Precios al consumidor

Manuel Toledo y Lucas Pescetto

2023-07-10

Introducción

Este proyecto surge a partir de la información provista por el SIPC (Sistema de Información de Precios al Consumidor) del Ministerio de Economía y Finanzas. Dicho organismo brinda información acerca de los precios de una serie de productos a través del tiempo y para distintos establecimientos en todos los departamentos de Uruguay.

Por otro lado, a raíz de la amplia diversidad de los productos de los cuales se tienen datos, se decidió tomar solamente aquellos que son parte de la CBA (Canasta Básica de Alimentos) de Uruguay.

El objetivo del trabajo es generar un análisis y visualizaciones que permitan ver las variaciones en los precios de dichos productos a lo largo del tiempo, en distintos lugares dentro del país y en establecimientos dentro de Montevideo. Esto resultaría de utilidad para ayudar a los consumidores a tomar mejores decisiones financieras a la hora de comprar alimentos.

Un objetivo adicional es el de utilizar modelos basados en series temporales para predecir los precios de los productos en el futuro.

El producto final que se busca proveer es una aplicación interactiva mediante el paquete *Shiny* (Chang et al., 2022) que para cada uno de los productos presentes en la CBA, despliegue una serie de visualizaciones descriptivas de su precio, en función de los aspectos mencionados anteriormente.

A lo largo del informe se usarán los datos de los precios de los fideos como referencia para explicar los datos y visaulizaciones, dado que no se encontró una forma de ponderar los precios de la canasta básica. Sin embargo, la aplicación de Shiny permitirá seleccionar los productos de manera individual.

Datos

El SIPC presenta los datos en tres datasets:

- Establecimientos: es una lista de los establecimientos de los cuales se obtienen los precios. Se obtiene en la web de Catálogo abierto de datos.
- Productos: es una lista de los productos de los cuales se tienen los precios. Se obtiene en la web de Catálogo abierto de datos.
- Precios: contiene la información acerca de los precios registrados en cada momento del tiempo, para cada producto en cada establecimiento. Si bien se puede obtener en la web de Catálogo abierto de datos, debido a su tamaño se extrae con una consulta SQL.

A continuación se muestra una imágen de la estructura de la base:

Además de las variables presentes en la imagen, el dataset de Establecimientos tiene la dirección, centro comunal zonal, barrio, cantidad de cajas, cadena, longitud y latitud, ciudad, departamento y superficie.

Se cuenta con datos a partir del año 2016 y hasta marzo del 2023. Estos contienen 363 productos dentro 766 establecimientos. De esos, solamente se usarán los 18 productos dentro de la CBA más 2 productos que se agregaron por decisión personal (para cada producto existen varias marcas).

Los productos son:

- Aceite de girasol 900 cc
- Aguja vacuna 1 kg (con y sin hueso)
- Arroz blanco 1 kg
- Arvejas en conserva 300 g
- Azúcar blanco 1 kg
- Carne picada vacuna 1 kg
- Cocoa 500 g
- Dulce de leche 1 kg
- Fideos secos al huevo 500 g
- Galletitas al agua 140 g
- Harina trigo común 0000 1 kg
- Huevos colorados 1/2 docena
- Manteca 200 g
- Pan flauta 215 g
- Papel higiénico hoja simple 4 rollos 30 mts

- Pollo entero fresco con menudos 1 kg
- Pulpa de tomate 1 L
- Sal fina yodada fluorada 500 g
- Yerba mate común 1 kg
- Café (agregado)
- Fideos secos de sémola 500 g (agregado)

A partir de los datasets, se construyó uno con el precio promedio mensual para cada producto (desagregándolo según las marcas y establecimientos), que incluyera parte de la información presente en Establecimientos (el nombre de la sucursal, cadena, coordenadas, barrio y departamento) y Productos (nombre y marca). Tanto la agregación como pegado de los datos se hizo a través de la consulta SQL, utilizando la variable de fecha y las *keys* respectivas, id.establecimientos e id_productos.

El objetivo del trabajo esta bien planteado, las diferentes alternativas que tenían en mente se explicaron bien. La descripción de los datos un poco pobre y hay que afinar el concepto “debido a su tamaño se extrae con una consulta SQL”. Esto no es así, capaz no quedó clara la diferencia entre SQL y el servidor de base de datos, consideraciones menores.

La enumeración de los productos quedaría mejor presentado en una tabla. Nuevamente, el proceso de agregación de los datos fue crucial y lo mencionan al pasar. Hubiera sido importante justificar de por qué lo agregaron (granularidad) y la dificultad de analizarlos con las técnicas vistas en el curso dada la gran cantidad de datos.

Análisis exploratorio

Para comenzar con el análisis de los datos, se intentaron visualizar a nivel general las características de los productos y establecimientos disponibles.

Para los establecimientos, resulta interesante ver cómo se distribuyen en el territorio del país. Para eso se graficó un mapa donde cada uno es representado por una cruz (figura 1). Además, a través de un gráfico de barras se presenta la cantidad de establecimientos por departamento (figura 2).

