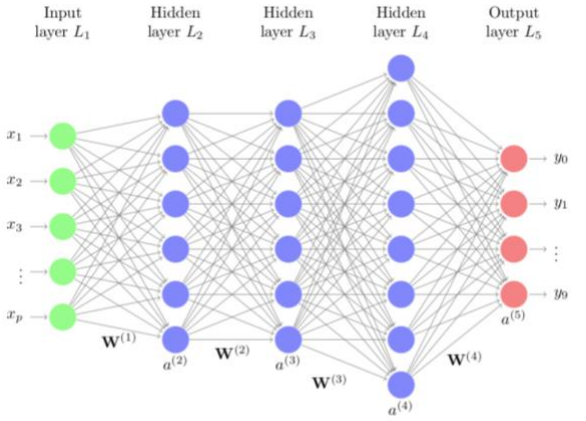
por la función de activación, que transforma el valor neto de entrada en un valor de salida. Si bien el valor que llega a la neurona siempre es una combinación lineal, gracias a la función de activación, se pueden generar salidas muy diversas representando el potencial de los modelos de redes para aprender relaciones no lineales.

Ilustración 2 Representación de una red neuronal mediante el concepto de capas

Panel 1. Modelo de una capa oculta

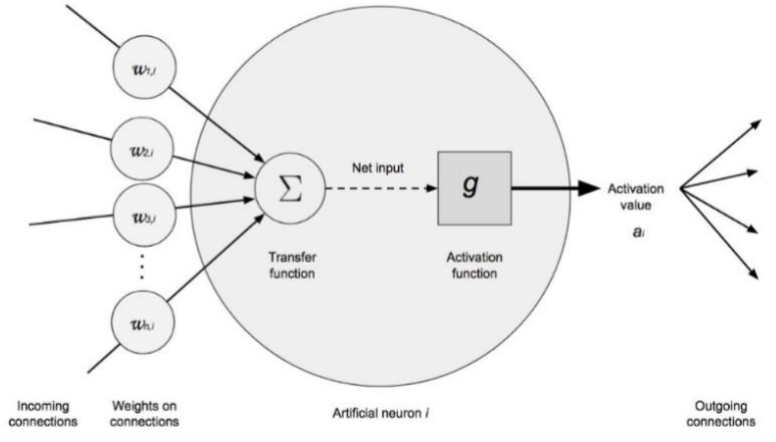


Panel 2. Modelo de varias capas ocultas



Fuente: (Hastie, 2021)

Ilustración 3 Representación del funcionamiento de una neurona en una RNA



Fuente: (Gibson, 2017)

Función de activación

La función de activación controla en gran medida qué información se propaga desde una capa a la siguiente. Esta convierte el valor neto de entrada a la neuronal (combinación de los input, pesos y sesgo) en un nuevo valor. Gracias a combinar funciones de activación no lineales con múltiples capas ocultas, los modelos de redes son capaces de aprender relaciones no lineales. La gran mayoría de convierten el valor de entrada neto de la neurona en un valor dentro del rango (0, 1) o (-1, 1). Cuando el valor de activación de una neurona es cero, se dice que la neurona está inactiva, ya que no pasa ningún tipo de información a las siguientes neuronas. A continuación, se describen las funciones de activación más empleadas.

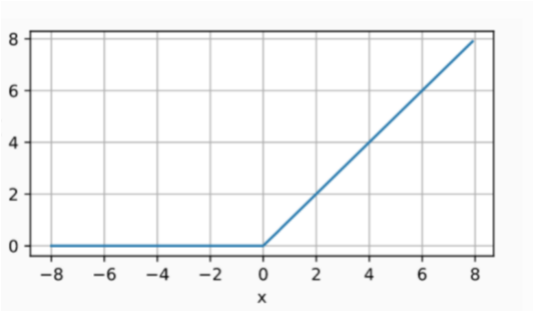
1. **Rectified linear unit (ReLU):** La función de activación ReLu aplica una transformación no lineal que activa la neurona solo si el input está por encima de cero. De tal manera que $ReLU(x) = \max(x, 0)$. ReLU es con diferencia la función más empleada por sus buenos resultados en aplicaciones diversas. La razón de esto reside en el comportamiento de su derivada, que es cero o constante. Gracias a esto se evita el problema de *vanishing gradients* que limita la capacidad de aprendizaje de los modelos de redes.

2. **Sigmoide:** La función sigmoide transforma valores en el rango de $(-\infty, \infty)$ a valores en el rango $(0,1)$ Aunque la función de activación sigmoide se utilizó mucho en los inicios de los modelos de redes, en la actualidad, suele preferirse la función ReLU. Un caso en el que la función de activación sigmoide sigue siendo la función utilizada por defecto es en las neuronas de la capa de salida de los modelos de clasificación binaria, ya que su salida puede interpretarse como probabilidad.

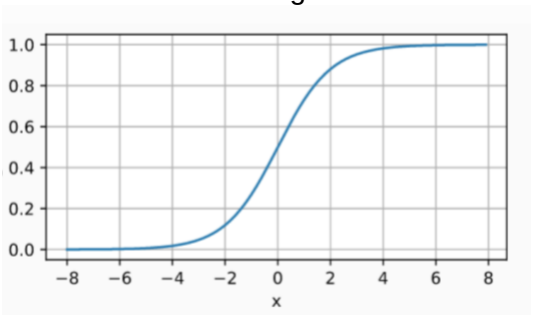
3. **Tangente hiperbólica (Tanh):** La función de activación Tanh, se comporta de forma similar a la función sigmoide, pero su salida está acotada en el rango (-1, 1).

Ilustración 4 Representación gráfica de las funciones de activación

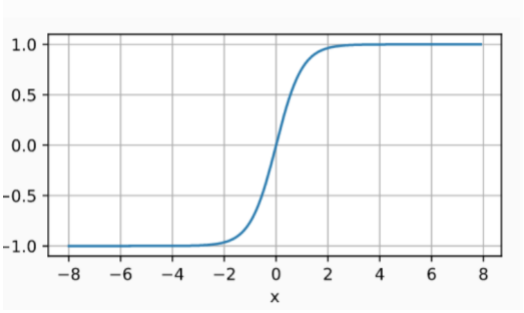
Panel 1. Rectified linear unit



Panel 2. Sigmoide



Panel 3. Tangente hiperbólica



Panel 4. Formas funcionales

ReLU: $\max(x, 0)$

Sigmoide: $\frac{1}{1+e^{x}}$

Tangente hiperbólica: $\frac{1-e^{-2x}}{1+e^{-2x}}$

Fuente: (Amat, 2021).

Función de costo

La función de coste, también llamada función de pérdida es la encargada de cuantificar la distancia entre el valor real y el valor predicho por la red, en otras palabras, mide cuánto se equivoca al realizar predicciones. Cuanto más próximo a cero es el valor de coste, mejor son las predicciones de la red (menor error), siendo cero cuando las predicciones se corresponden exactamente con el valor real. La función de coste puede calcularse para una única observación o para un conjunto de datos. Es el segundo caso el que se utiliza para dirigir el entrenamiento de los modelos. Dependiendo del tipo de problema, regresión o clasificación, es necesario utilizar una función de coste u otra. En problemas de regresión, las más utilizadas son el error cuadrático medio y el error absoluto medio. En problemas de clasificación suele emplearse la función log-loss.

1. **Error cuadrático medio:** el error cuadrático medio es la función de coste más utilizada en problemas de regresión. Para una determinada observación i, el error cuadrático se calcula como la diferencia al cuadrado entre el valor predicho $\hat{y}^{(i)}$ y el valor real $y^{(i)}$. Para cuantificar el error que comete el modelo todo un conjunto de datos, es común promediar el error de todas las N observaciones $L(w, b) = \frac{1}{n}\sum(\hat{y}^{(i)} - y^{(i)})^2$.

2. **Error medio absoluto:** el error medio absoluto es más robusto frente a atípicos que el error cuadrático medio. Cuando un modelo se entrena utilizando el error absoluto medio como función de coste, está aprendiendo a predecir la mediana de la variable respuesta $L(w,b) = \frac{1}{n}\sum|\hat{y}^{(i)} - y^{(i)}|$.
3. **Log loss, logistic loss o cross-entropy:** en problemas de clasificación, la red devuelve una serie de valores que pueden interpretarse como la probabilidad de que la observación predicha pertenezca a cada una de las posibles clases. Para problemas de clasificación con más de dos clases, esta fórmula se generaliza a: $L_{log}(Y,P) = -\frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=1}^{k-1} y_{i,j}\log(p_{i,j})$.

Para una serie de tiempo un modelo con p nodos de entradas, h nodos en la capa oculta y un nodo de salida puede ser modelado con una RNA p con variables de entrada y relacionarla con una variable de salida:

$$y_t = \alpha_0 + \sum_{j=1}^{h} \alpha_j\, G\left[\beta_{0j} + \sum_{i=1}^{p} \beta_{ij} y_{t-i}\right] + \epsilon_t$$

Donde:

y_t valor observado de la serie de tiempo en el instante t. Representa el valor para la variable de salida de la RNA.

y_{t-i} son los valores rezagados de la serie de tiempo en el instante t. Representan los valores de las p variables de entrada en la RNA.

α_j representan los pesos de la capa oculta a la capa de salida. Siendo el respectivo sesgo.

β_{ij} son coeficientes que representan los pesos de la capa de entrada a la capa oculta. Siendo el respectivo sesgo.

G representa la función de activación o transferencia de la capa de entrada, la cual determina la salida de la capa oculta. Las funciones de activación que generalmente son usadas son la función logística para la capa de entrada y la función identidad para la capa de salida.

h número de neuronas en la capa oculta.

p número de neuronas en la capa de entrada. Su valor determina el número de rezagos con que se analizará la serie de tiempo.

ϵ_t representa los errores aleatorios de modelo, los cuales se asumen que son independientes e idénticamente distribuidos con media cero y variancia constante.

Uso en la actividad económica

La aplicación realizada por (Sáenz, 2009) sobre el PIB en Colombia muestra relaciones no lineales en su proceso generador de datos, las cuales logran ser capturadas de manera exitosa por la RNA. Sin embargo, bajo este enfoque no es posible identificar la fuente de las relaciones no lineales. Los resultados muestran que el modelo ampliado incluyendo la tasa de interés de los CDT a 90 días mejora los pronósticos en por lo menos dos periodos[3], mostrando que esta variable resulta ser relevante para el comportamiento del PIB. Adicionalmente se demuestra la eficacia de la metodología Rolling al evaluar el desempeño de los modelos a la hora de pronosticar. Sin embargo, esfuerzos como los de (Iffat A. Gheyas, 2009) dan evidencia de que esta superioridad solo es significativa en el corto plazo.

