

Universitat Politècnica de Catalunya
Facultat de Matemàtiques i Estadística

Trabajo de final de máster

Predicción de precios de alquiler en la ciudad de Barcelona

Geraldo Gariza Gala

Director: Jesús Corral Lopéz

Estadística aplicada

Resumen

Palabras clave: Modelos bayesianos, API, web scrapping, precios hedónicos, Shiny.

La búsqueda de alquiler en las principales ciudades, tanto españolas como europeas, se ha intensificado en las últimas décadas. Este trabajo se centra en el caso de Barcelona e intenta, de forma empírica, encontrar un modelo estadístico que mejor se ajuste a los precios de alquiler por barrios. Se hace una comparativa entre diferentes modelos lineales, frequentistas y bayesianos, y se discuten sus ventajas y desventajas. Luego se desarrolla una aplicación web usando la librería Shiny de R, con dos secciones: la primera, donde se puede explorar los precios medios por barrio en un mapa interactivo; y la segunda, donde se puede predecir el precio medio de un alquiler dadas ciertas características como el barrio, metros cuadrados, número de habitaciones, etc. La aportación más significativa de este proyecto es la creación, desde el inicio hasta el final, de un producto basado en datos que los usuarios pueden usar para mejorar su búsqueda de vivienda.

Abstract

Keywords: Bayesian models, API, web scrapping, hedonic prices, Shiny.

The search for rental properties in major cities, both in Spain and across Europe, has intensified in recent decades. This work focuses on the case of Barcelona and seeks, empirically, to find a statistical model that best fits neighborhood rental prices. A comparison is made between different linear, frequentist, and Bayesian models, and their advantages and disadvantages are discussed. Subsequently, a web application is developed using the Shiny library in R, with two sections: the first, where users can explore average neighborhood prices on an interactive map, and the second, where the average rental price can be predicted given certain features such as neighborhood, square footage, number of rooms, etc. The most significant contribution of this project is the creation, from start to finish, of a data-driven product that users can utilize to enhance their housing search.

Índice general

1. Introducción	1
2. Recogida de datos	2
Datos Idealista	2
Datos Open Data Barcelona	2
Otras fuentes de datos	3
3. Tratamiento de los datos	4
Exploración de datos: Idealista	4
Exploración de datos: Open Data Barcelona	6
4. Modelización	7
Mínimos cuadrados ordinarios	7
5. Inferencia Bayesiana y Modelos Utilizados	10
Fórmula de Bayes e inferencia bayesiana	11
Información a priori	11
Modelo lineal bayesiano agrupado	12
Modelo lineal bayesiano no agrupado	13
Modelo jerárquico bayesiano	14
Modelo jerárquico bayesiano con covariable	15
6. Selección del modelo	16
Metricas de ajuste	17
7. Aplicación Shiny	19
Librería Shiny	19
Nuestra aplicación	20
8. Conclusiones	23
Bibliografía	25

1. Introducción

La búsqueda de alquiler en las principales ciudades, tanto españolas como europeas, se ha intensificado en las últimas décadas. Motivada principalmente por las preferencias de los ciudadanos a vivir en ciudades más grandes, con un mayor número de oportunidades laborales, académicas o de ocio. Esta situación se ha ido agravando aún más, después de la pandemia del covid, una elevada inflación y malestar por parte de los ciudadanos que ven cada vez más difícil acceder a una vivienda. Así a mediados del 2022 se aprobo en España la llamada "ley de la vivienda" que limita el precio de alquileres en zonas tensionadas del país].¹.

El objetivo principal de este trabajo es crear una aplicación web que ayude, ya sea a los demandantes o a los oferentes de vivienda, a determinar el precio de alquiler, dadas ciertas características, en Barcelona. Además, se quiere realizar un análisis empírico de los modelos bayesianos jerárquicos, comparando su potencia y utilidad frente a los modelos lineales.

El trabajo se divide en cuatro partes: La primera, extracción de datos de la página Idealista.com opendata-ajuntament.barcelona.cat/es usando métodos web scraping y utilización interfaces de programación de aplicaciones (API a partir de ahora).

La segunda se centra en la limpieza] y creación o modificación de variables.² Se decidió dedicar un bloque entero a esta tarea ya que el resultado de los modelos dependerá de la calidad de los datos usados.

La tercera parte agrupa la modelización e inferencia bayesiana. La primera es el uso de un modelo lineal utilizado para selección de variables y limpieza de valores atípicos, también será nuestro modelo baselina a comparar con el resto de modelos. El segundo bloque describe los modelos bayesianos y los diferentes tipos usados en este trabajo. Luego se compara todos los modelos y se argumenta la elección. Una consideración importante a la hora de seleccionar el modelo será la robustez de este y como se comportaría con nuevos datos o zonas.

En la cuarta parte se desarrolla un aplicación web donde se utiliza el modelo seleccionado para generar predicciones con las características dadas por el usuario, como por ejemplo: el barrio, número de habitaciones, metros cuadrados, etc.

Por último se discutirá los objetivos alcanzados, los trabajos futuros y una conclusión final.

¹Moncloa, gobierno de España: <https://www.lamoncloa.gob.es/serviciosdeprensa/notasprensa/transportes/Paginas/2023/040523-nueva-ley-vivienda-2023.aspx>

²El 80 % de la ciencia de datos es limpieza de datos: <https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=488794836f63>

2. Recogida de datos

La recogida de datos de este proyecto se divide en dos partes: web scrapping de la página web Idealista.com y recogida de datos de la página Open Data Barcelona.

Idealista es una web conocida que opera en España, Italia y Portugal. En este sitio web los usuarios pueden colgar anuncios de pisos para vender o alquilar pisos. A su vez, los demandantes de pisos pueden contactar con los propietarios para llegar a un acuerdo.

La segunda fuente de datos es una web del ayuntamiento de Barcelona donde se publican bases de datos de la ciudad. Estos se van actualizando mensual, semestral o anualmente. Como ejemplo podemos encontrar datos tan diversos como el número de árboles por barrio, el número de vehículos activos en la ciudad o datos de ingresos por unidad familiar.

Datos Idealista. Se optó por utilizar la técnica de web scrapping en Idealista porque su acceso mediante API era limitado a cincuenta peticiones mensuales con cincuenta observaciones cada una. Esto era una limitación ya que solo en Barcelona ciudad, en la web de Idealista, hay publicados alrededor de 3500 inmuebles destinados al alquiler.

Esta técnica consiste en extraer datos en un formato llamado lenguaje de marcado de hipertexto (html) de la página web en cuestión y guardarlos para su posterior uso. El lenguaje de programación usado para esta tarea fue Python por preferencia personal. Una de las dificultades encontradas en esta sección fue la diversidad de páginas web, esto hizo que el éxito de la extracción fuera a base de prueba y error.

Otra de las dificultades fue la limitación que marcan los sitios web dedicados al comercio. Esta limitación existe concretamente para evitar explotar gran cantidad de datos usando web scrapping. Por lo tanto, se aplicaron ciertas técnicas de pausa aleatoria en los códigos de extracción para evitar esta restricción.

La periodicidad de extracción fue mensual, empezando en noviembre de 2022 hasta agosto de 2023.³

Datos Open Data Barcelona. La página web del ayuntamiento de Barcelona mantiene alrededor de 600 conjuntos de datos abiertos a la ciudadanía. Abarca datos económicos, poblacionales, territoriales, entre otros. Su actualización varía dependiendo de la fuente, por ejemplo los datos de contaminación suelen ser diarios, los datos económicos mensuales y los de infraestructuras anuales.

El motivo de usar este recurso de datos se explica con la teoría de precios hedónicos.[HM10] Herath, Shanaka, and Gunther Maier. "The hedonic price method in real estate and housing market research: a review of the literature." (2010): 1. Esta teoría sostiene que el valor de los inmuebles no solo se rigen por el valor intrínseco de estos (como nº habitaciones, metros cuadrados...), sino también de la información

³Se pretende seguir extrayendo datos y tener el proyecto actualizado.

relativa a la ubicación del inmueble (nº de hospitales, zonas verdes...) o indicadores socioeconómicos de los territorios, como el nivel de renta de las familias.

Para la extracción se utilizó la API del sitio web. Se define las API como: "Las API son mecanismos que permiten a dos componentes de software comunicarse entre sí mediante un conjunto de definiciones y protocolos."⁴. En nuestro caso comunicarse se refiere a la extracción y posterior almacenamiento de datos. Esta es una forma sencilla de obtener información y altamente automatizable.

El incoveniente en este apartado fue el gran volumen de datos de diferente temática. Con más de 500 fuentes de datos, la labor más ardúa fue separar el grano de la paja. Se detectó primero de forma manual, los conjuntos más relevantes y luego de forma empírica en los modelos. Se descartaron los datasets que no mostraban una mejora significativa en la modelización.

Gran parte de la información proveniente del portal Open data es fija con actualizaciones menos frecuente. Esto a lo referente a parques, playas, bares o información económica que se actualiza de forma trimestral o anual. En cualquier caso, en este proyecto se utilizan los datos más recientes, automatizando las llamadas a la API.

Otras fuentes de datos. A parte de las dos mencionadas, también se necesitaba datos en formato ".shp" de la ciudad de Barcelona. El formato ".shp" shapefile por sus siglas en inglés[], es formas archivo de formas, desarrollada por la compañía ESRI⁵ y es utilizado en programas geográficos para definir formas o ubicaciones de unidades geográficas.

