Statihouse®: desarrollo tecnológico basado en ciencia de datos para explorar estadísticamente el sector inmobiliario

Statihouse®: technological development based on data science to statistically explore the real estate sector

Jorge Iván Pérez Rave ^{1*}

Recibido 13 de julio de 2017, aceptado 28 de febrero de 2018

Received: July 13, 2017 Accepted: February 28, 2018

RESUMEN

El sector inmobiliario requiere un mejor aprovechamiento de la ciencia de datos y los datos masivos, con el fin de favorecer la gestión informada y la toma de decisiones. El presente estudio tiene como objetivo describir un sistema (Statihouse®) que caracteriza estadísticamente propiedades en venta y predice su precio de oferta (en tiempo real o cercano), empleando datos masivos de la internet, ciencia de datos y una aplicación web. Se expone el desarrollo del sistema, enfatizando sobre sus características técnicas, los procesos y subprocesos, y la visualización para el usuario. Se presentan ejemplos de visualización, aportando resultados derivados de una muestra de 24.935 casas usadas, ofrecidas para la venta en Colombia (periodo enero 2017 – mayo 2017), los cuales fueron contrastados con algunos referentes a modo exploratorio. Este sistema resulta novedoso, pues no se encontraron antecedentes que ejecuten de forma automática desde la recolección de datos web, hasta la visualización (en tiempo real o cercano), con un alcance multitareas (descriptivo, comparativo, evolutivo, correlacional y predictivo) en el sector inmobiliario, con enfoque en la propiedad. Este sistema evidencia un caso de éxito de la ciencia de datos en el sector inmobiliario y sirve de guía y estímulo para nuevos desarrollos.

Palabras clave: Sector inmobiliario, ciencia de datos, statihouse.

ABSTRACT

The real estate sector requires better use of data science and the massive data, to favor informed management and decision making. The present study aims to describe a system (Statihouse®) that statistically characterizes properties for sale and predicts their offer price (in real time or near), using massive data from the internet, data science and a web application. It exposes the development of the system, emphasizing its technical characteristics, processes and subprocesses, and visualization for the user. We present examples of visualization, providing results derived from a sample of 24,935 used houses, offered for sale in Colombia (period January 2017 - May 2017), which were contrasted with some referents as an exploratory mode. This system is novel, since there were no antecedents that execute automatically from the collection of web data, to the visualization (in real time or near), with a multitasking scope (descriptive, comparative, evolutionary, correlational and predictive) in the Real estate sector, with a focus on property. This system evidences a case of success of data science in the real estate sector and serves as a guide and stimulus for new developments.

Keywords: Real estate, data science, statihouse.

¹ Grupo de investigación IDINNOV. IDINNOV S.A.S. Medellín, Colombia. E-mail: investigacion@idinno.com

INTRODUCCIÓN

El estudio del mercado inmobiliario es un tema de interés creciente para las naciones, por las implicaciones prácticas que este genera en diversos grupos organizacionales (valuadores, agremiaciones de bienes raíces, órganos de catastro, sectores de construcción,...). La principal variable de análisis en dicho contexto ha sido el precio, pues es reconocida como la proxy más razonable del constructo “valor” del inmueble [1]. El enfoque metodológico tradicional ha sido la regresión hedónica [2-4]. Su lógica es explicar el precio usando modelos econométricos cuyas covariables son atributos del bien o su entorno, para los cuales se asume que van adicionando “valor” al inmueble de forma individual. Este enfoque exige el cumplimiento de los supuestos básicos de la regresión (normalidad, homocedasticidad, independencia entre observaciones, ausencia de colinealidad), pero generalmente no se cumplen, exigiendo adecuaciones (Ej: transformaciones de variables) [4-5, 1].

El enfoque de regresión de precios hedónicos ha servido de base para proponer innumerables modelos, que en su mayoría se han centrado sobre atributos observables (no constructos) como covariables (Ej: [6-7]). Según Malpezzi, entre los predictores comúnmente usados, están: área construida, número de baños, número de habitaciones [8]. Estos predictores particulares se muestran correlacionados, lo cual puede generar riesgos de colinealidad. De darse colinealidad en una alta magnitud, no es confiable la estimación de los efectos individuales de los atributos sobre la variable respuesta. De hecho, Gujarati advierte que se inflan las varianzas de los estimadores, se obtienen intervalos de confianza más amplios y puede derivar en razones “t” no significativas para los parámetros, llevando a conclusiones erróneas [9]. Se trata de un problema comúnmente citado como parte de la crítica para justificar enfoques emergentes, basados en inteligencia artificial, como árboles de regresión (o de clasificación), redes neuronales, entre otros métodos [10, 5]. Sin embargo, cada uno de estos enfoques (basados en regresión hedónica o en inteligencia artificial) tiene sus ventajas y desventajas (véase [11]), y, a pesar de las inclinaciones hacia uno u otro, lo cierto es que a la fecha no hay consenso sobre el de mayor superioridad (incluso considerando aplicaciones en contextos no inmobiliarios).

Llama la atención que, a pesar de contar con el análisis de componentes principales para resumir información y, en consecuencia, mitigar problemas de colinealidad en este contexto, el uso de esta técnica no ha tenido el auge suficiente dentro del enfoque de regresión hedónica. No obstante, recientemente, trabajos como [12-14] han venido mostrando la utilidad de usar constructos derivados de análisis de componentes principales, no solo para el enfoque econométrico, sino también para aquellos basados en inteligencia artificial (ayuda a reducir atributos redundantes). Sin embargo, el poco uso de componentes principales, en este contexto de precios hedónicos, se ha ejecutado bajo método de regresión por componentes principales, en el cual la variable respuesta es explicada netamente por constructos (véase una descripción de este método, en [11]. Esto abre posibilidades, en nuevos estudios, para combinar, en un mismo modelo, constructos (atributos latentes) e indicadores observables del inmueble y su entorno.

Otro problema por destacar en el ámbito inmobiliario, aplicable a métodos econométricos, pero también a los intensivos en inteligencia artificial, es que no se están aprovechando suficientemente las tecnologías emergentes para la recolección, la limpieza, el procesamiento, el análisis y la visualización de datos [15-16]. Por el contrario, la mayoría de los trabajos emplean tecnologías tradicionales (Ej: software basado en botones, ejecución por pasos que depende de la manipulación humana, métodos de análisis limitados para datos masivos,...).

Adicionalmente, respecto a los tamaños de muestra, la tradición ha sido usar conjuntos de datos pequeños, tal como lo expresa [10]. En otras palabras, si bien se hacen grandes contribuciones teóricas y metodológicas, hay limitaciones para aprovechar la disponibilidad de datos masivos que diariamente se generan sobre la oferta inmobiliaria en la web. Esto puede deberse a la complejidad algorítmica y a la diversidad de disciplinas requeridas para la recolección, la limpieza, la estructuración de datos (estructurados, semiestructurados, no estructurados) de forma automática, eficaz y eficiente, entre otras actividades. De hecho, los datos, inmobiliarios son reconocidos como complejos y, con frecuencia, presentan observaciones atípicas e incorrectas [15]. Las actividades descritas y la integración de saberes de diversas disciplinas (Ej: estadística, computación,

matemática, gestión de procesos,...), son propias de un área emergente llamada ciencia de datos. Se estima que cerca del 70 % del tiempo de los proyectos en esta área es consumido por tareas de alistamiento de los datos, a las cuales se les reconoce gran parte de arte, y que se van dominando con la práctica [16].