3. Descripción del subsector

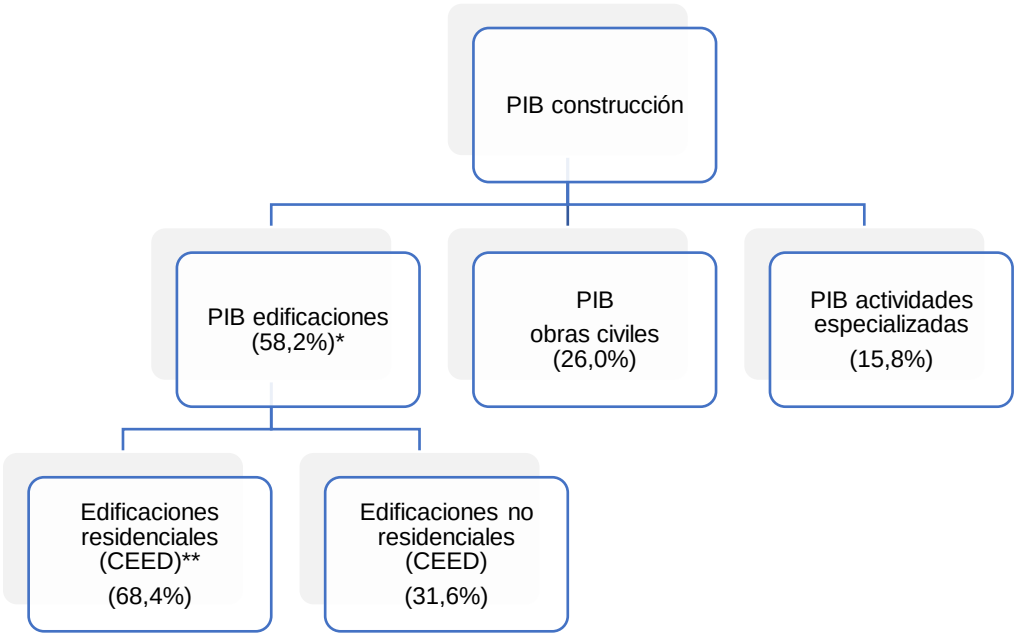
El subsector de edificaciones hace parte de la rama de la construcción, representando en promedio el 58,2% de su valor agregado. Asimismo, el subsector se desagrega en residencial,[4] que genera el 68,4% de la actividad y en no residencial[5] que aporta el 31,6% restante (ver ***Ilustración 5***). De esta tendencia se esperaría una gran influencia de los m^2 iniciados en vivienda sobre el valor agregado de edificaciones, sin embargo, se debe tener en cuenta el proceso de generación del valor agregado, ya que al usar el método de causación rezagada se minimiza el impacto de las iniciaciones en el momento exacto en que son medidas. Para entender esto se presenta el esquema de la ***Ilustración 6***, donde se destaca que solo el 10% del área en la primera etapa entraría en el proceso de causación del trimestre, etapa en que muy probablemente se encuentre el área iniciada; adicionalmente, el 55,2% del área iniciada en el 2021 lo hizo en el segundo semestre, lo cual, retrasa más su entrada al cálculo del valor agregado.

[3] Sin embargo, para realizar pronósticos con la RNA aumentada se necesitará construir un modelo para la tasa de interés de los CDT a 90 días, debido a que esta variable es exógena en el modelo; de lo contrario, no es posible realizar pronósticos del PIB con la red neuronal aumentada, pues no contamos con información sobre la tasa de interés en el futuro.

[4] El cual se puede desagregar entre vivienda y apartamentos.

[5] El que se puede desagregar entre: oficinas, comercio, bodegas, educación, hoteles, hospitales, administración pública y otros.

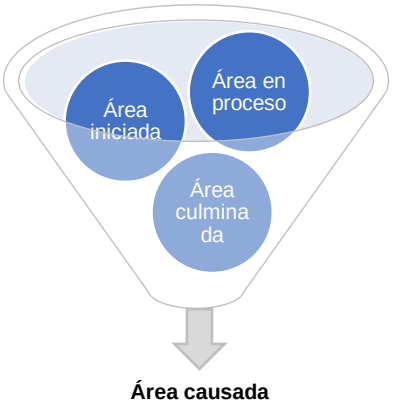
Ilustración 5. Estructura del PIB de la construcción y su respectiva distribución en Bogotá



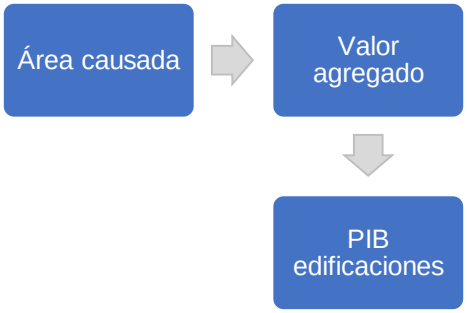
Fuente: DANE, CEED IV trimestre de 2021 y Cuentas Nacionales-PIB Bogotá III trimestre de 2021. Construcción SIS-SDHT.
* Se toma el promedio del peso anual de cada subsector en el valor agregado a precios corrientes de desde 2017 al III trimestre de 2021.
** Se toma el promedio trimestral del peso dentro del área causada en cada uno de los subsectores.

Ilustración 6. Esquema de generación del valor agregado en el subsector de edificaciones

Panel 1. Generación del área causada



Panel 2. Generación del PIB a partir del área causada



Panel 3. Proceso de causación, coeficientes de incidencia (%), por destino, según capitulo

Capitulo constructivo	Destino		
	Grupo 1 [1]	**Grupo 2** [2]	**Grupo 3** [3]
Preliminares excavación y cimentación	10	13	26
Estructura y cubierta	20	27	36
Mampostería, pañetes e impermeabilizantes	15	10	8
Acabados 1 [4]	45	43	24
Acabados 2 [5]	9	6	5
Acabados 3 [6]	1	1	1

Fuente: DANE, documento de diseño metodológico del PIB Bogotá, *https://www.dane.gov.co/index.php/estadisticas-por-tema/cuentas-nacionales/cuentas-nacionales-departamentales/cuentas-nacionales-departamentales-pib-trimestral-bogota-d-c.* Construcción SIS-SDHT.

[1] Vivienda unifamiliar y multifamiliar.

[2] Oficinas, locales, centros comerciales, centros de salud, hospitales, sedes institucionales y similares.

[3] Instalaciones industriales y bodegas.

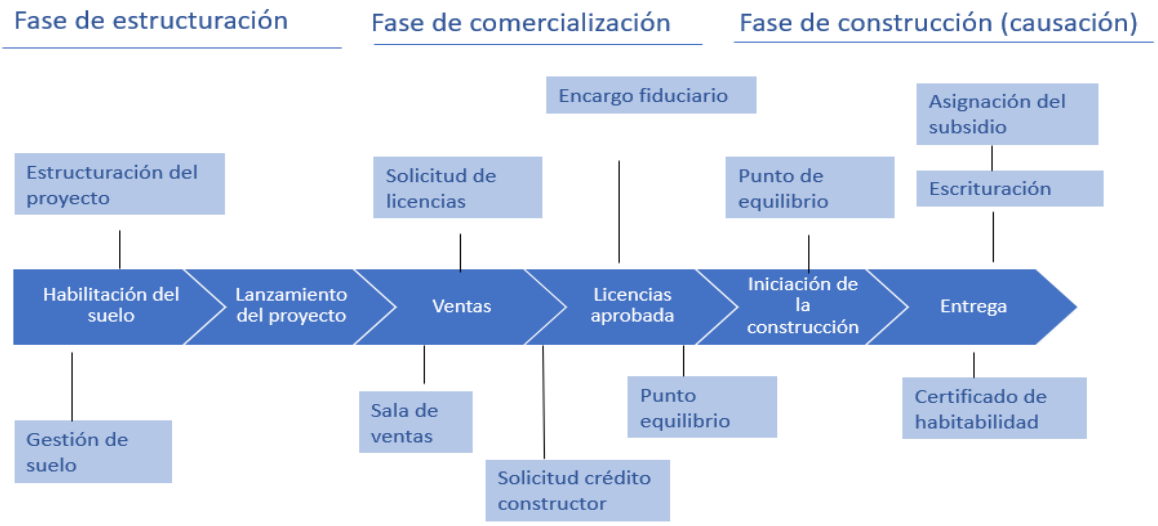
[4] Carpintería, metálica y de madera, pisos, enchapes, recubrimientos de muros y cielo raso.

[5] Pintura, instalación de equipos y alfombras, vidrios y espejos, instalación de apliques.

[6] Remates, aseo y limpieza.

Antes de analizar las relacionales temporales o rezagadas entre las variables es necesario presentar un esquema del funcionamiento teórico de un proyecto constructivo, a partir del cual se identifica como candidatas a predictores rezagados de la actividad a las ventas, el licenciamiento y las iniciaciones (ver ***Ilustración 7***). Adicionalmente, se puede apreciar que las iniciaciones son solo una pequeña parte del proceso de causación de cada trimestre, lo cual, permite ir dimensionando la magnitud del impacto y su papel como potencial futuro de causación y no fuente directa del valor agregado.

Ilustración 7. Fases del proceso constructivo del sector de edificaciones



Fuente: Ministerio de Vivienda, Ciudad y Territorio – MVCT, documento de coyuntura económica "Valor agregado de la construcción de edificaciones: edificando el futuro postpandemia" y Secretaría Distrital del Hábitat.

4. Aplicación

En esta seccipon se busca seleccionar el 10% de las especificaciones de RNA mejor ranqueadas, según su capacidad de predicción dentro de muestra, estos instrumentos vendrían a complementar los modelos clásicos y bayesianos ya utilizados por la SDHT en las perspectivas sectoriales, las cuales son usados en la toma de decisiones y la construcción del marco fiscal de mediano plazo de la Secretaría Distrital de Hacienda (SDH). Esta parte del trabajo incluye una mejora en los procesos de estimación de los modelos clásicos y bayesianos tanto univariados como multivariados, la cual consiste en generar estimaciones iteradas excluyendo paso a paso segmentos de los datos, esto con el fin de incrementar la estabilidad por modelo a través del uso de estadísticas de tendencia central como el promedio o la mediana y a la vez tener en cuenta la influencia de datos atípicos muy alejados en el tiempo (Castano & Melo, 1998). Para cada enfoque se excluyen aquellas especiaciones que no superan las pruebas sobre sus supuestos básicos.