En este proyecto se utilizó un formato .shp de la ciudad de Barcelona delimitada por barrios para la presentación de resultados y el desarrollo aplicación web.

⁴Información sobre API del Amazon web services: <https://aws.amazon.com/es/what-is/api/>

⁵Explicación de shapefiles por ESRI: <https://enterprise.arcgis.com/es/portal/latest/use/shapefiles.htm>

3. Tratamiento de los datos

Una vez creada las diferentes fuentes de datos pasamos a su limpieza y análisis. El primer paso en esta tarea es normalizar los nombres de los diferentes barrios. Las tres fuentes usan similares pero diferentes nombres para definir cada barrio. En algunos casos, como en el portal Idealista, agrupa territorios pequeños, formando una nueva unidad territorial. En el caso del portal Open Data sigue un patrón estandar y normbra cada barrio con su nombre oficial. Algunos de los ejemplos encontrados fueron: *besòs* siendo el nombre oficial a *besòs maresme* o *can peguera* a *can peguera turó peira*, el primero no cuenta con el nombre completo del barrio y en el segundo cuando no hay mucha oferta de inmuebles Idealista une barrios colindantes en uno solo. Se creo, por lo tanto, una nueva variable tipo mapeo, para poder relacionar las tres partes ente sí.

Realizada el mapeo de nombres, procedemos a unir las bases de datos de Idealista, Open Data y el shapefile en un solo dataset. Por ultimo filtramos las observaciones sin número identificación del barrio, cabe resaltar que los datos de Open Data son oficiales y cuentan con nombres únicos para cada barrio y con una identificación (*id*) facilitando exclusión de zonas no pertenecientes a Barcelona, como por ejemplo: *Teatinos, Playa de Palma, La Torrasa, Santa Eulàlia*, etc.

Exploración de datos: Idealista. Después de un tratamiento general y unificación de las fuentes de datos, pasamos a su análisis, comprensión y limpieza o filtrado, si cabe.

La ciudad de Barcelona cuenta con setenta y dos barrios de los cuales aproximadamente hay sesenta y cinco con observaciones en Idealista.com, esto debido a poblaciones o territorios mas pequeños o destiandos a la industria, como es el caso de La Marina del Prat Vermell, donde hay más actividad empresarial que vecinal.

Los dos datasets utilizados para el análisis son los extraidos en junio y julio de 2023, al ser los más recientes en el momento de escribir esta tesis. El primero fue destinado a entrenar (*training sample*) los modelos y el segundo al conjunto de testo (*test sample*). Se decidió focalizar los esfuerzos en los datos más actuales ya que han habido varios cambios en la legislación de la vivienda de alquiler, como topes de precios en áreas tensionadas, siendo así el análisis lo más actual posible.

A continuación se muestra una tabla con el resumen de los datos de Idealista.com.

Variable	Descripción	Tipo	Tipo de Dato
log_price	Logaritmo natural del precio	Dependiente	Numérica
barri	Barrios	Independiente	Categorica
log_smt	Logaritmo natural Metros Cuadrados	Independiente	Numérica
asc	Si hay ascensor	Independiente	Binaria
rooms2_0	Estudio, 0 hab.	Independiente	Binaria
rooms2_1	Pisos con 1 hab.	Independiente	Binaria
rooms2_2	Pisos con 2 hab.	Independiente	Binaria
rooms2_3	Pisos con 3 hab.	Independiente	Binaria
rooms2_4	Pisos con 4 hab. o más	Independiente	Binaria
new_planta	Variable creada con el número de planta	Independiente	Binaria
flag_planta	Dummy variable donde se creo "new_planta"	Independiente	Binaria
wc2_1	Pisos con 1 baño	Independiente	Binaria
wc2_2	Pisos con 2 baños	Independiente	Binaria
wc2_3	Pisos con 3 baños	Independiente	Binaria
wc2_4	Pisos con 4 baños o más	Independiente	Binaria
terraza	Si tiene Terraza	Independiente	Binaria
exterior	Si es exterior	Independiente	Binaria
amueblado	Si esta amueblada	Independiente	Binaria
lujo	Si precio es mayor a 5000 euros	Independiente	Binaria

TABLA 1. Descripción de los datos de Idealista.com

En la figura 1 vemos una gráfica box plot de los distintos barrios de Barcelona y el precio de los alquileres en logaritmo natural. Esta transformación es frecuente en economía y ayuda a la normalización de los datos, ya que los precios de alquiler suelen seguir una distribución asimétrica positiva, al arrastrar los valores extremos hacia la derecha.

También podemos observar que el número de registros por barrios es muy diferente. Barrios como Canyelles o Verdun cuentan con muy pocas observaciones, menos de diez. En cambio, barrios como la Eixample son lo que tienen más observaciones. Además el gráfico esta ordenado de mayor a menor precio de alquiler para ayudar a su compresión, teniendo la media más elevada Pedralbes y la menor Canyelles. Por último, vemos que la diversidad de precios cambia con los barrios. Por ejemplo, en Diagonal Mar i El Front Marítim el cincuenta por ciento de las observaciones se encuentran en una horquilla más amplia, indicando más dispersión, que barrios como Montbau.



FIGURA 1. Box plot por barrios y el precio del alquiler en logaritmo natural

Procesamiento de variables. Se realizaron diferentes ajustes al dataset inicial, mediante análisis exploratorio de datos se determinó:

- Solo usar observaciones con mas de 10 metros cuadrados. Se encontraron anuncios con 0 mt² debido, presumiblemente, a errores de la web o en la extracción.
- Se añadio la variable Casa o chalet como dummy extraida del nombre del anuncio.
- Se considero alquiler de lujo a partir de viviendas con precio mayor a 5000 euros al mes. Esta distinción también se encuentra en Idealista.
- Se considero como estudio si la variable nº de habitaciones era igual a zero.
- Se transformó la variable baños y nº habitaciones a factor, siendo 1,2,3,4 o más en ambos casos.
- Las variables precio y metros cuadrados se transformaron a logaritmo natural para mantener la normalidad.
- Se efectuaron más cambios, pero al no ser significativos, no se describieron.

Exploración de datos: Open Data Barcelona. En la página web Open Data encontramos diversos datasets con información de la ciudad de Barcelona. En este estudio utilizaremos los siguientes conjuntos de datos:

- Renta neta por unidad familiar Barcelona 2020.
- Tasa de paro mensual por barrios Barcelona 2023.
- Hospitales y centros de atención primaria.
- Arbolados (viarios, parques, plazas etc.) en la ciudad Barcelona.
- Espacios de música y bares de copas.

Estos fueron los principales conjuntos de datos extraídos de la web. Nuevamente, esta lista fue más extensa al principio del proyecto, pero a base de prueba y error se fueron añadiendo o descartando datasets.

Las principales variables extraídas fueron los totales por barrio y distrito de cada conjunto de datos. Por ejemplo, se añadio el número de hospitales, Caps y Cuaps por barrio y distrito.

4. Modelización

La modelización se divide en dos partes, en la primera se optó por un modelo de mínimos cuadrados ordinarios (OLS en inglés) añadiendo progresivamente variables hasta alcanzar el máximo R-cuadrado ajustado. La segunda parte se divide en los diferentes modelos lineales bayesianos utilizados:

- (1) Modelo lineal agrupado.
- (2) Modelo lineal no agrupado.
- (3) Modelo jerárquico con variación en el intercepto.
- (4) Modelo jerárquico con variación en el intercepto y covariable.

Mínimos cuadrados ordinarios. En la primera parte de la modelización se utilizó el método de Mínimos Cuadrados Ordinarios (*MCO*). Este enfoque suele ser el más utilizado para analizar datos de tipo económico, ya sea por su facilidad en la interpretación o por compensar su sencilla implementación con la potencia de los resultados. La formula general de un modelo lineal multiple usando mínimos cuadrados es:

$$y = \beta X + \epsilon$$

Donde la y es la variable respuesta o dependiente del modelo, la X los diferentes predictores o variables independientes, la β vector de coeficientes a calcular y ϵ el vector de errores.

En nuestro caso, la y son los precios de alquileres a predecir y la X podría ser el número de habitaciones de cada piso o los metros cuadrados. La facilidad en su interpretación yace $precio = \beta_1 + \beta_2 * metros_cuadrados + \epsilon$, siendo un modelo lineal especificado para nuestro caso. Un aumento en una unidad en los *metros_cuadrados* aumentará en β_1 el precio de la observación. Además, podemos hacer el modelo más complejo añadiendo nuevas variables $precio = \beta_1 + \beta_2 * metros_cuadrados + \beta_3 * n_habitaciones + \epsilon$ como por ejemplo el número de habitaciones. Esta linearidad en la interpretación hace que los modelos lineales fáciles de usar.

El marco teórico detrás de este enfoque es el método de precios hedónicos (*HPM* a partir de ahora). Los HPM no sólo cuantifican las características individuales de un bien, como en nuestro caso los metros cuadrados, número de habitaciones, etc., sino que también ayuda a comprender cuánto contribuye al precio cada una de estas variables. Además, mide las características externas del bien, en nuestro caso el barrio donde se encuentra el piso o el nivel de renta. Uniendo estas dos partes se obtiene una regresión hedónica que puede ayudar a cuantificar el valor de un bien. Por lo tanto, se hace conveniente el modelo lineal múltiple para analizar este tipo de activos.

