APLICACIONES ESTADÍSTICAS

6ª Socialización de Experiencias 2021

Duitama, Boyacá
UPTC, Facultad Seccional Duitama
4 de diciembre

ESPECIALIZACIÓN EN ESTADÍSTICA

Uptc
Universidad Pedagógica y Tecnológica de Colombia

ACREDITACIÓN INSTITUCIONAL DE ALTA CALIDAD
MULTICAMPUS
RESOLUCIÓN 3910 DE 2015 MEN / 6 AÑOS

Posgrados Duitama

Aplicaciones Estadísticas. Socialización de Experiencias
"ISSN: 2619-2888 (En línea)"
http://rdigitales.uptc.edu.co/memorias/

APLICACIONES ESTADÍSTICAS
Socialización de Experiencias

2021

ESPECIALIZACIÓN EN ESTADÍSTICA

Contacto

Universidad Pedagógica y Tecnológica de Colombia
Facultad Seccional Duitama
Escuela de Posgrados Sede Duitama
Teléfono: (57+8) 7624431
Conmutador (57 + 8) 7605306 Ext: 2838 - 2830
Carrera 18 Calle 22 Edificio Administrativo Piso 1
Duitama - Boyacá - Colombia

www.uptc.edu.co
posgrados.duitama@uptc.edu.co

APLICACIONES ESTADÍSTICAS
Socialización de Experiencias

2021

ESPECIALIZACIÓN EN ESTADÍSTICA

PRESENTACIÓN

En calidad de coordinadora académica de la Especialización en Estadística, quisiera celebrar con ustedes este ejercicio de divulgación de los trabajos de aplicación desarrollados por los graduados de la sexta cohorte. Se ha dispuesto de este espacio para intercambiar experiencias de la aplicación de técnicas estadísticas en ámbitos como la economía, educación, agronomía, administración, ingeniería, entre otros.

La información contenida en estas memorias es el fruto de un año de intenso trabajo por parte de nuestros estudiantes, agradecemos mucho por la confianza que depositaron en nuestra Institución, y confío en que con el paso del tiempo serán recompensados por decisión de cursar la Especialización. Expresamos nuestro reconocimiento a los profesores que aportaron en los trabajos de aplicación, gracias por el profesionalismo, dedicación y buena voluntad.

Es nuestro deseo que esta publicación sea fuente de consulta para profesionales que requieren el uso de técnicas estadísticas de dependencia e interdependencia para la solución de problemas en su área de trabajo.

Carmen Helena Cepeda Araque
Coordinadora Académica
Especialización en Estadística

APLICACIONES ESTADÍSTICAS
Socialización de Experiencias

2021

ESPECIALIZACIÓN EN ESTADÍSTICA

TABLA DE CONTENIDO

Construcción y validación de una prueba de matemáticas mediante el modelo de Rasch
ANDRÉS FELIPE CHAPARRO LÓPEZ
andres.chaparro04@uptc.edu.co
Universidad Pedagógica y Tecnológica de Colombia-Duitama
1

Modelo de regresión logística en el análisis del riesgo por consumo de sustancias psicoactivas en universitarios
Laura Camila Ramírez Ortiz
laura.ramirez02@uptc.edu.co
Universidad Pedagógica y Tecnológica de Colombia-Duitama
11

Aplicación de los modelos ARIMA en los precios de la cebolla junca (allium fistulosum) en el municipio de Aquitania
YESID FERNANDO MONTAÑA MONTAÑA
yesid.montana@uptc.edu.co
Universidad Pedagógica y Tecnológica de Colombia-Duitama
21

APLICACIONES ESTADÍSTICAS
Socialización de Experiencias

2021

ESPECIALIZACIÓN EN ESTADÍSTICA

CONSTRUCCIÓN Y VALIDACIÓN DE UNA PRUEBA DE MATEMÁTICAS MEDIANTE EL MODELO DE RASCH

Especialización en Estadística

ANDRÉS FELIPE CHAPARRO LÓPEZ[1,a]

[1]Escuela de Posgrados, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

Resumen

El presente articulo expone los resultados del diseño, construcción y validación de un examen de matemáticas para estudiantes de grado octavo del Colegio Salesiano Maldonado de la ciudad de Tunja. Se trabajó con una muestra de 66 estudiantes que fueron elegidos de manera no probabilística y por conveniencia. La prueba consta de 12 ítems divididos en tres niveles de dificultad (fácil, medio, difícil) según el criterio de docentes del área de matemáticas y resultados obtenidos. Para el análisis de la información se utilizaron los supuestos teóricos del modelo de Rasch, estimando los parámetros de los ítems y del nivel de habilidad de cada estudiante se logró evidenciar que más del 60 % de los entrevistados poseen un nivel de habilidad media y alta para responder a ítems de mayor dificultad que miden los procedimientos inductivos, así mismo, se observa un desajuste en los ítems clasificados a priori en cada una de las categorías.