Hoy día, diversos estudios [17-18, 15] llaman la atención sobre la necesidad de acercar esta nueva disciplina al sector inmobiliario. No obstante, también expresan lo incipiente que se encuentra dicho acercamiento. A través de la ciencia de datos, aplicada al sector inmobiliario, se posibilita obtener información más oportuna para la toma de decisiones (en tiempo real o cercano), basada en muestras de mayores tamaños y mediante actividades automáticas, derivadas del aprendizaje estadístico, entre otros. Esto resulta vital para la racionalidad del pensamiento, para generar conocimiento socialmente útil, y para obrar de forma más oportuna y eficaz ante los retos de las profesiones y el entorno.

Este trabajo se suma a esfuerzos recientes por cambiar de paradigma en el estudio del sector inmobiliario, y tiene como objetivo describir un sistema que caracteriza estadísticamente propiedades en venta y predice su precio de oferta (en tiempo real o cercano), empleando datos masivos de la internet, ciencia de datos y una aplicación web.

Este sistema resulta novedoso, en la medida que cumple una función multitareas (descriptiva, comparativa, evolutiva, correlacional y predictiva) en este sector, con enfoque en datos de la propiedad y su entorno, que comprende desde la recolección automática de datos web, hasta la visualización (en tiempo real o cercano). La articulación lógica y sistemática de métodos, tecnologías y capacidades cobra valor desde la teoría de recursos y capacidades [19], considerando que el uso de los recursos *per se* no necesariamente genera ventajas competitivas, por lo que un proceso vital para lograrlas es la “orquestración de recursos” [20]. En este caso concreto, tal proceso se materializa por medio de la integración inteligente y sistemática de los componentes, en un sistema que redunda en nuevas capacidades analíticas. Este estudio sirve de caso de éxito y motivación para futuros trabajos, ante varios de los retos planteados para la ciencia de datos en el sector inmobiliario. Este sistema ha recibido

el nombre de Statihouse®. El alcance de la versión estándar comprende netamente casas usadas, ofrecidas para la venta en Colombia a través de canales web. No obstante, este sistema puede escalarse fácilmente, bajo versiones no estándar, para otros inmuebles (Ej: apartamentos) y tipos de oferta (Ej: arrendamiento).

Este artículo describe el sistema Statihouse® como un todo, y su enfoque es de difusión tecnológica. En otras publicaciones se detallará en algunos de sus componentes.

El resto del documento se estructuró así: se expone el desarrollo del sistema, enfatizando sobre sus características técnicas, los procesos y subprocesos, y la visualización para el usuario. Esta última, aportando ilustraciones del formato de visualización para algunas de las preguntas de analítica que se busca resolver a través del sistema. Luego, viene el informe de discusión, tomando en consideración algunos referentes para comparación exploratoria. Seguido, se plasman las conclusiones, lo cual también comprende desafíos y limitaciones.

DESARROLLO

A continuación se describe las características técnicas, los procesos (y subprocesos) y las visualizaciones que aporta Statihouse®.

Características

Statihouse® articula métodos y herramientas de estadística (descripción de variables, análisis de segmentación, modelos de regresión, análisis de componentes principales), de ciencias de la computación (programación y algoritmia, recolección automática de datos, aprendizaje de máquinas, almacenamiento de datos, visualización), de ingeniería de calidad (cartas de control) y de negocios (enfoque por procesos). Todo esto, sobre la base de la metodología de la investigación científica, para: 1. Capturar, estructurar, limpiar y actualizar de forma periódica, datos de anuncios de internet sobre la oferta de venta de inmuebles. 2. Realizar procesamiento de dichos datos y proveer de forma oportuna (en tiempo real o cercano) información estadística de tipo descriptivo, comparativo, evolutivo, correlacional y predictivo. 3. Divulgar la información de forma amigable, clara y oportuna, de modo que posibilite generar conocimiento y favorecer el proceso racional de toma de decisiones.

El presente artículo describe la tecnología en general (Statihouse®), pero el caso de ejemplo corresponde a casas usadas, ofrecidas para la venta en Colombia por medio de canales web, con una muestra de 24.935 registros, recopilados en el periodo enero 02 de 2017 – mayo 31 de 2017. Se trata de un sistema híbrido, programado en R [21], con apoyo de Rstudio [22] y Shiny [23], que hace viable el monitoreo de variables en tiempo real o cercano y la predicción de precios, usando datos de anuncios web y ciencia de datos.

En la Figura 1 se presenta el modelo general de operación, el cual se compone de cuatro procesos. Para poder ejecutar los procesos en cuestión, solo se requiere un computador con acceso a internet y dos carpetas con archivos específicos, sobre los que se ha automatizado, en forma de algoritmos, las actividades inmersas en cada proceso.

Procesos y subprocesos

En la Figura 2 se presentan los procesos y subprocesos necesarios para la caracterización de los inmuebles y la predicción de precios.

El objetivo del primer proceso (“Alistamiento de materiales y procesos”) es preparar las carpetas, los archivos y los datos preliminares. Este se ejecuta una sola vez, al instalar el sistema. En cambio, los demás procesos son recurrentes. El segundo proceso (“Conformación y actualización de la base de trabajo”) tiene como objetivo captar, estructurar, limpiar y actualizar datos. Este se lleva a cabo según se requiera. En vista de que el mercado inmobiliario no es altamente cambiante a nivel de segundos o minutos, puede ser ejecutado una vez al día. Es decir, diariamente despliega sus actividades internas automatizadas y provee a los demás procesos con nuevos datos (Ej: base de datos del día), los cuales se integran a la base de trabajo. El objetivo del tercer proceso (“Procesamiento de datos y producción de resultados”) es justamente realizar el procesado de los datos y generar resultados estadísticos de naturaleza descriptiva, comparativa, evolutiva, correlacional y predictiva para 25 preguntas de analítica. El objetivo del cuarto y último proceso (“Visualización interactiva de resultados”) es divulgar las salidas del proceso anterior de forma clara y oportuna, según lo demande la audiencia. Para esto se ha desarrollado un entorno *dashboard* con tecnologías Shiny [23], que facilita la visualización desde medios web y la interacción por parte del usuario.

Figura 1. Modelo general del sistema Statihouse®.

Figura 2. Procesos y subprocesos del sistema Statihouse®

Figura 3. Ejemplo de ficheros de bases de datos del día, a través del uso del algoritmo Estrella®.

El subproceso denominado Estrella®, es la herramienta empleada por el proceso de “Conformación y actualización de la base de trabajo”. Se trata de un algoritmo a la medida, basado en recolección automática de datos desde la web, cartas de control y análisis de segmentación. Este algoritmo está programado en el computador para que se ejecute periódicamente de forma automática. En cada

ejecución, lleva a cabo sus tareas y actualiza la base de datos de trabajo, la cual es el insumo principal del proceso de “Procesamiento de datos y producción de resultados”.

En la Figura 4 se muestra un ejemplo de cómo, cada cierto periodo, Estrella® va generando archivos relacionales en formato “.csv”.