La intuición es que a más variables añadamos al modelo, más poder de predicción tendrá. El inconveniente con el método de MCO es que tiene unos supuestos muy estrictos:

- (1) La varianza de los errores debe ser homocedástica.
- (2) No colinealidad de las variables explicativas.
- (3) Los errores no pueden estar correlacionados.

Si analizamos estos supuestos y los aplicamos al problema planteado vemos que añadir nuevas variables explicativas infinitamente chocaría con el supuesto 2, algunas de las variables explicarían parcial o totalmente lo mismo. Entonces, añadir variables sin una métrica y análisis previo no es una buena práctica. Por otro lado, en este trabajo tratamos de predecir los precios de alquiler de pisos en Barcelona. Cabe esperar que los pisos próximos entre sí tengan cierta relación, por ejemplo del mismo barrio, esto incumpliría el supuesto 3 de correlación de errores.

Incumplir con algunos de estos supuestos, intuitivamente, comportaría por ejemplo: si la varianza no es homocedástica en los errores, es decir, la varianza de los errores va cambiando entre las observaciones, el valor medio del modelo \hat{y} será más preciso para algunas y tendrá más error para otras. Por lo tanto la inferencia o predicción no será precisa.

Por estos motivos, construimos un modelo lineal múltiple incremental, empezando desde el modelo más sencillo sólo con un predictor, añadiendo una nueva variable en cada iteración. Este modelo se usará de proxy para detectar outliers y calcular el Factor de Inflación de la Varianza (*VIF*)⁶[Jam+13], por sus siglas en inglés, que mide la multicolinealidad entre las variables. Posteriormente, después de esta selección de variables y detección de observaciones atípicas continuaremos con los modelos bayesianos.

El Factor de la inflación de la varianza mide el nivel de multicolinealidad entre las variables independientes.

$$VIF_i = \frac{1}{1 - R_i^2}$$

⁶ En este caso se crea un modelo por cada la variable independiente X_i . Mientras más alto sea el valor R^2 la ratio será mayor lo cual indicará alta colinearidad. Un valor de 1 indicará no colinearidad y normalmente se usa el umbral de 5 o 10 para

⁶ R_i^2 es la varianza explicada por el modelo.

alta colinealidad. Repetimos el proceso con el resto de variables y así podemos tener un idea de que variables añadir en el modelo.

La motivación para usar esta técnica como selector de variables se verá con más detalle en el apartado de modelización bayesiana, pero se puede adelantar que este método es computacionalmente intensivo y esta pre selección nos permite ahorrar tiempo y recursos.

La distancia de Cook detecta observaciones influyentes en el modelo[Coo77].

$$D_i = \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2}{ps^2}$$

Donde $\hat{y}_{j(i)}$ es la estimación de la j th variable respuesta quitando la observación i , \hat{y}_j la estimación de la observación j th. Por lo tanto, el numerador es la suma de la diferencia de las j th observaciones al cuadrado. La p es el número de variables independientes del modelo y la s^2 es el error cuadrático medio.

Las observaciones influyentes pueden distorsionar el valor de los coeficientes $\hat{\beta}$ obtenidos. Por lo tanto, usar esta técnica nos ayudará también en la selección de variables.

Módelo lineal multiple. En esta sección se mostrará un breve resumen y se analizará el modelo de regresión multiple elegido y se explicará los resultados obtenidos.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
aire	0.03	0.01	3.17	0.00	0.01	0.04
amueblado	0.19	0.01	23.33	0.00	0.18	0.21
asc	0.07	0.01	6.60	0.00	0.05	0.09
calef	0.06	0.01	7.05	0.00	0.04	0.08
exterior	-0.06	0.01	-5.74	0.00	-0.08	-0.04
flag_planta	0.05	0.01	4.23	0.00	0.03	0.07
lujo	0.65	0.02	34.36	0.00	0.61	0.68
rooms21	0.17	0.02	8.80	0.00	0.13	0.21
rooms22	0.27	0.02	13.96	0.00	0.23	0.31
rooms23	0.27	0.02	13.34	0.00	0.23	0.31
rooms24 o mas	0.24	0.02	10.22	0.00	0.19	0.28
square_mt	0.00	0.00	25.49	0.00	0.00	0.00
terraza	0.09	0.01	11.40	0.00	0.08	0.11
wc22	0.17	0.01	16.84	0.00	0.15	0.19
wc23	0.31	0.02	16.01	0.00	0.27	0.35
wc24 o mas	0.28	0.03	9.59	0.00	0.22	0.34

TABLA 2. Variables del modelo lineal

En la tabla 2 podemos observar el resultado del modelo. Vemos que todas las variables utilizadas son significativas al p valor <0.05 . También observamos el signo de cada una de las variables, esto es importante porque nos ayuda a entender que peso y que dirección aporta cada variable al modelo. Vemos que la única negativa es $exterior = -0.06$. Se puede interpretar que dependiendo de la altura del piso

y si exterior hay más ruido. Sin embargo, cuando agregamos una interacción con estas dos variables el resultado no es significativo. Al final se decidió dejar solo la variable exterior en el modelo. Otra variable interesante es $lujo = 0,65$ que captura las viviendas de más de 5000 euros. Por último, destacar la interpretación del modelo. Al estar la variable dependiente en logaritmos⁷ la interpretación es la siguiente⁷: Si tomamos la exponencial de la variable lujo $\exp(0,65) = 1.915$, quiere decir que de media si una vivienda es considerada de lujo aumenta su precio en un 91.5 % su precio.

variables	VIF	df	VIF ajustado
barri	3.37	64.00	1.01
square_mt	3.20	1.00	1.79
asc	1.30	1.00	1.14
rooms2	2.87	4.00	1.14
wc2	4.09	3.00	1.26
terraza	1.18	1.00	1.09
exterior	1.14	1.00	1.07
amueblado	1.24	1.00	1.11
flag_planta	1.11	1.00	1.05
aire	1.22	1.00	1.11
calef	1.17	1.00	1.08
lujo	1.69	1.00	1.30

TABLA 3. Factor de la inflación de la varianza

En la tabla 3 vemos el resultado del VIF. Vemos que ninguna variable supera el umbral de 5, moderada colinealidad. Solo algunas variables tipo factor tienen valores entre el 3 y el 4 pero para estas el VIF ajustado a los grados de libertad es bajo. La única variable que supera el 3 sin ser factor es *square_mt* siendo que las otras variables independientes explican relativamente bien los metros cuadrados. De igual manera se decidió dejarla en el modelo, siendo la variable explicativa principal junto con *barri*.

5. Inferencia Bayesiana y Modelos Utilizados

En esta sección haremos una breve introducción a la inferencia bayesiana. Luego, explicaremos los diferentes modelos utilizados. Nuestro propósito principal es alcanzar una modelización óptima que nos ayude a realizar buenas predicciones con datos futuros.

Esta sección sigue el siguiente orden: La selección de la distribución a priori de cada parámetro y modelo. Para pasar luego a los modelos: Modelo lineal bayesiano agrupado, Modelo bayesiano no agrupado, Modelo jerárquico bayesiano y Modelo jerárquico bayesiano con covariante. Como podemos observar los modelos incrementan en dificultad, empezando con el modelo más simple y con el objetivo de identificar el mejor trade-off entre potencia y complejidad.

⁷Interpretación de log log models: <https://data.library.virginia.edu/interpreting-log-transformations-in-a-linear-model/>

Por último, se escogerá el 'mejor' modelo siguiendo criterios de reducción error, mejor convergencia y sensatez de los coeficientes.

Fórmula de Bayes e inferencia bayesiana. La principal motivación para utilizar modelos bayesianos está en la capacidad de incorporar información previa o a priori, ya sea de estudios realizados anteriormente o basada en conocimiento experto en la materia. Esta idea se encapsula en la fórmula de Bayes[Gel+95]:

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)}$$

Donde θ representa los parámetros desconocidos, y son los datos observados, $p(\theta)$ es la distribución a priori de θ , $p(y|\theta)$ es la verosimilitud y $p(y)$ es la evidencia o la distribución marginal de los datos.⁸.

El software utilizado en este proyecto es Stan⁹, un programa desarrollado en C++ que emplea una variante de cadenas de Markov llamada Hamiltonian Monte Carlo. Este método extrae muestras que convergen hacia una distribución común. Si el modelo especificado converge, es una señal de que está bien especificado y se puede confiar en los resultados obtenidos.

Información a priori. En esta sección explicamos la metodología utilizada para escoger las distribuciones a priori y fuentes consultadas.

Una parte crucial del análisis bayesiano es la información a priori que el investigador o estudios anteriores puedan ofrecer. En este trabajo, se utilizaron distribuciones *Cauchy*(0, 10) para los interceptos y *Cauchy*(0, 2,5) para el resto de las variables. Esta decisión se basó en el documento ".^A Weekly Informative Default Prior Distribution For Logistic and Other Regression Models" de Andrew Gelman et al. (2008)[Gel+08]. En dicho estudio, tras analizar tres tipos de modelos logísticos de diferentes ámbitos, como política, ciencias médicas y sociales, se concluyó que una distribución a priori adecuada para modelos lineales es la Cauchy previamente descrita.