***Palabras clave*:** validación de prueba, modelo de Rasch, matemáticas, parámetros..

Abstract

This article presents the results of the design, construction and validation of a mathematics test for eighth grade students at Colegio Salesiano Maldonado in Tunja. The sample was conformed by 66 students who were chosen non-probabilistically and by convenience. The test consists of 12 items divided into three levels of difficulty (easy, medium, difficult) according to the criteria of mathematics teachers and the results obtained. For the analysis of the information, the theoretical assumptions of the Rasch model were used, estimating the parameters of the items and the level of ability of each student, it was possible to show that more than 60 % of the interviewees have a medium and high level of ability to respond to items of greater difficulty that measure the inductive procedures, likewise, a mismatch is observed in the items classified a priori in each of the categories.

***Key words*:** test validation, Rasch model, mathematics, parameters..

1. Introducción

Uno de los principales objetivos de la educación colombiana es garantizar una educación de calidad en todos sus niveles, lo cual se logra a partir de un proceso continuo y colectivo entre las Instituciones Educativas (IE) del país, por ende, se han desarrollado estrategias desde la de educación básica y media que buscan mejorar el desempeño académico en los resultados de las diferentes evaluaciones que realiza el Instituto Colombiano

[a]Estudiante de Especialización en Estadística. E-mail: andres.chaparro04@uptc.edu.co

para el Fomento de la Educación Superior (ICFES), pues si bien es cierto es la principal herramienta que se utiliza para estudiar el rendimiento de los estudiantes que aspiran a cursar un estudio universitario en el país (Moncayo, 2016).

En este sentido, la calidad de la educación vista desde el rendimiento académico es actualmente un tema primordial dentro de las propuestas educativas de cada IE, particularmente el Colegio Salesiano Maldonado (COLSAMA) de la ciudad de Tunja ha implementado un instrumento de evaluación que se realiza de manera trimestral en cada uno de los grados desde tercero hasta grado undécimo, esta propuesta se realiza de manera similar a las pruebas saber 11 que realiza el Ministerio de Educación Nacional (MEN), donde los docentes encargados ajustan los contenidos temáticos a una prueba de selección múltiple con una única respuesta compuestas por reactivos o ítems cerrados con varias opciones de respuesta.

En particular, se evidencia un bajo rendimiento académico en los resultados obtenidos en el área de matemáticas en cada uno de los exámenes trimestrales desarrollados en el presente año, que puede ser explicado teniendo en cuenta que no existe unos lineamientos específicos para la realización de los ítems en cada uno de los exámenes, esto quiere decir que la redacción se hace de manera empírica y sin tener en cuenta el nivel de dificultad de cada uno de los ítems, lo cual dificulta que se realice una evaluación estandarizada, de igual forma existe una discrepancia a la hora de analizar los resultados de las pruebas del Icfes con los resultados que se obtienen en los exámenes generales de cada trimestre.

Por esta razón, es de interés el desarrollo de la presente investigación, ya que, son pocos los estudios realizados en torno a la educación analizada desde la TRI, ya que como lo indica Cronbach citado por Abal las pruebas que plantean problemas para que el individuo muestre su capacidad de resolución se denominan test de ejecución máxima, los cuales han tenido una acogida en el grupo de investigadores, sin embargo existen pocos registros que estudian el nivel de constructo en pruebas académicas donde se evalúan habilidades, aptitudes y rendimiento académico (Abal et al., 2010), y aunque se están desarrollando estrategias desde las instituciones educativas del país, en el COLSAMA no se tiene establecido ningún tipo de norma para el diseño y la estandarización de los exámenes. Tampoco se aplica ningún tipo de análisis que permita comprobar la validez y la confiabilidad de la evaluación que se realiza a los estudiantes de tercer a undécimo grado.

Teniendo en cuenta esto, se evidencia la necesidad de conocer ¿cómo construir y validar un examen de matemáticas a partir del modelo de Rasch para mejorar los resultados obtenidos en las pruebas trimestrales que se realizan en el COLSAMA?