Ilustración 8. Ejemplo cálculo modelo VAR(1) de forma iterativa

Matriz de datos hasta eliminar k datos.	Especificación	Pronóstico
$\begin{bmatrix} x_{11} & \cdots & x_{p1} \\ \vdots & \ddots & \vdots \\ x_{1n} & \cdots & x_{pn} \end{bmatrix}$	$X_t = A_0 + A_1 Y_{t-1} + \varepsilon_t$	$[X_{1,t+1} \quad \dots\dots \quad X_{1,t+10}]$
$\begin{bmatrix} x_{12} & \cdots & x_{p2} \\ \vdots & \ddots & \vdots \\ x_{1n} & \cdots & x_{pn} \end{bmatrix}$	$X_t = A_0 + A_1 Y_{t-1} + \varepsilon_t$	$[X_{2,t+1} \quad \dots\dots \quad X_{2,t+10}]$
. . .	. . .	. . .
$\begin{bmatrix} x_{1k} & \cdots & x_{pk} \\ \vdots & \ddots & \vdots \\ x_{1n} & \cdots & x_{pn} \end{bmatrix}$	$X_t = A_0 + A_1 Y_{t-1} + \varepsilon_t$	$[X_{k,t+1} \quad \dots\dots \quad X_{k,t+10}]$
		$\frac{1}{k}\sum_{i=1}^{k} [X_{i,t+1} \quad \dots\dots \quad X_{i,t+10}]$

Fuente: SDHT-SIS.

El proceso plasmado en la ***Ilustración 8*** se realiza para cada una de las m especificaciones utilizadas previamente, de esta manera se tendrían m estimaciones de la trayectoria del PIB, información que puede ser resumida a través de una medida de tendencia central o un proceso de reducción de dimensiones como un Análisis de Componentes Principales-ACP (WICHERN, 2007), contar con este abanico de posibilidades facilita la generación de escenarios posibles dados la historia del sector.

Modelo de clasificación

La idea en esta parte del documento es combinar la metodología de RNA y la oferta de indicadores sectoriales de actividad, que por su periodicidad y velocidad de publicación permiten dilucidar el posible resultado del valor agregado en el corto plazo, permitiendo complementar los pronósticos realizados en este periodo. El primer paso para esto es construir una variable de tipo multinomial con base en la variación anual del valor agregado con base en los siguientes criterios:

1 para variaciones entre el mínimo registrado + un 3%.
2 Rango 1 + 3%.
3 Rango 2 + 3%.

N Máximo -3%.

Esto permitiría ver a corto plazo la senda de variación más probable del valor agregado, de tal manera que el modelo a estimar vendría dado por la siguiente expresión

5. Resultados

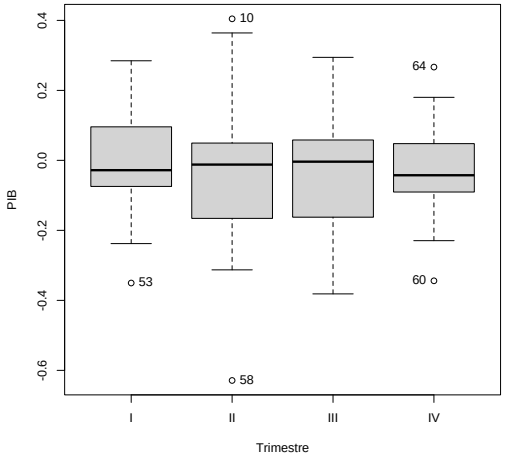
A continuación, se enlistan los principales resultados de estimación por metodología, es importante recordar que estos tienen un carácter temporal debido a cambios en los datos por actualizaciones metodológicas o ajustes hechos por el DANE, adicionalmente las mejoras en las actuales metodologías tantos de estimación como de comparación pueden generar cambios en los resultados, por lo cual se recomienda actualizar este trabajo de manera continua y espacios de dos años.

Modelos de una sola variable

Antes de estimar los modelos de una variable es necesario remover los datos atípicos[6], para esto se combinan técnicas gráficas y analíticas, en el panel 1 de la ***Ilustración 9*** encontramos que el primer trimestre de 2019, el segundo de 2008, el segundo de 2020 y el cuarto de 2019 y 2020 son sospechosos de ser atípicos, adicionalmente, las pruebas analíticas arrojan que el segundo trimestre de 2021 es candidato a atípico. Estos puntos son remplazados por la mediana de cada trimestre.

Ilustración 9 Resultados análisis atípicos PIB de edificaciones en Bogotá

Panel 1. Diagrama de caja del PIB según trimestre

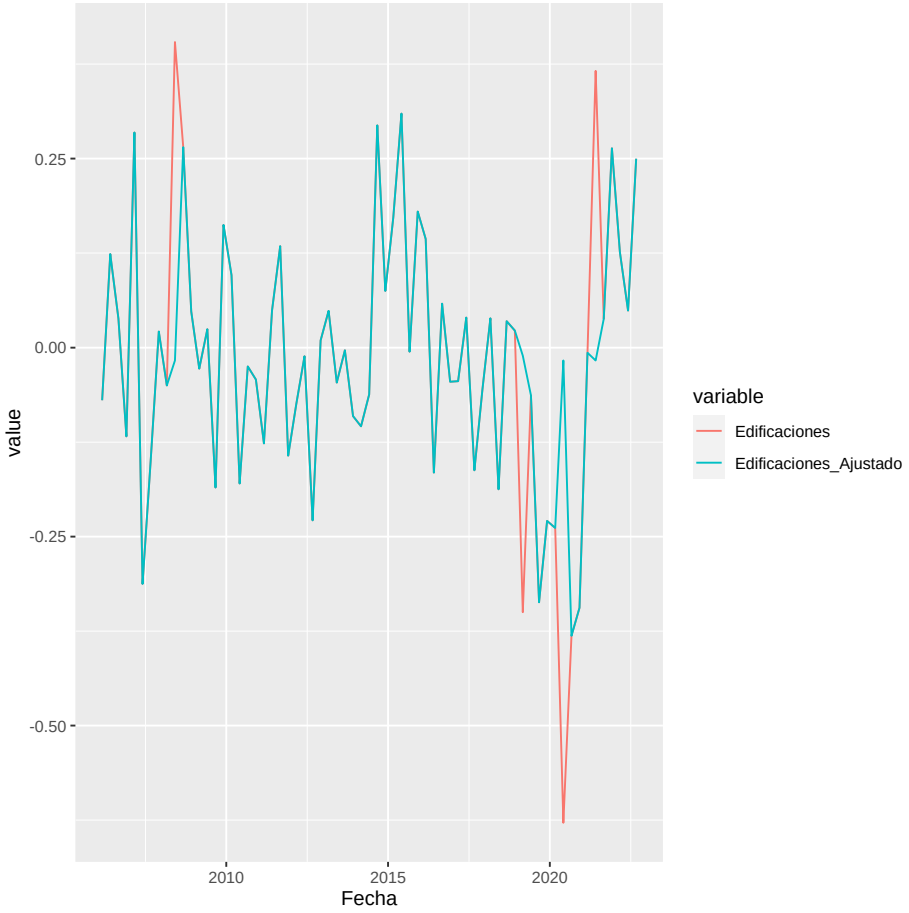


Panel 2. Resultados pruebas analíticas.

Prueba sobre el mínimo de la serie	Se rechaza la hipótesis nula de que el segundo trimestre de 2020 no es un atípico.
Prueba sobre el máximo de la serie	Se rechaza la hipótesis nula de que el segundo trimestre de 2008 no es un atípico.
Prueba general	Se rechaza la hipótesis nula de que el segundo trimestre de 2021 no es un atípico.

[6] Se considera relevante ya que los modelos univariados sirven para capturar la tendencia central de la serie, para los los modelos multivariados no se hace este paso, ya que al ser un modelo auto explicado es posible capturar caídas y subidas muy fuertes por encima del atractor de la serie.

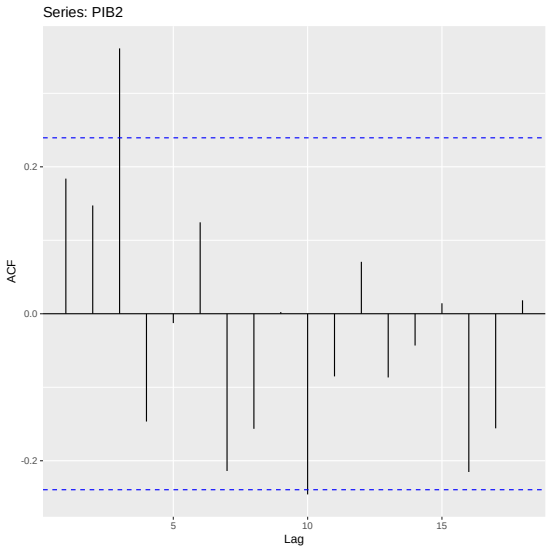
Panel 3. Resumen proceso de imputación de atípicos



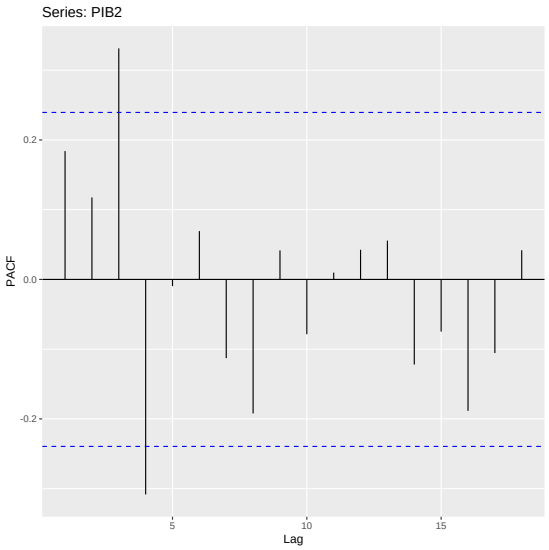
Fuente: SDHT-SIS.

Ilustración 10 Gráficos de resumen del PIB de edificaciones de Bogotá

Panel 1. ACF PIB



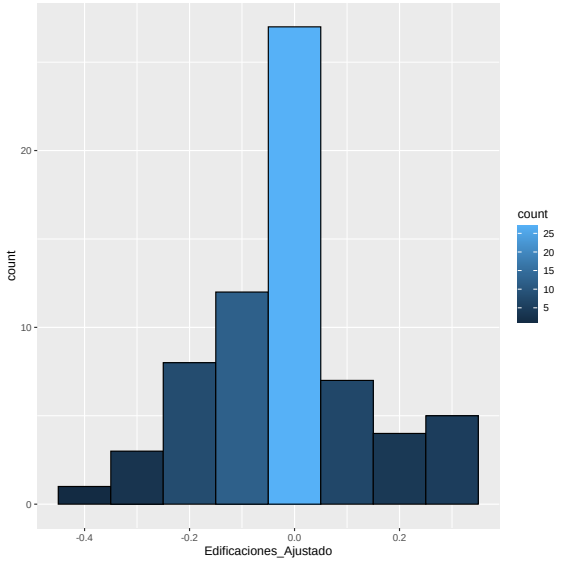
Panel 2. PACF PIB



Panel 4. Estadísticas de resumen

Estadístico	Valor
Promedio	-0.01510837
Mediana	-0.0172703
Máximo	0.3094524
Mínimo	-0.3812816
Desviación	0.1437239
Coeficiente de variación	-9.512867

Panel 4. Histograma PIB ajustado



Fuente: SDHT-SIS.