Una distinción importante entre el documento consultado y nuestro análisis es el tratamiento de las variables numéricas y categóricas. Gelman sugiere estandarizar las variables numéricas continuas y la dependiente a una media de 0 y una desviación típica de 0.5. Asimismo, propone transformar las variables binarias para que tengan una proporción en el rango (-1,1). En nuestro estudio, solo hemos transformado las variables numéricas y la dependiente usando el logaritmo natural, dejando las variables binarias sin cambios. La transformación sugerida por el autor no generaba una mejora significativa en la predicción, por lo que decidimos optar por un modelo más sencillo y con la menor cantidad de modificaciones posibles.

⁸Información extraída del libro: Bayesian Data Analysis, Andrew Gelman, 3rd edition

⁹Para más información: <https://mc-stan.org/users/documentation/>

term	estimate	std.error	rhat	ess
b0	4.98	0.0574	1.00	1526
log_mt	0.495	0.0146	1.00	1401
rooms2_1	0.135	0.0218	1.00	1568
rooms2_2	0.0930	0.0220	1.00	1417
rooms2_3	-0.00433	0.0237	1.00	1417
rooms2_4	-0.0673	0.0272	1.00	1430
asc	0.0821	0.0105	1.00	4259
wc2_2	0.177	0.0109	1.00	2354
wc2_3	0.332	0.0209	1.00	2008
wc2_4	0.370	0.0310	1.00	1984
terraza	0.0538	0.00884	1.00	3860
amueblado	0.230	0.00851	1.00	3592
lujo	0.714	0.0197	1.00	3224
sigma_y	0.293	0.00282	1.00	3736

TABLA 4. Covariables utilizadas en el modelo agrupado

Modelo lineal bayesiano agrupado. Como primer paso, se planteará un modelo de regresión lineal múltiple bayesiano, a partir de ahora modelo agrupado:

$$y \sim \text{normal}(\alpha + \beta * X_n, \sigma)$$

¹⁰.

Siendo y el valor inferido, α la constante, X_n la matriz de diseño, es decir las variables usadas, β el vector de coeficientes a estimar y σ la desviación típica del error del modelo. El subíndice n indica que habrá n observaciones para $n \in N$.

Las variables utilizadas para ajustar el modelo agrupado fueron las que se muestran en la tabla 1, se muestra una salida del modelo a continuación:

Este modelo se denomina 'agrupado' porque no hace distinciones entre los diferentes barrios como covariables. Los modelos bayesiano usando la función `stan_model()` y `sampling()` del paquete de R `rstan` necesitan como parametros el modelo especificado en formato texto, una lista de datos con las variables y termino dependiente, las cadenas con la que se corra el sampleo Hamiltonian Monte Carlo y el número de iteraciones. Ejemplo de código de este modelo agrupado: `sampling(model, data = data_list, chains = 4, iter = 2000)`.

En la tabla 4 podemos observar los coeficientes obtenidos por el modelo agrupado, usa la media de los barrios como $b0$. Vemos que la mayoría de los coeficientes tienen sentido, excepto para $rooms2_3$ y $rooms2_4$ que presentan coeficientes negativos, sobretodo los pisos con 4 habitaciones o más. La intepretación sería a más número de habitaciones menor precio, lo cuál no parece lógico. Esto podria deberse a una mala especificación del modelo o podría tener alguna explicación estadística, ya que al agregar también la variable $lujo$ esta podría estar absorviendo el signo positivo

¹⁰Para su forma compacta de matriz siguiendo una distribución normal con error σ .

term	estimate	std.error	rhat	ess
log_smt	0.424	0.0137	1.02	263
rooms2_1	0.112	0.0198	1.00	3011
rooms2_2	0.130	0.0203	1.00	3047
rooms2_3	0.0849	0.0221	1.00	2880
rooms2_4	0.0371	0.0253	1.00	2306
asc	0.0390	0.00980	1.00	8595
wc2_2	0.127	0.0103	1.00	1811
wc2_3	0.252	0.0198	1.00	1536
wc2_4	0.301	0.0289	1.00	1768
terraza	0.0747	0.00805	1.00	8209
amueblado	0.205	0.00796	1.00	10059
lujo	0.670	0.0184	1.00	3476

TABLA 5. Covariables utilizadas en el modelo no agrupado

de $rooms2_4$. Pisos más lujosos suelen ser más grandes y por lo tanto con más habitaciones.¹¹

Por otro lado, el modelo parece haber convergido bien, al tener los $rhat$ iguales a 1. El $rhat$ es una métrica que nos indica que el coeficiente ha convergido bien en las cadenas y el número de iteraciones cuando mas cercano a 1 sea. Además las ess muestras efectivas, que se recomiendan ser mayor al diez por ciento del número de iteraciones, es alto para todas las variables. Esto es esperable al haber especificado un modelo relativamente sencillo y con muchas observaciones.¹²

Modelo lineal bayesiano no agrupado. La principal diferencia con el modelo anterior es que en el modelo 'no agrupado' se plantea una regresión por barrio, en total 65, ya que se removieron los barrios con menos de 10 observaciones para reducir el ruido o error de las predicciones futuras. Además, como veremos más adelante los modelos jerárquicos permiten generar predicciones en grupos no presentes en los datos pero si en la población.

La especificación del modelo es la misma que el aparato anterior solo que en vez de una constante $b0$ se usan β_{n} con $n \in N$ igual al numero de barrio modelado. En código r: `sampling(model, data = data_list, chains = 4, iter = 4000, verbose = True, seed = 123)`. Se ha aumentado el número de iteraciones ya que es un modelo más complejo y para algunos zonas hay pocas observaciones. También se ha puesto una semilla 'seed' ya que al ser un modelo probabilista los resultados puede cambiar ligeramente, además de aporta reproducibilidad.

A continuación se muestra una salida de los coeficientes del modelo no agrupado:

Para la representación del modelo no agrupado se han removido con coeficientes de barrio, ya que dificulta la lectura al tener el modelo en total 79 covariables. En la tabla 5 podemos observar dos cosas importantes, primero el número de muestras efectivas para log_smt es bajo, ya que al tener 4000 iteraciones esperamos al menos

¹¹Este es una simple hipótesis y sera contrastada más adelante con otros modelos.

¹²Gráficos con convergencias del modelo se muestran en el apéndice.

400 *ess*. Segundo, los coeficientes de *rooms2_3* y *rooms2_4* ahora son positivos. El bajo número de muestras efectivas podría deberse al menor número de observaciones para ciertos barrios. Por otro lado, seguiremos el comportamiento de las variables *rooms2_3* y *rooms2_4* en los modelos posteriores.

Modelo jerarquico bayesiano. En esta sección se presenta el modelo jerarquico y su definición, se explica la metodología seguida y por último se analizan los resultados.

Esta sección sigue la metodología descrita en GELMAN, Andrew. Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics*, 2006, vol. 48, no 3, p. 432-435 [Gel06]. En este artículo Gelman describe el problema de gas radon en los condados de Estados Unidos. Altos niveles de concentración de este gas pueden llegar a ser cancerigenos. El estudio se centra en usar modelos jerarquicos para predecir el valor de radon en los hogares. Los parecidos con el problema que intentamos modelar son evidentes.

Primero, tenemos datos geograficos, en nuestro caso barrios, en los que queremos predecir el nivel de precios de alquiler. Segundo, el número de observaciones varia según la región y se decide utilizar una covariable que modele el efecto mixto de las variables barrios. Por último, hay regiones que no están entre los datos o se han decidido no modelar por falta de muestras/ no muestras. Por estos motivos, hemos decidido adoptar la metodología usada por Gelman como nuestro modelo base.

La expresión matematica del modelo es:

$$y_{i,j} \sim N(\beta_0 + b_{0,j} + \beta_1 x_{i,j}, \sigma^2) \quad b_{0,j} \sim N(0, \tau^2)$$

Dónde:

- $y_{i,j}$ es la observación i en el grupo j .
- N denota la distribución normal.
- β_0 es el intercepto global.
- $b_{0,j}$ es el efecto aleatorio del grupo j .
- β_1 es el coeficiente de la variable $x_{i,j}$.
- σ^2 es la varianza del error.
- τ^2 es la varianza del efecto aleatorio.

Y donde $j \in J$ representa a los 66 barrios utilizados en el modelo. La motivación para utilizar este modelo, es que se encuentra a medio camino del modelo agrupado y no agrupado. Es decir, no utiliza el mismo intercepto para todos los grupos como en el primer modelo, ni realiza un modelo por cada grupo, con la perdida de precisión que supone para los barrios con menos observaciones o *overfitting* sobre ajuste de los barrios más poblados.

A continuación se muestra la salida de este tercer modelo:

La especificación del modelo en R es: `sampling(model, data = data_list, chains = 4, iter = 5000, verbose = TRUE, seed = 1568)`. Se han aumentado ligeramente el número de iteraciones al ser un modelo más complejo. Lo que nos sorprende de

term	estimate	std.error	rhat	ess
log_smt	0.419	0.0145	1.02	205
rooms2_1	0.0502	0.0114	1.00	2665
rooms2_2	-0.0213	0.0168	1.00	2581
rooms2_3	-0.116	0.0268	1.00	2266
rooms2_4	-0.226	0.0410	1.00	2378
asc	0.0440	0.00974	1.00	8808
wc2_2	0.132	0.0103	1.00	1840
wc2_3	0.253	0.0195	1.00	1458
wc2_4	0.286	0.0281	1.00	1434
terraza	0.0756	0.00824	1.00	8911
amueblado	0.207	0.00791	1.00	10117
lujo	0.670	0.0181	1.00	2759

TABLA 6. Covariables utilizadas en el modelo jerarquico

esta salida es que 3 de los 4 coeficientes para *rooms2* son negativos. También las muestras efectivas para *log_smt* siguen siendo bajas. Por lo general, parece que el modelo converge bien.