El objetivo general del presente trabajo es construir y validar un instrumento de evaluación de matemáticas a partir del modelo de Rasch para determinar si existen mejoras en el rendimiento académico de los estudiantes de grado octavo del COLSAMA. Otros objetivos son determinar el nivel de dificultad de cada uno de los ítems mediante el modelo de teoría TRI de Rasch y así segmentar el examen por niveles de dificultad (fácil, medio, difícil), así mismo, identificar el grado de habilidad en cada uno de los estudiantes.

Con el fin de dar respuesta a los objetivos planteados se realiza la construcción de un examen que consta de 15 ítems de selección múltiple con única respuesta que fueron validados por docentes del área de matemáticas del COLSAMA, del mismo modo, haciendo uso del modelo de Rasch uno de los modelos univariados de la TRI que comúnmente se utiliza para la validación y construcción de evaluaciones en un grupo heterogéneo de personas.

2. Referente Conceptual

2.1. Teoría de respuesta al ítem (TRI)

Es una teoría de investigación psicométrica que nace en 1960 por parte de Rasch y Birnbaum. El factor común de estos desarrollos es que establecen una relación entre el comportamiento de un sujeto frente a un ítem y los constructos no latentes (características no observadas), donde se considera al ítem como unidad básica de medida (Cortada de Kohan, 2004). Para ello, recurren a funciones matemáticas que describen la probabilidad de dar una determinada respuesta al ítem para cada nivel de habilidad medido por este, en este sentido, la principal aportación de la TRI es colocar en la misma escala la dificultad del ítem y la habilidad del examinado (Attorresi et al, 2009).

El principal objetivo de la Teoría de Respuesta al Ítem (TRI) es garantizar la invarianza entre las poblaciones, es decir, las características de un reactivo no deben depender de la muestra en la cual se aplica, lo cual implica que dos individuos con el mismo nivel de habilidad tienen la misma probabilidad de responder correctamente a un ítem independientemente de la población de pertenencia, además las puntuaciones de los alumnos deben describir su nivel de habilidad sin depender del examen (Attorresi et al, 2009). Como lo resalta Hidalgo y French (2016) el análisis usando TRI, a través de modelos matemáticos unidimensionales y multidimensionales, proporcionan una visión de la relación entre el nivel del rasgo latente de un individuo que incluyen características como ansiedad, motivación, depresión entre otros tipos de rasgos psicológicos y las características de cada uno de los ítems (dificultad, discriminación, adivinación). Según Cortada de Kohan (2004) la TRI se apoya en dos supuestos fundamentales:

1. La unidimensionalidad del rasgo latente. Es decir que los ítems que constituyen un test deben medir sólo una aptitud o rasgo. 2. La independencia. Es decir que las respuestas de un examinado a cualquier par de ítem son independientes y no existe relación entre las respuestas de un examinado a diferentes ítems. Así, las aptitudes especificadas en el modelo son los mismos factores que influyen sobre las respuestas a los ítems del test. De esta manera la probabilidad del tipo de respuesta a un conjunto de ítem es igual al producto de las probabilidades asociadas con las respuestas del examinado a los ítems individuales.

Al igual que la Teoría Clásica de los Test (TCT) la TRI busca estimar los parámetros de cada uno de los ítems, para ello se apoya de modelos matemáticos en los cuales se intenta describir la probabilidad de que el ítem sea contestado correctamente, en este sentido las funciones de distribución de probabilidad y la logística cumplen con el supuesto de monotonía creciente. Ambas curvas adoptan una forma de S suavizada con un punto de inflexión en el que se encuentra el valor máximo de su pendiente que se analiza mediante la discriminación de cada uno de los ítems; para la descripción de esta curva se han generalizado tres modelos, en el modelo de un parámetro se estima la dificultad del ítem, en el modelo de dos parámetros se agrega la discriminación de cada ítem y finalmente en el modelo de tres parámetros se estudia la pseudo adivinación (Attorresi et al, 2009).

2.2. Modelo de Rasch

El modelo propuesto por George Rasch en 1960 conocido como el modelo logístico de un parámetro permite analizar la medición conjunta en una misma escala, de las personas y de las puntuaciones obtenidas en una prueba dada (Martín, Díaz, Córdoba y Picquart, 2011). Una de las ventajas que tiene este modelo es que permite analizar las interacciones entre el nivel de dificultad de un ítem y el nivel de habilidad de una persona, según Salas y Montero (2011) este modelo se basa en los siguientes supuestos:

El atributo que se desea medir puede representarse en una única dimensión en la que se sitúan conjuntamente las personas y los ítems El nivel de habilidad de la persona en el atributo y la dificultad del ítem determinan la probabilidad de que la respuesta sea correcta.