Para la primera parte se toma los modelos arrojados por la función “auto.arima” de R, la cual utiliza una combinación de pruebas de raíz unitaria, procesos de minimización de AIC y MLE para determinar probables modelos óptimos tipo ARIMA para la serie. Con el fin de garantizar la pertinencia de usar el marco metodológico Box-Jenkins se realizan pruebas de raíz unitaria tipo Dickey-Fuller, las cuales confirman que la serie de crecimiento anual del PIB sigue un comportamiento

estacionario. A continuación, se enlistan los modelos mejor punteados en este segmento.

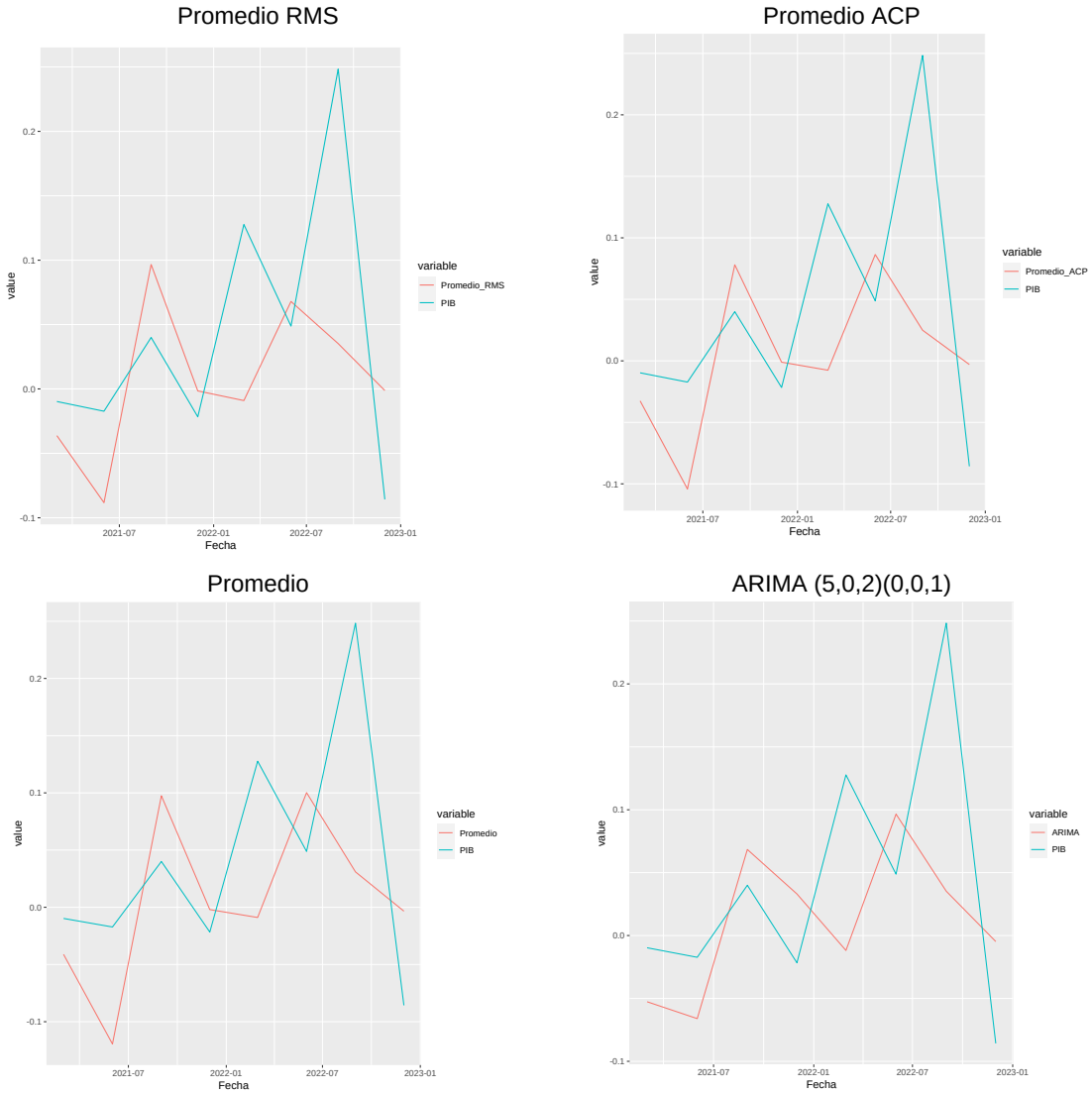
Ilustración 11 Resultados mejores modelos según desempeño dentro de muestra

Modelo	RMSE	MAE	Tasa de falla	Promedio
Promedio RMS[1]	55,5	80,0	33,3	56,3
Promedio ACP[1]	59,3	82,5	33,3	58,4
Promedio[1]	62,0	89,0	33,3	61,4

Fuente: SDHT-SIS.

[1] Se toman los modelos Arima (5,0,2)(0,0,1), Arima (5,0,4)(0,0,1), Arima (5,0,5)(0,0,1), Arima (5,0,3)(0,0,1), Arima (2,0,2)(0,0,1) y Arima (1,0,2)(0,0,1)

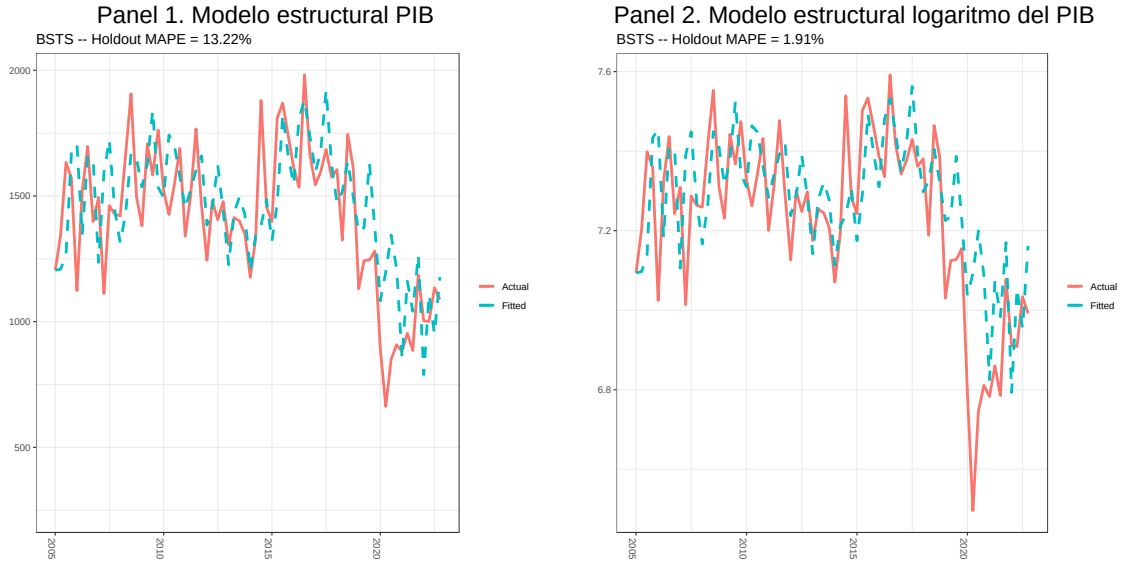
Ilustración 12 Resultados esquema ARIMA pronóstico dentro de muestra 8 pasos adelante



Fuente: SDHT-SIS.

En cuanto a la parte bayesiana se toman dos enfoques el primero es estimar dos modelos estructurales sobre la base original[7] y su logaritmo natural, para la segunda parte se toman los modelos ARIMA seleccionados en el apartado anterior y se estima su contraparte bayesiana. Lo cual arroja como los mejores candidatos a los modelos ARIMA (1,0,2)(0,01), ARIMA (2,0,2)(0,01) y ARIMA (5,0,2)(0,01). Sin embargo, Estos modelos parecen subestimar la capacidad de recuperación del subsector en la ciudad.

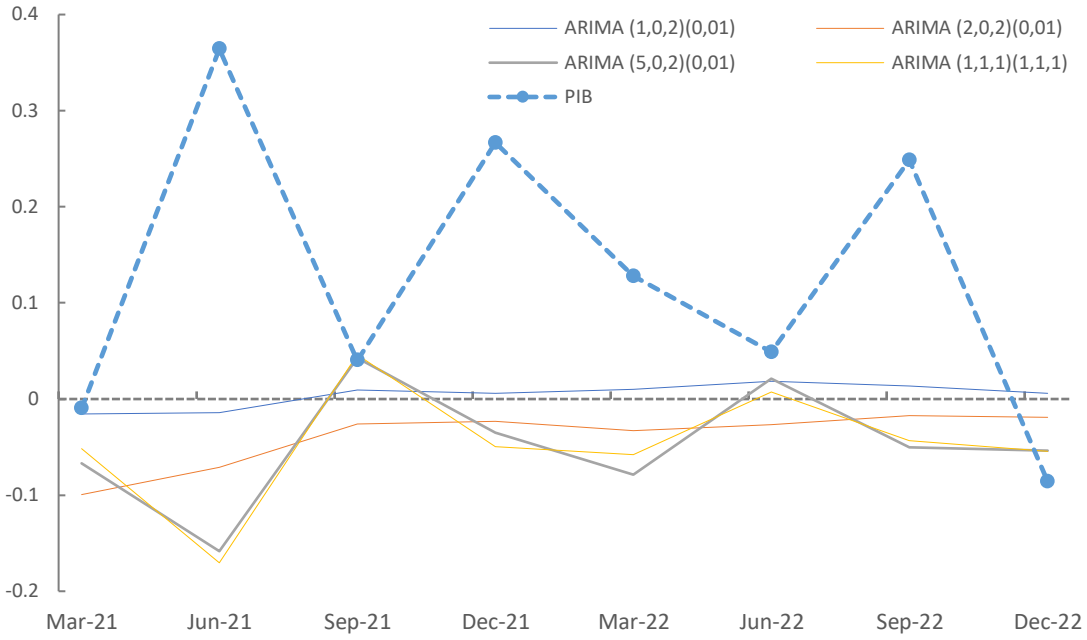
Ilustración 13 Resultados modelos estructurales



Fuente: SDHT-SIS.

[7] Esto implica remplazar los atípicos ya identificados.

Ilustración 14 Modelos bayesianos seleccionados



Fuente: SDHT-SIS.

Para resumir en esta parte se seleccionan las siguientes especificaciones clásicas Promedio RMS, ACP[1], Promedio[1] y desde el enfoque bayesiano se seleccionan ARIMA (1,0,2)(0,01), ARIMA (2,0,2)(0,01) y ARIMA (5,0,2)(0,01).