En general, los modelos analizados hasta ahora nos dan resultados diferentes para la variable *rooms2_x* y similares para el resto de variables.

Modelo jerarquico bayesiano con covariable. Hasta este punto solo hemos utilizado información intrínseca del bien, características de los pisos en alquiler. En este modelo agregaremos una covariable al modelo que genera los diferentes interceptos para cada barrio. Usaremos los datos extraídos del portal Open Data Barcelona, ya que creemos que información de los grupos (barrios) ayudaran a mejorar la inferencia del modelo.

Matemáticamente:

$$\begin{aligned} y_{i,j} &\sim N(\beta_0 + b_{0,j} + \beta_1 x_{i,j}, \sigma^2) \\ b_{0,j} &\sim N(\gamma_0 + \gamma_1 z_j, \tau^2) \end{aligned}$$

Dónde:

- $y_{i,j}$: Es la respuesta observada para el individuo i en el grupo j .
- β_0 : Es el intercepto fijo global.
- $b_{0,j}$: Es el efecto aleatorio del intercepto para el grupo j .
- β_1 : Es la pendiente fija asociada a la covariable x .
- $x_{i,j}$: Es el valor de la covariable para el individuo i en el grupo j .
- σ^2 : Es la varianza de los errores.
- γ_0 : Es el intercepto de la relación lineal entre $b_{0,j}$ y la covariable z_j .
- γ_1 : Es la pendiente de la relación lineal entre $b_{0,j}$ y la covariable z_j .
- z_j : Es una covariable a nivel de grupo que afecta el intercepto aleatorio $b_{0,j}$.
- τ^2 : Es la varianza de los efectos aleatorios $b_{0,j}$.

Como podemos observar este modelo tiene muchos más parametros que los modelos anteriores. Intuitivamente, este modelo sugiere que existe una variable dentro del

term	estimate	std.error	rhat	ess
log_smt	0.422	0.0141	1.00	357
rooms2_1	0.112	0.0197	1.00	3580
rooms2_2	0.129	0.0204	1.00	3659
rooms2_3	0.0824	0.0220	1.00	3533
rooms2_4	0.0336	0.0255	1.00	3192
asc	0.0402	0.00976	1.00	8905
wc2_2	0.131	0.0104	1.00	2317
wc2_3	0.258	0.0195	1.00	1944
wc2_4	0.308	0.0283	1.00	2067
terraza	0.0741	0.00813	1.00	7762
amueblado	0.207	0.00805	1.00	9270
lujo	0.674	0.0178	1.00	3417

TABLA 7. Covariables utilizadas en el modelo jerarquico con covariable

modelo que se relaciona linealmente con el valor del intercepto para cada barrio. Esta inferencia a nivel de barrio, que asumimos sigue una distribución normal, e influencia la variación en el intercepto del modelo general.

Con la introducción de esta variable, que discutiremos más adelante, pretendemos conseguir una estimación más precisa para los precios de los pisos en alquiler de Barcelona, ya que añade información que antes no habíamos considerado.

En nuestro dataset de variables externas de los barrios de la ciudad, decidimos usar el *nivel de renta medio por barrio* como única covariable en el nivel superior del modelo jerarquico de dos niveles. Se probaron diferentes covariables y esta resultó la más efectiva a nivel de poder de precisión. La utilización de por ejemplo la covariable *playa* si el barrio tiene playa (Poblenou,Brcloneta, etc) tanto en el nivel superior como de variable en el modelo principal aumentaba el error del modelo.

En la tabla 7 vemos que este modelo a pesar de ser más complejo tiene un mayor número de muestras efectivas para casi todas las variables. Además, la estimación puntual de *rooms24* es ahora positiva aunque muy pequeña, en el percentil 2,5% es negativa con valor -0.02. A pesar de este inconveniente este modelo es el que más sentido tiene y el que mejor converge.¹³

6. Selección del modelo

El propósito de este trabajo es encontrar el modelo más sencillo y con potencia predictora suficiente para que sea útil. En este caso, tenemos que definir una métrica/s que nos permitan escoger entre los 4 (5 si consideramos el modelo lineal de la sección 5) modelos que hemos especificado.

En esta sección presentaremos y argumentaremos las métricas utilizadas para la selección del modelo final y los intervalos de credibilidad de las diferentes variables.

¹³Sin tener en cuenta el modelo agrupado, que al ser mas sencillo no tiene problemas de convergencia.

Model	R^2	R_{adj}^2	RMSE
lm	0.763	0.759	933.935
pooled	0.746	0.746	965.386
no_pooled	0.706	0.702	1039.901
hierarchical	0.623	0.618	1176.690
hierarchical_cov	0.712	0.707	1029.722

TABLA 8. Métricas de rendimiento de los modelos

Metricas de ajuste. La dos métricas utilizadas en este trabajo serán la raíz del error cuadratico medio (a partir de ahora RMSE) por sus siglas en ingles y el R_{adj}^2 R cuadrado ajustado como metricas de bondad del ajuste.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

El RMSE mide la diferencia entre el valor real de las observaciones y_i y el valor a predecir \hat{y}_i , elevado al cuadrado obteniendo la varianza de esta diferencia, luego tomando la raíz cuadrada poniendo el error en terminos de desviación típica y en la misma unidad que la variable dependiente. Mientras menor sea el error mejor será el modelo especificado.

$$R_{adj}^2 = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - k - 1} \right)$$

El R cuadrado, intuitivamente, mide la varianza explicada por los regresores del modelo entre la varianza total del modelo con una sola variable, la media. Cuando lo ajustamos, tenemos tambien en cuenta el tamaño muestral n y el número de covariables k . Sus valores están comprendidos entre 1 y 0, siendo 1 el máximo ajuste.

Estas dos medidas fueron escogidas por su simplicidad en la interpretación y por us amplia utilización en modelos lineales.

En la tabla 8 podemos observar estas dos metricas para cada modelo especificado. Vemos tambien que el modelo con mayor R_{adj}^2 y RMSE es el modelo lineal múltiple. Sin embargo, debemos tener en cuenta algunas consideraciones:

Este modelo no es capaz de agregar nueva información e inferir nuevos datos. Por ejemplo, al predecir los nuevos valores tuvimos que quitar: $df = filter(bassi! = "CiutatMeridiana")$ del conjunto de datos, ya que en el conjunto de entrenamiento no teniamos datos para este barrio. Esto no ocurre en los modelos jerarquicos.

Otra consideración, es que hay mucha disparidad en los tamaños muestrales entre barrios. Uno de nuestros objetivos poder predecir con fiabilidad, independientemente de la zona donde se quiera comparar los precios. Si podemos ajustar bien un barrio con muchos datos, pero no uno que actualmente esta en crecimiento o tiene poca oferta de viviendas, el objetivo de esta análisis resultaría poco útil. Al

Model	R^2	R^2_{adj}	RMSE	WRMSE
lm	0.763	0.759	933.935	756.850
pooled	0.746	0.746	965.386	781.716
no_pooled	0.706	0.702	1039.901	870.026
hierarchical	0.623	0.618	1176.690	1005.060
hierarchical_cov	0.712	0.707	1029.722	864.167

TABLA 9. Métricas de rendimiento de los modelos

repetirse el error de estos barrios en menor medida (al tener menos observaciones) se menosprecia su presencia en las metricas finales. Por estos motivos hemos decidido implementar un raíz del error cuadrático medio ponderada (a partir de ahora WRMSE) dando más peso a los grupos con menos observaciones.

En la tabla 9 vemos que los modelos más simples siguen desempeñándose mejor. El modelo jerárquico con covariable se mantiene en la misma posición, tercer lugar. Sin embargo, el error entre el *lm* y el *hierarchical_cov* es menor.

En la figura 2 podemos observar los cuatro barrios con mayor oferta (más observaciones): Sant Gervasi - Galvany, l'Antiga Esquerra de l'Eixample, el Raval, la Dreta de l'Eixample; y los cuatro con menos observaciones la Trinitat Nova, la Vall d'Hebron, Can Baró y la Trinitat Vella. Vemos que los intervalos de credibilidad ¹⁴ en los cuatro con menos oferta son más amplios, esto tiene sentido al tener menos observaciones para validar el modelo. Con los cuatro barrios más grandes ocurre lo contrario. También cabe resaltar que en los barrios más pequeños el intervalo del modelo jerárquico con covariable *hierarchical_cov* se encuentra a medio camino del modelo no agrupado y el jerárquico. Al tener presente que el modelo jerárquico con covariable es el que tiene menor RMSE y WRMSE de los tres, nos indica que es el mejor candidato.

¹⁴2 desviaciones estandar equivale aproximadamente al 95 % de credibilidad.