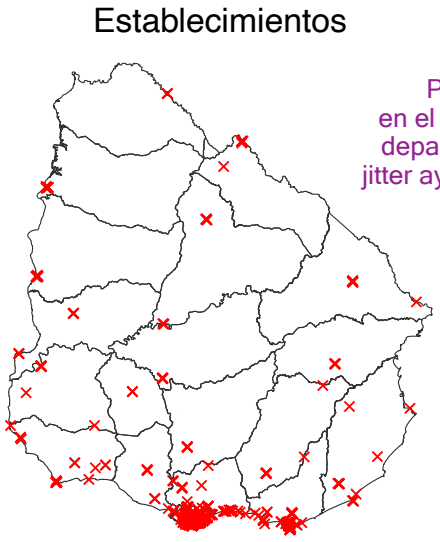
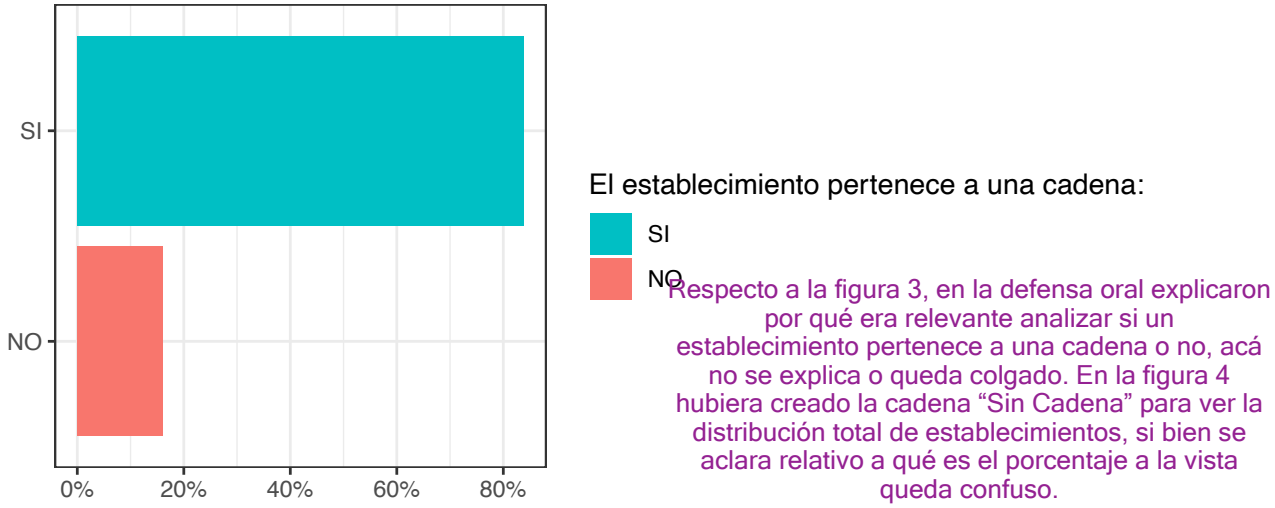


Figure 1: Mapa de los establecimientos disponibles. Se concentran principalmente en la capital y el sur del país.

Figure 2: Cantidad de establecimientos por departamento. Mas del 60% se encuentran en Montevideo

Como es posible ver en ambos gráficos, los establecimientos encuestados se distribuyen principalmente sobre la costa y en la capital. Algunos departamentos casi no están representados, dada la baja cantidad de establecimientos relevados.

Siguiendo con los establecimientos, también interesa saber qué proporción pertenece a una cadena, y cuáles son las cadenas que tienen la mayor cantidad de establecimientos



Respecto a la figura 3, en la defensa oral explicaron por qué era relevante analizar si un establecimiento pertenece a una cadena o no, acá no se explica o queda colgado. En la figura 4 hubiera creado la cadena "Sin Cadena" para ver la distribución total de establecimientos, si bien se aclara relativo a qué es el porcentaje a la vista queda confuso.

Figure 3: Proporción de establecimientos que pertenecen a una cadena de supermercados. Más del 80% pertenecen a alguna de las cadenas

Cómo se puede observar en la figura 3, la mayoría de los establecimientos forman parte de una cadena. A continuación se presentan las diez cadenas con la mayor cantidad de establecimientos (figura 4).

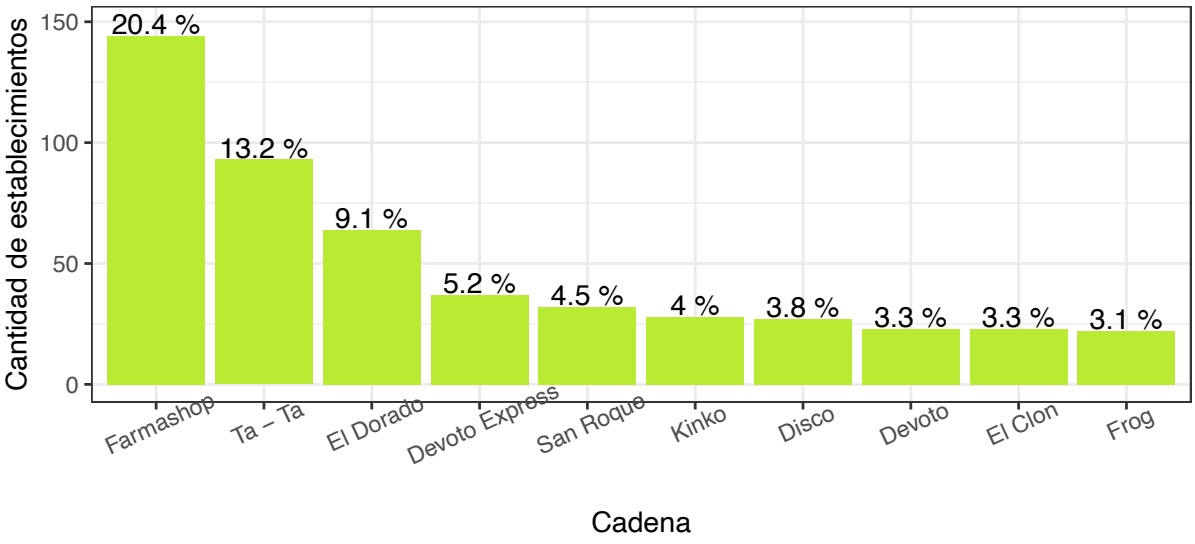


Figure 4: Cadenas con la mayor cantidad de establecimientos. Los porcentajes son con respecto a la cantidad total de establecimientos.

Este gráfico permite dar cuenta que más de la mitad de los establecimientos relevados corresponden solamente a cinco cadenas. Sumando las diez con mayor cantidad, ese porcentaje alcanza el 70%.