Esquema multivariado

A partir de lo expuesto en la Descripción del subsector se identifica como candidatas a predictores de la actividad a las ventas, el área causada, el licenciamiento y las iniciaciones (ver ***Ilustración 7***). Adicionalmente, se puede apreciar que las iniciaciones son solo una pequeña parte del proceso de causación de cada trimestre. La ***Tabla 1***, contiene los indicadores involucrados en el análisis multivariado, variables escogidas según el proceso constructivo y la literatura disponible sobre el tema (CAMACOL, 2023), (MVCT, 2018) y (FEDESARROLLO, 2004).

Tabla 1. Indicadores contenidos en el trabajo

Nombre	Fuente	Corte
Área iniciada residencial + no residencial	DANE	I trimestre 2023
Área causada residencial + no residencial	DANE	I trimestre 2023
Área licenciada	DANE	I trimestre 2023
Unidades vendidas	GI	I trimestre 2023
Tasa de interés [2]	Ban Rep[1]	I trimestre 2023
Producto interno bruto del subsector de edificaciones	DANE	I trimestre 2023
Área en proceso*	DANE	I trimestre 2023
Desembolsos hipotecarios	DANE	I trimestre 2023
PIB por persona	DANE	I trimestre 2023
Formación de hogares [3]	DANE	NA

Fuente: SIS-SDHT.

* Para la variable área en proceso, se toma la resta entre el área total en proceso reportada en el anexo DANE y el área iniciada en cada trimestre, información que puede ser consultada en el siguiente enlace: *https://www.dane.gov.co/index.php/estadisticas-por-tema/construccion/censo-de-edificaciones.*

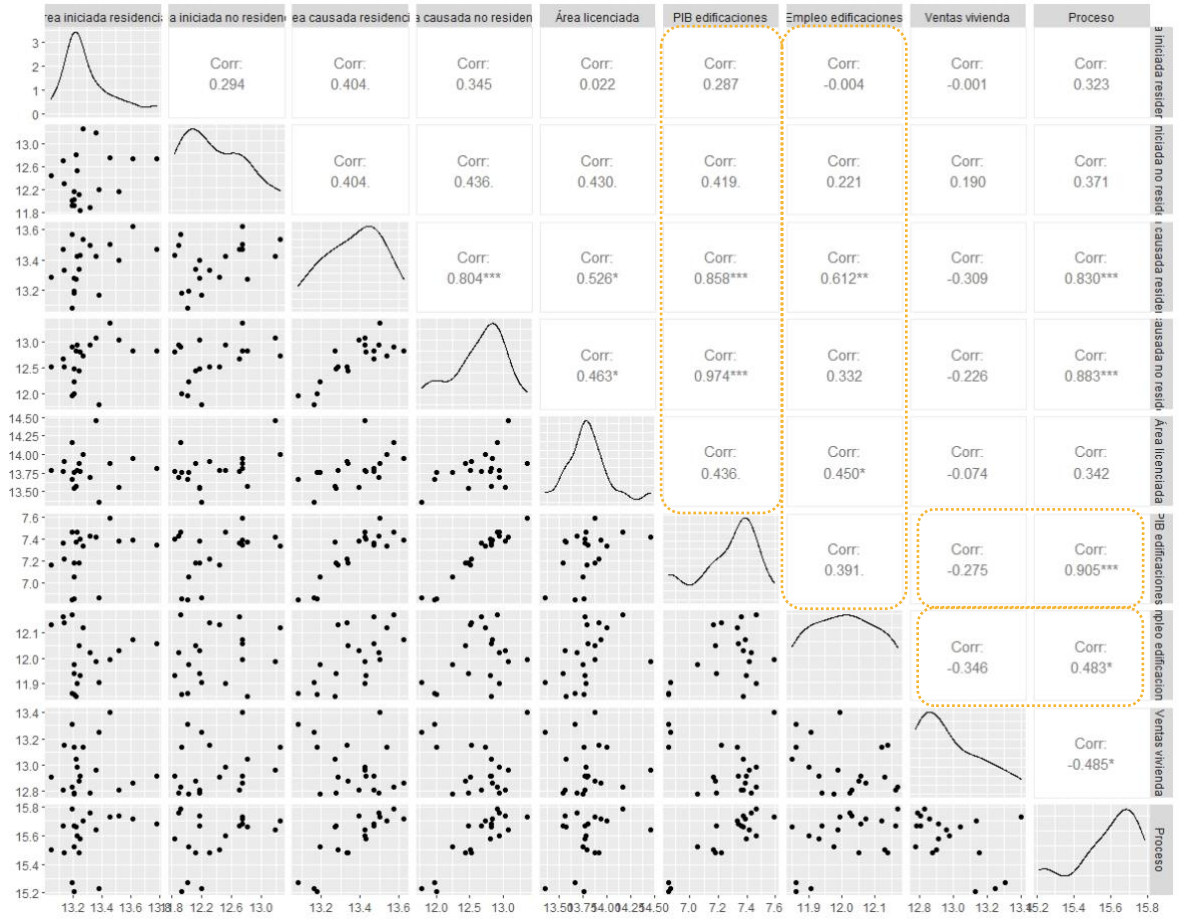
[1] Banco de la Republica.

[2] Tasa de colocación sin tesorería incluye créditos de consumo, ordinario y preferencial.

[3] Esta información viene de forma anual, para efectos del documento se toma el valor de cada año y se divide en cuatro obteniendo de esta manera el valor de cada trimestre, esto permite controlar los efectos demográficos sobre la demanda.

La ***Ilustración 15***, la cual contiene la matriz de correlaciones de los indicadores, evidencia la influencia positiva del área iniciada, causada y en proceso sobre el valor agregado, sin embargo, el área causada y el área en proceso son las únicas significativas y por lo tanto los mejores predictores contemporáneos del PIB. En otras palabras, pueden asociarse altos niveles de causación y área en proceso durante un trimestre con altos niveles del PIB con mayor certeza que la que se le podría asociar al área iniciada y altos niveles de PIB, lo cual, tiene sentido si se tiene en cuenta el proceso de causación. En cuanto a los niveles de ocupación parecen ser influenciados de manera contemporánea por el área en proceso y el área licenciada, pero no en la misma proporción como ocurre con el valor agregado ya que el coeficiente de correlación solo llega al 48,3%.

Ilustración 15. Relaciones contemporáneas entre las variables analizadas



Fuente: DANE, cuentas nacionales – PIB Bogotá III trimestre de 2021, estadísticas de empleo – Gran Encuesta Integrada de Hogares (GEIH), corte octubre de 2021, ELIC con corte a diciembre de 2020 y GI con corte a diciembre de 2021. Construcción SIS-SDHT.
Nota: se excluyen del análisis los datos de 2020, ya que se considera que el tema pandemia pude afectar la racionalidad del mercado y afectar el proceso constructivo.

En conclusión, se evidencia que la variable de área causada es la de mayor correlación, iniciaciones no tiene un peso superior al 10% en el PIB de edificaciones y presenta un rezago temporal entre 1 y 2 trimestres, finalmente se evidencia una estrecha relación entre las ventas y los niveles de iniciación.

Para determinar la influencia temporal de la variable iniciaciones residenciales se propone la estimación de un modelo de Vectores Autorregresivos (VAR) en su forma clásica y bayesiana, en su forma clásica son una forma sencilla y flexible de capturar complejas interacciones entre un gran número de variables macroeconómicas. Sin embargo, su sobre-parametrización puede causar problemas al momento de realizar inferencia especialmente en los pronósticos realizados con el modelo. Una solución a este problema es crear modelos restringidos a través del uso de

información previa de los fenómenos de análisis [8]. Un modelo VAR estándar de orden p con m variables puede ser representado como:

$$y_t = \Phi + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + \varepsilon_t \text{ con } p < t \text{ y } \varepsilon_t \sim N(0, \Sigma)$$

Donde y_t representa un vector de variables endógenas tomadas en el tiempo $t \in [1, T]$, ε_t es un vector de choques exógenos distribuidos normalmente y $\Phi, \beta_1, \beta_2, \ldots, \beta_p$ y Σ son matrices de dimensiones adecuadas que representan los parámetros desconocidos del modelo. Según Geweke (2005) y Canova (2007) este modelo puede ser escrito de manera compacta como:

$$y = (I_m \otimes Z)\alpha + \epsilon$$

Con $Z_{(Tx(1_c+mp))}$, $\alpha = vec(\beta)$[9], $y_{((Txm)x1)} = vec(Y_{(Txm)})$ y $\epsilon_{((Txm)x1)} \sim (0, \Sigma \otimes I_T)$ donde 1_c toma el valor de 1 si hay intercepto y cero en otro caso. A partir de esto y bajo el supuesto de normalidad en ϵ se tiene que la función de verosimilitud para el modelo viene dada por:

$$\mathcal{L}(\alpha, \Sigma | y, Z) = (2\pi)^{-m\frac{T}{2}} |\Sigma \otimes I_T|^{-\frac{1}{2}} exp\left\{-\frac{1}{2}[y - (I_m \otimes \mathrm{Z})\alpha]^T [\Sigma^{-1} \otimes I_T][y - (I_m \otimes \mathrm{Z})\alpha]\right\}$$

Al definir $\hat{\alpha} = (\Sigma^{-1} \otimes Z^T Z)^{-1}(\Sigma^{-1} \otimes Z)^T$ y $\mathfrak{F} = I_m \otimes Z$; y utilizar las propiedades de las matrices simétricas, del producto Kronocker y de las formas cuadrática matriciales, se tiene que el logaritmo de la función de verosimilitud viene dado por:

$$\ln\left(\mathcal{L}(\alpha, \Sigma | y, Z)\right) \propto ln\{N(\alpha | \alpha, \Sigma, \mathfrak{F}, y).W(\Sigma | \hat{\alpha}, \mathfrak{F}, y)\}$$

Donde $N(.)$ denota la distribución normal y $W(.)$ la distribución Wishart de tal manera que la función de log-verosimilitud para un modelo VAR viene dada por el logaritmo natural del producto entre la distribución condicionada de α y de Σ. Este hecho es importante ya que describe las distribuciones canónicas a priori para α y Σ.

El sistema descrito anteriormente posee $(k + pk2)$ coeficientes para estimar. Este tamaño es considerable dada la longitud promedio de las series empleadas, lo cual a su vez puede resultar en estimaciones poco significativas, de poca precisión y con problemas de correlación serial (Quilis, 2002). Con este contexto, para este trabajo se retoman los trabajos de autores como Todd (1984), Doan, Litterman & Sims (1984) y Litterman (1986), quienes buscaron superar estos inconvenientes a través de la inclusión de información previa en las estimaciones, haciendo referencia a los posibles valores que podrían tomar los coeficientes, independientemente de la información derivada de los datos muestrales.