En la Tabla 1 se resume la caracterización general de los subprocesos de “Procesamiento de datos y producción de resultados”. A modo de ilustración para procesamiento de precios (Tabla 1), se parte de dos preguntas de analítica. Una tendiente a describir el precio total (ofrecido para

venta) y la otra el precio de metro cuadrado. También se especifican las variables a estudiar para responder dichos interrogantes, las actividades automatizadas en R (en este caso univariadas), y los formatos y estadísticos de visualizaciones para el usuario.

Tabla 1. Caracterización general de los subprocesos del “Procesamiento de datos y producción de resultados”.

Sub-procesos	Preguntas de analítica	Variables respuesta	Actividades	Salidas
Precios	P1-P2. ¿Cuál es el precio (total y del m ²) de las casas?	1. Precio total de las casas ofrecidas 2. Precio del m ² de las casas ofrecidas	Resumir medidas y gráficos descriptivos de naturaleza univariada.	Cuadros resumen, con: media, mediana, media recortada (2.5%), mín, máx,..., número de registros, departamentos y municipios. Histograma, densidad y Boxplot, en escalas originales y log.
Atributos	P3. ¿Qué atributos presentan las casas usadas?	Frecuencia de casas con cada atributo: Dicotómicos (Ej: con o sin jardín), politómicos (estrato), discretos (número de baños) y continuo: área construida.	Resumir las frecuencias o descriptivos univariados	Cuadros con % de casas ofrecidas para la variable categórica dicotómica. Gráfico de barras para variables categóricas politómicas o discretas. Histograma para variable continua.
Zonas	P4-P7. ¿Cuál es el precio (total y del m ²) de las casas, según departamento / municipio P8. ¿Cuáles son los mapas de precios, áreas y frecuencia de casas?	1. Precio total de las casas ofrecidas por zona. 2. Precio del m ² de las casas ofrecidas por zona. 3. Área construida en m ² de las casas ofertada por zona. 4. Muestra de casas por zona.	Segmentar estadísticos descriptivos, según zonas. Mapear el país, diferenciando niveles de las variables respuesta.	Cuadro con descriptivos (media, mediana,...) con segmentación por: departamentos o municipios. Para variables 1-3, Boxplot con opciones de segmentación. Mapa de colores del país. Con opción de gráfico para la media, la mediana o la desviación estándar de la variable elegida.
Series	P9-P10. ¿Cómo se viene comportando el precio (total y del m ²) de las casas? P11-16. ¿Cómo se viene comportando el precio (total y del m ²) y el área construida de las casas, según departamento / municipio?	1. Evolución de precios (total y del m ²) de las casas. 2. Evolución del precio del m ² de las casas. 3. Evolución del área construida de las casas.	Resumir la evolución de las variables por periodo, según: global, y zonas.	Gráfico de serie combinado o individual, para media o mediana según se elija, con opciones de segmentación por: global, por departamentos (todos o principales) y por municipios (todos o principales).
Segmento	P17-P19. ¿Cómo se distribuyen los atributos de las casas usadas, ofrecidas para la venta, según estrato, departamento y municipio? P20-P21. ¿Cómo se comporta el precio (total y del m ²) de las casas usadas, ofrecidas para la venta, según atributos de la vivienda?	A. % de casas con el atributo, segmentada. Atributos. Dicotómicos (Sí, No): Piso en baldosa mármol. Patio. Zona de paqueo. Red de gas,... B. Precio total y del m ² según atributos.	Resumir variables respuesta tipo “A”, segmentada por estrato y zona. Resumir variables respuesta tipo “B”, segmentada por atributos.	Gráficos (barras, líneas) y tablas para % de casas con el atributo, según: estrato, departamento, municipio. Gráficos de precio medio, con intervalos de confianza, según atributos. Con opción de conversión a log y filtro según % de registros. Tabla con descriptivos del precio medio de casas ofrecidas según atributos.
Grandeza	P22. ¿Qué métrica permite resumir la amplitud de las casas usadas, ofrecidas para la venta, a la vez que correlacione razonablemente con su precio de oferta?	Grandeza: se trata de una combinación lineal de atributos estructurales de la propiedad (área, habitaciones,...) usando Análisis de Componentes Principales.	Proveer un indicador latente del inmueble (grandeza), que además correlacione con el precio total	Tabla de resultados de análisis de componentes principales. Tablas descriptivas (media, mediana, media recortada (2.5%), mín, máx, Q1, Q3,...) de los puntajes de grandeza. Histograma.
Modelo	P23-P24. ¿Qué factores del inmueble y de su entorno ayudan a explicar razonablemente el precio de las casas usadas, ofrecidas para la venta? Y ¿Cuáles son sus efectos?	Log. (precio total de las casas usadas, ofrecidas para la venta)	Proveer un modelo de regresión, a partir de características del inmueble y de su entorno	Resumen de la regresión (predictores, efectos, errores, significancia, R ² ,...). Gráficos de ajuste y de diagnóstico, VIF para multicolinealidad y Test de normalidad.
Predice	P25. ¿Cuál se esperaría que fuese el precio medio de una casa usada para la venta, con determinadas características?	Predicción del precio total medio condicional de casas usadas, ofrecidas para la venta	Predecir el precio medio de la casa, usando el modelo, datos desde el usuario y métricas internas del sistema.	Tabla con valores para la predicción media y para límites superior e inferior de intervalo de confianza al 95% para la media, e indicando la cantidad de registros usados.

Visualización

Se diseñó una aplicación web, que emula cada uno de los subprocesos descritos (Figura 2). En la Figura 4 puede verse el lado izquierdo del *dashboard*, con los nombres de cada subproceso. Estos dan lugar a las salidas explicadas en la caracterización de los subprocesos (Tabla 1).

Figura 4. Ilustración del módulo de visualización interactiva de resultados: *dashboard* de Statihouse®.

EJECUCIÓN DE LOS SUBPROCESOS

La ilustración de los resultados particulares cubre cada subproceso. En la mayoría de los casos, a excepción grandeza, modelo y predice, los resultados abordan al menos una de las preguntas de analítica, expuestas en la caracterización de subprocesos (véase Tabla 1). Recuerdese que tales resultados fueron obtenidos a partir de una muestra de 24.935 registros, obtenidos entre enero 02 de 2017 – mayo 31 de 2017, pero el sistema tiene la capacidad de actualizarse diariamente. Para los subprocesos restantes (grandeza, modelo y predice), se presentan algunas evidencias de la visualización.

Procesado de precios

¿Cuál es el precio (total y del m²) de las casas usadas, ofrecidas para la venta en Colombia? (véase la Figura 5).

Procesado de atributos

¿Qué atributos presentan las casas usadas, ofrecidas para la venta en Colombia? (Véase la Figura 6).

Figura 5. Visualización de “Procesado de precios” de Statihouse®. Precios de casas usadas (en millones de pesos), ene.2016-may.2017.

Figura 6. Visualización de “Procesado de atributos” de Statihouse®. Casas usadas, ene.2016-may.2017.

Procesado de series

¿Cuál es el precio del m² de las casas usadas, Departamento? (véase la Figura 7).
ofrecidas para la venta en Colombia, según

Figura 7. Visualización de “Procesado de series” de Statihouse®, opciones de precio del m², departamento y gráficas separadas. Casas usadas, ene.2016-may.2017.

Procesado de zonas

¿Cuáles son los mapas de precios, áreas y frecuencia Colombia? (véase la Figura 8).
las casas usadas, ofrecidas para la venta en

Figura 8. Visualización de “Procesado de zonas” de Statihouse®, mediana y mapas. Casas usadas, ene.2016-may.2017.