Según Luzardo (2011) la presentación matemática del modelo se basa en una matriz de datos bidimensional obtenida al administrar n ítems a N individuos examinados. Las respuestas obtenidas son puntuadas en forma dicotómica uij = 0, 1 donde i indica el ítem y j al sujeto, de igual forma 1 simboliza que el individuo respondió correctamente a lo planteado en cada ítem y 0 simboliza que respondió erróneamente, en este contexto Rasch denota a θ_s el nivel de habilidad del sujeto y θ_i nivel de dificultad de cada ítem.

Rasch utilizó una función biyectiva entre los números reales $\mathbf{R}$ y [0,1] conectando así la variable e_{ij} la respuesta correcta propuesta por una persona j ante un ítem de habilidad i con una función de probabilidad, lo cual se resume en:

$$\begin{aligned} P\left(u_{ij}=1 \mid \xi_{ij}\right) &= \frac{\xi_{ij}}{1+\xi_{ij}} \\ P\left(u_{ij}=0 \mid \xi_{ij}\right) &= 1-\frac{\xi_{ij}}{1+\xi_{ij}}=\frac{1}{1+\xi_{ij}} \end{aligned} \tag{1}$$

Donde 1: éxito y 0: fracaso, entiendase éxito como responder correctamenta a un determinado ítem i.

De igual forma, el parámetro se puede definir de manera general si $\xi_{ij}=\frac{\eta_j}{\delta_i}$ e interpretar en términos del Odds ratio de la siguiente manera:

$$\xi_{ij} = \frac{P\left(u_{ij} = 1 \mid \eta_j, \delta_i\right)}{P\left(u_{ij} = 0 \mid \eta_j, \delta_i\right)} = \frac{\frac{\eta_j}{\delta_i + \eta_j}}{\frac{\delta_i}{\delta_i + \eta_j}} = \frac{\eta_j}{\delta_i} \tag{2}$$

2.3. Estimación de los parámetros

Prieto y Delgado (2003) destacan que el modelo de Rasch proporciona una solución para calibrar instrumentos de evaluación utilizando la función logística que estima la probabilidad de que un reactivo i tenga una respuesta satisfactoria, para un nivel de habilidad θ_s del sujeto s, a partir de las características y el nivel de habilidad que tiene el examinado versus el nivel de dificultad de cada uno de los ítems.

$$\begin{aligned} &\log\left(\frac{\pi}{1-\pi}\right) = \theta_s - \theta_i \\ &\theta_s - \theta_i = 0 \\ &\theta_s - \theta_i > 0 \\ &\theta_s - \theta_i < 0 \end{aligned} \tag{3}$$

De igual manera, definen cada una de las propiedades en función de la probabilidad de responder acertadamente a un ítem, lo cual favorece la construcción de pruebas de selección en un grupo de personas con un nivel de dificultad variante en cada una de las preguntas.

(1) Si la habilidad de la persona es igual a la dificultad del ítem, la persona tiene un 50 % probabilidad de acertar el ítem. (2) Si la habilidad de la persona es superior a la dificultad del ítem, la persona tiene un 75 % probabilidad de acertar el ítem. (3) Si la habilidad de la persona es menor a la dificultad del ítem, la persona tiene un 25 % probabilidad de acertar el ítem.

Una formulación más conocida del modelo de Rasch se deriva de la predicción de la probabilidad de responder correctamente al ítem a partir de la diferencia en el atributo entre el nivel de la persona θ_s y el nivel del ítem θ_i (Luzardo, 2013). En este caso,

$$P\left(u_{ij} \mid \theta_j \beta_i\right) = \frac{e^{u_{ij}(\theta_j - \beta_i)}}{1 + e^{(\theta_j - \beta_i)}} \tag{4}$$

En este sentido, la probabilidad de que un sujeto j tenga un vector de respuestas U_{ij} aplicando el método de máxima verosimilitud condicional esta dado por:

$$L = P\left((u_{ij}) \mid \theta_j \beta_i\right) = \prod_{i=1}^{n} P\left(u_{ij} \mid \theta_j \beta_i\right) = \prod_{i=1}^{n} \frac{e^{u_{ij}(\theta_j - \beta_i)}}{1 + e^{(\theta_j - \beta_i)}} = \frac{e^{\left(\sum_{i=1}^{n} u_{ij}\theta_j - \sum_{i=1}^{n} u_{ij}\beta_i\right)}}{\prod_{i=1}^{n}\left[1 + e^{(\theta_j - \beta_i)}\right]} \tag{5}$$

