



UNIVERSIDAD
AUSTRAL | INGENIERÍA

MAESTRÍA EN EXPLOTACIÓN DE DATOS Y GESTIÓN DEL CONOCIMIENTO

ANÁLISIS DE SERIES TEMPORALES

TRABAJO PRÁCTICO N°2

ALUMNOS: **DEL VILLAR, JAVIER**
OTRINO, FACUNDO DAMIÁN
PISTOYA, HAYDEÉ SOLEDAD
ROJAS, MARIANO ARTURO
SORZA, EDWIN ANDRÉS
VAILLARD, LEANDRO CARLOS

FECHA: **22 DE AGOSTO DE 2021**

Resumen Ejecutivo

Este estudio fue realizado con la finalidad de poner en práctica los conocimientos adquiridos en la materia, para la realización de este trabajo se utilizaron datos de la producción, precio de venta y consumo de carne en Argentina, más específicamente series temporales en las cuales se registró:

- “Serie Faena” correspondiente la cantidad de cabezas de ganado faenadas.
- “Serie Consumo” correspondiente al consumo interno per cápita en el mercado argentino.
- “Serie Precio” correspondiente al precio en dólares de kilo de novillo vivo.

Para el desarrollo del trabajo se aplicaron dos tipos de aproximaciones para el modelado de dichas series que son la aplicación de:

- VAR que consiste en un modelo lineal de n-ecuaciones y n-variables en donde cada una de las variables se explica por sus propios valores anteriores. La ventaja de este modelo es que permite analizar los efectos de cualquier variable sobre otra, y medir el tiempo en que se tarda en estabilizar la variable después de una perturbación.
- AutoML (Auto Machine Learning): Esta aproximación es el resultado de la estandarización de soluciones de machine learning, esta metodología permite el entrenamiento de múltiples modelos con el conjunto de datos correspondiente a este trabajo práctico, posteriormente poder comparar los resultados y seleccionar los modelos que más se ajusten a las necesidades de análisis.

Los resultados obtenidos de este trabajo nos brindaron la posibilidad de conocer las distintas interacciones entre las variables de las series analizadas mediante la aplicación de distintos modelos predictivos, y con estas interacciones poder realizar predicciones que podrían brindar ayuda en la toma de decisiones en distintos campos de acción.

Este análisis realizado abre la puerta a una exploración más profunda sobre la realidad de producción, venta y consumo de carne en Argentina y su contexto en el mundo.

Abstract

The following study was carried out with the purpose of putting into practice the knowledge acquired in the “Temporal Series Analysis” course. The data used corresponds to the production, sales price, and consumption of meat in Argentina.

- “Serie Faena” shows the data corresponding to cattle heads that were butchered.
- “Serie Consumo” shows the data corresponding to the internal per capita consumption of meat.
- “Serie Precio” shows the data corresponding to the price expressed in USD per kilogram of live cattle.

For the development of the work, two types of approximations were applied for the modeling of said series, which are the application of:

- VAR consisting of a linear model of n-equations and n-variables where each of the variables is explained by its own previous values. The advantage of this model is that it allows us to analyze the effects of any variable on another and measure the time it takes to stabilize the variable after a disturbance.
- AutoML (Auto Machine Learning): This approach is the result of the standardization of machine learning solutions, this methodology allows the training of multiple models with the data set corresponding to this practical work, subsequently being able to compare the results and select the models that best suit the needs of analysis.

The results obtained from this work gave us the possibility of knowing the different interactions between the analyzed variables by applying different predictive models, and with these interactions to be able to make predictions that could help to decision-making in different action field.

This analysis carried out opens the door to a more in-depth exploration of the reality of meat production, sale and consumption in Argentina and its context in the world.

Índice

Resumen Ejecutivo.....	1
Abstract.....	2
Consigna.....	4
Origen de los Datos.....	4
Código Empleado para el Desarrollo del Presente Trabajo	4
Introducción (Punto 1).....	5
Marco Teórico (Punto 1).....	6
Modelo VAR	6
Cointegración	8
Auto Machine Learning (AutoML).....	9
Análisis de Resultados.....	12
Prueba de Ljung-Box	12
Modelo de Vectores Autorregresivos (Punto 2).....	13
Análisis de estacionariedad de las series.....	13
Pruebas de causalidad en el sentido de Granger.....	15
Identificación del orden del VAR	18
Estabilidad del modelo.....	18
Pruebas de especificación.....	19
Pruebas de autocorrelación en los residuos.....	19
Prueba Ljung-Box	19
Pruebas de Normalidad de los residuos	22
Prueba de homocedasticidad de los residuos	22
Estabilidad estructural de los residuos	23
Análisis Impulso Respuesta y Análisis de Descomposición de la Varianza	23
Análisis impulso respuesta.....	24
Análisis de descomposición de la varianza	25
Modelo Alcanzado	26
Otros modelos (Punto 3).....	27
Pronósticos Alcanzados (Punto 4).....	30
Conclusiones	32
Bibliografía	34

Consigna

A continuación, se presenta la consigna del presente trabajo práctico:

1. Buscar tres series de tiempo y exponer la problemática de interés analítico (es importante detallar cuáles fueron los motivos de elección y situar al lector en el contexto adecuado).
2. Construir un modelo de Vectores Autorregresivos (VAR) que será utilizado para realizar las predicciones. Justificar la elección del modelo con todo lo visto en clase. Es posible complementar con otros análisis.
3. Para cada serie empleada en el modelo VAR fittear un modelo mediante la utilización de alguna técnica de Machine Learning (Redes Neuronales, SVM, AutoML, Darts, Prophet, etc.).
4. Pronosticar con el modelo seleccionado para cada serie para una ventana temporal razonable, en función de la periodicidad y el comportamiento de las mismas.

Origen de los Datos

El conjunto de datos a ser analizado proviene de la página web de datos del Ministerio de Agricultura, Ganadería y Pesca. Se puede acceder a ellos por medio del siguiente vínculo:

<https://datos.agroindustria.gob.ar/dataset/indicadores-mensuales-sector-bovino/archivo/7afe10d1-e9bc-4383-9e3c-c8066bc21f65>

Para el desarrollo del presente trabajo, se ha denominado:

- “Serie Faena” al análisis correspondiente a la faena de ganado vacuno durante el período bajo análisis.
- “Serie Consumo” al análisis correspondiente al consumo per cápita de carne vacuna durante el período bajo análisis.
- “Serie Precio” al análisis correspondiente al precio de los novillos por kilos expresados en dólares estadounidenses (USD).

Código Empleado para el Desarrollo del Presente Trabajo

El código empleado para el desarrollo del presente trabajo se encuentra almacenado en los siguientes repositorios de Kaggle:

- VAR: <https://www.kaggle.com/haydysole/carnes-var-vectores-auto-regresores-final>
- AutoML: <https://www.kaggle.com/fotrino/carnes-automl>

Introducción (Punto 1)

En el Trabajo Práctico N°1 fueron analizadas dos dimensiones importantes en la industria de la carne vacuna (bovina) tales como la cantidad de cabezas faenadas y el consumo de carne per cápita. Dado que dichas series arrojaron resultados interesantes, se ha optado por continuar profundizando el análisis al agregar la serie de tiempo que indica la evolución del precio por kilogramo de novillo expresado en dólares estadounidenses. La decisión de utilizar esta serie radica en que resulta de interés conocer cómo fue la variación del precio del kilogramo de novillo vivo utilizando una moneda de referencia dado que durante el período de análisis (1998-2019) el peso argentino (ARS) ha sufrido grandes cambios en su cotización (pasando de un tipo de cambio 1 ARS = 1 USD en 1998, a 63.12 ARS = 1 USD a fines de agosto de 2019) y adicionalmente que dicha serie permite agregar una dimensión adicional como ser el precio el cual en la transaccionalidad es el vínculo que permite complementar a la oferta (faena) con la demanda (consumo) ya sea de índole interna (consumo interno) o externa (Exportaciones).

Según el artículo “Caída del consumo de carne vacuna en Argentina” publicado en el portal de la Bolsa de Comercio de Rosario, “en 2010 se consumían -de acuerdo con las estadísticas oficiales- 58 kg de carne vacuna, 35 kg de pollo y 8 kg de cerdo, es decir 101 kg totales. En 2020, el consumo de carne de vacuna cayó a 50 kg, el pollo pasó a 45 kg y el cerdo a unos 14 kg per cápita. Claramente, el consumidor también ha experimentado cambios en sus hábitos de consumo que lo llevan a incorporar otras opciones proteicas, tanto de origen animal como vegetal. En este sentido, no todo es precio en materia de consumo, también intervienen aspectos menos tangibles que paulatinamente van definiendo el perfil del consumidor.” (Della Siega, M., 2021). Si bien el autor de la cita expone que el precio no es el único factor que puede llegar a explicar la baja en el consumo, se ha optado por llevar a cabo el análisis por ser una variable de carácter cuantitativo mensurable. En el caso de los cambios en el gusto de los consumidores, esto puede ser considerado como una apreciación por parte del autor al no contar con datos que permitan explicar en forma fehacientemente que ese es el motivo de reducción en el consumo de carne vacuna y no factores económicos.

El presente trabajo tiene como finalidad el análisis de las series temporales que se mencionaron en la sección “Origen de los Datos” utilizando para ello un modelo de Vectores Auto Regresivos. Además, se desarrollará un modelo que permita predecir el comportamiento futuro de alguna/s de las series temporales elegidas analizando su autocorrelación y con las demás series de tiempo. Dicho modelo se empleará para el análisis y desarrollo de pronósticos para las tres series. Por último, se emplearán distintas herramientas de Auto Machine Learning (Auto ML) para el desarrollo de modelos óptimos para las series seleccionadas.

Marco Teórico (Punto 1)

Modelo VAR

Según (Stock, J. y Watson, M., 2001) una regresión univariada es un modelo de una ecuación y una variable en donde el valor actual de una variable se explica por los valores anteriores. Un Vector Auto Regresivo (VAR) es un modelo lineal de n-ecuaciones y n-variables en donde cada una de las variables se explica por sus propios valores anteriores, además de los valores actuales y pasados de las n-1 variables remanentes. En cambio, (Del Rosso, 2021) define al modelo VAR como “un modelo de ecuaciones simultáneas formado por un sistema de ecuaciones de forma reducida sin restringir. Que sean ecuaciones de forma reducida quiere decir que los valores contemporáneos de las variables del modelo no aparecen como variables explicativas en ninguna de las ecuaciones. Por el contrario, el conjunto de variables explicativas de cada ecuación está constituido por un bloque de rezagos de cada una de las variables del modelo.” El uso del modelo VAR permite analizar los efectos de cualquier variable sobre otra, y medir el tiempo en que se tarda en estabilizar la variable después de una perturbación. Estos modelos tienen las siguientes propiedades (Del Rosso, 2021):

- Parte de un enfoque ateórico; y,
- Es capaz de separar los efectos pasados que explican al vector de las variables endógenas a través de su pasado o mediante variables autoregresivas.

Según (Stock, J. y Watson, M., 2001), existen tres variedades de los modelos VAR:

- **Forma reducida:** un VAR expresado en forma reducida es una función lineal de sus propios valores pasados y los valores pasados de todas las otras variables consideradas y una serie de términos de error no correlacionados. Los términos de error de estas regresiones son los movimientos sorpresivos en las variables luego de considerar los valores pasados. Si las diferentes variables se encuentran correlacionadas, entonces los términos de error estarán correlacionados en todas las ecuaciones.
- **Recursivo:** un VAR recursivo construye los términos de error en cada una de las ecuaciones regresivas serán no correlacionados con el error de las ecuaciones precedentes. Esto se logra al incluir en forma juiciosa algunos valores contemporáneos como regresores.
- **Estructural:** un VAR estructural utiliza la teoría económica para ordenar los vínculos contemporáneos entre las variables. Estos modelos requieren identificar los supuestos que permiten interpretar la causalidad de las correlaciones. Estos supuestos pueden involucrar a todo el VAR, por lo tanto, todos los vínculos causales en el modelo pueden mostrarse en una única ecuación que permita identificar un vínculo casual específico. La cantidad de modelos de VAR estructural depende de la inventiva del investigador.