Procesado de segmento

¿Cómo se distribuyen los atributos de las casas usadas, ofrecidas para la venta en Colombia, según estrato? (véase la Figura 9).

Figura 9. Visualización “Procesado de segmento” Statihouse®, atributos, opciones de estrato (de 1 a 6) y gráficas de barras. Casas usadas, ene.2016-may.2017.

¿Cómo se comporta el precio total de las casas usadas, ofrecidas para la venta en Colombia, según atributos? (véase la Figura 10).

Figura 10. Visualización de “Procesado de segmento”, precio total según atributos. Casas usadas, ene.2016-may.2017.

Procesado de grandeza

¿Qué métrica permite resumir la amplitud de las casas usadas, ofrecidas para la venta en Colombia, a la vez que correlacione razonablemente con su precio de oferta?

El subproceso de “procesado de grandeza” gira en torno a un nuevo indicador propuesto, llamado “grandeza”, el cual es una combinación lineal de diversas variables del inmueble. A modo de ejemplo, en la Figura 11 se presenta la distribución de frecuencias de este nuevo indicador (lado izquierdo), y, cómo este indicador correlaciona con el precio total de casas usadas (lado derecho).

Procesado de modelo

¿Qué factores del inmueble y de su entorno ayudan a explicar razonablemente el precio de las casas usadas, ofrecidas para la venta en Colombia? Y ¿Cuáles son sus efectos?

Figura 11. Visualización de “Procesado de Grandeza” de Statihouse® y correlación Grandeza-Precios. Casas usadas, ene.2016-may.2017.

A modo de ejemplo, en la Figura 12 se presenta solo una parte de la visualización de resultados de procesamiento del modelo, pues el detalle teórico y empírico de este y de otros componentes (grandeza y predice) forma parte de otras publicaciones.

Figura 12. Algunos de los elementos de la visualización de “Procesado de modelo” de Statihouse®. Casas usadas, ene.2016-may.2017

Mediante pruebas en diferentes muestras de registros de casas, el coeficiente de determinación siempre fue superior al 80%. Asimismo, no se hallaron desviaciones extremas de los supuestos básicos para el análisis de regresión (normalidad, homocedasticidad, etc.).

Procesado de predice

¿Cuál se esperaría que fuese el precio medio de oferta de una casa usada para la venta, con determinadas características?

Otro subproceso importante dentro del procesamiento y análisis es el de predicción. Este, justamente, permite predecir el precio medio de la casa y sus intervalos de confianza al 95% (véase la Figura 13).

Figura 13. Algunos de los elementos de la visualización de “Procesado de predice” de Statihouse®. Casas usadas, ene.2016-may.2017.

DISCUSIÓN

Con relación a los precios de las casas usadas, ofrecidas para la venta por medio de anuncios web, en la Figura 5 se destaca la forma distribucional. Tanto el precio total como el precio del metro cuadrado presentan una distribución asimétrica. En ambos casos se observa un sesgo hacia la derecha, mucho más pronunciado en el precio total que en el precio del metro cuadrado. Esto es de esperar, ya que según Krause & Lipscomb, los datos de precios inmobiliarios tienden a presentar algunas observaciones con precios extremos [15]. Asimismo, se trata de variables y objetos complejos, no homogéneos y para los que existe multiplicidad de factores del inmueble y de su entorno, que generan diferencias significativas. De ahí la estrategia común, en modelos de precios hedónicos, de tomar el logaritmo natural a los precios de los inmuebles [1, 4-5]. Asimismo, en vista de la asimetría de las distribuciones y de limitaciones de datos en el sector inmobiliario, otra práctica común es emplear la mediana como medida de tendencia central, en lugar de la media. En esta oportunidad, la visualización no solo reporta la mediana del precio (300 millones de pesos para precio total; 2.06 millones de pesos para precio del m²), sino también la media (421 millones de pesos para precio total; 2.28 millones de pesos para precio del m²), la media recortada al 2.5% (395.8 millones de pesos, precio total), y los cuartiles. Cabe recordar que la muestra constó de 24.935 registros. Otro aspecto por señalar, es que la patente US5361201A, de Jost et al., (1994) en United States [24], que ofrece, entre otros, un modelo para avalúo basado en redes neuronales, presenta una gráfica con tres histogramas de frecuencias. Uno de ellos comprende el precio de venta total, otro el área (en pies cuadrados) y el restante, el precio del pie cuadrado. Precisamente, las distribuciones expuestas en la Figura 5 son consistentes con las que se presentan en el documento de la patente citada.

Por otro lado, la consistencia de las frecuencias de los atributos de los inmuebles o de su entorno (Figura 6), con respecto a reportes de otras fuentes no fue tan directa como con los precios. Esto, debido a que los informes secundarios disponibles o estudios previos no acostumbran a reportar estas estadísticas. No obstante, fue posible localizar algunas métricas para el atributo (del entorno) estrato socioeconómico. En la Figura 14 se presentan tres diagramas de barras. Los de los

Figura 14. Comparación indirecta – exploratoria entre el comportamiento del atributo estrato socioeconómico arrojado por Stathiouse® y el construido con información de referencia, capturada desde el reporte “Análisis inmobiliario 2016-2017” de Catastro Bogotá [25].

extremos fueron construidos a partir del reporte de “Análisis inmobiliario 2016-2017” publicado por Catastro-Bogotá [25]. La gráfica del medio corresponde a la distribución de las casas usadas, ofrecidas para la venta en Colombia, arrojada por Stathiouse®, como se mostró en la Figura 6.

Si bien las variables que resumen los tres gráficos expuestos en la Figura 14 son diferentes, se espera que estén relacionadas y sirvan de medio de exploración indirecta de la consistencia de los resultados arrojados por Stathiouse® para este atributo (estrato socioeconómico). Por lo mismo, la distribución del total de predios residenciales en la ciudad capital (Bogotá) y el área construida de nuevos predios residenciales en tal ciudad, pueden servir como un posible mecanismo exploratorio de validación de criterio. Haciendo una analogía con elementos de la psicometría, este tipo de validación consiste en estudiar la asociación entre la medida sometida a prueba y criterios con los que se espera que esta se relacione. Nótese la similitud entre la distribución arrojada por Stathiouse® (Colombia) y las dos de referencia (Bogotá) (Figura 14). En las gráficas de los extremos, los porcentajes de casas en el estrato 1 (cerca del 5%), aunque son bajos en comparación con los de los estratos 2-4 (similar a lo arrojado por Stathiouse®), son mucho más altas que el porcentaje para estrato 1 ofrecido por Stathiouse® (cerca del 2%). Al respecto, vale señalar que un aspecto es ser dueño de una propiedad (distribución de predios) y otra es decidir colocarla en venta (Ej: ofrecimiento web). Asimismo, debe tomarse en cuenta que la unidad de análisis es la

casa usada ofrecida por internet y el estrato 1 puede tener menores posibilidades de acceso a internet, en comparación con los demás (si bien el 67% de los hogares en Colombia presenta (reporte al 2017) acceso a internet, el “alto costo” es el principal motivo para que la proporción restante no cuente con dicho servicio [26]). A pesar de las distinciones entre las dos variables de referencia (total de predios y área construida de nuevos precios; ambos residenciales y en Bogotá) y la variable objeto de análisis, en general se encuentran resultados consistentes, tomando en cuenta las formas de las distribuciones. Nótese que entre los estratos 2-4 el porcentaje de casas usadas, ofrecidas para la venta en Colombia desde la web, se ubicó cerca de 77%. Asimismo, el porcentaje para el total de predios residenciales en Bogotá (gráfico izquierdo), para dicho rango de estratos, está alrededor de 83%, y para el área construida de nuevos predios residenciales (gráfico de la derecha) está cerca del 71%.