De lo cual se obtiene:

$$L = P\left((u_{ij}) \mid \theta_j \beta_i\right) = \frac{e^{\left(r\theta_j - \sum_{i=1}^{n} u_{ij}\beta_i\right)}}{\prod_{i=1}^{n}\left[1 + e^{(\theta_j - \beta_i)}\right]} \tag{6}$$

Particularmente si se calcula la verosimilitud de un patrón de respuestas se tiene:

$$L = P\left(u_1, u_2, \ldots, u_n \mid \theta\right) = \prod_{i=1}^{n} P\left(u_i \mid \theta\right) \tag{7}$$

Donde U_i puede tomar los valores de 0 y 1 debido a que la variable respuesta es de tipo dicotómico, luego:

$$L = \prod_{i=1}^{n} P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i} = \prod_{i=1}^{n} \left(\frac{P_i(\theta)}{1 - P_i(\theta)}\right)^{u_i} Q_i(\theta) \tag{8}$$

2.4. Ajuste del modelo

Además de estimar los parámetros de los sujetos y de cada uno de los ítems del examen es importante estudiar el ajuste del modelo a los datos encontrados, en este sentido Martín, Díaz, Córdoba y Picquart, (2011) destacan dos medidas importantes para evaluar la bondad de ajuste del modelo de Rasch:

> ...el INFIT que se interpreta como ajuste interno, es un valor sensible al comportamiento inesperado que afecta a los reactivos cuya dificultad está cerca del nivel de habilidad de una persona y el OUTFIT que se interpreta como ajuste externo, es un valor sensible al comportamiento inesperado que afecta a los reactivos cuya dificultad está lejos del nivel de habilidad de la persona. (pág. 6).

Los cuales se interpretan como medida de los errores cuadráticos medios y como los errores estandarizados, El rango de los valores del infit que indican un buen ajuste en el modelo dependen del tamaño de la muestra como se muestra en la tabla 1.

Tamaño de la muestra	**Valores**
Menores de 500 casos	Entre 0,7 y 1,3
Entre 500 y 1000 casos	Entre 0,7 y 1,2
Más de 1000 casos	Entre 0,7 y 1,1

Tabla 1: Rangos valores de infit

De la misma forma, el outfit es un estadístico que evalúa el nivel de discriminación de cada uno de los ítems y su rango debe encontrarse entre 0,5 y 1,5 (Solorzano y Montero, 2011). Una vez que se ha verificado el ajuste del modelo y las estimaciones de los parámetros producidas por el modelo de Rasch propuesto ofrecerán la medida de cada sujeto en el rasgo latente. La precisión y el análisis de esta medida suelen representarse mediante la curva característica de cada ítem que suele describirse en forma de una S alargada que relaciona el nivel del rasgo latente θ de un determinado sujeto y el patrón de respuesta esperado en el examen.

La curva característica de un examen (CCI) es la suma de las curvas características de cada uno de los items, es decir, para obtener un determinado nivel de θ se suman los valores de la probabilidad $P(\theta)$ de cada ítem del examen que se esta interpretando para ese nivel (fácil, medio, díficil) (Pérez, 2014). Esta definicón puede expresarse matemáticamente como:

$$CCT = \sum_{i=1}^{k} P_i(\theta) \tag{9}$$

En la CCI se denota $\theta = 0$ como el nivel de habilidad media en cada uno de los ítems como se muestra en la figura 1.

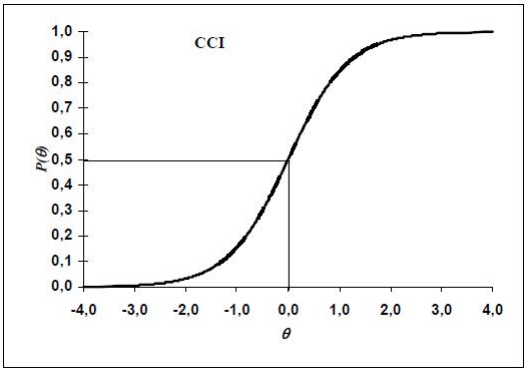

Figura 1: Curva Caracteristica del ítem. Tomada de: Pérez 2014

3. Metodología

El enfoque de investigación es cuantitativo con un enfoque de tipo exploratorio que según Hernández y Torres (2018), el propósito de los diseños transeccionales exploratorios es comenzar a conocer una variable o un conjunto de variables, una comunidad, un contexto, un evento, una situación. Se trata de una exploración inicial en un momento específico. Por lo general, se aplican a problemas de investigación nuevos o poco conocidos; además, constituyen el preámbulo de otros diseños (no experimentales y experimentales). El procesamiento de los datos se realizó en el software R, mediante el entorno de desarrollo integrado RStudio.