El modelo VAR estructural postulado por Sims en 1980 consiste en la modelación conjunta de las series:

$$y_t = b_{10} - b_{12}z_t + \gamma_{11}y_{t-1} + \gamma_{12}z_{t-1} + \varepsilon_{y_t}$$

$$z_t = b_{20} - b_{21}y_t + \gamma_{21}z_{t-1} + \gamma_{22}y_{t-1} + \varepsilon_{z_t}$$

Donde se asume (Montes-Rojas, G., s.f.):

- y_t y z_t son estacionarias;

- $\varepsilon_{y_t} \sim (0, \sigma_y)$ y $\varepsilon_{z_t} \sim (0, \sigma_z)$ son ruido blanco; y,
- ε_{y_t} y ε_{z_t} no están autocorrelacionados. Estos se definen como “shocks” que son cambios exógenos con sentido económico.

El modelo VAR, en su versión más simple, se puede expresar en forma estándar como: $x_t = A_0 + A_1 x_{t-1} + e_t$, donde:

$$e_t = \begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix} = \begin{pmatrix} \frac{\varepsilon_{y_t} - b_{12}\varepsilon_{z_t}}{1 - b_{12}b_{21}} \\ \frac{\varepsilon_{z_t} - b_{21}\varepsilon_{y_t}}{1 - b_{12}b_{21}} \end{pmatrix}$$

Los elementos del vector e_t cumplen con tener valor esperado igual a cero, varianzas constantes y están individualmente no correlacionados (Del Rosso, 2021). La varianza de estas perturbaciones se calcula mediante el momento de segundo orden:

$$E(e_{1t}^2) = \frac{\sigma_y^2 + b_{12}^2 \sigma_z^2}{(1 - b_{12}b_{21})^2} \quad E(e_{2t}^2) = \frac{\sigma_z^2 + b_{21}^2 \sigma_y^2}{(1 - b_{12}b_{21})^2}$$

Las autocorrelaciones de orden j valen 0, con lo que se demuestra que e_t es un proceso estacionario. Cabe destacar que en el caso que e_{1t} y e_{2t} estén correlacionados, por tanto, los dos shocks están correlacionados. Si $b_{12} = b_{21} = 0$, entonces no hay efectos contemporáneos de y_t sobre z_t ni al revés, los dos shocks están no correlacionados: $\sigma_{21} = \sigma_{12}$.

Según (Del Rosso, 2021), un modelo VAR de orden n se especifica de la siguiente manera:

$$x_t = A_0 + A_1 x_{t-1} + A_2 x_{t-2} + \dots + A_p x_{t-p} + e_t$$

De estos modelos, (Novales, 2017) plantea las siguientes consideraciones para su estimación:

- Todas las variables del modelo deben ser tratadas simétricamente, siendo explicadas por el pasado de todas ellas. El modelo tiene tantas ecuaciones como variables, y los valores retardados de todas las ecuaciones aparecen como variables explicativas en todas las ecuaciones.
- Una vez estimado el modelo, se pueden excluir algunas variables explicativas en función de su significancia estadística, pero no es recomendable. Esto se debe a que la poca significancia de variables puede deberse a la colinealidad inherente al modelo y no tanto a la falta de contenido informativo de las variables.
- En el modelo VAR puede estimarse con bastante precisión los elementos globales del modelo, como el R2, la derivación típica residual, y los mismos residuos, o el efecto global de una variable sobre las otras. Sin embargo, no cabe hacer interpretaciones de coeficientes individuales en distintos retardos, ni llevar a cabo contrastes de hipótesis sobre coeficientes individuales.
- A mayor orden del modelo VAR (mayor cantidad de ecuaciones), mayor cantidad de parámetros a estimar. Este crecimiento no es lineal, por tanto, se vuelve un proceso laborioso cuando se intenta aplicar a modelos VAR de orden alto.

(Novales, 2017) indica que un modelo VAR no se estima para hacer inferencia acerca de coeficientes de variables individuales. Dada la baja precisión en su estimación, se desaconseja cualquier análisis de coeficientes individuales y tiene mucho sentido el análisis conjunto de los coeficientes asociados a un

bloque de retardos en una determinada ecuación. Una estrategia para encontrar el orden del modelo VAR consiste en examinar los denominados criterios de información, que son determinadas correcciones sobre el valor muestral de la función logaritmo de verosimilitud. Los más conocidos son los de Akaike (AIC) y Schwartz (SBC o BIC):

$$AIC = -2 \frac{l}{T} + 2 \frac{P}{T}$$

$$BIC = -2 \frac{l}{T} + p \frac{\ln(T)}{T}$$

$$Hannan - Quinn = -2 \frac{l}{T} + 2 \frac{k \ln(\ln(T))}{T}$$

Siendo $l = -\frac{Tk}{2} (1 + \ln 2\pi) - \frac{T}{2} |\hat{\Sigma}|$, y $p = k(d + nk)$ el número de parámetros estimados en el modelo VAR, siendo d el número de variables exógenas, n el orden del VAR y k el número de variables endógenas.

Cointegración

Según (Peña, 2010), es posible que, aunque las dos series x e y sean no estacionarias, una de ellas explique totalmente el comportamiento de la otra. Si las series están relacionadas por $y_t = \beta_0 + \beta_1 x_t + n_t$ donde n_t es estacionario, la variable x_t es capaz de explicar totalmente el comportamiento no estacionario de y_t , entonces las variables están cointegradas.

Un serie temporal es integrada de orden d ($I(d)$), si es necesario aplicar d diferencias para transformarla en estacionaria, o $I(0)$. Las series $I(d)$ integradas, x_t, y_t , están cointegradas si existe una combinación lineal entre ellas que es de orden de integración menor de d . Es decir, si se puede construir una serie $n_t^* = \alpha_1 y_t + \alpha_2 x_t$ que es $I(d_1)$, donde $d_1 < d$. A la combinación (α_1, α_2) se la denomina relación de cointegración. Esta relación no es única, ya que cualquier relación del tipo $(c\alpha_1, c\alpha_2)$ es también de cointegración para cualquier $c \neq 0$. Cuando existe cointegración, una de las dos variables explica parte de la tendencia de la otra. Al dividir la ecuación anterior por α_1 y reordenando los términos se puede escribir $y_t = \beta x_t + n_t$ donde $\beta = -\frac{\alpha_2}{\alpha_1}$ y $n_t = \frac{n_t^*}{\alpha_1}$ es un proceso $I(d_1)$.

En caso de que exista cointegración al construir el modelo de función de transferencia encontraremos que la variable independiente o exógena explicará totalmente la no estacionariedad de la variable dependiente, y la perturbación seguirá un proceso estocástico.

Tal como menciona (Del Rosso, 2021), desde el punto de vista del fenómeno a estudiar, se dice que dos o más series están cointegradas si las mismas se mueven conjuntamente a lo largo del tiempo y las diferencias entre ellas son estables (es decir estacionarias), aún cuando cada serie en particular contenga una tendencia estocástica y sea por lo tanto no estacionaria. Por lo tanto, la integración refleja la presencia de un equilibrio a largo plazo hacia el cual converge el sistema en el tiempo. Las diferencias (o términos de error) en la ecuación de cointegración se interpretan como el error de desequilibrio para cada punto particular de tiempo.

Existen dos métodos para estudiar la cointegración:

- El contraste de causalidad de Granger resulta interesante, donde se supone que se está explicando el comportamiento de una variable y utilizando su propio pasado. Se dice que una

variable x no causa a la variable y si al añadir el pasado de x a la ecuación anterior no añade capacidad explicativa. El contraste consiste en analizar la significación estadística del bloque de retardos de x en la ecuación mencionada, y la hipótesis nula es que la variable x no causa, en el sentido de Granger, a la variable y . (Novales, 2017)

- El modelo de Johansen que es aplicable a sistemas de ecuaciones y se encuentra basado en modelos VAR. Es un test de máxima verosimilitud que requiere muestras grandes (100 o más datos) y prueba la existencia de múltiples vectores de cointegración entre las variables, mediante la prueba de la Traza y el Eigenvalor máximo. Este modelo descansa fuertemente en la relación entre el rango de la matriz y sus raíces características. (Del Rosso, 2021).

(Montes-Rojas, s.f.) enuncia que para poder hacer el contraste de causalidad de Granger se deben tener en cuenta las siguientes hipótesis:

- $H_0: y_t, x_t$ no están cointegradas.
- $H_1: y_t, x_t$ están cointegradas

El proceso de contraste de hipótesis consiste en los siguientes pasos:

- Correr la regresión de y_t en x_t .
- Construir residuos $\hat{e}_t = y_t - \beta x_t$.
- Usar el contraste Dickey-Fuller (o Dickey-Fuller aumentado) a la serie \hat{e}_t , donde los valores críticos deben ser ajustados porque estamos estimando β .

Auto Machine Learning (AutoML)

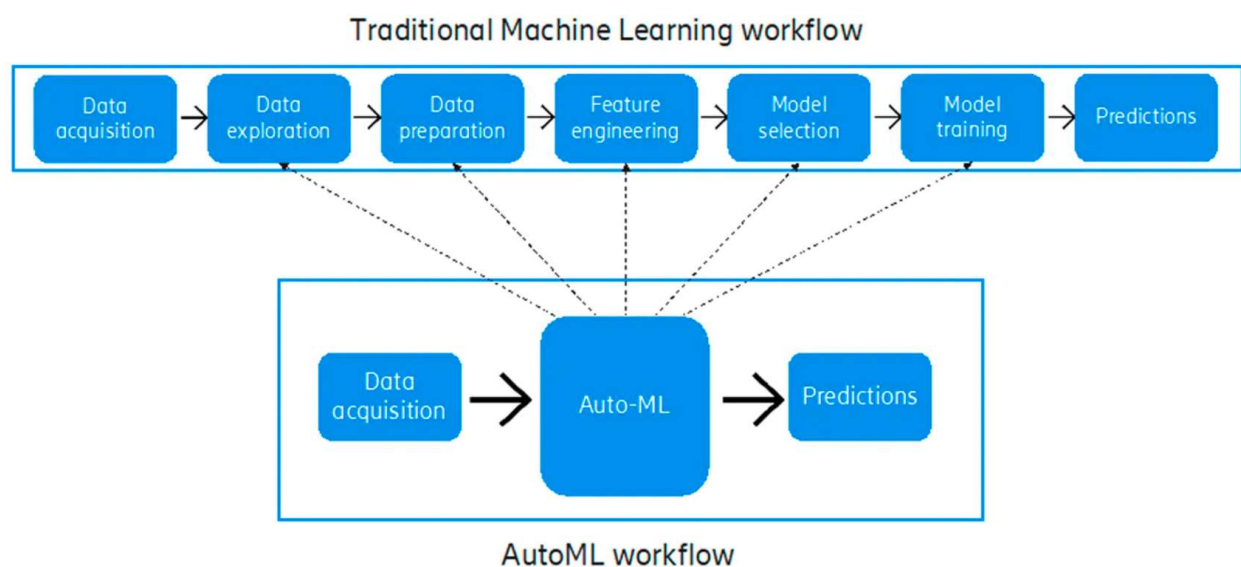
En (Sarafanov, M., 2021) se menciona que en la actualidad los científicos de datos (*data scientists*) recolectan y filtran los datos antes de alimentar a los modelos de *machine learning*. Luego, entrenan dichos modelos para poder seleccionar el que arroje los mejores resultados. Según (TechyTacos, 2019), el *pipeline* para un modelo de *machine learning* tradicional involucra los siguientes pasos: adquisición de datos, exploración de datos, preparación de datos, *feature engineering*, selección del modelo, entrenamiento del modelo, ajuste de los hiperparámetros y predicción. En cambio, según (Sarafanov, M., 2021) el AutoML fue desarrollado con la finalidad de ahorrar tiempo a los científicos de datos en las tareas de poco valor agregado. Para ello se desarrollaron *pipelines* que permiten automatizar los pasos de recolección, pre-proceso, *feature engineering*, filtrado y remoción de valores atípicos. Existen numerosas herramientas para generar estos *pipelines* automáticos, tales como TPOT, AutoGluon, MLJAR y H2O. Lo que ocurre con estos *pipelines* es que todos funcionan en forma similar y lo único que varían son los hiperparámetros de los modelos aplicados.