Otro atributo para el que es viable explorar referentes, es el de posesión de red de gas (véase Figura 6). Al respecto, vale decir que al cierre del año 2015, cerca de 7.7 millones de hogares residenciales (7.9 millones en total) contaban con servicio de gas natural (Con base en informe de MINMINAS-Colombia) [27]. Y tomando en cuenta que se reporta un referente de cerca de 14 millones de hogares para dicho periodo (con base en [28]), se tendría un cubrimiento de alrededor de 55%. Según la “Encuesta Nacional de Calidad de Vida 2015” del DANE, “la cobertura de gas natural en el total nacional fue 62,5%, en las cabeceras la cobertura fue 76,1%, mientras que en

los centros poblados y rural disperso se presentó una cobertura de 11,7% de los hogares” [29]. Ahora bien, el porcentaje de casas usadas reportadas por Statihouse®, ofrecidas para la venta en Colombia por medio web entre enero 2017 – mayo 2017, que cuentan con red de gas, fue de 51% (véase Figura 6). Para ambos referentes el porcentaje proporcionado por Statihouse® fue inferior al de estos (11.5 % menos respecto al informe del DANE [29]; y 4% menos respecto a MINMINAS [27]). No obstante, también es de esperar que la cobertura haya aumentado de 2015 a 2017, por lo que las diferencias pueden ser mayores. Estas diferencias parecen razonables, tomando en cuenta que las cifras de los referentes están considerando todos los tipos de hogares residenciales, lo que también incluye apartamentos (nuevos y usados) y casas (nuevas y usadas). En cambio, el caso de ejemplo solo comprende casas usadas. Los apartamentos, al tender a ser más nuevos, modernos y ubicados en zonas más desarrolladas, se espera que tengan mayor cobertura de gas, en comparación con las casas usadas. Otra eventual explicación, es que para una parte de los anuncios de venta de casas se haya omitido este atributo (red de gas), bien sea por desconocimiento del que lo publica, o por simple omisión intencional. Sin embargo, se esperaría que estos últimos motivos fueran una proporción mínima de casos, ya que el mismo oferente por la web estaría restándole “valor” a su oferta. En sí sería útil nuevos referentes que permitan segmentar casas usadas, con el fin de realizar validaciones más pertinentes para la ocurrencia del atributo “red de gas”.

Pasando al procesado de series, una de las visualizaciones ofrece el comportamiento global de

la mediana del precio del m², cada mes, para ocho departamentos específicos y una categoría adicional que representa a los demás (véase Figura 7). El periodo de ilustración es de solo cinco meses, por lo que puede ser poco para notar cambios sustanciales. No obstante, se nota un comportamiento más estable en algunos (Ej: Antioquia y Cundinamarca) que en otros (Ej: Valle del Cauca y Santander). Asimismo, sobresalen valores más altos en Cundinamarca, Antioquia y Atlántico, en comparación con los de Valle del Cauca y Norte de Santander. La identificación de datos públicos que cubran un periodo tan actualizado (ene.2017 - may.2017) son escasos en Colombia, por lo que los contrastes con referentes se dificultan (al momento del estudio). No obstante, se sabe que el precio de la vivienda en Bogotá (ciudad capital, localizada en Cundinamarca) tiende a ser el más alto del país. Para explorar contrastes adicionales, se recurrió a la visualización por municipio (Statihouse®), pero ampliando el periodo a enero 2016 – mayo 2017 (Figura 15).

De acuerdo con una publicación del diario El Tiempo, el 08 de febrero de 2016, titulado “Más oferta hará que precios de vivienda nueva sigan cediendo” [30], se reportó que “la ciudad en la que más se vio la caída fue en Medellín, en donde los precios de los inmuebles nuevos bajaron 2,6 por ciento real, seguidos por los de Bogotá, que tuvieron una variación negativa real del 2,41 por ciento.”. Si bien este segmento de bienes (nuevos) no corresponde a los de exposición en este artículo (casas usadas), ni tampoco al periodo de interés (ene.2017 – may.2017), ante limitantes de informes actualizados y abiertos,

Figura 15. Visualización de “Procesado de series”, con opciones de precio del m², mediana, gráficas separadas y municipio, pero para un periodo de mayor tamaño, Casas usadas, ene.2016 – may.2017.

sirve de referente exploratorio. Nótese, en la Figura 16, la inclinación a la baja para el precio del metro cuadrado en Bogotá, Medellín y Cúcuta, en comparación con los demás casos. Se reitera que estos contrastes son solo a modo exploratorio, pues los casos, unidades de análisis y periodos no son necesariamente comparables.

Con relación a los mapas de precios, áreas y frecuencias (Figura 8), el interior del país tiende a presentar la mayor concentración de precios altos, en comparación con las zonas restantes. Asimismo, como es de esperar, Cundinamarca (que incorpora Bogotá, ciudad capital) reúne una cantidad de oferta de casas usadas muy superior a los demás departamentos.

Enfatizando sobre procesado de segmento, los resultados son consistentes con la lógica del constructo “estrato socioeconómico”. Véase, en la Figura 9, cómo a medida que aumenta el estrato, tiende a aumentar la ocurrencia de atributos que infieren “valor” al inmueble. Entre ellos: red de gas, patio, cocina integral, pago de administración, garaje y jardín. Todos estos, con una tendencia a la alza, a medida que aumenta el estrato. En ese mismo subproceso de segmento, pero pasando a explorar el precio según atributos (Figura 10), véase que ante la presencia de atributos, el precio del inmueble tienden a ser mayor (Ej: a medida que aumenta el estrato, el precio total también lo hace). Asimismo, el precio total se muestra mayor en las ciudades principales (Bogotá, Medellín, Cali, Barranquilla), que fuera de estas. Estos análisis no deben confundirse con que las variaciones en los precios son causadas por dichos atributos, ya que pueden deberse a otros factores con los que estos correlacionan. Es decir, la asociación no debe tomarse como causalidad.

Respecto al procesado de grandeza, considere la forma de la distribución de este indicador (Figura 11), dando lugar a una enorme cantidad de aplicaciones, gracias a que se cumplen las propiedades de la distribución normal y a que resume, en una sola métrica, diversas variables del inmueble (74 % de la información de estas). Cabe anotar que la grandeza típica de una casa usada en Colombia, ofrecida para la venta por medio web, es de ocho (8) puntos. En general, estas propiedades oscilan aproximadamente entre 5 puntos y 11 puntos. Dicha métrica puede emplearse para clasificar, comparar y caracterizar

las casas, los mercados, los ámbitos geográficos, por mencionar algunos. Sobre todo, esta métrica resulta de utilidad, también por su alta asociación con el precio total del inmueble (Figura 11). Se esperaría que un modelo de regresión, con solo grandeza como predictora, explique cerca del 59 % del logaritmo del precio del inmueble (correlación al cuadrado).