En primer lugar se realizará la construcción de un examen de matemáticas que consta de 12 ítems, los cuales se clasificaron según la dificultad entre fácil, medio y difícil, esta clasificación y validación se realizó con ayuda de los docentes del área de matemáticas del COLSAMA, posterior se aplico la prueba en los 66 estudiantes de grado octavo de esta institución educativa, finalmente con ayuda del software estadístico R se verificaron los niveles de dificultad de cada ítem y se analizaron las curvas características de los ítems, resaltando así la dificultad media y la habilidad necesaria en cada uno de ellos.

Etapas de investigación

Para el dar solución a los objetivos planteados el estudio se va a realizar en cuatro etapas como se muestra en la tabla 2.

Etapa	**Procedimiento**
Selección unidades de estudio	Muestreo no probabilístico
Validación	Prueba piloto
Análisis mediante el modelo de Rasch	Análisis mediante el software R
Comparar resultados	Comparación de medias

Tabla 2: Etapas de la investigación

En la selección de las unidades de estudio se utilizó un método no probabilístico por conveniencia como lo indica Hernández y Torres (2018) esta técnica de investigación es útil es estudios de tipo exploratorio donde se busca validar pruebas pre test y pruebas piloto, en la etapa dos se realizó el proceso de construcción y validación del examen, en la cual se recurrió a docentes expertos en el área de matemáticas del COLSAMA quienes determinaron el nivel de habilidad en la que se ubicaban cada uno de los ítems (fácil, medio, difícil), para el análisis y la comparación de los datos mediante el modelo de Rasch condicional se utilizo el software estadístico R que facilitó la estimación e interpretación de los parámetros.

4. Resultados y conclusiones

A continuación se muestra el análisis el análisis de la construcción y validación de un examen de matemáticas mediante el modelo de Rasch, el cual se aplicó a una muestra de 66 estudiantes que cursaban el grado octavo en el COLSAMA de la ciudad de Tunja, el examen fue construido con 12 ítems, cuatro para cada nivel de dificultad fácil, medio, difícil, los cuales fueron clasificados teniendo en cuenta la clase de contenido que se estaba evaluando declarativo, conceptual y procedimental respectivamente. Las respuestas obtenidas fueron codificadas dicotómicamente y analizadas mediante el software estadístico R. En primer lugar en la tabla 3 se muestra la estimación de los parámetros θ_i del modelo de Rasch, se observa que el nivel de dificultad del ítem 8 es muy bajo se aconseja replantearlo o cambiarlo, de igual forma, los ítems 3, 4, 5, 10 y 12 tiene un nivel de dificultad fácil; los ítems 1 y 7 poseen un nivel de dificultad medio; finalmente los ítems 2, 6 y 11 poseen un nivel de dificultad que se ubica en la escala difícil según el criterio establecido, lo cual evidencia la falta de estructuración en el examen dado que no existen 4 ítems para cada nivel de dificultad.

Item Easiness Parameters (beta) with 0.95 CI:				
	Estimate	Std. Error	lower CI	upper CI
beta Item 1	-0,073	0,268	-0,597	0,451
beta Item 2	1,620	0,371	0,893	2,347
beta Item 3	-0,669	0,263	-1,184	-0,153
beta Item 4	-0,149	0,266	-0,671	0,372
beta Item 5	-0,816	0,264	-1,334	-0,298
beta Item 6	1,215	0,331	0,566	1,864
beta Item 7	0,413	0,282	-0,14	0,967
beta Item 8	-1,118	0,270	-1,647	-0,588
beta Item 9	-0,073	0,268	-0,597	0,451
beta Item 10	-0,816	0,264	-1,334	-0,298
beta Item 11	0,987	0,314	0,372	1,602
beta Item 12	-0,522	0,263	-1,037	-0,007

Tabla 3: Estimación parámetros de los ítems

De la misma manera, en la tabla 4 se muestran las estimaciones de los parámetros θ_s que corresponden a los parámetros por persona del modelo de Rasch, para facilitar la lectura solamente se muestran los valores iniciales obtenidos, se evidencia que los individuos 8, 12, 16, 26 y 63 tienen una habilidad muy baja; en el mismo sentido, los individuos 20, 27, 40, 51, 53, 61, 62, 64 y 65 poseen un nivel de habilidad bajo para responder a ítems de dificultad media y difícil; además los individuos 3, 5, 6, 10, 13, 14, 17, 23, 31, 39, 49, 52, 56, y 59 poseen gran habilidad para responder a ítems que requieren un nivel de dificultad alto.