Según (Sarafanov, M., 2021), como regla general, las librerías de AutoML sólo sirven para tareas generalizadas, como la clasificación y regresión de información tabulada. Con mucha menor frecuencia se utiliza el AutoML para el procesamiento de texto, imágenes y predicciones de series de tiempo. Uno de los motivos por los que el AutoML no se emplea en forma generalizada a la predicción de series de tiempo es la dificultad en adaptar la funcionalidad de las librerías existentes para predecir series temporales sin antes ajustarlas para las demás tareas (clasificación y regresión). Esto se debe a que la predicción de series de tiempo difiere de los problemas genéricos de regresión. Por ejemplo, sería ilógico reordenar la serie en forma aleatoria para validar el modelo de series temporales.

(Liang, C. y Lu, Y., 2020) plantean que la predicción de series temporales utilizando *machine learning* presenta algunas dificultades. Existe la incertidumbre donde se busca predecir el futuro en base a datos pasados. A diferencia de otros modelos de *machine learning*, el conjunto de pruebas puede tener una distribución diferente a la de la serie original. Además, las series temporales reales pueden tener intermitencias, esto es, que los datos se encuentren incompletos. Por otra parte, para poder desarrollar una herramienta automatizada para el pronóstico series temporales para todo propósito ya que esta debe ser aplicable para una gran cantidad de conjunto de datos.

(Liang, C. y Lu, Y., 2020) hacen alusión a la metodología que ellos aplicaron al desarrollar su AutoML, pero se pueden generalizar sus hallazgos dado que, en todos los casos, los modelos de AutoML buscan la mejor combinación de hiperparámetros de distintos modelos como XGBoost, Gradient Boosting, LightGBM, TensorFlow, entre otros, para lograr la mejor predicción de series de tiempos.

En resumen, tal como explica (Martínez F., 2021), la comparación de los procesos tradicionales de *machine learning* y AutoML se puede resumir por la imagen a continuación:



En (Martínez, F., 2021) se indica que para la selección del mejor modelo de AutoML se toman en cuenta las medidas de precisión, entre las que se encuentran:

- **Desvío Residual Medio:** en caso de que la distribución sea gausiana, entonces es igual al Mean Square Error (MSE), y cuando no lo es, provee una estimación más útil del error.
- **Root Mean Square Error (RMSE):** evalúa qué tan bien un modelo puede predecir un valor continuo. Las unidades de medida son las mismas que la variable objetivo, lo que es útil para comprender el tamaño del error. Cuanto más pequeña sea el RMSE, mejor será el rendimiento del modelo. Dicha métrica es sensible a valores atípicos.

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

Donde y_i es la variable objetivo, \hat{y}_i es la predicción y N es el total de filas (observaciones de la base).

- **Mean Square Error (MSE):** es la métrica que mide el promedio de los cuadrados de los errores o desviaciones. MSE toma las distancias desde los puntos hasta la línea de regresión (estas distancias son los errores) y las eleva al cuadrado para eliminar cualquier signo negativo. Además, incorpora tanto la varianza, como el sesgo del predictor y es sensible a valores atípicos. El MSE también da más peso a diferencias más grandes. Cuando mayor es el error, más se penaliza. Cuanto más pequeño sea el MSE, mejor será el rendimiento del modelo.

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

- **Mean Absolute Error (MAE):** es un promedio de los errores absolutos. Las unidades del MAE son las mismas que la variable objetivo, lo que es útil para comprender si el tamaño del error es significativo. Cuando menor sea el MAE, mejor será el rendimiento del modelo. Además, es robusto frente a los valores atípicos.

$$MSE = \frac{1}{N} \sum_{i=1}^N |x_i - \hat{x}_i|$$

Donde $|x_i - \hat{x}_i|$ son los errores absolutos.

- **Root Mean Square Logarithmic Error (RMSLE):** evalúa qué tan bien un modelo puede predecir un valor continuo. Las unidades de la métrica son las mismas que la variable objetivo. Cuanto más pequeño sea este error, mejor será el rendimiento del modelo. Es sensible a valores atípicos.

$$RMSLE = \sqrt{\frac{1}{N} \sum_{i=1}^N \left(\ln \left(\frac{y_i + 1}{\hat{y}_i + 1} \right) \right)^2}$$

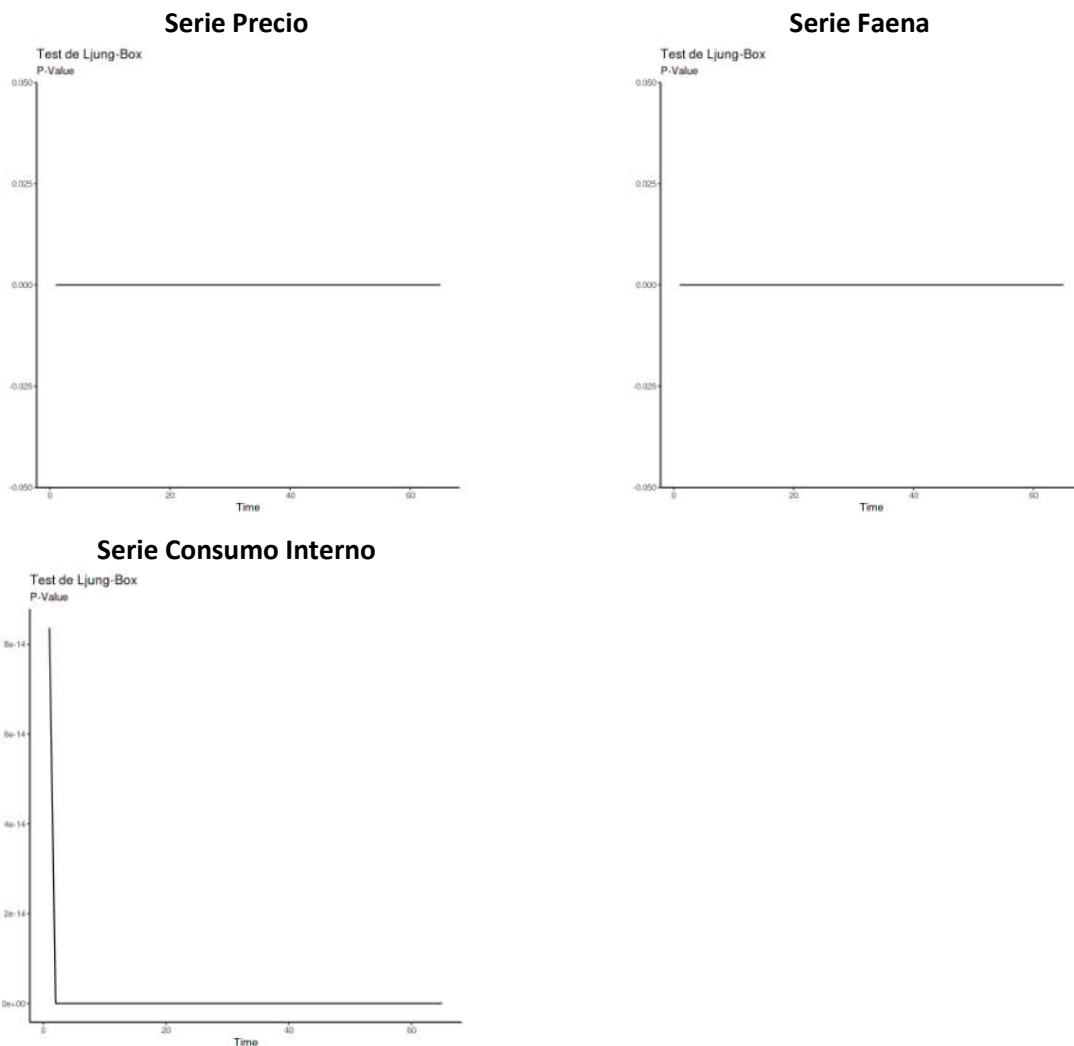
Análisis de Resultados

Prueba de Ljung-Box

Según (Glen, S., 2018), la prueba de Ljung-Box (también conocida como prueba de Box-Pierce) modificada es una forma de comprobar si hay ausencia de autocorrelación hasta una cantidad determinada de períodos. Esto quiere decir que la prueba busca que los errores no estén correlacionados (conformen una serie de ruido blanco). La hipótesis nula de esta prueba es que el modelo no muestra una falta de ajuste a una serie temporal, mientras que la hipótesis alternativa es que sí existe una falta de ajuste a una serie temporal. Esto es importante de notar ya que en caso de que la serie muestre una falta de ajuste a una serie temporal, se dirá que se trata de una serie conformada por ruido blanco y no podrá aplicarse al análisis posterior.

Un p-valor significativo (mayor a 0.05) significa que se debe rechazar la hipótesis nula, y, por tanto, se debe seleccionar otra serie para realizar el análisis.

Al aplicar la función denominada “Incorrelación” (Martinez, F., 2021), se obtuvieron los siguientes resultados:



Luego de realizada la prueba de Ljung-Box se observó que las series “KgVivo” y “Faena” arrojaron un p-valor igual a 0 para todos los retraso. En cambio, la serie “Consumo Interno” arrojó un valor distinto de cero para el primer período (8×10^{-14}). Dicho valor sigue siendo significativamente menor a 0.05, por tanto no se puede rechazar la hipótesis nula, lo que significa que no hay evidencia para descartar que los datos corresponden a una serie temporal.

Modelo de Vectores Autorregresivos (Punto 2)

Con el objeto de desarrollar esta sección lo primero a analizar será la condición de estacionariedad de cada una de las series de tiempo en cuestión, en el caso que no la cumpliesen se realizarán ajustes sobre ellas para alcanzar tal condición.