A continuación se presenta el análisis focalizado en los productos. Como se mencionó anteriormente, se tomaron los fideos como producto de referencia.

La figura 5 presenta una comparativa de los precios de fideos a lo largo del tiempo entre las distintas marcas y

Deberían usar referencias cruzadas para referir a las figuras en el texto

tipos (de sémola o al huevo). Se observó que todos los productos aumentan a un mismo ritmo con excepción de los fideos Puritas, que mantuvieron su precio durante años, lo cual resulta bastante extraño.

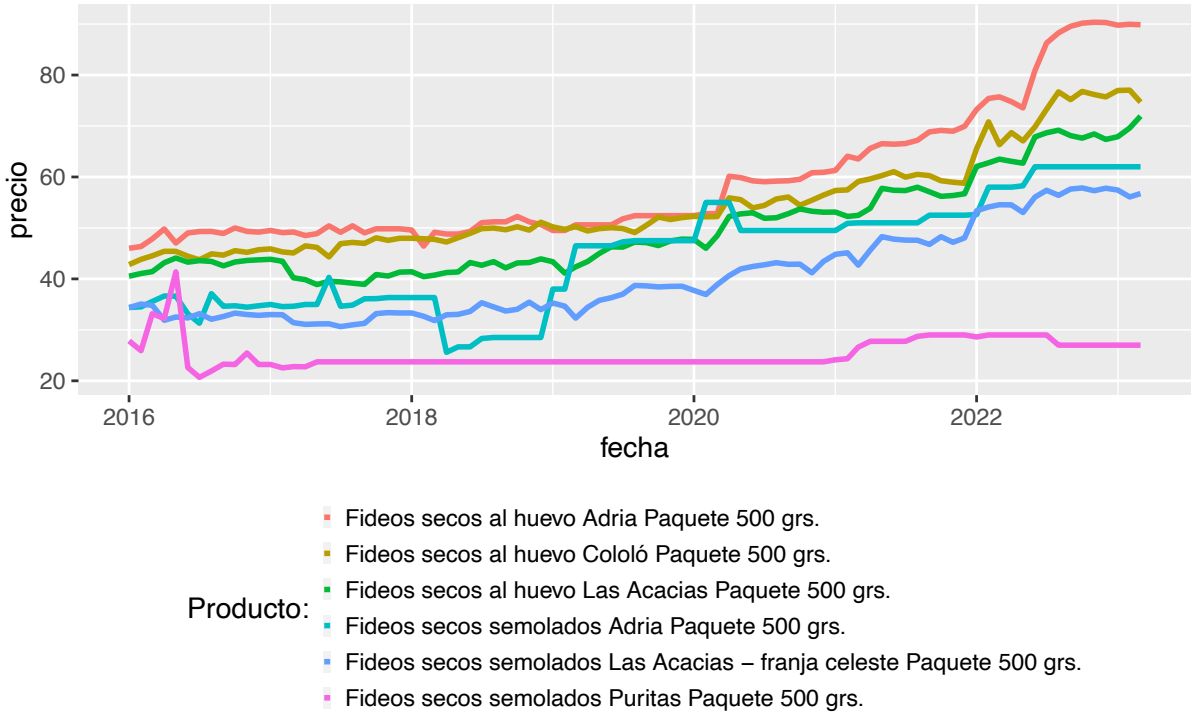


Figure 5: Precio de los fideos a lo largo del tiempo. Se observa una tendencia al aumento a través de los años

tendrían que haber trabajado con los precios a precios constantes, la tendencia es explicada principalmente por la inflación

Analizando los precios de los fideos por región (figura 6) se notó bastante homogeneidad a nivel departamental, con algunas excepciones que también se podrían explicar por el bajo tamaño de muestra o la ausencia de marcas baratas/caras en algunos departamentos.

Dentro de Montevideo (figura 7) se notaron más diferencias. Los barrios de mayor poder adquisitivo tienen los precios más elevados y las áreas rurales los precios más bajos.

Mejorable, analizando con precios deflactados y viendo si hay diferencias anuales

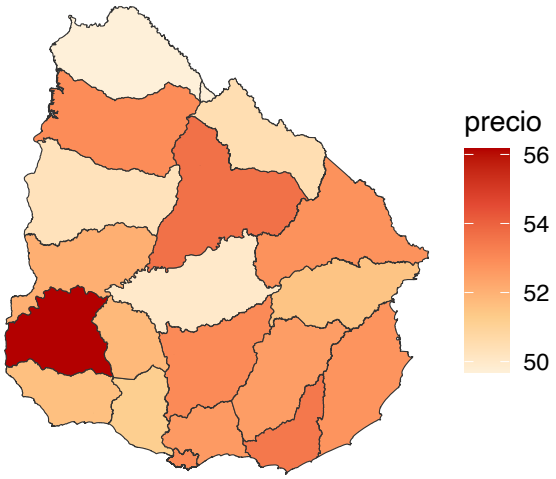


Figure 6: Mapa de Uruguay según el precio promedio de los fideos en cada departamento para todo el período.

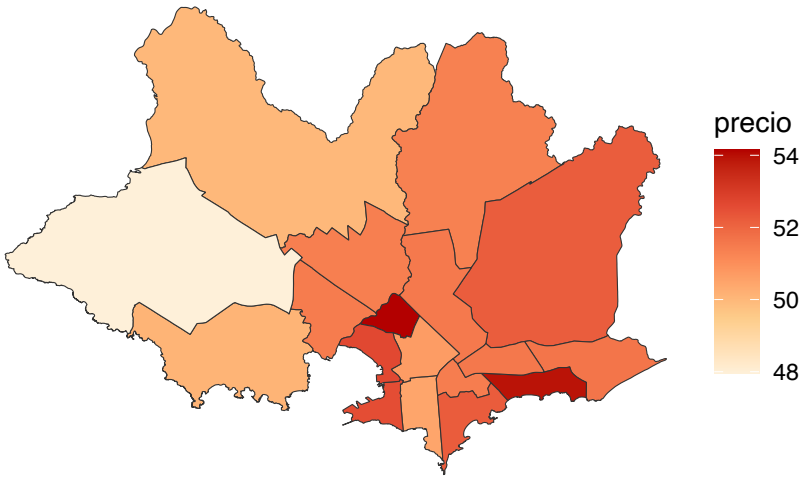


Figure 7: Mapa de Montevideo según el precio promedio de los fideos en cada CCZ para todo el período.