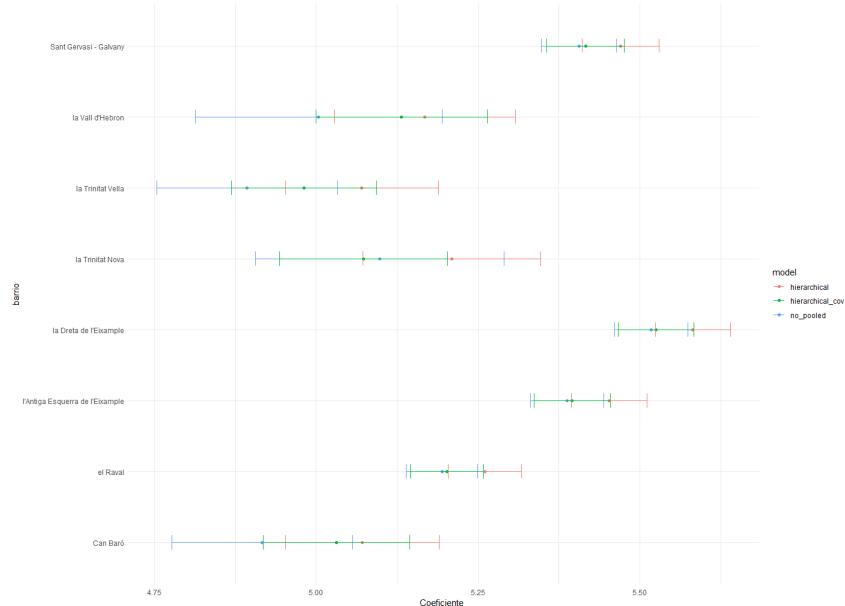


FIGURA 2. Coeficientes de los barrios con mayor y menor oferta

Un último punto a tener en cuenta es que los modelos jerarquicos son los únicos que pueden agregar nueva información. Por lo tanto, la decisión es subjetiva a los objetivos del estudio. En este caso se decide seguir con el modelo jerarquico con covariante para hacer la predicciones de la siguiente sección, ya que el objetivo de estudio es poder predecir precios incluso en zonas que no hay una oferta consolidada, además de las zonas ya consolidadas.

7. Aplicación Shiny

En esta sección explicaremos, sin entrar en profundidad, como funciona una aplicación Shiny¹⁵. Luego expondremos la aplicación Shiny¹⁵ creada para este proyecto y pondremos algunos ejemplos de como utilizarla.

Librería Shiny. Shiny es una librería del lenguaje de programación R en el cual se puede crear cuadros de mando enfocado a ciencia de datos. Al estar integrado en R no es requisito saber desarrollo web ya que las funciones básicas vienen en lenguaje R importando la librería *import(shiny)*.

Toda aplicación Shiny consta de tres partes[Wic21]: interfaz de usuario (UI) por sus siglas en inglés. En este apartado añadiremos todos los objetos visuales que queramos mostrar en nuestra aplicación. El segundo apartado en el servidor, aquí expondremos la lógica, es decir, todas las operaciones, funciones, etc que necesita nuestra app para funcionar y mostrarlo en la UI. La tercera parte es correr el

¹⁵Página web de shiny: <https://shiny.posit.co/>

comando `shinyApp(ui = ui, server = server)` que es donde se ejecuta nuestros dos bloques anteriores.¹⁶.

Nuestra aplicación. Nuestra aplicación, creada siguiendo la lógica del apartado anterior, consta de dos partes:

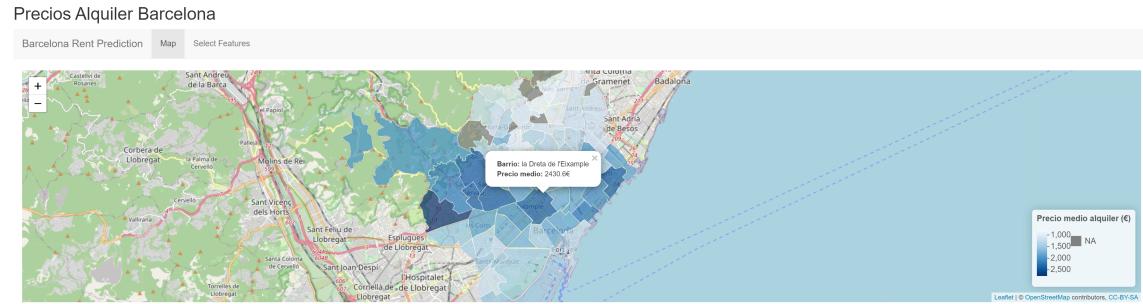


FIGURA 3. Mapa de Barcelona con precios medios por barrios.

Primero, como podemos observar en la figura 3, de un mapa interactivo de la ciudad de Barcelona que muestra los 72 barrios de la ciudad. Si interactuamos con el mapa podemos observar el nombre del barrio y su precio medio de alquiler. También vemos que hay barrios que aparecen en gris, con NA, con ningún valor. Son zonas que no se han podido recopilar datos, como se comentó en el apartado de modelización este inconveniente se resuelve utilizando modelos jerárquicos bayesianos en la predicción.

¹⁶Para más información consultar: Wickham, Hadley. Mastering shiny. .O'Reilly Media, Inc.", 2021.

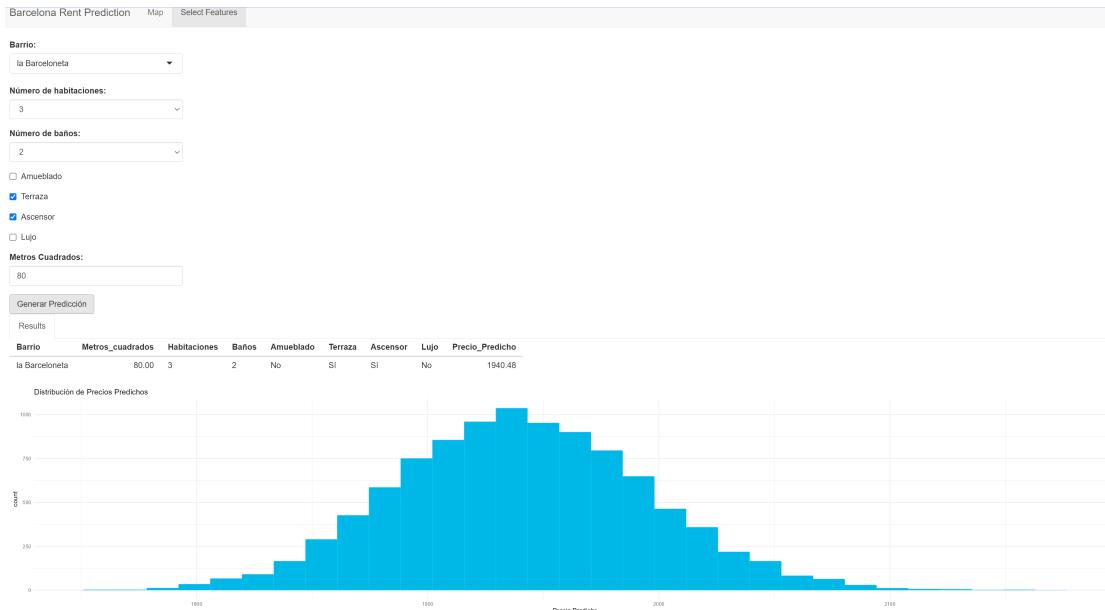


FIGURA 4. Predicción de precios de alquiler.

En la segunda pestaña de nuestra app, como muestra la figura 4, podemos observar un cuadro de mando que permite seleccionar mediante una casilla desplegable el barrio donde queremos predecir el precio de una vivienda.¹⁷

Además, hay una serie de botones seleccionables que nos permiten ajustar nuestra predicción con, por ejemplo, el número de habitaciones, baños, si tiene terraza, etc. Esta aplicación es reactiva, cada vez que cambiamos una opción se actualiza el histograma de precios generados con las opciones dadas. Cabe resaltar que si una opción no es seleccionada en el modelo, cuenta como 0, es decir su coeficiente no computa en el cálculo.

Como vemos en nuestro ejemplo, con las opciones requeridas, el precio medio es de 1940 euros.

Para validar nuestro ejemplo vamos a buscar un piso con estas características en la página Idealista. Como podemos observar en la figura 5, solo encontramos 1 apartamento con estas características o similares, tuvimos que relajar la condición de 2 a 1 baño y de 80 a 75 metros cuadrados. Además, si leemos más a detalle la descripción vemos que es un alquiler de corta estancia, menos de un año.

Por último, cabe resaltar que ningún modelo es perfecto, por lo tanto la utilización de esta aplicación es más bien un guía con tal de poder tomar una decisión mas acertada a la hora de alquilar, donde podemos comparar los precios medios actuales y el valor del piso que deseamos encontrar. También permite a propietarios dar una idea del mercado y los precios de alquiler sin tener que acudir a un profesional.

¹⁷En nuestro caso hemos escogido un barrio aleatorio "la Barceloneta": 3 habitaciones, 2 baños, terraza y ascensor, 80 metros cuadrados.

1 Piso en alquiler en La Barceloneta, Barcelona

Nuevos anuncios en tu email

Guardar búsqueda

Comprar **Alquilar** Compartir

Unstado Mapa

Ordenar Relevancia Baratos Recientes Más ▾

Piso en paseo de Joan de Borbó, La Barceloneta, Barcelona

1.950 €/mes

3 hab. 75 m² Planta 4^a exterior con ascensor

(DISPONIBLE A PARTIR DE ENERO DE 2024 PARA ALQUILERES MENSUALES DE 3 A 11 MESES) Este apartamento con impresionantes vistas se encuentra en el quinto piso en un edificio equipado con ascensor. Tiene un ampl...