ML estimated ability parameters (without spline interpolated values):				
	Estimate	Std. Error	2.5 %	9.5 %
beta Individuo 1	0.0258535	0.6236973	-1.19657073	1.2482777
beta Individuo 2	0.4157334	0.6277502	-0.81463435	1.6461013
beta Individuo 3	1.8252128	0.8003327	0.25658949	3.3938362
beta Individuo 4	0.8215493	0.6497625	-0.45196186	2.0950605
beta Individuo 5	1.2723955	0.6986277	-0.09688958	2.6416806
beta Individuo 6	1.2723955	0.6986277	-0.09688958	2.6416806
beta Individuo 7	0.0258535	0.6236973	-1.19657073	1.2482777
beta Individuo 8	-1.8439688	0.8207250	-3.45256012	-0.2353774
beta Individuo 9	0.0258535	0.6236973	-1.19657073	1.2482777
beta Individuo 10	2.6527305	1.0624718	0.57032395	4.7351370

Tabla 4: Estimación parámetros de los individuos

Para evaluar el ajuste del modelo propuesto se analizaron los estadísticos infitmnsq y outfitmnsq los cuales evidenciaron que, de los 12 ítems planteados, 11 ajustan muy bien al modelo, dado que los valores se encuentran entre 0,7 y 1,3, sin embargo, se recomienda replantear o cambiar el ítem 11 porque presenta un desajuste, de igual forma, los ítems 3 y 5 ajustan perfectamente al modelo dado que su valor estimado es muy cercano a 1.

ítems					Infit MSQ	Outfit t	Infit t	Discrim
Item 1	84.885	60	0.019	1.392	1.219	2.195	1.778	0.041
Item 2	48.162	60	0.864	0.790	0.896	-0.372	-0.364	0.374
Item 3	65.017	60	0.306	1.066	1.076	0.471	0.738	0.229
Item 4	49.792	60	0.824	0.816	0.868	-1.186	-1.162	0.547
Item 5	62.592	60	0.384	1.026	1.028	0.215	0.298	0.232
Item 6	51.177	60	0.784	0.839	0.977	-0.376	-0.051	0.279
Item 7	52.840	60	0.732	0.866	0.947	-0.590	-0.341	0.379
Item 8	59.005	60	0.512	0.967	1.040	-0.116	0.388	0.282
Item 9	68.588	60	0.209	1.124	1.097	0.787	0.838	0.203
Item 10	60.179	60	0.469	0.987	1.004	-0.031	0.070	0.348
Item 11	38.078	60	0.988	0.624	0.746	-1.342	-1.504	0.567
Item 12	47.162	60	0.886	0.773	0.818	-1.596	-1.779	0.588

Tabla 5: Valores ajuste del modelo

La imagen 2 muestra la curva característica de los ítems en la cual se analiza los índices de dificultad, se puede evidenciar que tienen un comportamiento uniforme cumpliendo con los supuestos del modelo de Rasch, de igual forma, se observa que a mayor dificultad del ítem mayor habilidad necesita el sujeto para responder correctamente, además no existe ninguna intersección entre las curvas lo cual hala bien del modelo planteado para la construcción y validación del examen.

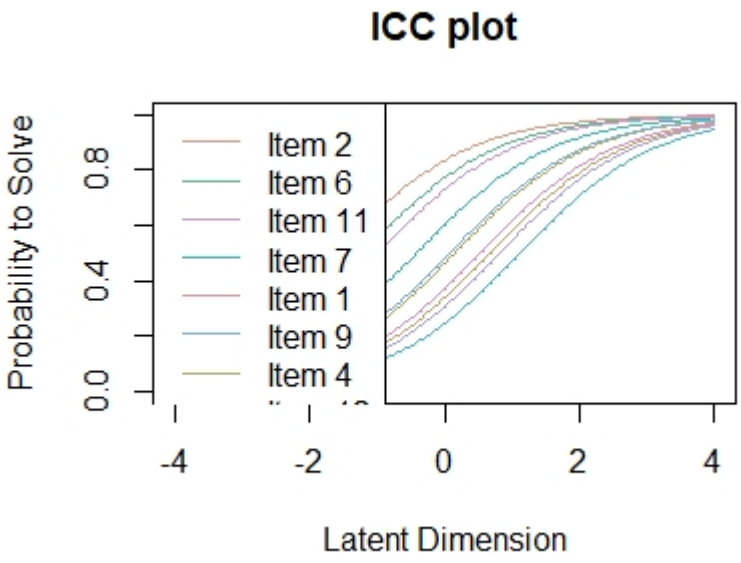

Figura 2: Curvas características de los ítems

Finalmente es importante analizar el comportamiento del ítem 11, dado que como se mencionaba anteriormente muestra un desajuste significativo en el modelo propuesto.