---

[8] Ejemplos muy populares de este tipo de información son: el hecho que la gran mayoría de las series macroeconómicas son series integradas de orden 1 y que muchas de ellas guardan relaciones de cointegración de largo plazo entre ellas.

[9] *vec* representa el operador apilamiento de columnas, para más información ver Macedo, H. D.; Oliveira, J. N. (2013).

El enfoque bayesiano generalmente asume que tanto α como Σ son variables aleatorias, donde α sigue una distribución normal multivariada y Σ una distribución Wishart. Sin embargo, algunos autores como Litterman (1986) trabajan con información obtenida de hechos estilizados donde, tal vez el más importante de ellos, es que las series macroeconómicas pueden ser descritas de manera eficiente por una caminata aleatoria de la forma $Y_{i,t} = \mu_t + Y_{i,t-1} + \varepsilon_{i,t}$ con $\varepsilon_i \sim N(0, \sigma^2)$. Bajo esta especificación, los rezagos más próximos tienen mayor importancia y la información de las demás variables no es tan relevante como la propia. Estos dos hechos se incluyen en la metodología al aceptar los siguientes supuestos:

Desde el enfoque clásico destacan los siguientes modelos:

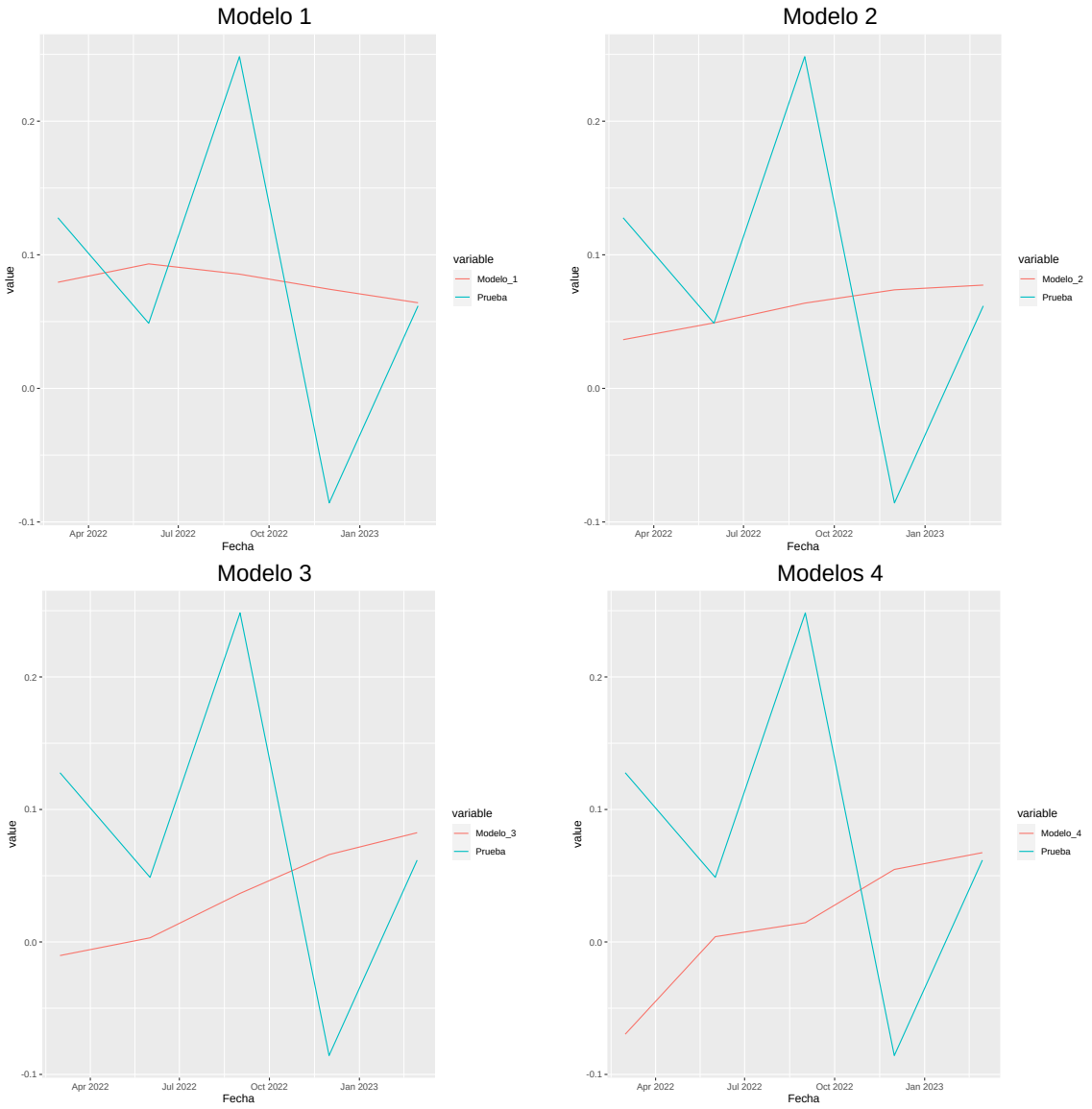
1. Modelo 1: VAR con un rezago sin constante ni tendencia, utilizando las variables PIB edificaciones, formación de hogares y número de desembolsos.
2. Modelo 2: VAR con un rezago sin constante ni tendencia, utilizando las variables PIB edificaciones, área en proceso, licencias de construcción, formación de hogares, número de desembolsos e iniciaciones.
3. Modelo 3: VAR con un rezago sin constante ni tendencia, utilizando las variables área en proceso, licencias de construcción, PIB por persona, desembolsos, PIB de edificaciones, formación de hogares e iniciaciones.
4. Modelo 4: promedio RMS, este promedio contiene las estimaciones de los tres anteriores modelos, más los resultados de un VAR(1) con constante sin tendencia con las variables (área en proceso, licencias, PIB por persona, desembolsos, PIB de edificaciones, formación de hogares e iniciaciones) y las estimaciones de una VAR(1) sin constante ni tendencia, el cual utiliza las variables (área en proceso, licencias, ventas, PIB por persona, desembolsos, PIB de edificaciones, formación de hogares e iniciaciones).

Tabla 2 Resultados mejores modelos según desempeño dentro de muestra

Modelo	RMSE	MAE	Tasa de falla	Promedio
Modelo 1	11,7	32,5	25	23,0
Modelo 2	14,1	35,1	25	24,7
Modelo 3	18,5	44,2	50	37,5
Modelo 4	23,9	48,5	50	40,8

Fuente: SDHT-SIS.

Ilustración 16 Resultados esquema VAR pronóstico dentro de muestra 8 pasos adelante



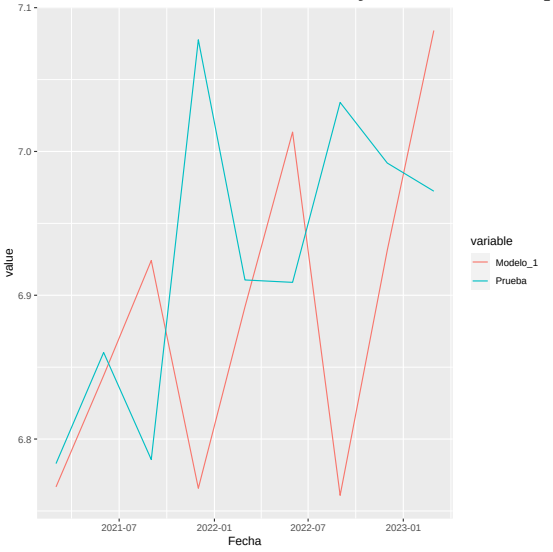
Fuente: SDHT-SIS.

Desde la perspectiva bayesiana destacan las siguientes especificaciones un modelo VAR(1) con una distribución previa tipo Minnesota y las variables en logaritmo PIB de edificaciones, área causada y los desembolsos de créditos hipotecarios, un VAR (1) tipo Minnesota con las variables PIB de edificaciones, área causada e

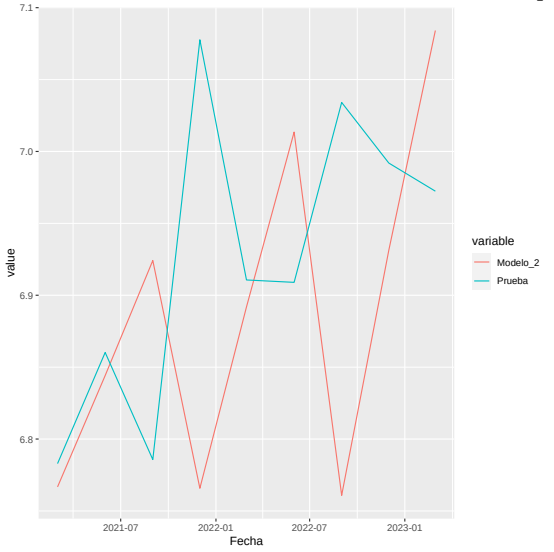
iniciaciones, el promedio en términos de RMS y un índice construido a partir de un análisis de ACP[10].

Ilustración 17 Resultados esquema VAR bayesiano pronóstico dentro de muestra 8 pasos adelante

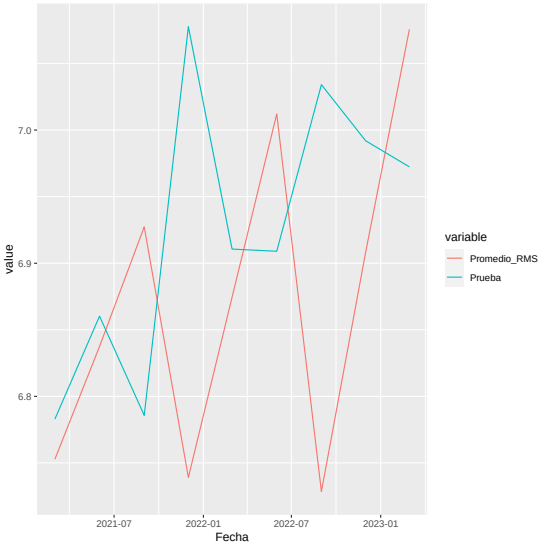
Panel 1. Modelo VAR (1) [PIB de edificaciones, área causada y desembolsos]



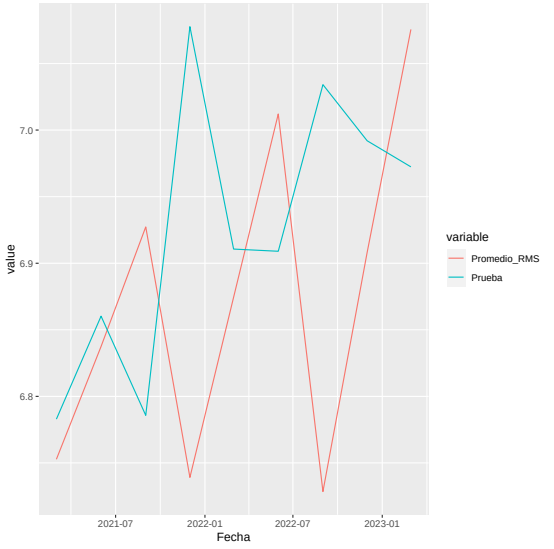
Panel 2. Modelo VAR (1) [PIB de edificaciones, área causada e iniciaciones]



Panel 3. Promedio RMS



Panel 4. Promedio ACP



Fuente: SDHT-SIS.