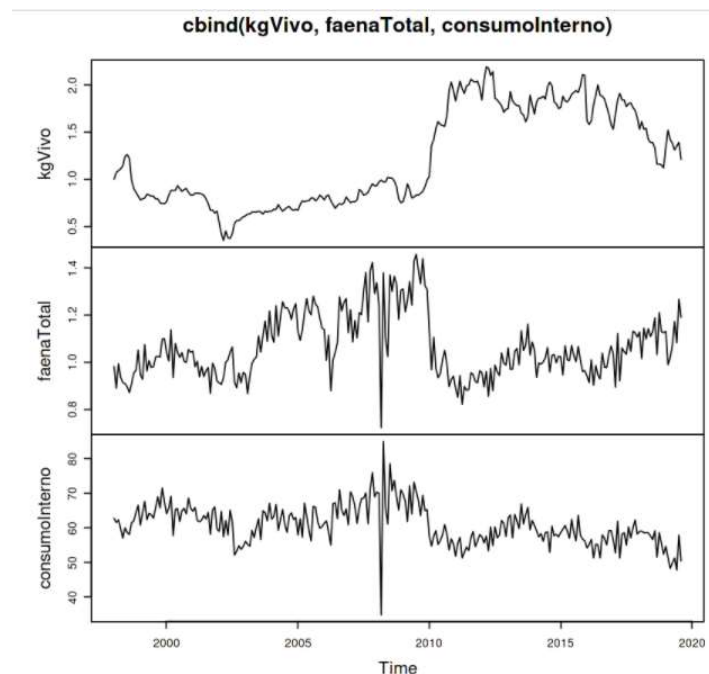
Validados los primeros pasos se aplicará el test de Granger con el objetivo de detectar si existe o no causalidad en el sentido de Granger esto es, identificar cual es la variable dependiente y cuales las variables explicativas

Si las variables verifican la causalidad de Granger, se procederá a identificar el orden del VAR y se realizarán los tests necesarios a fin de examinar el modelo alcanzado, en este sentido, se analizará su estabilidad, y sus residuos para verificar que éstos no presenten autocorrelación, sean homocedásticos y se distribuyan de manera normal.

Si el modelo alcanzado verificase lo anterior se procederá a aplicar uno de los usos más comunes del VAR, esto es, ver cómo cada variable afecta y es afectada por las otras variables del modelo.

Análisis de estacionariedad de las series

Se procede a analizar la estacionariedad de las series a utilizar.



A fin de analizar la condición de estacionariedad se aplicó el test de Dickey-Fuller.

Dicho test tiene las siguientes hipótesis:

- $H_0: \varphi_1 - 1 = 0$ es decir, el proceso es no estacionario, la serie tiene una raíz unitaria.
- $H_1: \varphi_1 - 1 < 0$, es decir, hay evidencia suficiente para suponer que el proceso es estacionario, la serie no tiene una raíz unitaria.

Los resultados alcanzados por las series bajo análisis son las siguientes:

	P - value	Status
Serie Faena	0.4123	No Rechazo H0
Serie Consumo	0.249	No Rechazo H0
Serie Precios	0.5604	No Rechazo H0

Siendo que los p-values fueron superiores a 0.05, no se rechaza la hipótesis nula, lo que significa que las series originales no cumplen con el requisito de estacionariedad.

Dado que ninguna de las 3 series verifica que sean estacionarias se realizaron una serie de transformaciones con el objeto de alcanzar dicha condición. En este sentido, se han aplicado los siguientes ajustes a la serie, algunos de los cuales se encuentran descritos por (Athanasopoulos & Hyndman, 2018):

- **Transformaciones:** por medio de logaritmos que pueden ayudar a estabilizar la varianza de una serie de tiempo y suavizar sus picos.
- **Diferenciaciones:** las cuales se utilizan para intentar eliminar la tendencia y la estacionalidad en la media de la serie.

En el caso de la diferenciación simple esta ayuda a eliminar o reducir el impacto de la tendencia de una serie, y puede ser escrita de la siguiente manera

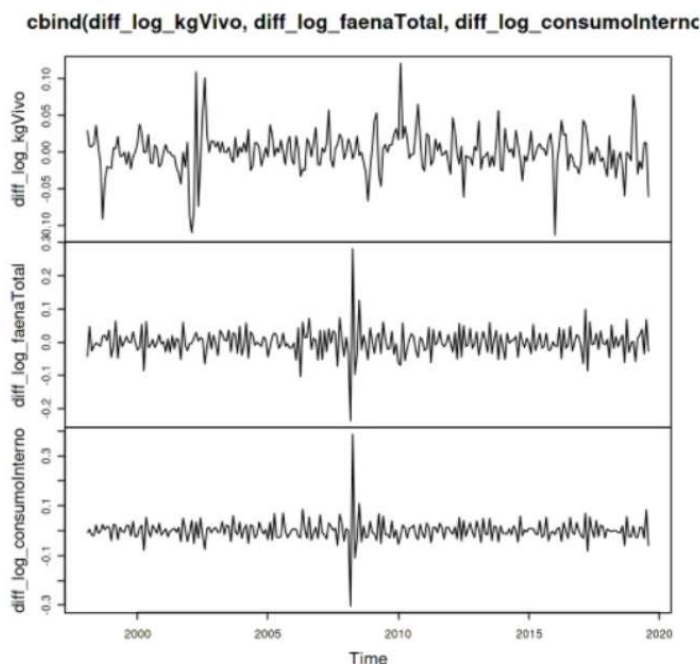
$$y'_t = y_t - y_{t-1}$$

En el caso de la diferenciación estacional, busca eliminar o mitigar el impacto que puede tener la presencia de estacionalidad en la serie, en este sentido se refiere a la diferencia entre la observación y la observación previa referida a la misma estación, esto es, por ejemplo, en el caso que la estacionalidad tuviera una duración de 12 meses, este ajuste debiera restar al valor de enero 2020, el valor de enero 2019.

En términos formales:

$$y_t = y_{t-m} + \varepsilon_t$$

Una vez aplicados los ajustes se obtuvieron las siguientes series:



Al aplicar los tests de estacionariedad se verifica que por medio de dichas transformaciones se alcanza la estacionariedad.

	P – value	Status
Serie Faena	0.01	Rechazo H0
Serie Consumo	0.01	Rechazo H0
Serie Precios	0.01	Rechazo H0

Con p-values inferiores a 0.05 se rechaza la hipótesis nula, por ende, se rechaza la hipótesis de no estacionariedad.

Pruebas de causalidad en el sentido de Granger

Se aplica la prueba de causalidad en el sentido de Granger con el objeto de analizar si los rezagos de una o varias variables ayudan a explicar el valor actual de otra.

El test de Causalidad de Granger, en su hipótesis nula, postula que los coeficientes de los rezagos de las demás variables en la ecuación de regresión son iguales a 0, es decir, que no afectan a la variable endógena (y).

Según lo indicado por (Kirchgässner, Wolters, 2007), asumiendo un proceso estacionario definido de forma débil, si I_t corresponde al total de la información disponible en el tiempo t , dicho total de la información incluye principalmente las dos series de tiempo 'x' e 'y'. Si \bar{x}_t corresponde a la totalidad de todos los valores de x por ejemplo, $\bar{x}_t = \{x_t, x_{t-1}, \dots, x_{t-k}\}$ y lo mismo aplica para y, por otro lado, $\sigma^2(\cdot)$ es la varianza del error de estimación.

Bajo tales supuestos, Granger propone la siguiente definición de causalidad entre 'x' e 'y'. 'x' causa en el sentido de Granger a 'y', si y solo si, de la aplicación de una predicción lineal la función alcanza lo siguiente:

$$\sigma^2(y_{t+1}|I_t) < \sigma^2(y_{t+1}|I_t - \bar{x}_t)$$

Esto significa que, los valores futuros de 'y' pueden ser estimados de mejor manera, es decir, con una menor varianza, si los valores de 'x' son introducidos en el modelo.

El primer término de la igualdad indicaría que la varianza del modelo que incluye ambas variables (x e y) es inferior a la varianza alcanzada por medio de un modelo que excluye a la variable x. De verificarse lo indicado, se puede decir que la presencia de 'x' genera un mejor modelo que la inexistencia de este en el modelo.

En función de lo comentado, se analizó si existía causalidad en el sentido de Granger entre cada una de las variables por medio del test "grangertest". Los resultados fueron los siguientes:

Test	p-value	Resultado	Conclusión	
diff_log_kgVivo ~ diff_log_faenaTotal	0.20532	No Rechazo H0	"Faena Total" no causa en el sentido de Granger a "Kg Vivo"	
diff_log_kgVivo ~ diff_log_consumoInterno	0.73700	No Rechazo H0	"Consumo interno" no causa en el sentido de Granger a "Kg Vivo"	
diff_log_faenaTotal ~ diff_log_kgVivo	0.13500	No Rechazo H0	"Kg Vivo" no causa en el sentido de Granger a " Faena Total"	
diff_log_faenaTotal ~ diff_log_consumoInterno	0.00085	RechazoH0	"Consumo interno" causa en el sentido de Granger a "Kg Vivo"	
diff_log_consumoInterno ~ diff_log_kgVivo	0.05147	No Rechazo H0	"Kg Vivo" no causa en el sentido de Granger a " Consumo Interno"	*
diff_log_consumoInterno ~ diff_log_faenaTotal	0.53540	No Rechazo H0	"Faena Total" no causa en el sentido de Granger a " Consumo Interno"	

Tal como se observa, en principio, la única variable que causa en el sentido de Granger a otra es el caso de "Consumo interno" respecto a "Kg Vivo". Sin embargo, cabe señalar que, si bien no se rechazó la hipótesis nula que "Kg Vivo" no causa en el sentido de Granger a "Consumo Interno", el p-value estuvo cercano al límite de la zona de rechazo.

En función de ello también se aplicó el test de Granger por medio de la función "causality" del paquete vars, el cual, a diferencia del anterior, toma el VAR estimado para desarrollar el test.

Test	p-value	Resultado	Conclusión
diff_log_faenaTotal ~ diff_log_kgVivo & diff_log_consumoInterno	0.03785	Rechazo H0	"Kg Vivo" & "Consumo Interno" causan en el sentido de Granger a "Faena Total"
diff_log_consumoInterno ~ diff_log_kgVivo & diff_log_faenaTotal	0.0365	Rechazo H0	"Kg Vivo" & "Faena Total" causan en el sentido de Granger a "Consumo Interno"
diff_log_kgVivo ~ diff_log_faenaTotal & diff_log_consumoInterno	0.7335	No Rechazo H0	"Consumo Interno" & "Faena Total" no causan en el sentido de Granger a "Kg Vivo"

Según de observa, con un p-value de 0.03785, se rechaza la hipótesis nula que “Kg Vivo” y “Consumo Interno” no causan en el sentido de Granger a Faena Total, por ende, tales variables ayudarían a explicar Faena Total.

Por otro lado, también se rechaza la hipótesis nula que "Kg Vivo" & "Faena Total" no causan en el sentido de Granger a "Consumo Interno”.

Finalmente, se analizaron cada uno de los modelos especificados en el VAR para el caso de Faena Total, Consumo Interno y Kg Vivo.

El caso de Faena total es el caso que muestra mayor significatividad en los coeficientes de sus propios rezagos, así como también de los rezagos de las otras variables.