Modelo estadístico

En esta sección se describe el proceso de construcción de un modelo del precio de los productos en función del tiempo, que permita hacer predicciones sobre su variación en el futuro.

Para la confección del modelo, el producto seleccionado fue el paquete de fideos. Como se vio en la sección anterior, entre los productos trabajados, existen dos variedades de fideos: al huevo y semolados. A su vez, para cada variedad hay tres marcas.

En primer lugar se filtraron los datos para trabajar solamente con los precios del producto elegido. Para cada período (mes) se cuenta con precios por variedad, marca y establecimientos. Como todas las variedades y marcas registran el precio de paquetes de fideos del mismo tamaño (500 gramos), tuvo sentido utilizar como variable respuesta el promedio de todas las marcas y variedades presentes en los datos.

En cuanto a las herramientas utilizadas, se exploraron varias alternativas hasta llegar a una satisfactoria. En primer lugar se escogieron los paquetes **caret** (Kuhn & Max, 2008) y **xgboost** (Chen et al., 2023) para construir el modelo; y el paquete **forecast** (R. J. Hyndman & Khandakar, 2008) para visualizar sus predicciones. El modelo se construyó en base a las instrucciones brindadas en un blog (Alice, 2020). El problema de esta opción fue que el modelo creado no reconocía que la variable provista era el tiempo. Como indica la guía, la variable de tipo *Date* utilizada como explicativa se separa en dos: una indica el año y la otra el mes de la fecha correspondiente. Sin embargo, para el modelo esas dos variables son simplemente numéricas, y lo que es peor, son tratadas como independientes entre sí.

Este problema se reflejó en predicciones que mostraban un patrón de variaciones que se repetía año a año, pero a pesar de la clara tendencia de aumento de los precios a lo largo del tiempo en los datos, los predichos se mantenían en un mismo rango. A continuación se muestra el resultado de las predicciones.

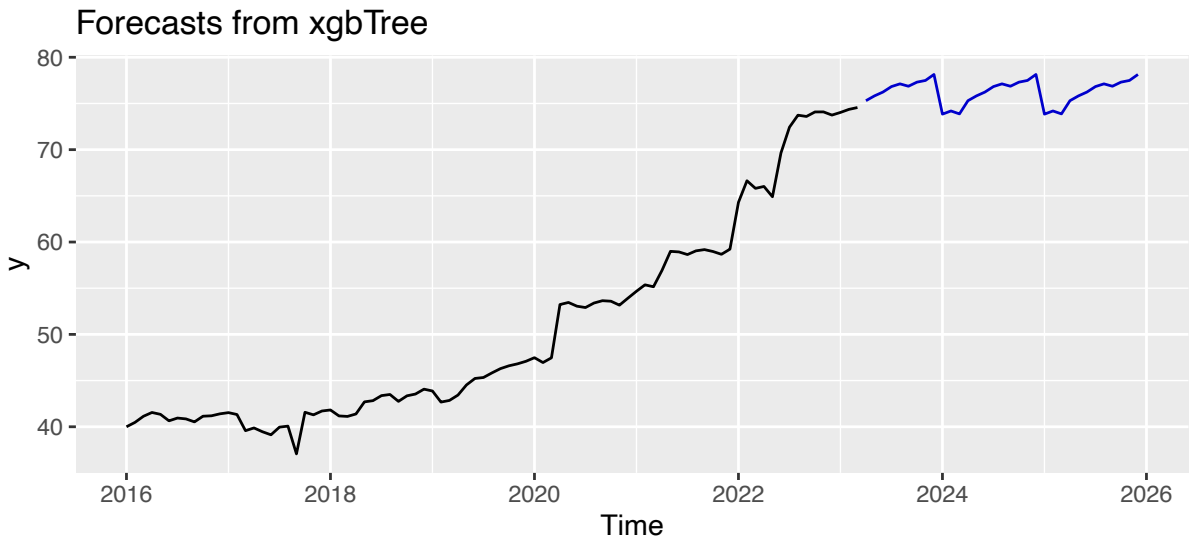


Figure 8: Prediccion del precio promedio de los fideos. El tramo en azul corresponde al período predicho.

En segundo lugar se optó por los paquetes **modeltime** (Dancho, 2023) y **timetk** (Dancho & Vaughan, 2023) para construir las series de tiempo. También se utilizó **tidymodels** (Kuhn & Wickham, 2020), que permite utilizar la gramática de tidyverse en la confección de modelos. Estos paquetes difieren de los utilizados anteriormente en el sentido de que están específicamente diseñados para modelar con series de tiempo. A modo de ejemplo, en este caso la variable explicativa puede utilizarse en su clase natural (*Date*), en vez de transformarla a una o más numéricas.

Se realizaron 3 regresiones: ARIMA, GLMNET y Prophet. Para ello se seleccionaron los conjuntos de entrenamiento (*training*) y prueba (*test*). Para el primero se tomaron los datos desde el primer mes disponible

hasta diciembre de 2022. Para el segundo se tomaron los primeros 3 meses de 2023 (los últimos períodos para los que hay datos). Los demás parámetros de cada regresión se dejaron con sus valores por defecto. Una vez creados los modelos, se utilizó la función *modeltime_table* de **modeltime** para integrarlos en una tabla y así llevar a cabo un análisis de los tres en conjunto.