[10] Se incluye un modelo VAR(1) tipo Minnesota con el PIB de edificaciones, área causada y los desembolsos hipotecarios, un VAR (1) con las variables PIB de edificaciones, área causada e iniciaciones, un modelo VAR(2) con PIB de edificaciones, área causada y los desembolsos, modelo VAR(3) con una con las variables PIB de edificaciones, área causada e iniciada y un VAR (1) con PIB de edificaciones, área causada y desembolsos.

Redes neuronales:

En este capítulo se explicará el proceso para elaborar la red neuronal con base la información del PIB de edificaciones, realizando una clasificación de las mejores combinaciones de neuronas y capas ocultas según su capacidad de predicción dentro de muestra. Se plantea una predicción usando como variables explicativas tres rezagos. Según la ***Tabla 3*** el mejor modelo de los 100 analizados es aquel con seis neuronas y dos capas ocultas, los modelos restantes entran a hacer parte de la batería usada por la SDHT para estimar el comportamiento futuro de la serie del PIB.

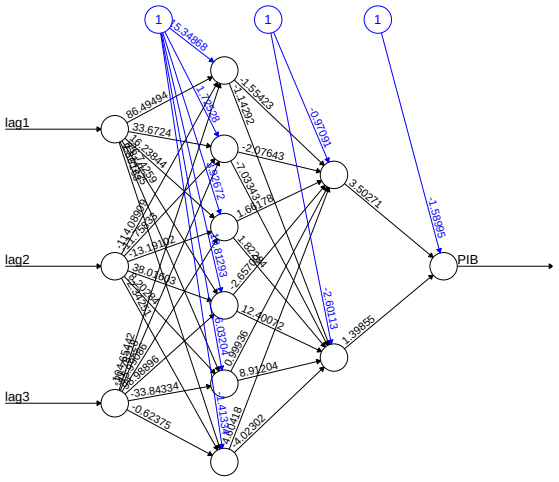
Tabla 3 Top 10 modelos de redes neuronales

Número de neuronas	Capas ocultas	MSE	RMSE	MAE	Tasa de falla	Promedio
6	2	0,035	0,187	0,141	0,397	0,19
1	7	0,036	0,19	0,142	0,397	0,191
6	6	0,037	0,194	0,142	0,397	0,193
4	3	0,039	0,197	0,147	0,397	0,195
2	2	0,04	0,199	0,147	0,397	0,196
2	7	0,038	0,196	0,144	0,413	0,198
2	4	0,042	0,205	0,151	0,397	0,199
5	1	0,04	0,2	0,149	0,413	0,201
9	1	0,039	0,198	0,148	0,429	0,204
6	3	0,04	0,2	0,149	0,429	0,204

Fuente: SDHT-SIS.

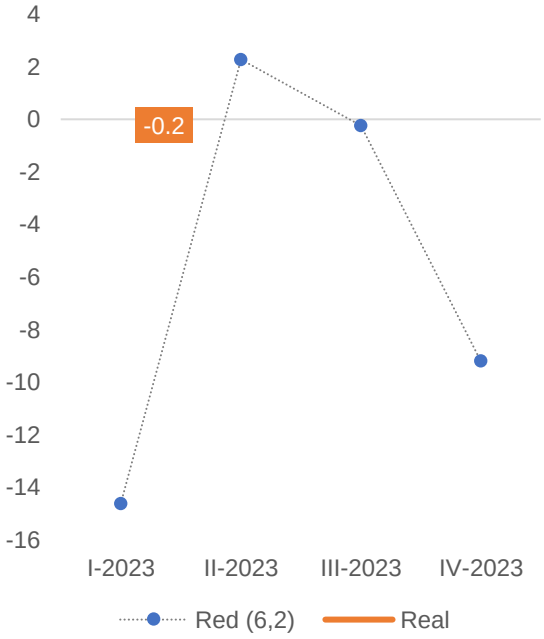
Ilustración 18 Ejemplo de dos redes y su pronóstico para 2023 dada la información hasta el cuarto trimestre de 2022.

Red (6,2)

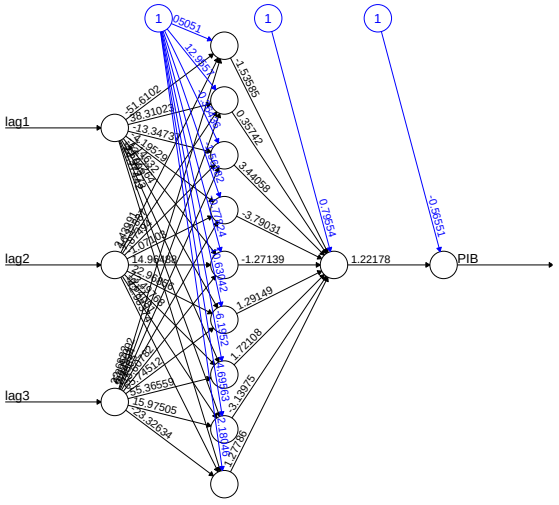


Error: 0.204667 Steps: 4650

Pronóstico red (6,2) 2023

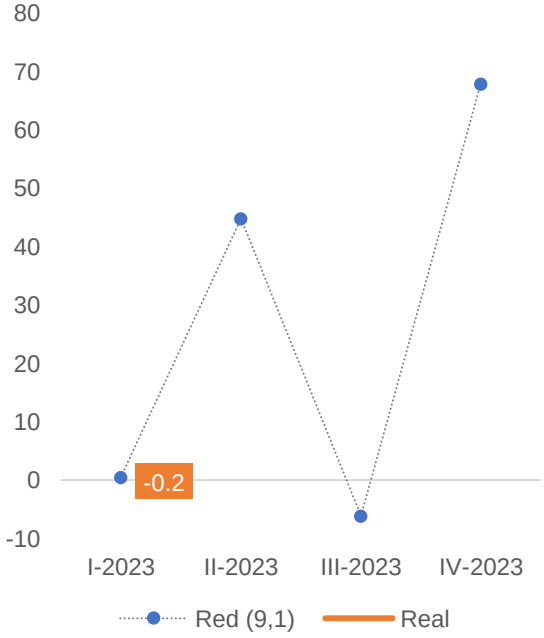


Red (9,1)



Error: 0.265742 Steps: 976

Pronóstico red (9,1) 2023

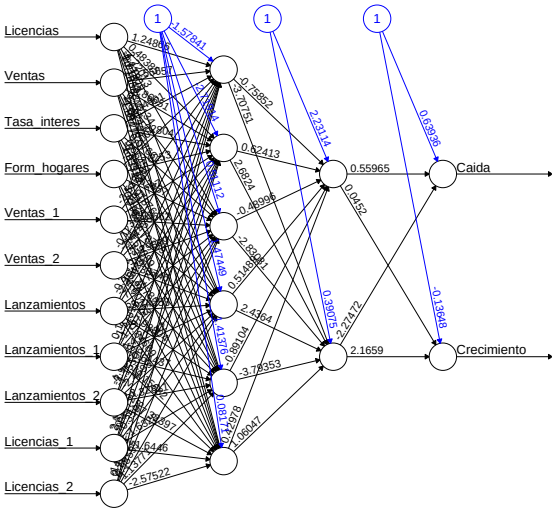


Fuente: SDHT-SIS.

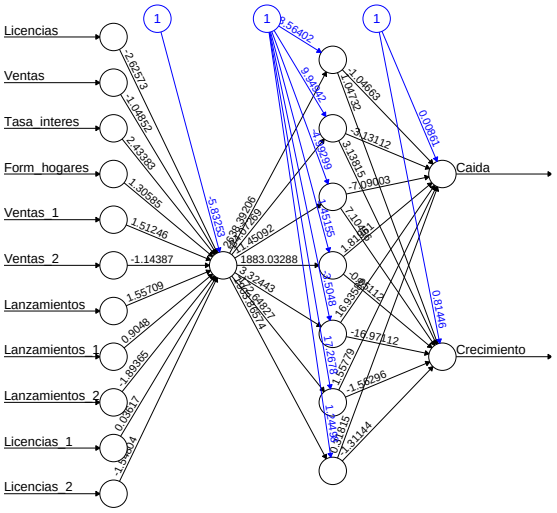
Adicionalmente, se entregan los resultados de las redes neuronales como modelo de clasificación, debido al bajo volumen de datos se decide desechar la idea de un modelo multinomial y optar por un modelo dicotómico, con dos posibles valores “Crecimientos” y “Caídas”, se estiman los cuatro mejores modelos de la parte anterior, Los cuatro modelos con los datos disponibles para el cuarto trimestre de 2022 pronostican una caída anual del PIB de edificaciones, información que puede ser combinada con los pronósticos puntuales trimestre a trimestre.

lustración 19 Redes neuronales estimadas para calcular la posible tendencia dados los datos de ventas, lanzamientos, licenciamiento, formación de hogares y tasa de interes durante el primer trimestre de 2023.

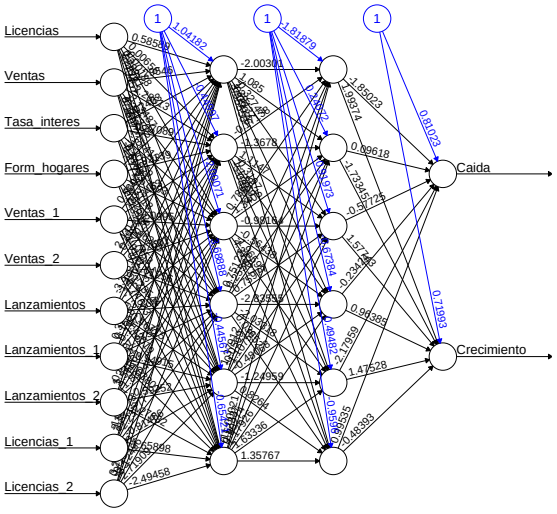
Red (neuronas=6, capas ocultas=2)



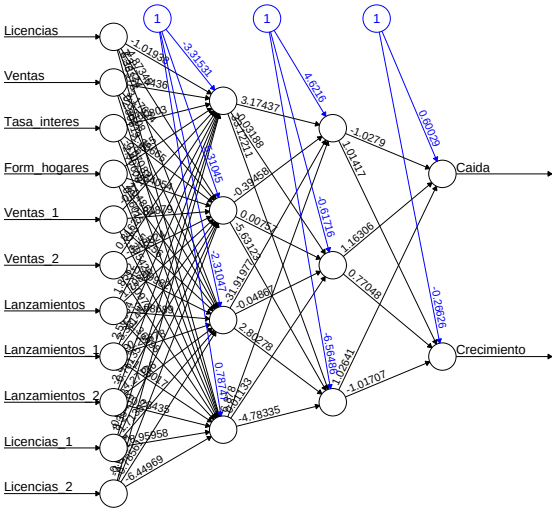
Red (neuronas=1, capas ocultas=7)



Red (neuronas=6, capas ocultas=6)



Red (neuronas=4, capas ocultas=3)



Fuente: SDHT-SIS.