	Estimate	Std. Error	t value	Pr(> t)	
KGVivo.l1	-0.207618	0.077765	-2.67	0.008183	**
FaenaTotal.l1	-0.302169	0.128343	-2.354	0.019476	*
ConsumoInterno.l1	-0.401133	0.125308	-3.201	0.001581	**
KGVivo.l2	-0.096372	0.080409	-1.199	0.232066	
FaenaTotal.l2	-0.10057	0.129835	-0.775	0.43945	
ConsumoInterno.l2	-0.271528	0.146707	-1.851	0.065599	.
KGVivo.l3	0.053243	0.082233	0.647	0.518037	
FaenaTotal.l3	0.031369	0.129894	0.241	0.809407	
ConsumoInterno.l3	-0.126272	0.158431	-0.797	0.426342	
KGVivo.l4	-0.075427	0.082266	-0.917	0.360267	
FaenaTotal.l4	-0.179476	0.13231	-1.356	0.176401	
ConsumoInterno.l4	-0.03042	0.167241	-0.182	0.855842	
KGVivo.l5	-0.061756	0.080773	-0.765	0.445394	
FaenaTotal.l5	-0.206995	0.136595	-1.515	0.131176	
ConsumoInterno.l5	0.063457	0.173704	0.365	0.715244	
KGVivo.l6	0.04631	0.080386	0.576	0.565168	
FaenaTotal.l6	-0.233063	0.136738	-1.704	0.089777	.
ConsumoInterno.l6	0.141167	0.175127	0.806	0.421105	
KGVivo.l7	0.022631	0.082131	0.276	0.783162	

	Estimate	Std. Error	t value	Pr(> t)	
FaenaTotal.l7	-0.300407	0.134094	-2.24	0.026121	*
ConsumoInterno.l7	0.183529	0.170665	1.075	0.28344	
KGVivo.l8	0.024833	0.082916	0.3	0.764855	
FaenaTotal.l8	-0.264377	0.133529	-1.98	0.049019	*
ConsumoInterno.l8	0.241678	0.166351	1.453	0.147767	
KGVivo.l9	0.050615	0.083372	0.607	0.544443	
FaenaTotal.l9	-0.059034	0.133155	-0.443	0.657969	
ConsumoInterno.l9	0.150657	0.162934	0.925	0.356211	
KGVivo.l10	-0.144675	0.083966	-1.723	0.086356	.
FaenaTotal.l10	-0.195594	0.126788	-1.543	0.124412	
ConsumoInterno.l10	0.292162	0.152041	1.922	0.056011	.
KGVivo.l11	-0.025691	0.083485	-0.308	0.758586	
FaenaTotal.l11	-0.325047	0.119848	-2.712	0.007239	**
ConsumoInterno.l11	0.439361	0.130614	3.364	0.000914	***
KGVivo.l12	-0.024862	0.081341	-0.306	0.760176	
FaenaTotal.l12	0.496999	0.123317	4.03	7.79E-05	***
ConsumoInterno.l12	-0.238942	0.116911	-2.044	0.042223	*
const	0.00117	0.001948	0.6	0.548853	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Considerando que una de las 3 variables bajo estudio, sí sería explicada por las otras variables, es que se ha optado por mantener todas ellas dentro del presente análisis.

Identificación del orden del VAR

Según los criterios de información, por medio de la función 'VARselect' se obtiene que el VAR de orden 12 es aquel que corresponde a aplicar. Cabe señalar que, siendo que el orden 12, era el valor máximo asignado por la función, es que también se probó aumentar tal valor. El resultado alcanzado también indicaba que un modelo de 12 podía resultar adecuado. Debajo se detallan el número de orden en función de distintos criterios de información.

AIC	HQ	SC	FPE
12	12	2	12

Estabilidad del modelo

Una vez obtenido el orden del VAR se procedió a su estimación, y a fin de verificar la estabilidad del modelo, se han analizado las raíces del polinomio las cuales para garantizar estabilidad deben ser en valores absolutos inferiores a 1.

Es preciso que dichos valores decaigan hacia cero pues de no ocurrir lo indicado, el futuro lejano tendría efectos sobre el presente, lo cual es lo contrario a lo que ocurre en un proceso estacionario donde hay una rápida amortiguación temporal de los efectos del pasado sobre el presente, es decir que sus efectos

se disipan en el tiempo. La condición que las raíces del polinomio sean diferentes a 1, es análogo a lo que se busca en un proceso autorregresivo de una sola variable (Novales, 2017).

Según los resultados alcanzados, se verifica la estabilidad del modelo siendo que ninguno de sus valores es igual a 1:

0.9925	0.9925	0.9867	0.9656	0.9656	0.9571	0.9571	0.9504	0.9504
0.9439	0.9439	0.9133	0.9133	0.8936	0.8901	0.8901	0.886	0.886
0.8745	0.8745	0.8723	0.8723	0.8533	0.8533	0.8519	0.8519	0.8453
0.8453	0.8201	0.8201	0.8179	0.8099	0.8099	0.6264	0.6264	0.2581

Pruebas de especificación

Pruebas de autocorrelación en los residuos

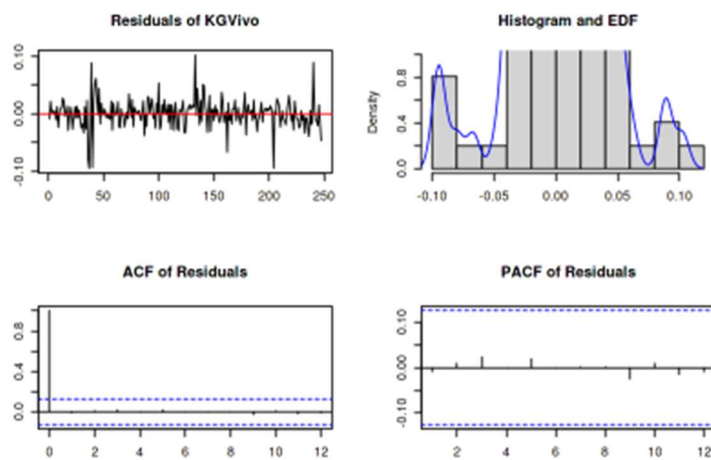
Como un primer paso derivado de la ejecución del VAR se obtuvo la matriz de correlación de los residuales de las diferentes variables analizadas. De este resultado se indica que existe correlación entre la variable Consumo Interno y Faena Total, es decir entre más suba la Faena Total aumentara de manera similar el Consumo Interno de la carne.

```
Correlation matrix of residuals:
      KGVivo FaenaTotal ConsumoInterno
KGVivo      1.00000    -0.1103    -0.09517
FaenaTotal  -0.11031     1.0000     0.89274
ConsumoInterno -0.09517    0.8927     1.00000
```

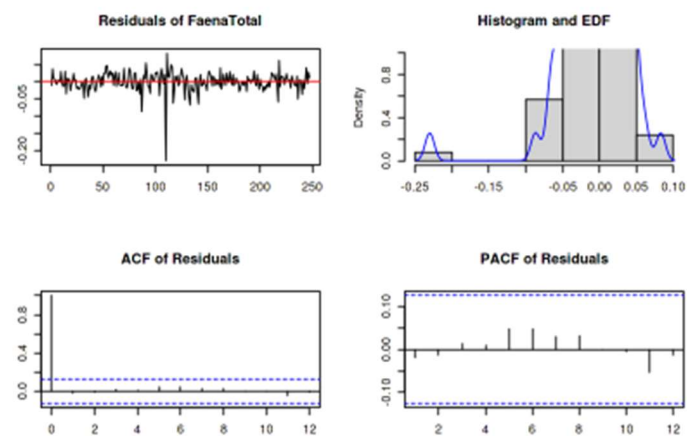
Prueba Ljung-Box

Realizando la prueba de Ljung-Box, se pudo identificar que los residuales para cada una de las series tiene un comportamiento no correlacionado, es decir, distribuidos a lo largo del tiempo con una media 0. Se realizó el grafico de los residuos cercanos a la Media 0 y el ACF de los mismos no superan los intervalos de confianza.

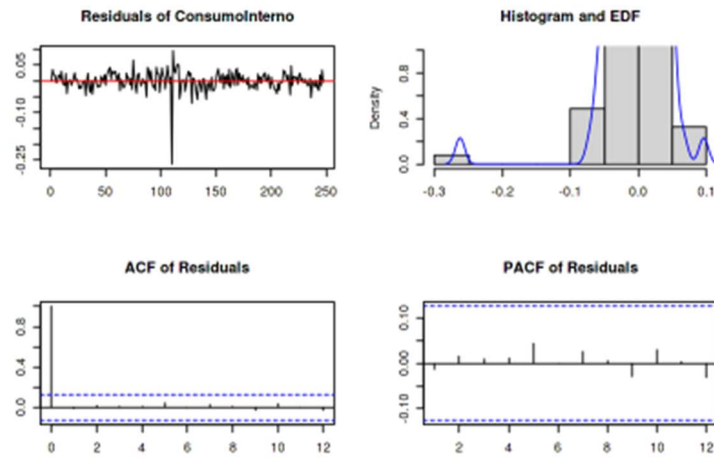
- **Serie Precio**



- **Serie Faena**



- **Serie Consumo**



Para corroborar lo mencionado anteriormente se realiza el test de Portmanteau con el cual se valida si existe autocorrelación en los residuales (H_0 : No hay autocorrelación). La prueba está dando como resultado un p-value inferior a 0.05 por lo cual se puede indicar que los residuales están correlacionados.

Portmanteau Test (asymptotic)

```
data: Residuals of VAR object bv.est  
Chi-squared = 38.866, df = 0, p-value < 2.2e-16
```

Pruebas de Normalidad de los residuos

Para la prueba de normalidad de los residuales multivariados se realizaron la ejecución del test de Jarque-Bera combinado con la prueba de skewness y kurtosis de las cuales se puede identificar que los datos no siguen una distribución simétrica, indicando que los residuos no son normales, llegando afectar la predicción del modelo desarrollado.

```
JB-Test (multivariate)

data:  Residuals of VAR object bv.est
Chi-squared = 4444.2, df = 6, p-value < 2.2e-16

$Skewness

Skewness only (multivariate)

data:  Residuals of VAR object bv.est
Chi-squared = 259.93, df = 3, p-value < 2.2e-16

$Kurtosis

Kurtosis only (multivariate)

data:  Residuals of VAR object bv.est
Chi-squared = 4184.3, df = 3, p-value < 2.2e-16
```

Prueba de homocedasticidad de los residuos

Se realiza la prueba de homocedasticidad de los residuos para validar si los estimadores pueden llegar a estar sesgados, mediante la prueba realizada se observa que no se rechaza la hipótesis nula, indicando que los datos son heterocedásticos a un p-value inferior al 0.05.

```
ARCH (multivariate)

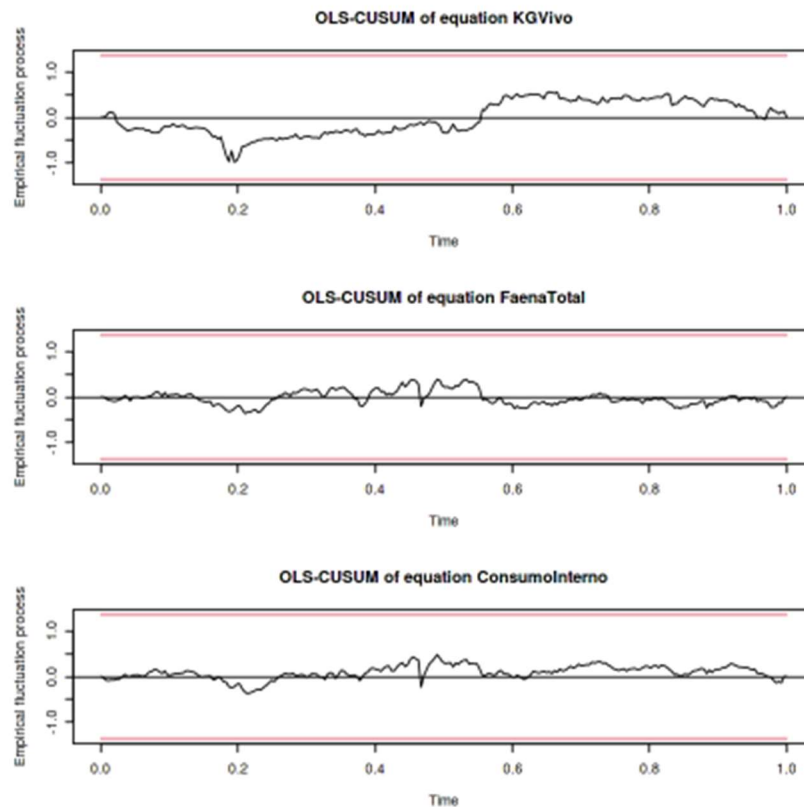
data:  Residuals of VAR object bv.est
Chi-squared = 504.82, df = 432, p-value = 0.008828
```

Estabilidad estructural de los residuos

Cuando se ajusta una regresión de series de tiempo se está asumiendo que los coeficientes son estables en el tiempo. El test estat sbcusum tiene como finalidad probar dicha suposición. Este, basa su resultado en si la serie de tiempo cambia abruptamente de formas no predichas por su modelo. Dicho de manera más técnica, prueba las roturas estructurales en los residuos.