Ver teléfono Contactar

Precio medio 26,00 eur/m²

¿Buscas un profesional inmobiliario? **Immobiliarias en La Barceloneta**

Podría interesarle también cerca de La Barceloneta

Barrios

- 368 Sant Pere - Santa Caterina i la Ribera
- 343 El Raval
- 279 El Gòtic

Playas

Calles cerca

- 45 Calle Mar
- 34 Calle Sant Miquel
- 27 Calle Sant Elm

Ver más ▾

Transportes

FIGURA 5. Ejemplo portal Idealista.

8. Conclusiones

El objetivo principal de esta tesis era construir un aplicación que sea de utilidad para las personas que demandan o ofertan pisos. El segundo objetivo era comparar la potencia de predicción de modelos lineales con los modelos bayesianos y analizar empíricamente los resultados del sector del alquiler en Barcelona.

Teniendo en cuenta los diferentes modelos que han sido validados y comparados, con sus pros y contras, se concluye que se ha conseguido el objetivo, pero parcialmente. La aplicación es todavía un prototipo y no está abierta, de momento, al público general. Además, el número de variables a considerar e investigar es muy grande y aunque las métricas alcanzadas son buenas, por ejemplo R² mayor a 0.70 en 4 de los 5 modelos, cabe posibilidad de mejora. Teniendo en cuenta las dos partes, el proyecto tiene mucho potencial.

Para investigaciones futuras se pueden crear nuevos modelos usando otras variables y actualizando los modelos ya entrenados, gracias a los modelos bayesianos, considerando los coeficientes obtenidos como distribuciones apriori de los nuevos datos. También, se cree que información de geolocalización sería muy beneficiosa para el proyecto, pero no posible de obtener con el método de web scrapping por privacidad de los anunciantes.

Para terminar, la motivación personal de este proyecto era aprender más sobre métodos bayesianos y se considera que esta meta se ha alcanzado satisfactoriamente.

Bibliografía

- [Coo77] R Dennis Cook. *Detection of influential observation in linear regression*. Consultado el 25 de Junio de 2023. 1977. URL: <http://www.stat.ucla.edu/~nchristo/statistics100C/1268249.pdf>.
- [Gel+95] Andrew Gelman et al. *Bayesian data analysis*. Consultado el 5 de Julio de 2023. 1995. URL: <http://www.stat.columbia.edu/~gelman/book/BDA3.pdf>.
- [Gel06] Andrew Gelman. *Multilevel (hierarchical) modeling: what it can and cannot do*. Consultado el 12 de Julio de 2023. 2006. URL: <http://www.stat.columbia.edu/~gelman/research/published/multi2.pdf>.
- [Gel+08] Andrew Gelman et al. *A weakly informative default prior distribution for logistic and other regression models*. Consultado el 10 de Julio de 2023. 2008. URL: <http://www.stat.columbia.edu/~gelman/research/published/priors11.pdf>.
- [HM10] Shanaka Herath y Gunther Maier. *The hedonic price method in real estate and housing market research: a review of the literature*. Consultado el 10 de Junio de 2023. 2010. URL: <https://ro.uow.edu.au/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1977&context=buspapers>.
- [Jam+13] Gareth James et al. *An introduction to statistical learning*. Consultado el 22 de Junio de 2023. 2013. URL: https://www.stat.berkeley.edu/users/rabbee/s154/ISLR_First_Printing.pdf.
- [Wic21] Hadley Wickham. *Mastering shiny*. Consultado el 25 de Julio de 2023. 2021. URL: <https://mastering-shiny.org/>.
- [] *Amazon API definición*. Fecha de Consulta: 02 de Junio de 2023. URL: <https://aws.amazon.com/es/what-is/api/>.
- [] *Forbes Data Cleaning*. Fecha de Consulta: 05 de septiembre de 2023. URL: <https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/>.
- [] *Interpretación modelos log-log*. Fecha de Consulta: 26 de Junio de 2023. URL: <https://library.virginia.edu/data/articles/interpreting-log-transformations-in-a-linear-model>.
- [] *Moncloa nota de prensa*. Fecha de Consulta: 05 de septiembre de 2023. URL: <https://www.lamoncloa.gob.es/serviciosdeprensa/notasprensa/transportes/Paginas/2023/040523-nueva-ley-vivienda-2023.aspx>.

- || *Página web de Shiny.* Fecha de Consulta: 25 de Julio de 2023. URL:
<https://shiny.posit.co/>.
- || *Shapefile definición.* Fecha de Consulta: 03 de Junio de 2023. URL:
<https://enterprise.arcgis.com/es/portal/latest/use/shapefiles.htm>.

Apéndice

September 21, 2023

1 Gráficos de convergencia de modelos

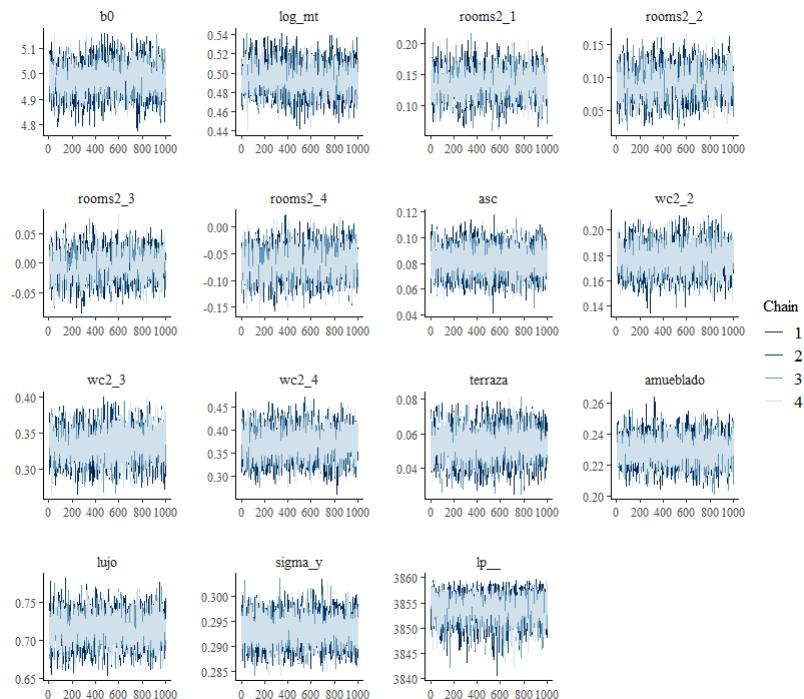


Figure 1: Gráfico de convergencia de cadenas. Modelo agrupado.

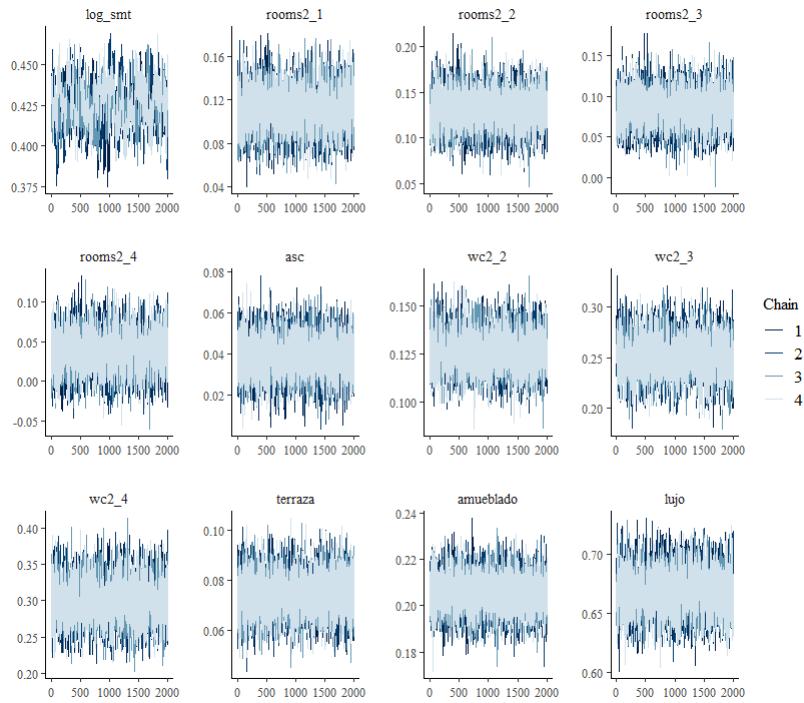


Figure 2: Gráfico de convergencia de cadenas. Modelo no agrupado.