Pronósticos depurados

En este apartado se presenta el resultado final del ejercicio. Gracias a que para el momento de construcción de este documento se dispone de datos para el primer trimestre de 2023 es posible evaluar las proyecciones construidas a partir de la información disponible con corte a 2022 ejercicio que es útil para ir ajustando de manera gradual los pesos de cada modelo dentro de la proyección final para los trimestres siguientes. A continuación, se presenta el esquema simplificado de uso para la herramienta que acompaña este ejercicio:

1. Se actualizan el último corte cada una de las variables utilizadas.

2. Se estiman los pronósticos de cada uno de los siguientes modelos.
    - Promedio RMS
    - Promedio ACP
    - Promedio
    - ARIMA (1,0,2) (0,01)
    - ARIMA (2,0,2) (0,01)
    - ARIMA (5,0,2) (0,01)
    - VAR (rezago=1) sin tendencia ni constante [PIB de edificaciones, formación de hogares y desembolsos hipotecarios]
    - VAR (rezago=1) sin tendencia ni constante [PIB de edificaciones, área en proceso, área licenciada, formación de hogares, desembolsos hipotecarios e iniciaciones]
    - VAR (rezago=1) sin tendencia ni constante [área en proceso, licencias de construcción, PIB por persona, desembolsos hipotecarios, PIB de edificaciones, formación de hogares e iniciaciones]
    - Contiene los tres anteriores modelos, más los resultados de un VAR(rezago=1) con constante sin tendencia con las variables [área en proceso, licencias, PIB por persona, desembolsos, PIB de edificaciones, formación de hogares e iniciaciones] y las estimaciones de una VAR(rezago=1) sin constante ni tendencia con las variables [área en proceso, licencias, ventas, PIB por persona, desembolsos, PIB de edificaciones, formación de hogares e iniciaciones],
    - VAR(rezago=1) con las variables PIB de edificaciones, área causada y los desembolsos de créditos hipotecarios,
    - VAR (rezago=1) con las variables PIB de edificaciones, área causada e iniciaciones
    - Se incluye un modelo VAR(rezago=1) con el PIB de edificaciones, área causada y los desembolsos hipotecarios; un VAR (rezago=1) con las variables PIB de edificaciones, área causada e iniciaciones; un VAR(rezago=2) con PIB de edificaciones, área causada y los desembolsos, un VAR(rezago=3) con las variables PIB de edificaciones, área causada e iniciada y un VAR (rezago=1) con PIB de edificaciones, área causada y desembolsos,
    - Neuronas (n)=6 & capas ocultas (co)=2
    - n=1 & co=7
    - n=6 & co=6
    - n=4 & co=3
    - n=2 & co=2

- n=2 & co=7
- n=2 & co=4
- n=5 & co=1
- n=9 & co=1
- n=6 & co=3

2. Se sugiere construir a futuro la suma ponderada de estos pronósticos, esto implica usar el desempeño de cada uno por trimestre como ponderador, es decir, el peso de un modelo i dentro del pronóstico final para el trimestre j depende del inverso de la diferencia entre el último pronóstico realizado por dicho modelo para el trimestre de referencia.

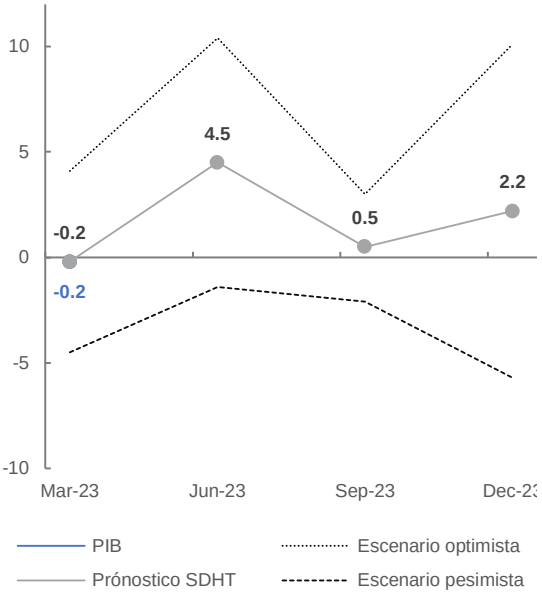
3. Para construir los escenarios se propone utilizar como banda el 0,5 de la desviación estándar de los pronósticos de cada trimestre. Este parámetro puede ser ajustado según las perspectivas de volatilidad del mercado.

La ***Ilustración 20*** muestra las proyecciones del ejercicio para 2023, de las cuales van de acuerdo con al dato real del primer trimestre en tendencia y precisión[11], reafirmando que el uso de varios tipos de modelos incluidos las redes neuronales aportan en la construcción de las proyecciones, sin ser necesariamente mejores unos que otros. Las expectativas de la SDHT para todo 2023 están 2% por debajo de la planteada por la entidad más importante de investigación del sector CAMACOL, la cual esperan un crecimiento del 9%.

[11] La precisión va a depender de los datos disponibles y la coyuntura del momento.

Ilustración 20 Resultados proyección 2023 con información a 2022 y desempeño por modelo para el primer trimestre de 2023

Panel 1. Pronóstico PIB de edificaciones SDHT para 2023 con información al cuarto trimestre de 2022 Vs PIB publicado primer trimestre de 2023



Panel 2. Desempeño por modelo frente al primer trimestre de 2023

Metodología	Enfoque	Modelo	Diferencia primer trimestre de 2023
Diferencia anual en logaritmo del PIB de edificaciones	Clásicos	Promedio RMS	ooooooooo
		Promedio ACP	ooooooo
		Promedio	ooooooo
	Bayesiano	ARIMA (1,0,2)(0,01)	ooooo
		ARIMA (2,0,2)(0,01)	oooo
		ARIMA (5,0,2)(0,01)	o
Multivariado	Clásicos diferencia anual de las variables en logaritmo	VAR (rezago=1) sin tendencia ni constante [PIB de edificaciones, formación de hogares y desembolsos hipotecarios]	ooooooooo
		VAR (rezago=1) sin tendencia ni constante [PIB de edificaciones, área en proceso, área licenciada, formación de hogares, desembolsos hipotecarios e iniciaciones]	ooooooo
		VAR (rezago=1) sin tendencia ni constante [área en proceso, licencias de construcción, PIB por persona, desembolsos hipotecarios, PIB de edificaciones, formación de hogares e iniciaciones]	ooooooo
		Contiene los tres anteriores modelos, más los resultados de un VAR(rezago=1) con constante sin tendencia con las variables [área en proceso, licencias, PIB por persona, desembolsos, PIB de edificaciones, formación de hogares e iniciaciones] y las estimaciones de una VAR(rezago=1) sin constante ni tendencia con las variables [área en proceso, licencias, ventas, PIB por persona, desembolsos, PIB de edificaciones, formación de hogares e iniciaciones],	ooooooo
	Bayesiano familia Minnesota variables en logaritmo	VAR(rezago=1) con las variables PIB de edificaciones, área causada y los desembolsos de créditos hipotecarios,	o
		VAR (rezago=1) con las variables PIB de edificaciones, área causada e iniciaciones	ooooo
		Se incluye un modelo VAR(rezago=1) con el PIB de edificaciones, área causada y los desembolsos hipotecarios; un VAR (rezago=1) con las variables PIB de edificaciones, área causada e iniciaciones; un VAR(rezago=2) con PIB de edificaciones, área causada y los desembolsos, un VAR(rezago=3) con las variables PIB de edificaciones, área causada e iniciada y un VAR (rezago=1) con PIB de edificaciones, área causada y desembolsos,	o
Redes neuronales		Neuronas=6 & capas ocultas=2	oooooooooooooo
		Neuronas=1 & capas ocultas=7	ooooooooo
		Neuronas=6 & capas ocultas=6	ooooooooooooooooo
		Neuronas=4 & capas ocultas=3	ooooooooo
		Neuronas=2 & capas ocultas=2	ooooooooo
		Neuronas=2 & capas ocultas=7	oooooo
		Neuronas=2 & capas ocultas=4	ooooooooo
		Neuronas=5 & capas ocultas=1	ooooooo
		Neuronas=9 & capas ocultas=1	
		Neuronas=6 & capas ocultas=3	ooooooooo

Fuente: SDHT-SIS

Conclusiones

Este ejercicio muestra que los enfoques analizados aportan información relevante sobre el posible comportamiento futuro de la actividad edificadora en la ciudad, se revelan las limitaciones por falta de volumen de datos a que se enfrenta el enfoque de redes neuronales a la hora de hacer pronósticos puntuales, el cual no puede desplegar todo su potencial dada la poca disponibilidad de escenarios de aprendizaje pero aportando a la hora de marcar la tendencia de corto plazo, por lo cual los esquemas bayesianos ganan fuerza, en cuanto al enfoque multivariado clásico este parece describir bastante bien la tendencia de la serie, esto debido a la similitud de funcionamiento que tiene con el proceso de causación usado por el DANE para genera el PIB. Es importante aclarar que este documento no agota todas las posibilidades del enfoque de redes neuronales como herramienta de pronóstico, este es solo un trabajo exploratorio que busca aportar en el tema.

6. Bibliografía

Amat, J. (2021). *Ciencias de Datos*. Retrieved from https://www.cienciadedatos.net/documentos/68-redes-neuronales-r.html

Gibson, J. P. (2017). *Deep Learning A Practitioner's Approach.* United States of America: O'Reilly Media, Inc.

Hastie, B. E. (2021). *Computer Age Statistical Inference:Algorithms, Evidence and Data Science.* Cambridge: Cambridge University Press.

Iffat A. Gheyas, L. S. (2009). A Neural Network Approach to Time Series Forecasting. *Proceedings of the World Congress on Engineering 2009 Vol II*.

Pitts, W. S. (1943). *A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics.* Springer.

Sáenz, J. M. (2009). Evaluación de pronóstico de una red neuronal sobre el PIB en Colombia. *Borradores de ECONOMÍA* , 1-55.

WICHERN, R. A. (2007). *Applied Multivariate Statistical Analysis* . New Jersey: Prentice Hall.