El test estat sbcusum utiliza la suma acumulada de residuos recursivos o la suma acumulada de residuos MCO para determinar y probar si hay una ruptura estructural. Bajo la hipótesis nula, la suma acumulada de residuos tendrá una media igual a cero.

Como resultado de la prueba se ha podido identificar que los residuos son constantes en el tiempo para las tres variables, tal como muestra la imagen que se encuentra a continuación.



Análisis Impulso Respuesta y Análisis de Descomposición de la Varianza

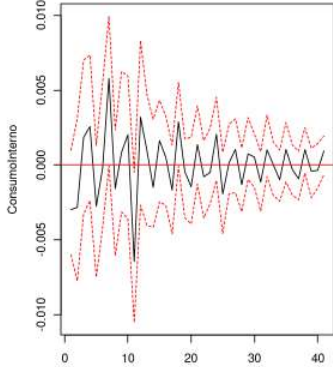
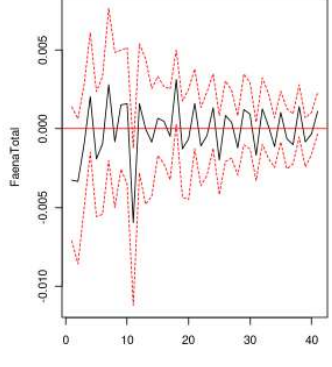
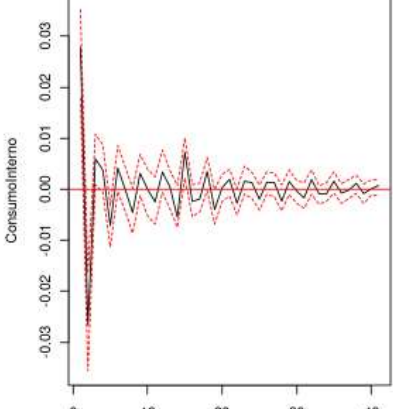
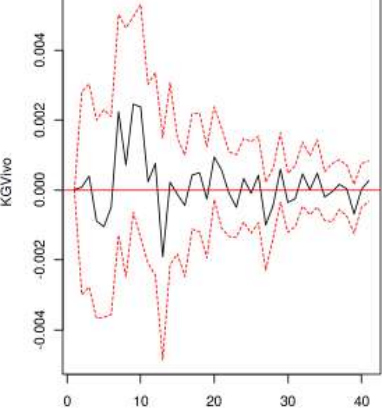
Uno de los motivos por los cuales los modelos VAR han tenido éxito dentro del ámbito económico es que éstos han introducido dos métodos de análisis: el análisis del impulso respuesta y la descomposición de la varianza. Estos proveen de mayor conocimiento en lo referido a las relaciones dinámicas entre las variables de un sistema.

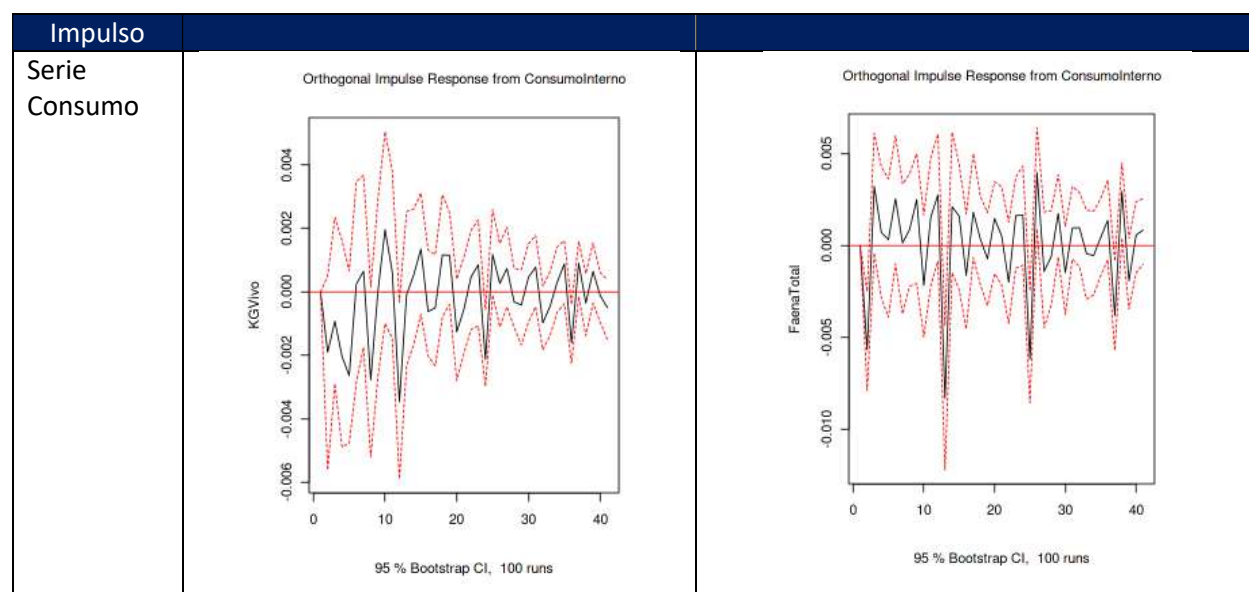
Análisis impulso respuesta

El análisis impulso respuesta se lleva a cabo a fin de observar las trayectorias de las variables de estudio. El objetivo consiste en detectar cómo un cambio en los residuales, innovaciones, influyen los componentes del vector de una variable.

Por medio de los análisis de impulso respuesta se pretende medir el efecto de un impulso unitario, por ejemplo, un shock de un desvío estándar del error de una variable en el momento 0 sobre otra variable en momentos posteriores a 0 (Kirchgässner & Wolters, J., 2007).

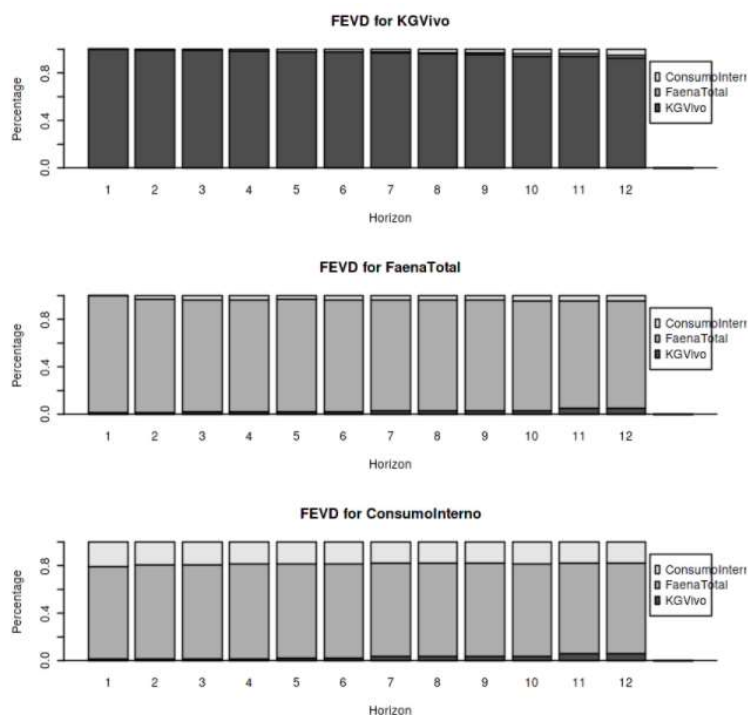
En el único caso donde se observan impactos significativos es en el caso cuando el impulso es generado por Faena Total sobre la variable Consumo Total, donde se observa un rango de impacto directo entre el impulso y la respuesta que oscila entre -0.03 y 0.03, impacto superior al registrado en los otros gráficos. Asimismo, se observa que de todas las gráficas esta es la única que presenta intervalos de confianza más acotados lo cual nos permite inferir una mayor precisión en la estimación.

Impulso		
Serie Precio	<p>Orthogonal Impulse Response from KGVivo</p>  <p>95 % Bootstrap CI, 100 runs</p>	<p>Orthogonal Impulse Response from KGVivo</p>  <p>95 % Bootstrap CI, 100 runs</p>
Serie Faena	<p>Orthogonal Impulse Response from FaenaTotal</p>  <p>95 % Bootstrap CI, 100 runs</p>	<p>Orthogonal Impulse Response from FaenaTotal</p>  <p>95 % Bootstrap CI, 100 runs</p>



Análisis de descomposición de la varianza

Según lo indicado por (Novales, 2017), la descomposición de la varianza permite dividir la varianza del error de predicción de cada variable en los componentes que son atribuibles a los distintos shocks que puede experimentar el sistema. La descomposición de la varianza se obtiene a partir de la función de respuesta al impulso, y ambas se obtienen a partir de la representación de medias móviles del proceso. A continuación, se realiza el análisis de descomposición de la varianza.



En el caso de “Kg Vivo” y “Faena Total”, según se observa, no se encuentran significativamente afectados por los shocks de las otras variables. Tal como se aprecia, su variabilidad se encuentra principalmente explicada por la propia variable.

Esto es diferente en el caso de “Consumo interno”, donde se observa que, aproximadamente el 78% de la variación de la variable “Consumo interno” se encuentra explicada por shocks en la variable “Faena Total” y el porcentaje restante corresponde en su mayoría a “Consumo Interno”

Modelo Alcanzado

El modelo fue entrenado con una ventana temporal comprendida entre enero 1998 y diciembre 2017, y evaluado en el período comprendido entre enero 2018 y diciembre 2018.