Pasando a la etapa de evaluación, se analizaron las diferencias entre el conjunto de prueba y las predicciones halladas. Luego se reajustó el modelo considerando la totalidad de los datos y se calcularon las predicciones para los 9 meses restantes de 2023. Además de gráficos interactivos con los valores predichos para cada regresión, el paquete **modeltime** cuenta con la función *modeltime_accuracy*, que muestra una serie de medidas para evaluar sus desempeños.

A continuación se presentan imágenes con los resultados obtenidos para cada regresión en el caso del precio promedio general(figuras), y la tabla con las medidas (figura):

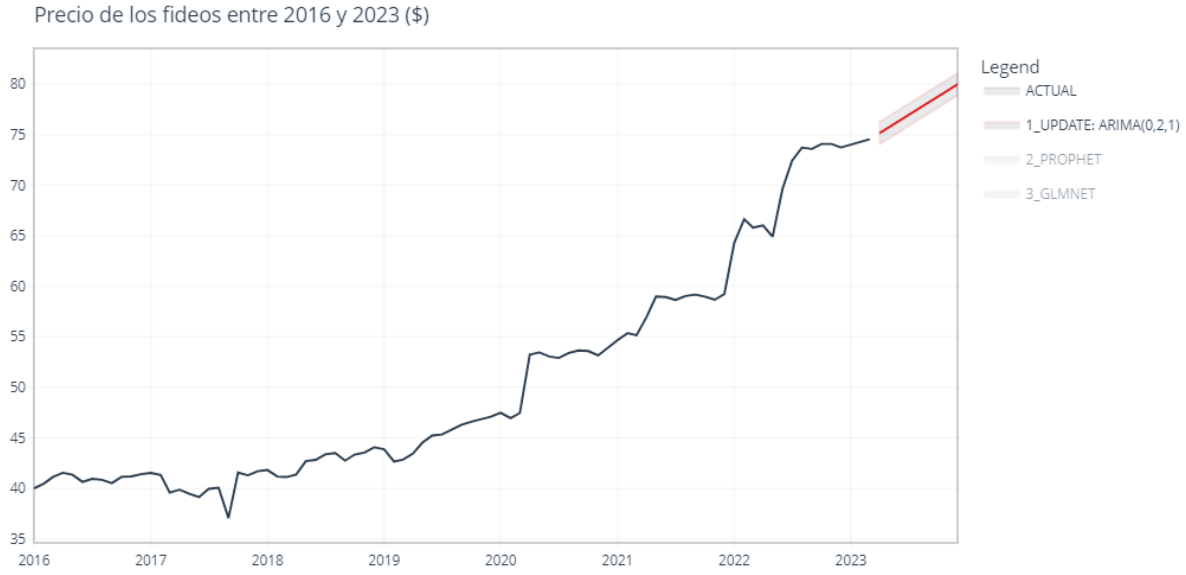


Figure 9: Predicción realizada por la regresión ARIMA.

Sobre el modelo estadístico, explican bien las desventajas de incluir el tiempo de esta forma (determinística) en el modelo inicial de XGBoost y la resolución con modelos específicos de series de tiempo donde el tiempo se modela de forma estocástica.

La explicación de la división del conjunto de training y test para no romper la estructura temporal fue muy bien explicada. Las figuras 9, 10 y 11 (las que tienen los intervalos de predicción de cada modelo) las hubiera puesto en uno solo, las interpretaciones fueron correctas.

La explicación de los diferentes tipos de métricas para medir el error fue acorde a su nivel, me hubiera gustado un poco más que ahondaran en los errores escalados, por ejemplo: el Prophet y el GLMNET eran modelos malísimos, era mejor predecir con la media o el valor anterior (modelo Naive). Esta conclusión es bastante fuerte, los modelos clásicos (ARIMA) funcionan siempre mejor en este caso. De todas maneras, no hay que olvidar que utilizan los parametros por defecto y los supuestos de cada modelo no fueron profundizados dado el contexto del curso.

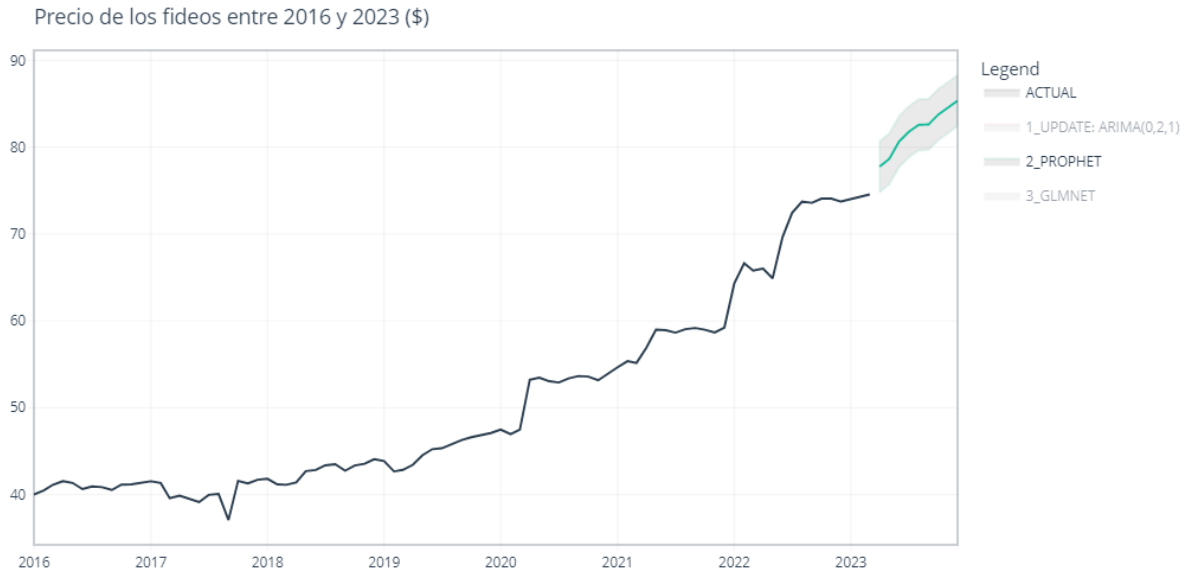


Figure 10: Predicción realizada por la regresión prophet.