Pasando a los dos últimos subprocesos (modelo y predice), solo se proveen algunas visualizaciones a modo de ejemplo (Figuras 12-13), pues el detalle será compartido en otras publicaciones. Nótese el ajuste razonable, de cerca del 82% de explicación de la varianza (R^2) y esto lo logra con solo cuatro predictores, uno de los cuales es la grandeza. Atendiendo a oportunidades detectadas en la literatura previa, el modelo combina un constructo y variables observables. Esto ayuda a superar eventuales problemas de colinealidad y favorece una mejor explicación de la respuesta. Este mismo modelo es dispuesto internamente para predecir el precio medio de casas usadas, condicionadas por los valores de los predictores, lo cual puede apoyar los procesos de avalúo. Fruto de exploraciones adicionales, se destaca la consistencia del R^2 cuando se consideran otros periodos: enero 01 - 08 de 2018 (84.3%; 1.214 obs); 01 de julio 2017 – 08 de enero de 2018 (81.79%; 33.140 obs). También los demás resultados del modelo de regresión son consistentes con lo ya expuesto en el artículo.

A modo de discusión general sobre el sistema en su conjunto, vale señalar que en la revisión de literatura realizada no se ha encontrado un sistema multitareas (descriptivas, comparativas, evolutivas, correlacionales y predictivas) que ejecute de forma automática, para el sector inmobiliario, desde la captura hasta la visualización de los datos, usando insumos de anuncios web y ciencia de datos. Se han encontrado aportes metodológicos muy particulares, en su mayoría con datos de encuestas o de publicaciones de entidades (fuentes secundarias de datos estructurados). Los pocos trabajos localizados, que se acercan a la atarea de caracterizar el sector o sus precios (bienes raíces) a nivel de la propiedad, en su mayoría son de China y Estados Unidos. No obstante, la mayoría de estos trabajos apunta a aspectos específicos, como una ecuación con alguna técnica de inteligencia artificial o regresión para el estudio de efectos o predicciones; o estudios de caso para modelar alguna variable específica. De hecho,

el acercamiento de la ciencia de datos al sector inmobiliario, usando datos masivos de internet es declarado en infancia.

Ahora bien, pasando, por ejemplo, a aplicaciones web que existan en Colombia, la mayoría se enfocan en proveer medios para que los usuarios encuentren viviendas o hagan ofertas de estas. Es decir, el alcance y foco de atención no es caracterizar el sector y predecir precios en tiempo real o cercano, usando datos masivos de anuncios web y ciencia de datos para apoyar otro tipo de decisiones estratégicas.

Por otro lado, el 13/02/2017 se hizo una exploración sobre patentes, bajo el siguiente algoritmo: [https://patents.google.com/?q=TI%3d\"real+estate\"&q=\"big+data\", \"data+mining\", \"data+science\"&type=PATENT](https://patents.google.com/?q=TI%3d\)

Se recuperaron 40 patentes (en proceso, aprobadas o vencidas). De estas, 5 se relacionaron de forma directa con el tópico. Las demás se enfocaron sobre sistemas o métodos para apoyar la comercialización de bienes raíces, o para programar citas, o para intercambio de información. Esta búsqueda se complementó con otra más flexible, no restringida al título, y se logró introducir tres resultados más, para un total de 8 patentes de referencia [24, 31-37]. En el anexo 1 se resumen dichos desarrollos y se comparan con el sistema propuesto.

CONCLUSIONES

Se ha desarrollado un sistema (Statihouse®) que caracteriza estadísticamente propiedades en venta y predice su precio de oferta (en tiempo real o cercano), empleando datos masivos de la internet, ciencia de datos y una aplicación web. Statihouse® brinda la posibilidad de aprovechar grandes cantidades de datos que diariamente se generan en medios web, por lo que se distingue de aplicaciones tradicionales, delimitadas en su mayoría a formatos estructurados y menores conjuntos de datos. Este sistema resulta novedoso, pues no se encontraron antecedentes que ejecuten de forma automática desde la recolección de datos web, hasta la visualización (en tiempo real o cercano), con un alcance multitareas de tipo descriptivo, comparativo, evolutivo, correlacional y predictivo en el sector inmobiliario, con enfoque en datos de la propiedad.

Para la tecnología Statihouse® se ilustró los formatos de visualización, aportando resultados derivados de una muestra de 24.935 casas usadas, los cuales fueron contrastados con algunos referentes a modo exploratorio (no confirmatorio), debido a la escasez de informes lo suficientemente actualizados (ene.2017-may.2017). Este sistema puede escalarse, bajo versiones no estándar, para otros inmuebles (Ej: apartamentos) y tipos de oferta (Ej: arrendamiento).

El sistema desarrollado articula: 1) Un nuevo método de recolección automática de datos de anuncios Web. 2) Un nuevo método de limpieza de datos, basado en cartas de control y segmentación. 3) Una nueva métrica latente, que resume variables clásicas, observadas en la vivienda, propias de la amplitud, y que presenta alta correlación con el precio de oferta de los inmuebles. 4) Un nuevo modelo de regresión de precios hedónicos, que incorpora el constructo grandeza, pero además, variables directamente capturadas a partir del bien inmueble y de su entorno. 5) Una nueva interfaz de visualización de resultados. Vale destacar que el sistema puede actualizarse diariamente de forma automática; es decir, no emplea una muestra estática, sino que provee estadísticas periódicamente y según el lapso que elija el usuario. Puede verse que Statihouse © es consistente con las demandas de procesos de “orquestración de recursos”, mediante la articulación lógica, sistemática y novedosa de capacidades, tecnologías, datos y métodos, en este caso, basados en ciencia de datos [20].

Respecto a las implicaciones prácticas del sistema Statihouse®, es de señalar el uso de los procesos en cuestión para ayudar en problemas de: 1) Altos costos en la captura de los datos (encuestadores, entrenamiento, tabulación, revisión, procesamiento, entre otras tareas). 2) Demora entre la obtención del dato y la publicación de información, lo cual afecta la toma de decisiones oportuna. 3) Captura de datos basada en muestreos, que no alcanzan a aprovechar la diversidad y riqueza de la información disponible. 4) Alta subjetividad en los procesos de avalúos.

Como oportunidades de estudio, están: validar los resultados arrojados por cada subproceso, con métricas más directas y actualizadas. Publicar con detalle teórico y empírico, cada uno de los componentes emergentes de Statihouse® (método de control de

calidad, indicador grandeza, modelo de precios hedónicos). Escalar el presente desarrollo tecnológico al caso de otros tipos de bienes (apartamentos, terrenos, locales, entre otros). Incorporar nuevos métodos, de machine learning (árboles de regresión, random forest,...) para la función predictiva y comparar con respecto a la regresión hedónica. El sistema también arroja resultados en formato de tablas (Ej: precios y diferentes estadísticas por departamento y municipio), los cuales pueden ser usados como datos de entrada para la aplicación de diversas técnicas multivariantes (Ej: análisis clúster), con el fin de proveer nueva información y/o practicar en la docencia con casos reales. Este sistema, a nivel de estructura general, características técnicas y componentes, se constituye en un caso de éxito de la ciencia de datos en el sector inmobiliario, y sirve de guía y estímulo para estudiantes, docentes, profesionales e investigadores, de cara a nuevos desarrollos. Así, se contribuye al necesario cambio de paradigma en cuanto a información, tecnología y toma de decisiones en dicho sector.