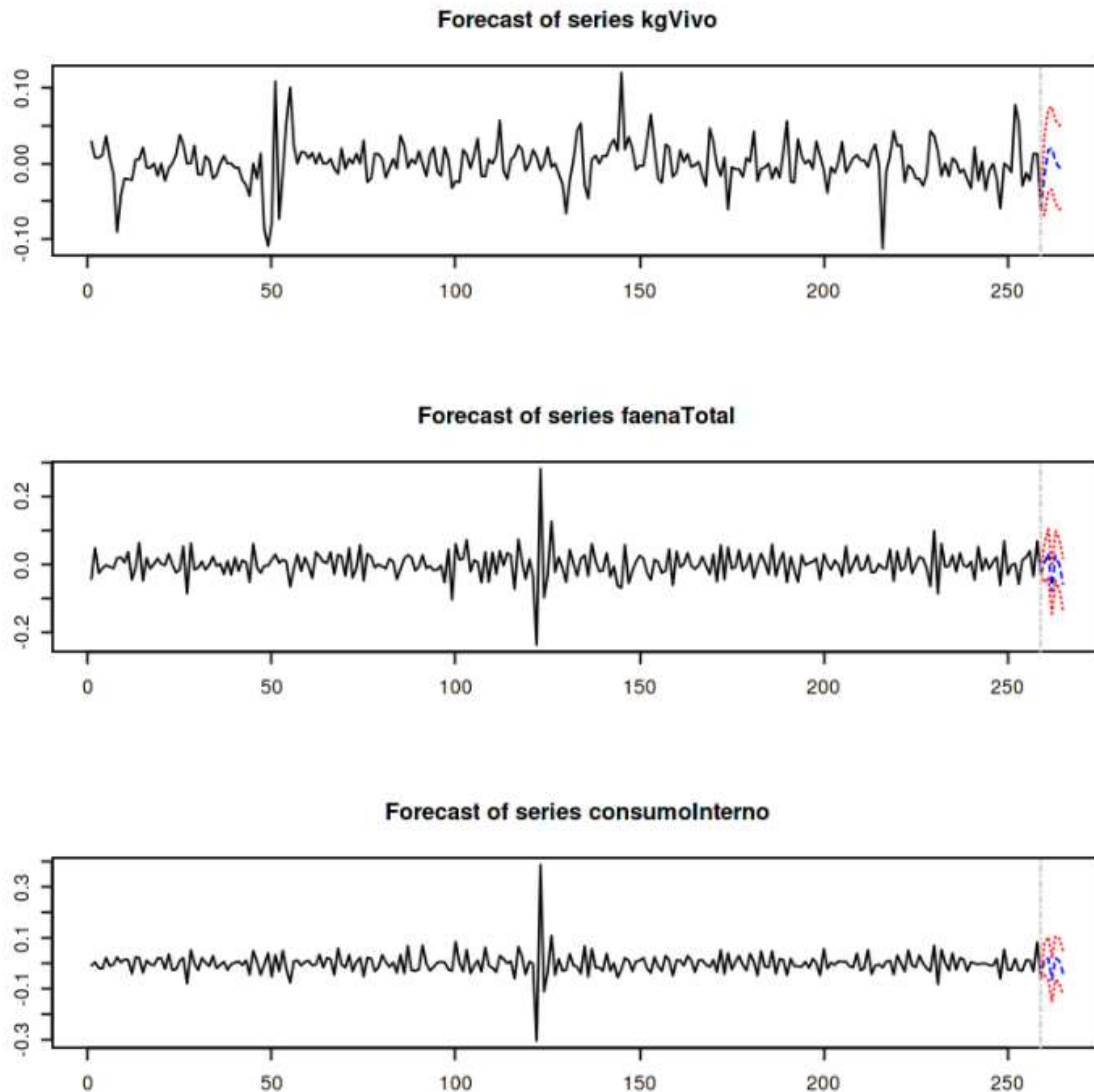
El modelo fue evaluado por medio de la métrica MASE (Error Escalado Absoluto Medio), se consideró que esta era la medida más indicada ya que es una medida escalada de los errores lo cual genera que sea independiente de las unidades de medida de las series bajo análisis. Siendo que en el presente modelo se aplicaron 2 transformaciones que cambian los valores originales de la serie, resultó necesario tomar una métrica que fuese independiente de los valores originales de las series para luego ser comparada con otros modelos que pudieran no compartir las mismas unidades de medida.

Según lo indicado por (Athanasopoulos & Hyndman, 2018), si el resultado es inferior a 1 entonces este sería mejor que los resultados alcanzados por una predicción basada en un pronóstico ingenuo promedio calculado sobre los datos de entrenamiento.

Los resultados del MASE fueron los siguientes:

	VAR		
	KG Vivo	Faena Total	Consumo Interno
MASE	0.590402769252551	0.277512930893411	0.546550104932048

A continuación, se exponen gráficamente los resultados alcanzados. Cabe señalar que, siendo que los residuos no verificaron la condición de homocedasticidad ni la normalidad de los mismos es que las predicciones podrían no resultar precisas.



Otros modelos (Punto 3)

Basándose en el material provisto por (Martinez, F., 2021), se optó por desarrollar el modelo de Auto Machine Learning (AutoML) para encontrar los modelos que produzcan los mejores resultados para las tres series temporales bajo análisis.

Para desarrollar los modelos de cada una de las series, se utilizaron los datos de entrenamiento con rezagos de la propia variable.

El AutoML de la librería H2O arrojó los siguientes resultados, indicando los tres mejores modelos para cada una de las series, fueron:

- 'GLM' para el caso de Kg. Vivo;
- 'DeepLearning' para el caso de Faena Total;

- 'XGBoost' para el caso de Consumo Interno.

Top 3 de Modelos para Kg Vivo

model_id	mean_residual_deviance	rmse	mse	mae	rmsle
GLM_1_AutoML_20210819_182137	0.01093	0.10456	0.01093	0.08090	0.04456
StackedEnsemble_BestOfFamily_AutoML_20210819_182137	0.01256	0.11209	0.01256	0.09039	0.04755
StackedEnsemble_AllModels_AutoML_20210819_182137	0.01387	0.11777	0.01387	0.10174	0.04975

Top 3 de Modelos para Faena Total

model_id	mean_residual_deviance	rmse	mse	mae	rmsle
DeepLearning_grid_2_AutoML_20210819_190201_model_3	5.03E+09	70908.6	5.03E+09	57,959.10	0.06371
DeepLearning_grid_2_AutoML_20210819_190201_model_7	5.06E+09	71140.6	5.06E+09	55,093.90	0.06449
StackedEnsemble_BestOfFamily_AutoML_20210819_190201	5.53E+09	74370.7	5.53E+09	64,492.90	0.06733

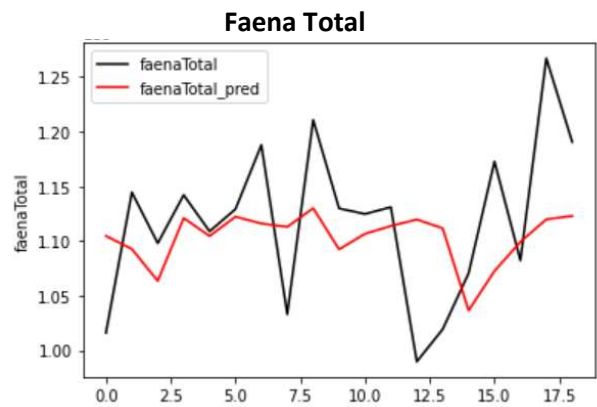
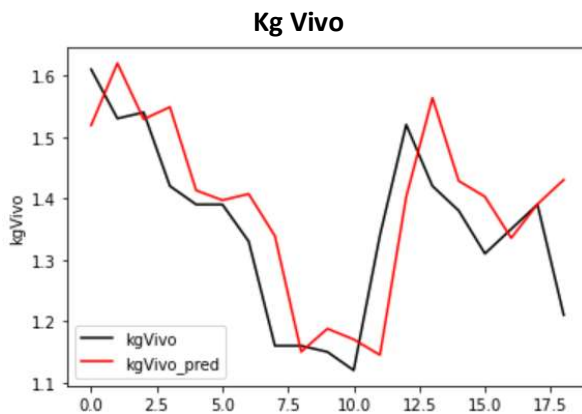
Top 3 de Modelos para Consumo Interno

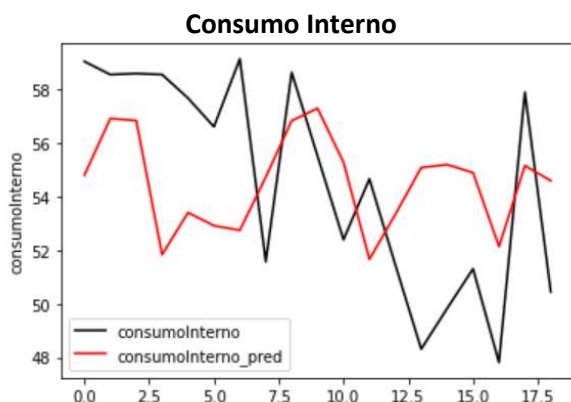
model_id	mean_residual_deviance	rmse	mse	mae	rmsle
XGBoost_grid_1_AutoML_20210819_193913_model_4	16.37	4.04556	16.3666	3.70	0.07375
XGBoost_grid_1_AutoML_20210819_193913_model_2	17.40	4.1711	17.3981	3.16	0.07639
DeepLearning_1_AutoML_20210819_193913	17.62	4.19817	17.6246	3.22	0.07717

Sus respectivos MASE fueron los siguientes:

Modelo	GLM	Deep Learning	XGBoost
	KG Vivo	Faena Total	Consumo Interno
MASE	1.055165904	0.728245501	1.135256826

Las predicciones alcanzadas por medio de GLM, Deep Learning y XGBoost se exponen en los gráficos a continuación:





En el caso de Kg Vivo, se observa que el GLM logra capturar su comportamiento sin embargo se encuentran desfasadas en el tiempo. En el caso de Faena Total, este exhibe el menor MASE de ser comparado con los otros 2 modelos alcanzados por medio de AutoML. Se observa que dicho modelo logra capturar las oscilaciones de la serie, lo hace en manera suavizada. En el caso de Consumo Interno, también se observa que no logra capturar el comportamiento de la serie donde hay marcadas caídas.

Con el objeto de evaluar los resultados alcanzados por medio del modelo VAR frente a aquellos alcanzados por medio de Auto Machine Learning, se procedió a comparar sus respectivas métricas de MASE. los resultados obtenidos se muestran a continuación:

KG Vivo

	VAR	GLM
MASE	0.590402769	1.055165904

Faena Total

	VAR	Deep Learning
MASE	0.277512931	0.728245501

Consumo Interno

	VAR	XGBoost
MASE	0.546550105	1.135256826

Según se observa siendo el MASE del VAR inferior a aquel alcanzado por medio de los otros modelos se determinó que el modelo VAR resultó ser el mejor en todos los casos. Restaría ampliar la investigación dentro de AutoML para analizar la posibilidad de realizar ajustes de hiperparámetros para obtener mejores resultados.

Pronósticos Alcanzados (Punto 4)

Como fuese indicado en la sección anterior, al comparar los resultados alcanzados por el VAR frente a aquellos modelos que surgieron de la aplicación de AutoML es que se eligió el VAR como aquel que captaba de mejor manera el comportamiento de las series en cuestión.

A fin de generar las predicciones se procedió a entrenar el modelo con la totalidad de los datos disponibles.

Considerando que a las series originales se les había aplicado una transformación por medio de logaritmos y una diferenciación de orden 1, es que, una vez obtenidas las predicciones se procedieron a aplicar tales ajustes a la inversa para poder alcanzar los valores de predicción en la misma unidad que las series originales.

Las predicciones abarcan un período de 6 meses, septiembre 2019 – marzo 2020, se eligió tal período considerando dos aspectos, por un lado, cuanto más lejano en el tiempo se realicen las predicciones estas presentarían más inestabilidad por la propia lógica del modelo, y, por otro lado, en el caso del modelo utilizado, la falta de homocedasticidad de los residuos y la falta de normalidad de estos hacen que exista mayor inestabilidad de los resultados. Por lo expuesto, se consideró apropiado limitar la ventana de predicción a 6 meses.