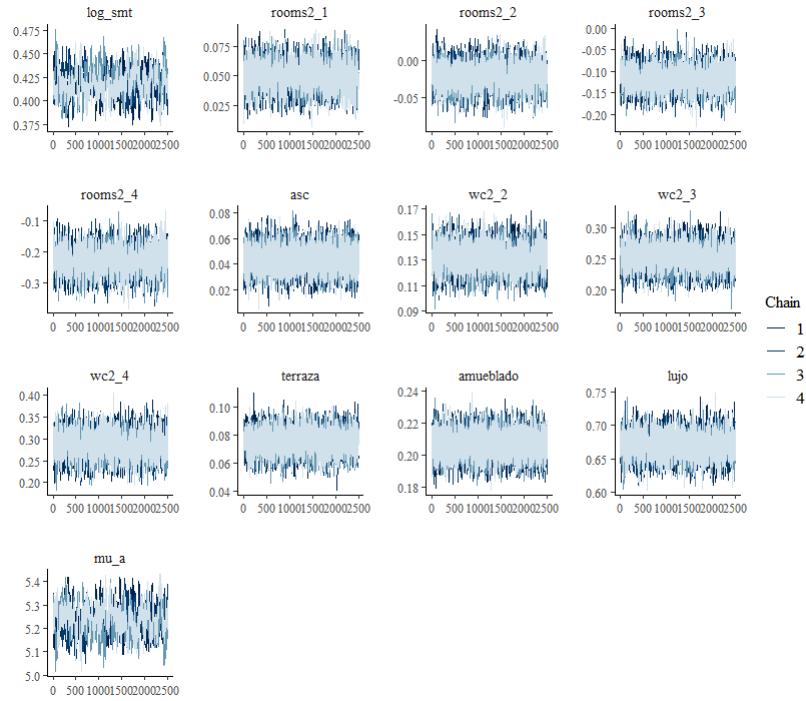


Figure 3: Gráfico de convergencia de cadenas. Modelo jerárquico.

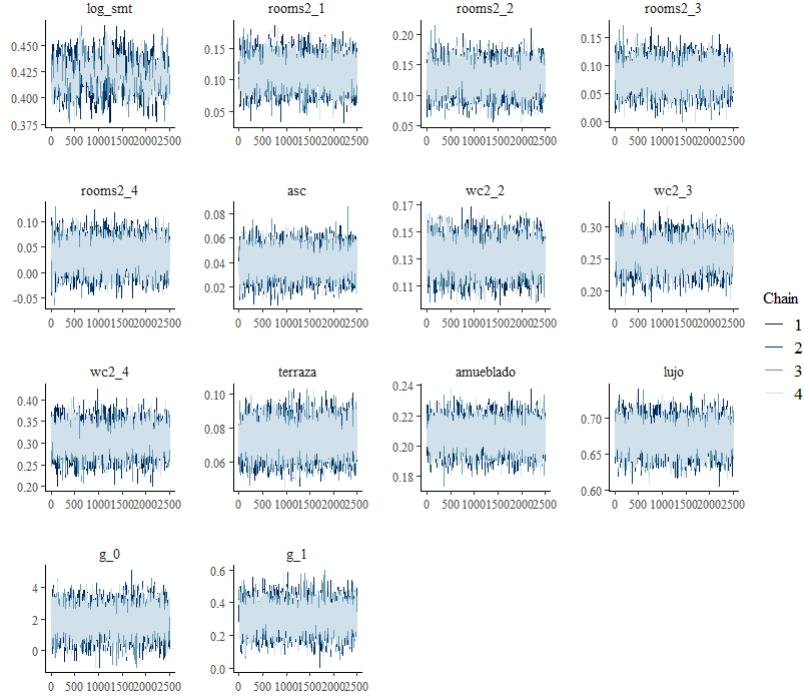


Figure 4: Gráfico de convergencia de cadenas. Modelo jerárquico con covariable.

En los gráficos de mezcla de cadenas se puede observar que apartir del modelo agrupdo, el más sencillo, la variable metros cuadrados no se mezcla bien, debido a posibles problemas de correlacion y poca información. Se alcenza el mejor resultado visual en el modelo jerárquico con covariable (sin tener en cuenta el modelo agrupado).

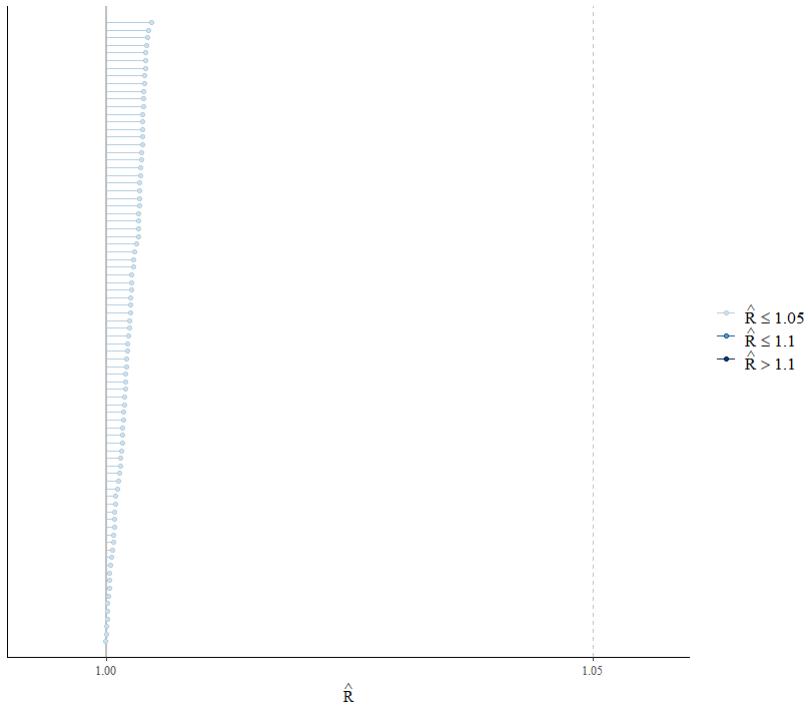


Figure 5: Gráfico Rhats. Modelo jerárquico con covariable.

Ningún Rhat llega a ser mayor de 1.05.

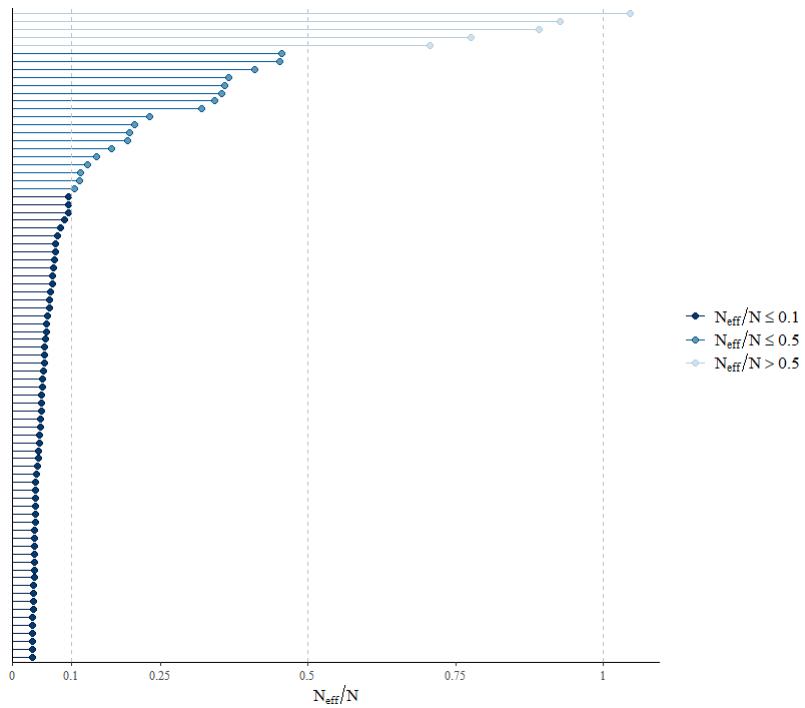


Figure 6: Gráfico muestras efectivas. Modelo jerárquico con covariable.

Se observa pocas muestras efectivas para los barrios en general y para los más pequeños en particular.

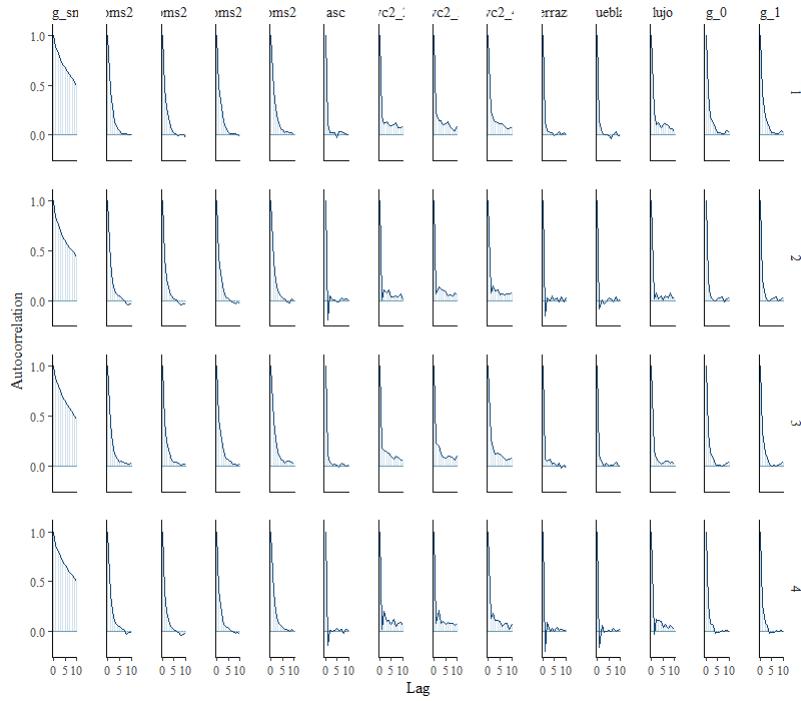


Figure 7: Gráfico autocorrelación de variables. Modelo jerárquico con covariante.

Observamos que la única variables autocorrelacionada son los metros cuadrados. Esto es debido a el poco número de observaciones en ciertos barrios. Una posible solución es ir mejorando el modelo con muestras futuras.