Entre las limitaciones, están: el caso de ejemplo solo cubrió casas usadas, esto dificulta las comparaciones con métricas que incluyen otros tipos de inmuebles. Otra limitante es el posible sesgo de anuncios por internet, ya que una proporción de casas usadas para la venta, puede no estar siendo ofrecida por este medio, aunque en plena sociedad de la información y conocimiento, se esperaría que fueran mínimas. Adicionalmente, los precios de las transacciones finales de compra – venta de las propiedades no están disponibles a la misma velocidad y abundancia que los anuncios por internet. Futuros trabajos deberían buscar la posibilidad de estudiar las diferencias entre el precio ofertado por internet y el precio de la transacción, y establecer algún factor de corrección. El escalamiento de esta tecnología precisa conocimiento avanzado en campos de ciencia de datos, a la medida para este contexto específico; por tanto, es posible que al trasladarse a otros países, sectores, entre otros contextos, amerite adaptaciones especializadas, sobre todo para la captura, estructuración y limpieza de los datos.

REFERENCIAS

- [1] k. Adekunle and P. Omoruyi. “A Review of Real Estate Valuation and Optimal Pricing Techniques”. *Asian Economic and Financial Review*. Vol. 4 N° 12, pp. 1878-1893. 2014.
- [2] A. Court. “Hedonic price indexes with automotive examples. The dynamics of automobile demand”. New York. General Motors. 1939.
- [3] K. Lancaster. “A new approach of consumer theory”. *Journal of Political Economy*. Vol. 74 N° 2, pp. 132-157. 1966.
- [4] H. Ishijima and A. Maeda. “Real estate pricing models: theory, evidence, and implementation”. *Asia-Pacific financial markets*. Vol. 22 N° 4, pp. 369-396. 2015.
- [5] J. Núñez, J. M. Caridad y Ocerin and F. Rey. “Artificial neural networks for predicting real estate prices”. *Revista de Métodos Cuantitativos para la Economía y la Empresa*. N° 15, pp. 29-44. 2013.
- [6] H. Donner, H. Song and M. Wilhelmsson. “Forced sales and their impact on real estate prices”. *Journal of Housing Economics*. N° 34, pp. 60-68. 2016.
- [7] A. Krause. *Reproducible Research in Real Estate: A Review and an Example*. *Journal of Real Estate Practice and Education*. Vol. 19 N° 1, pp. 69-85. 2016.
- [8] S. Malpezzi. “Hedonic Pricing Models: A Selective and Applied Review”. In *Housing Economics and Public Policy* (eds T. O’Sullivan and K. Gibb). Blackwell Science Ltd, Oxford, UK. 2002.
- [9] D. Gujarati. *Econometría*. 2da, ed. McGraw-Hill, Bogotá, 1990.
- [10] J. Guan, D. Shi, J. Zurada, & A. Levitan. *Analyzing Massive Data Sets: An Adaptive Fuzzy Neural Approach for Prediction, with a Real Estate Illustration*. *Journal of organizational computing and electronic commerce*. Vol. 24 N° 1, pp. 94-112. 2014.
- [11] G. James, D. Witten, T. Hastie and R. Tibshirani. “An Introduction to Statistical Learning: With Applications in R”. 6th Ed., Springer. New York. 2015.
- [12] T. Oladunni and S. Sharma. “Hedonic Housing Theory—A Machine Learning Investigation”. In *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on IEEE*, pp. 522-527. 2016.
- [13] X. Wang and J. Zhang. “Principal Component Analysis of Influencing Factors of the Development of China’s Real Estate Market”. In *ICCREM 2013: Construction and Operation*

- in the Context of Sustainability, pp. 1027-1035. 2013.
- [14] H. Shi. "Determination of Real Estate Price Based on Principal Component Analysis and Artificial Neural Networks". In Intelligent Computation Technology and Automation, 2009. ICICTA'09. Second International Conference on IEEE. Vol. 1, pp. 314-317. 2009.
- [15] A. Krause and C. Lipscomb. "The Data Preparation Process in Real Estate: Guidance and Review". Journal of Real Estate Practice and Education. Vol. 19 N° 1, pp. 15-42. 2016.
- [16] H. Varian. "Big data: New tricks for econometrics". The Journal of Economic Perspectives. Vol. 28 N° 2, pp. 3-27. 2014.
- [17] A.J. Holland. "Integrating Data Science and Commercial Real Estate". NAL partners, 2016, Fecha de consulta: 5 de junio de 2017. URL: http://www.naipartners.com/Portals/248/DATA_SCIENCE_BRO.pdf
- [18] J. Sharfman. "Big Data and Real Estate". Part I. California Association of Realtors. Executive Report, 2015. Fecha de consulta: 5 de junio de 2017. URL: http://www.car.org/ccre/pdf/Big_Data_Executive_Report.pdf
- [19] J. Barney. "Firm resources and sustained competitive advantage". Journal of Management. Vol. 17, pp. 99-120, 1991.
- [20] D. Teece, G. Pisano and A. Shuen. "Dynamic capabilities and strategic management". Strategic Management Journal. pp. 509-533. 1997.
- [21] R. Project. "A language and environment for statistical computing". R Foundation for Statistical Computing, Austria. 2008.
- [22] R Studio Team, R. Studio: Integrated Development for R. RStudio, Inc., Boston, USA. 2015.
- [23] RStudio, Shiny. "A web application framework for R" URL: <http://shiny.rstudio.com>
- [24] A. Jost, J. Nelson, K. Gopinathan. C. Smith. "Real estate appraisal using predictive modeling". US Patent & Trade Mark Office. US5361201A. United States. 1994-11-01.
- [25] Unidad Administrativa Especial de Catastro Digital. "Análisis Inmobiliario 2016-2017". Fecha de consulta: 14 de junio de 2017. URL: https://www.catastrobogota.gov.co/sites/default/files/Resultados_Censo_2017%20version%20final.pdf
- [26] Agencia EFE. "¿Qué porcentaje de hogares en Colombia tiene hoy acceso a internet? Esto afirmó MinTIC". Discurso del Ministro de las Tecnologías de la Información y las Comunicaciones, David Luna, en la inauguración en Cartagena del Congreso Internacional TIC, Andicom 2017. 23 de agosto de 2017. Fecha de consulta: 9 de enero de 2018. URL: <http://www.wradio.com.co/noticias/economia/colombia-alcanza-un-67--de-cobertura-en-internet-segun-encuesta-tic/20170823/nota/3557964.aspx>
- [27] MINMINAS. "Cobertura del servicio de gas natural - I trimestre de 2015". Fecha de consulta: 14 de junio de 2017. URL: <https://www.minminas.gov.co/documents/10180/1157884/CoberturasGasNaturalCombustible-I+T+r+i+m+e+s+t+r+e+2+0+1+5+.pdf/f392f86a-c13f-45b6-be6b-ee83aaf5a919>
- [28] M. Sánchez. "Colombianos no cambian el cable por Netflix". Publicado el 23 de abril de 2016 en La República. Fecha de consulta: 13 de junio de 2017. URL: http://www.larepublica.co/colombianos-no-cambian-el-cable-por-netflix_371826
- [29] DANE. "Encuesta Nacional de Calidad de Vida – ECV- 2015". Fecha de consulta: 15 de junio de 2017. URL: <https://www.dane.gov.co/index.php/estadisticas-por-tema/salud/calidad-de-vida-ecv/encuesta-nacional-de-calidad-de-vida-ecv-2015>
- [30] O. Ahumada. "Más oferta hará que precios de vivienda nueva sigan cediendo". Publicado el 08 de febrero de 2016 en El Tiempo. Fecha de consulta: 13 de junio de 2017. URL: <http://www.eltiempo.com/archivo/documento/CMS-16503596>
- [31] N. Packes, E. Benoit. F. Talos. "Real estate evaluating platform methods, apparatuses, and media". US Patent & Trade Mark Office. US20150242747A1. United States. 2015-08-27.
- [32] C. Smith. "Automated Real Estate Valuation System". US Patent & Trade Mark Office. US20160292800A1. United States. 2016-10-06.
- [33] L. Mo, Y. Li and Q. Wu. "Real estate value estimation method based on artificial neural network statistic model". State Intellectual Property Office. CN103578057A. China. 2014-02-12.
- [34] R. Feenan, J. Kelly and M. Popova. "System and Method for Visualizing Real Estate Markets". US Patent & Trade Mark Office. US20150262202A1. United States. 2015-09-17.
- [35] M. Robbins. "Process for automated real estate valuation". US Patent & Trade Mark Office. US20010039506A1. United States. 2001-11-08.