Las estimaciones para cada una de las variables bajo análisis por medio del VAR son las siguientes:

Faena Total *

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2018	1.128422	1.016032	1.144461	1.097949	1.142051	1.108796	1.129144	1.187806	1.03295	1.210767	1.129775	1.124678
2019	1.130946	0.98968	1.019251	1.070475	1.172685	1.082169	1.267109	1.190601	1.202352	1.2964	1.082094	1.149413
2020	1.150754	1.006885										

* Valores expresados en millones

Predicciones del modelo

Kg Vivo

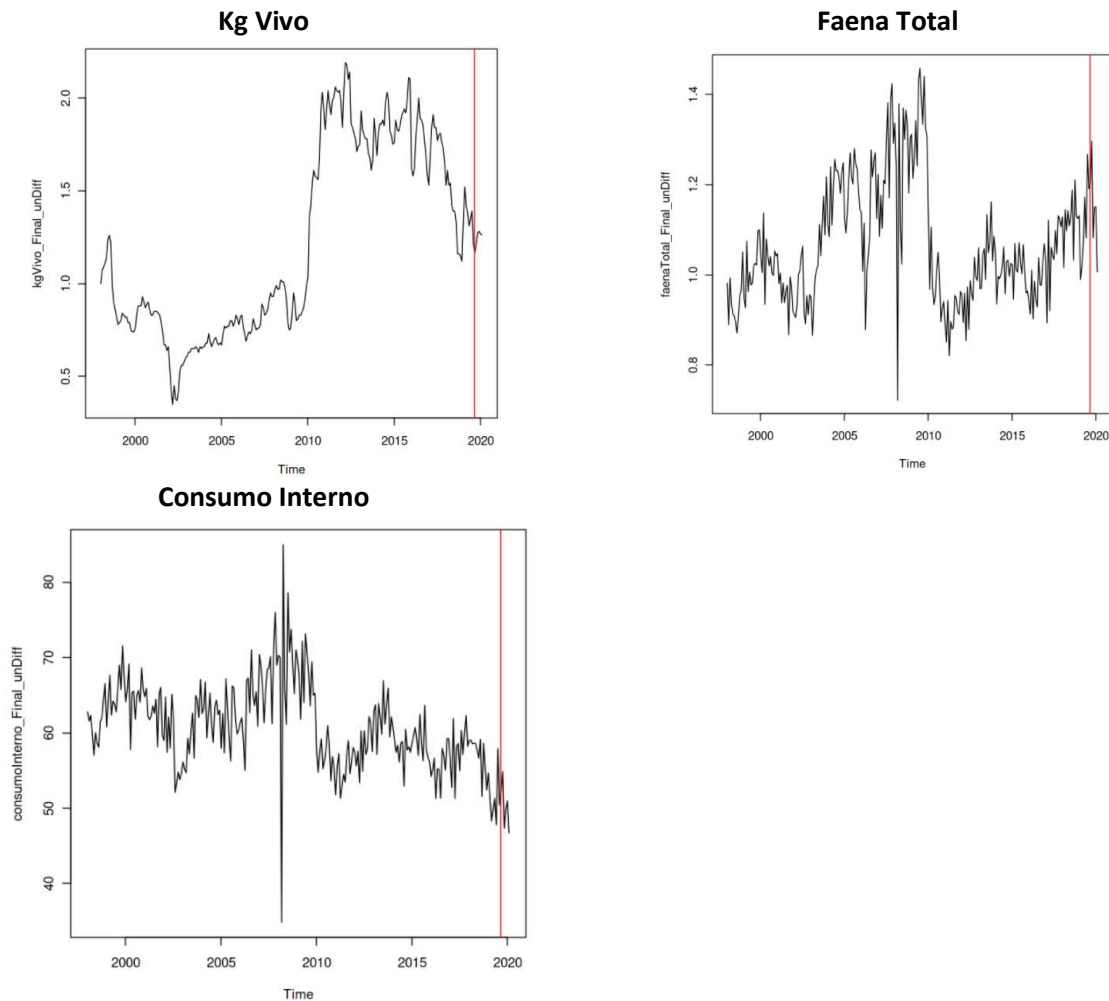
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2018	1.53	1.61	1.53	1.54	1.42	1.39	1.39	1.33	1.16	1.16	1.15	1.12
2019	1.34	1.52	1.42	1.38	1.31	1.35	1.39	1.21	1.166181	1.214956	1.273885	1.281184
2020	1.265749	1.263206										

Consumo Interno

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2018	59.01	59.06	58.57	58.61	58.57	57.69	56.62	59.16	51.58	58.65	55.52	52.4
2019	54.67	51.46	48.31	49.83	51.31	47.81	57.91	50.45	52.39791	54.88462	47.35099	49.75844
2020	50.99554	46.74296										

Predicciones del modelo

Sus representaciones gráficas son las que se exponen a continuación, la línea en colorado indica el inicio de la predicción.



La línea vertical roja indica la separación entre los datos reales y el pronóstico realizado con las tres series. En líneas generales, se puede concluir lo siguiente:

- **Serie Kg Vivo:** el precio del kilogramo vivo en dólares parece haber encontrado una estabilidad en torno a los 1.25 USD/kg.
- **Serie Faena Total:** los valores se encuentran expresados en mínimos, y se muestra una tendencia estable entorno al millón de cabezas de ganado mensuales.
- **Serie Consumo Interno:** en esta serie se observa que continuará la caída en el consumo per cápita de carne vacuna como se observa en el gráfico. Esto puede deberse a factores relacionados a cambios de hábitos de la sociedad o a la pérdida del poder adquisitivo de los salarios.

Conclusiones

Históricamente, Argentina ha sido conocida por ser un mercado de alta producción y consumo interno de carne vacuna. Por este motivo, se tomó la decisión de analizar las series de “Faena Total”, “Consumo Interno” y “Kg Vivo” para entender su comportamiento y de esa manera pronosticar cómo se comportarían en el futuro cercano.

Las series fueron analizadas para determinar si correspondían a series temporales, estacionarias y que la distribución de los residuos cumpliera con todos los requisitos para poder hacer las predicciones necesarias.

En una primera instancia se llevó a cabo el análisis VAR de las series y luego se lo contrastó con los resultados obtenidos por medio de AutoML. Como resultado del análisis se obtuvo que los modelos generados con VAR permitieron desarrollar mejores modelos que los arrojados por AutoML.

Utilizando los resultados obtenidos por medio del modelo VAR para cada una de las series se obtuvieron los resultados que se muestran a continuación:

Faena Total *

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2018	1.128422	1.016032	1.144461	1.097949	1.142051	1.108796	1.129144	1.187806	1.03295	1.210767	1.129775	1.124678
2019	1.130946	0.98968	1.019251	1.070475	1.172685	1.082169	1.267109	1.190601	1.202352	1.2964	1.082094	1.149413
2020	1.150754	1.006885										

* Valores expresados en millones

Predicciones del modelo

Kg Vivo

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2018	1.53	1.61	1.53	1.54	1.42	1.39	1.39	1.33	1.16	1.16	1.15	1.12
2019	1.34	1.52	1.42	1.38	1.31	1.35	1.39	1.21	1.166181	1.214956	1.273885	1.281184
2020	1.265749	1.263206										

Consumo Interno

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2018	59.01	59.06	58.57	58.61	58.57	57.69	56.62	59.16	51.58	58.65	55.52	52.4
2019	54.67	51.46	48.31	49.83	51.31	47.81	57.91	50.45	52.39791	54.88462	47.35099	49.75844
2020	50.99554	46.74296										

Predicciones del modelo

Estos resultados se encuentran expresados en las unidades originales de las series, por lo que se han revertido las transformaciones realizadas a las series para poder realizar los pronósticos con las unidades originales.

Los valores pronosticados para “Faena Total” se encuentran en línea con los obtenidos para los últimos períodos, en torno al millón de cabezas de ganado faenadas mensualmente. Por otro lado, el precio del kilogramo de novillo vivo muestra un comportamiento similar a períodos anteriores, sin embargo, se observa una disminución respecto al período 2018 y 2019. Esto se puede deber a la depreciación del peso argentino respecto del dólar estadounidense durante dicho período.

Por último, la serie “Consumo Interno” muestra un marcado descenso desde el enero 2018 hasta febrero de 2020. Esto se puede deber a cambios en el gusto de los consumidores o al deterioro en el ingreso de

las familias argentinas en el período bajo análisis. Dichos factores no fueron considerados como parte del presente trabajo, pero quedarían como posibles puntos para analizar en trabajos complementarios.

Muchas veces los análisis que se realizan basándose únicamente en las métricas dejan de lado los análisis contextuales que pueden explicar algunas relaciones que no pueden explicarse simplemente con los datos de las series. Para analizar los cambios en los gustos de los consumidores se requieren encuestas en hogares y personales con la finalidad de conocer la situación, requiriendo muestras que sean estadísticamente representativas. El deterioro del nivel de ingresos se puede analizar desde las estadísticas publicadas por los gobiernos nacionales y provinciales.

Para finalizar, es importante destacar, que el desarrollo de estos modelos siempre debe ser interpretados y analizados en el complejo contexto de la economía Argentina y tener en cuenta el fin para el cual van a ser utilizados, ya que la industria ganadera es compleja y abarca múltiples dimensiones de la micro y macroeconomía. La industria ganadera posee una estrecha relación con otros factores como ser la industria agrícola, situación que también sería un interesante campo a ser explorado e incorporar a los modelos VAR de manera de dotarlos de mayor potencia para su predicción. A su vez la cadena de valor también vislumbra relaciones complejas entre los distintos actores económicos, los cuales demuestran tensiones entre los productores, cadena de comercialización, exportadores y Estado, situación que le agregan mayor complejidad al proceso de predicción. De esta forma podría seguir agregando mayor dimensionalidad como ser: series de tiempo del clima, precios internacionales de otros bienes, regulaciones internas y externas etc. Es aquí donde el conocimiento experto en la materia bajo análisis y la utilidad de los modelos VAR han tenido tanto éxito en distintos campos de acción, como ser el campo de las finanzas, la microeconomía, políticas públicas etc., ya que permiten identificar el impacto y relaciones del shock de una decisión adoptada o impacto de una variable sobre las demás, permitiendo facilitar la toma de decisiones o tener una metodología robusta de predicción que permitan mitigar diferentes tipos de riesgos.

Bibliografía

1. Del Rosso, R. (2021, junio). *Análisis de Series Temporales* [Diapositivas de Clase]. Maestría en Ciencia de Datos, Universidad Austral.
2. Della Siega, M. (2021, abril 15). Caída del consumo de carne vacuna en Argentina. Bolsa de Comercio de Rosario. <http://www.bcr.com.ar/es/mercados/investigacion-y-desarrollo/informativo-semanal/noticias-informativo-semanal/caida-del-0>.
3. Enders, W. (2014). *Applied Econometric Time Series* (4th Edition). Wiley Series.
4. Erica, C (2021). The Intuition Behind Impulse Response Functions and Forecast Error Variance Decomposition. APTECH. Accedido: 16 agosto, 2021. [Online]. Disponible en: <https://www.aptech.com/blog/the-intuition-behind-impulse-response-functions-and-forecast-error-variance-decomposition/>.
5. Glen, S. (2018, septiembre 7). *Ljung box test: Definition*. Statistics How To. <https://www.statisticshowto.com/ljung-box-test/>
6. Kirchgässner, G. Wolters, J. (2007) *Introduction to Modern Time Series Analysis* (1st Edition). Springer.
7. Kotzé, Kevin. *Vector autoregression models*. Accedido: 18 agosto, 2021. [Online]. Disponible en: <https://kevinkotze.github.io/ts-7-var/>
8. Liang, C., & Lu, Y. (2020, diciembre 4). *Using automl for time series forecasting*. Google AI Blog. <http://ai.googleblog.com/2020/12/using-automl-for-time-series-forecasting.html>.
9. Martínez, F. (2021, junio). *Análisis de Series Temporales* [Código de Clase]. Maestría en Ciencia de Datos, Universidad Austral.
10. Montes-Rojas, G. (s. f.). *Procesos Multivariados* [Diapositivas].
11. Novales, A. (2017) *Modelos vectoriales autoregresivos (VAR)* (Versión Preliminar). Universidad Complutense.
12. Peña, D. (2010). *Análisis de Series Temporales* (1ra Edición). Alianza Editorial.
13. Sarafanov, M. (2021, julio 12). *AutoML for time series: Definitely a good idea*. Medium. <https://towardsdatascience.com/automl-for-time-series-definitely-a-good-idea-c51d39b2b3f>.
14. Stock, J., & Watson, M. (2001). *Vector Autoregressions*. https://faculty.washington.edu/ezivot/econ584/stck_watson_var.pdf.
15. TechyTacos. (2019, julio 7). *Automl vs traditional machine learning | plaforms to perform automl | thingstoknow*. https://www.youtube.com/watch?v=YA_dYTNa9tc.
16. Uriel, E. (1985). *Análisis de Series Temporales: Modelos ARIMA* (1ra Edición). Paraninfo.