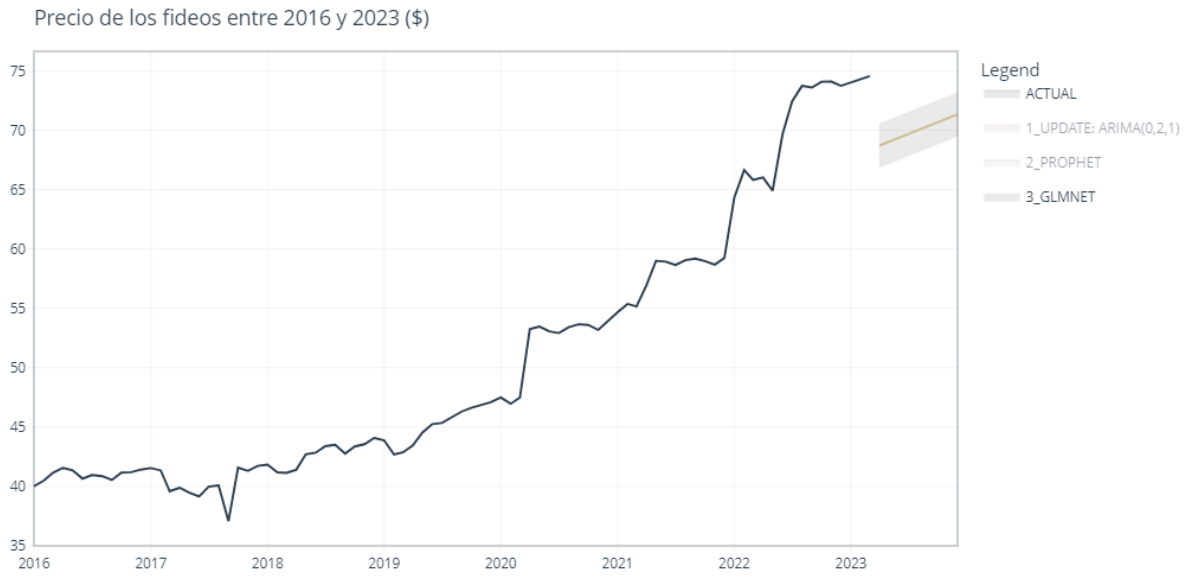


Figure 11: Predicción realizada por la regresión GLMNET.

	modelo	mae	mape	mase	smape	rmse
1	ARIMA(0,0,0) WITH NON-ZERO MEAN	0.47	0.63	1.77	0.63	0.52
2	PROPHET	1.28	1.72	4.83	1.74	1.37
3	GLMNET	0.83	1.11	3.11	1.12	0.85

Como se puede ver, los tres modelos predicen que el precio de los fideos va a aumentar en los próximos 9 meses.

En cuanto a la tabla, las medidas corresponden a distintas formas de agregar el error de predicción, que se define como la diferencia entre los datos del conjunto de prueba y las predicciones del modelo. Según el libro *Forecasting: Principles and Practice* (R. Hyndman & Athanasopoulos, 2018), se pueden clasificar en tres tipos: - Dependientes de la escala (de la variable respuesta): mae y rmse - Errores porcentuales: mape y smape - No dependientes de la escala: mase

El primer tipo sirve exclusivamente para comparar modelos en los que la variable respuesta esté medida en una unidad común, ya que el error se mide en las unidades de la misma. Los otros dos sirven para comparar el desempeño de modelos más diversos.

En este caso, como los tres modelos están construidos con el mismo conjunto de datos, se pueden considerar todos los estadísticos. En todos los casos el error más pequeño lo tiene la regresión de ARIMA, por lo que no es necesario llevar a cabo una mayor discusión y se concluye que es la mejor alternativa para esta serie de tiempo.

En cuanto a su intervalo de confianza, a un 90% tiene un rango de aproximadamente 2 pesos, que teniendo en cuenta el problema que se está estudiando, no presenta mayores inconvenientes.

A modo de conclusión, el modelo predice que al final del año 2023 el precio de los fideos habrá experimentado un aumento de 7.80% con respecto al año anterior. Este aumento está alineado con la inflación esperada para el año, que es de un 8.00% según la Encuesta de Expectativas Empresariales de junio ((INE), 2023).

Aplicación Shiny

La aplicación cuenta con la opción de seleccionar un producto y las marcas que se deseen para investigar su precio dentro de Montevideo mediante un mapa de *leaflet* (Cheng et al., 2023) con los establecimientos o para observar las diferencias entre departamentos.

Se implementó también la posibilidad de generar modelos de predicción temporal con los precios de cada producto, pero esto no pudo ser introducido a la aplicación por no tener suficientes recursos computacionales en el servidor gratuito de Posit como para generarlo.