- [36] X. Hao. “Real estate price calculating method and system”. State Intellectual Property Office. CN105590239A. China. 2016-05-18.
- [37] P. Lecomte. “Framework for modeling real estate assets based on genetics”. US Patent & Trade Mark Office. US20080249797A1. United States. 2008-10-09

ANEXOS

Anexo 1. Exploración de algunas patentes de referencia y comentarios respecto a las diferencias con Stathiouse®

ID	Título	Observaciones
[31] US20150242747A1 Fecha de publicación: 2015-08-27. Estados Unidos.	“Real estate evaluating platform methods, apparatuses, and media”	Determinar precios de bienes inmobiliarios a partir de redes neuronales. No corresponde a un sistema que automatice desde la recolección hasta la visualización, ejecutando además multitareas (descriptiva, comparativa, evolutiva, correlacional y predictiva). Se enfoca básicamente hacia lo predictivo.
[32] US20160292800A1 Fecha de publicación: 2016-10-06. Estados Unidos.	“Automated Real Estate Valuation System”	Es un sistema de avalúo de bienes raíces para propiedades específicas. Busca propiedades con características similares a la propiedad específica, reportadas por el usuario, y forma grupos. A partir de ellos, genera el valor del precio de la vivienda y describe varios atributos de las casas y zonas vecinas. Usa “Random Forest Regression Method”. El sistema es operable desde celulares, etc., y aprovecha tecnologías móviles y GPS. Alcance: procesamiento, análisis y visualización. No se enfoca en la captura de datos desde la web, limpieza y estructuración automática. Los predictores del modelo difieren de los empleados en Stathiouse®, a excepción de algunos relacionados con la amplitud del inmueble. No combina, en un mismo modelo, constructos y atributos observables.
[33] CN103578057A Fecha de publicación: 2014-02-12. China.	“Real estate value estimation method based on artificial neural network statistic model”	Método de estimación de avalúos de precios basado en redes neuronales. Alcance: predicción y visualización. No captura de forma automática datos de anuncios web, ni ejecuta procesos automatizados de limpieza. Tanto el enfoque como las etapas, el alcance y el tipo de técnica usada difieren del sistema postulado.
[34] US20150262202A1 Fecha de publicación: 2015-09-17. Estados Unidos.	“System and Method for Visualizing Real Estate Markets”	Un sistema y método para coleccionar, procesar, analizar y visualizar datos masivos del mercado de bienes raíces. Alcance: el enfoque tiende a ser macroeconómico, e incluye precios, tasas de interés, índice de transacciones comerciales, entre otros atributos, y genera rankings, los cuales también normaliza de forma automática y los clasifica para facilitar la comparación por geografía, entre otras variables. Este desarrollo dista considerablemente del postulado. No tienen el mismo objetivo. Si bien provee una enorme riqueza de datos de internet, el enfoque es macroeconómico (Ej: mercados de países, regiones). No se enfoca en actividades automáticas de recolección, limpieza, procesamiento, análisis y visualización para caracterizar atributos de las propiedades en venta ni modelar el precio de oferta según atributos, ni en actividad predictiva para propiedades específicas.
[35] US20010039506A1 Fecha de publicación: 2001-11-08. Estados Unidos.	“Process for automated real estate valuation”	Se enfoca en el avalúo de propiedad raíz, usando modelación predictiva. Usa regresión múltiple. Alcance: captura, procesamiento, análisis y resultados. La recolección de datos no es automática ni proviene de anuncios por internet (“comienza con la recopilación de datos mediante la obtención de registros, preferiblemente una copia del expediente de registro público (tal como los recogidos por el evaluador), de parcelas de bienes raíces dentro de un área definida.”). Tampoco aprovecha las bondades de la visualización de datos que existe con las tecnologías actuales. Si bien usa regresión múltiple, los modelos son diferentes. No consideran los mismos predictores ni la forma de la variable respuesta.
[36] CN105590239A Fecha de publicación: 2016-05-18. China.	“Real estate price calculating method and system”	Predicción de precios usando método y corrección de precios de referencia. No tiene el alcance del trabajo. Se enfoca básicamente hacia lo predictivo.
[24] US5361201A Fecha de publicación: 1994-11-01. Estados Unidos.	“Real estate appraisal using predictive modeling”	Modelo de predicción de precios basado en redes neuronales, que incorpora características de la propiedad, de las vecindades y niveles de especificaciones geográficas. Alcance: limpieza de datos, predicción y visualización de reportes de descripción de tendencias y de efectos de los factores. No incluye un sistema automático multitareas para captura, limpieza, estructuración, procesamiento, análisis y visualización de datos/información en tiempo real o cercano, usando datos de anuncios por internet. Tampoco usa los mismos métodos cuantitativos del trabajo postulado. Los modelos de predicción difieren en predictores y forma. El módulo de visualización también difiere significativamente del postulado. Presenta un interesante método de limpieza de registros (conforme, no conforme), consistente en la mediana ± 2 sigma para el precio del m ² de las propiedades vecinas. No obstante, este se diferencia sustancialmente del que emplea Stathiouse® en cuanto a técnicas usadas, parámetros, tipo de segmentación, variable respuesta, entre otros.

<p>[37] US20080249797A1 Fecha de publicación: 2008-10-09. Estados Unidos.</p>	<p>“Framework for modeling real estate assets based on genetics”</p>	<p>Es un modelo de predicción basado en algoritmos genéticos. Alcance: solo predicción. No incluye, de forma automática, captura, limpieza, estructuración, procesamiento, análisis y visualización de datos/información en tiempo real o cercano, usando datos de anuncios por internet. Tampoco usa los mismos métodos cuantitativos del trabajo postulado. Al comparar el alcance predictivo de dicho trabajo con el postulado, también se encuentran notables diferencias. El trabajo de referencia usa algoritmos genéticos, pero en el trabajo postulado es regresión múltiple con un constructo y variables observables para el log del precio total. Asimismo, hay diferencias entre los predictores incluidos en los modelos.</p>
---	--	--