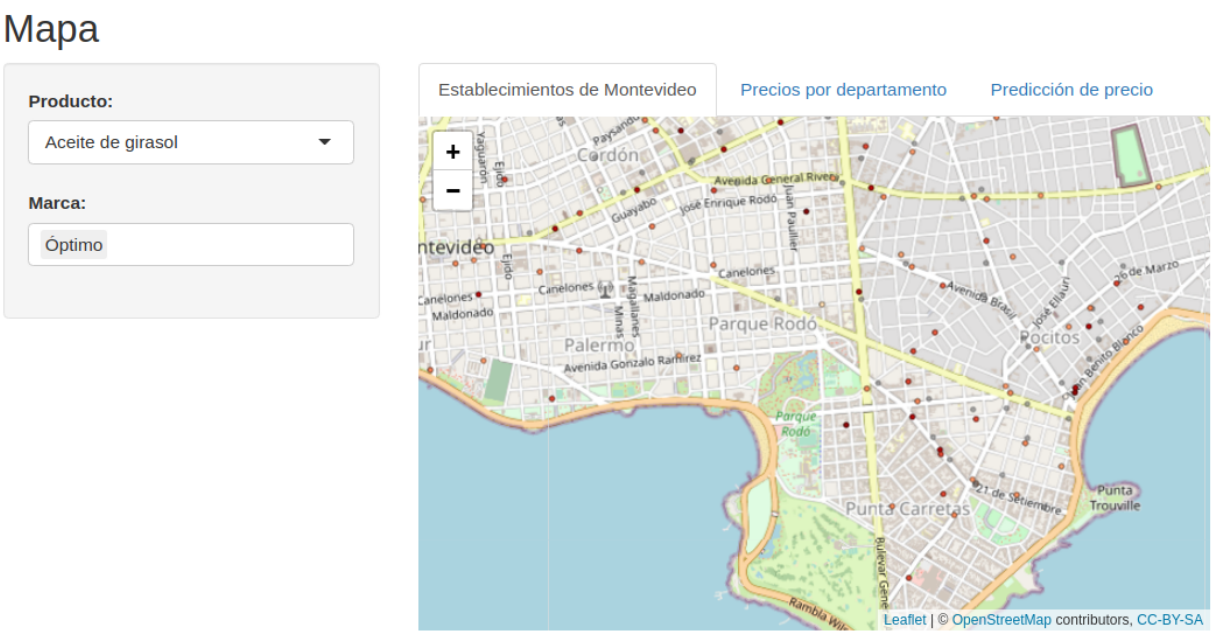


Figure 12: mvd_mapa

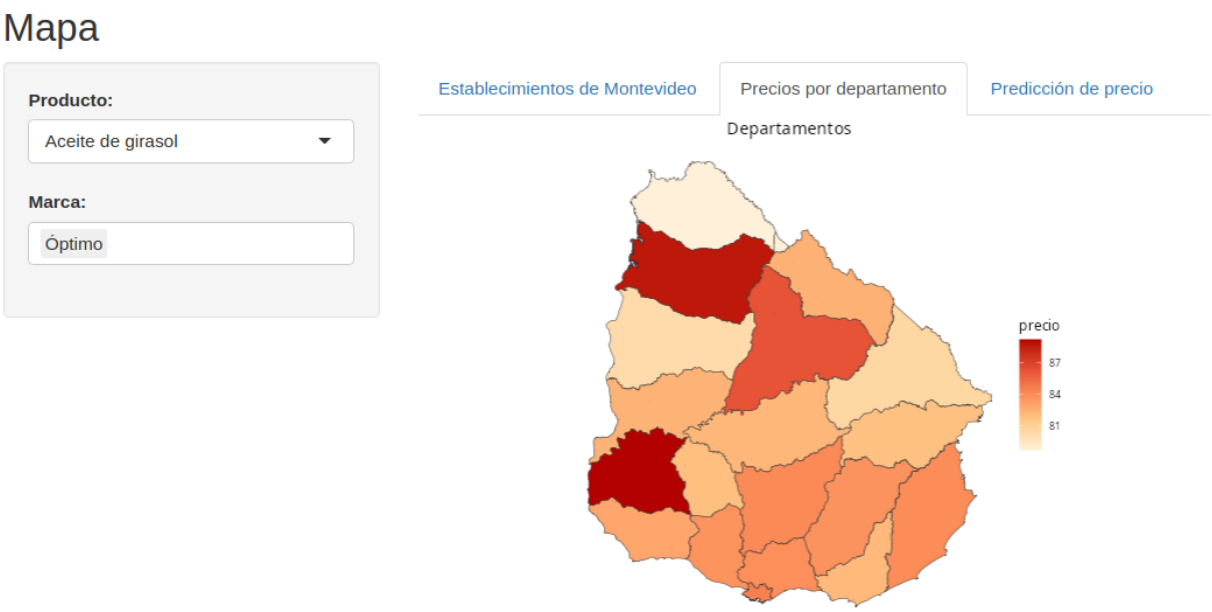


Figure 13: dpto_mapa

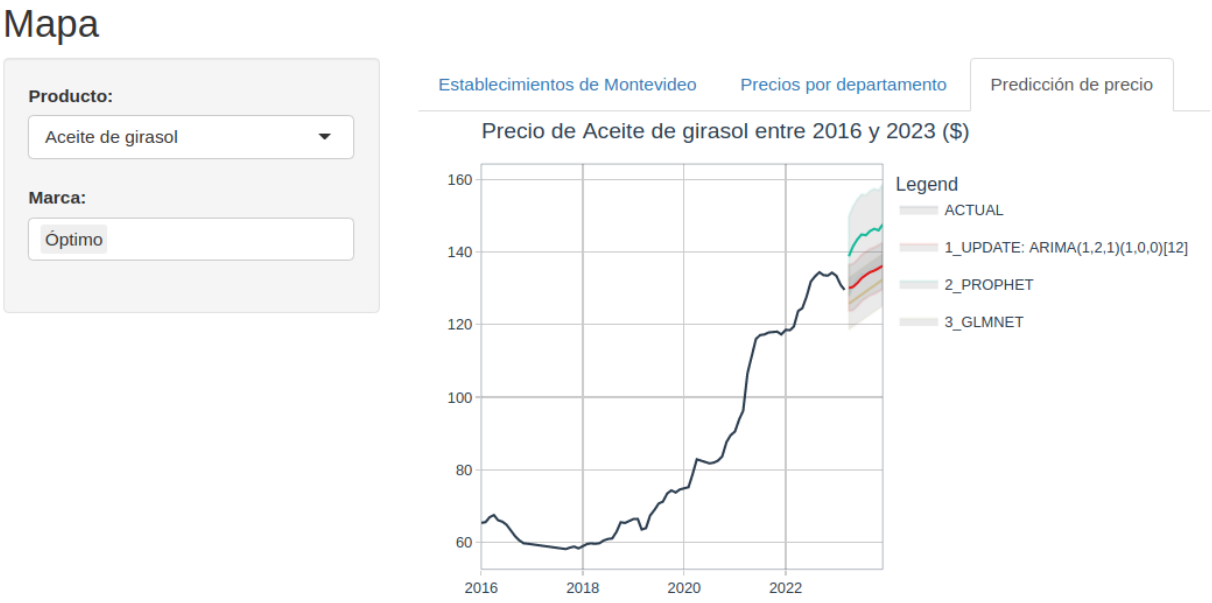


Figure 14: modelo

La aplicación se encuentra en https://my5poz-manuel0toledo.shinyapps.io/canasta_basica/

Conclusiones y Perspectiva a futuro

Los datos del Sistema de Información de Precios al Consumidor tienen mucho potencial de análisis si se trabaja sobre el marco correcto. La cantidad de dimensiones y variables con las que cuenta permite estudiar una amplio rango de problemas relacionados al consumo.

El modelo estadístico constituye un ejemplo de lo útil que es tener un registro de los precios de los productos a lo largo del tiempo, ya que estos pueden ser utilizados para construir predicciones que permiten una buena planificación tanto de políticas públicas como de las decisiones que toman las empresas y los consumidores.

Otro punto a resaltar es la amplia variedad de herramientas que provee R para fácilmente y sin ser expertos, poder armar distintos tipos de modelos y entender sus componentes y características para poder interpretarlos correctamente.

Por otro lado, sólo con un análisis superficial ya se encontraron varios datos erróneos o mal ingresados que deberían de evaluarse más a profundidad. En particular se encontraron varios datos de ubicación erróneos y establecimientos que no tienen los productos que dicen tener si se los visita físicamente.

También resultaría enriquecedor para el análisis que se incluyera una mayor representación de establecimientos en el interior del país, ya que la gran mayoría están ubicados en la capital; y una mayor cantidad de productos, especialmente aquellos que son más habituales para el consumo de la población, como por ejemplo las frutas y verduras.

Un análisis a futuro podría incluir distintos productos y la utilización de ponderadores para la canasta básica, como los que generó la escuela de nutrición (Prof. Agdo. Mag. Gabriela Fajardo, 2020). De esa forma sería posible analizar los productos de manera agregada en vez de individualmente.

Me gustó el contexto que le dieron al problema que pretendían resolver, una vez hecho el modelo predictivo el trabajo no termino allí ya que dado su pronóstico hicieron referencia a datos actuales y reales. La Shiny quedo bien una lástima que no puedan mostrarse los modelos.
Creo que un problema es el análisis de los precios sin deflactar y usar el promedio de eso para resumir.
Creo que aprendieron mucho en este trabajo, que enfrentaron muchos desafíos y que los resolvieron bien y que avanzaron en modelado que no llegamos aver en el curso y varias cosas para mejorar que van a ir aprendiendo en otros cursos de la Licenciatura.
Muy buen trabajo 95/100

Referencias

Alice. (2020). *Xgboost time series forecast in r*. http://datasideoflife.com/?p=1009

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2022). *Shiny: Web application framework for r*. https://CRAN.R-project.org/package=shiny

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., & Yuan, J. (2023). *Xgboost: Extreme gradient boosting*. https://CRAN.R-project.org/package=xgboost

Cheng, J., Karambelkar, B., & Xie, Y. (2023). *Leaflet: Create interactive web maps with the JavaScript 'leaflet' library*. https://CRAN.R-project.org/package=leaflet

Dancho, M. (2023). *Modeltime: The tidymodels extension for time series modeling*. https://CRAN.R-project.org/package=modeltime

Dancho, M., & Vaughan, D. (2023). *Timetk: A tool kit for working with time series*. https://CRAN.R-project.org/package=timetk

Defensa del Consumidor, I. de. (2021). *Tabla comparativa de precios de canasta básica de alimentos al 15 de setiembre de 2021*. https://www.gub.uy/ministerio-economia-finanzas/comunicacion/noticias/tabla-comparativa-precios-canasta-basica-alimentos-15-setiembre-2021

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, *26*(3), 1–22. https://doi.org/10.18637/jss.v027.i03

Hyndman, R., & Athanasopoulos, G. (2018). 3.4 evaluating forecast accuracy. In *Forecasting principles and practice*. OTEXTS.

(INE), I. nacional de estadística. (2023). *Boletín técnico: Encuesta de expectativas empresariales*. https://www5.ine.gub.uy/documents/Estad%C3%ADsticasecon%C3%B3micas/HTML/EEE/2023/Informe_Expectativas-jun23.html

Kuhn, M., & Wickham, H. (2020). *Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles*. https://www.tidymodels.org

Kuhn, & Max. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, *10*(1), 439–446. https://doi.org/10.32614/RJ-2018-009

Pebesma, E., & Bivand, R. (2023). *Spatial Data Science: With applications in R* (p. 352). Chapman and Hall/CRC. https://r-spatial.org/book/

Prof. Agdo. Mag. Gabriela Fajardo, Prof. MSc. M. B. (2020). *Canasta básica de alimentos con enfoque nutricional para la población uruguaya*. http://canastas.nutricion.edu.uy/wp-content/uploads/2022/06/Canasta-CBAN-Uruguay-1.pdf

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686