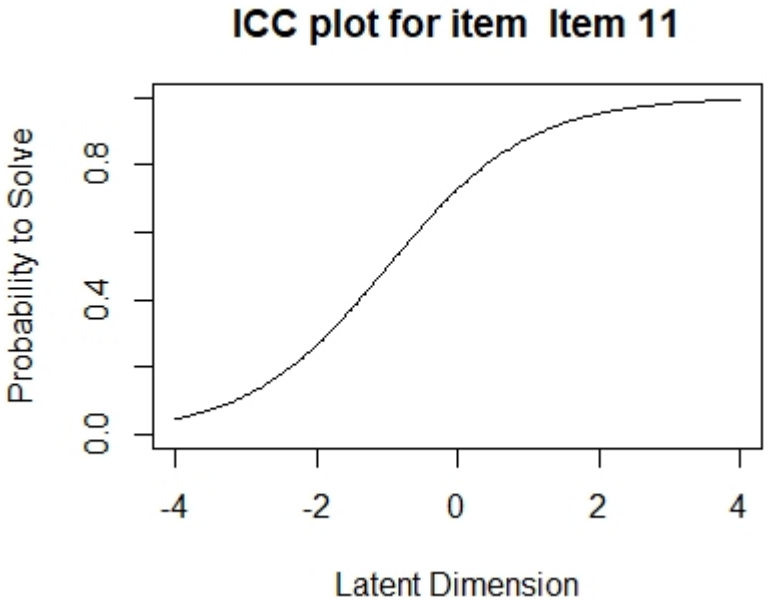

Figura 3: Curvas característica ítem 11

Se puede evidenciar que si el sujeto no posee habilidad para responder al ítem este tendra una probabilidad de aproximadamente el 10 % de responder correctamente, de igual forma si posee una habilidad media, es decir habilidad 0 tiene una posibilidad de aproximadamente el 70 % de responder correctamente a este ítem, lo que se manifiesta en desajuste completo del modelo, dado que es un ítem demasiado facíl en el que los entrevistados pueden responder utilizando elm azar, por esta razón se recomienda recalibrarlo o cambiarlo para posteriores pruebas.

5. Conclusiones

A través del modelo de Rasch que se propuso para la validación y construcción del examen fue posible conocer la calidad de los ítems planteados. Se logró verificar que sólo el ítem 11 no se ajustaba al modelo, sin embargo, se recomienda recalibrar y analizar con mayor profundidad, dado que pueden aparecer aspectos importantes como el pseudoazar que brinda herramientas útiles para futuros estudios de investigación.

Es importante señalar que la aplicación del método de Rasch es importante para evaluar la calidad de los exámenes, además brinda aspectos que permiten realizar una evaluación más confiable de los estudiantes a partir del nivel de dificultad de cada uno de los ítems que se plantean en un examen, lo cual se debe tener en cuenta por los profesionales encargados de la educación básica, media y superior del país.

La construcción y validación de la prueba de matemáticas brinda aspectos que pueden contribuir a la mejora del rendimiento académico de los estudiantes en esta área, dado que, se esta iniciando un proceso de estandarización que va a permitir un análisis detallado de las fortalezas y debilidades que se presentan en cada uno de los exámenes que se realizan de manera trimestral en el COLSAMA de la ciudad de Tunja, buscando mejorar índices de calidad en las pruebas que realiza el ICFES.

Finalmente es importante resaltar la importancia que tiene pensarse una educación de calidad que rompa con los paradigmas tradicionalistas en el aula de clase, esto empieza desde la manera en la cual se evalúa a los estudiantes, por esta razón es importante seguir realizando estudios en los cuales se tenga como referente la Teoría de Respuesta al Ítem y el modelo de Rasch que permiten una calibración más precisa de lo que cada profesional de la educación desea evaluar en cada uno de los instrumentos que utiliza.

Referencias Bibliográficas

Abal, F. J. P., Lozzia, G. S., Aguerri, M. E., Galibert, M. S. & Attorresi, H. F. (2010), 'La escasa aplicación de la teoría de respuesta al ítem en tests de ejecución típica', *Revista Colombiana de Psicología* **19**(1), 111–122.

Attorresi, H. F., Lozzia, G. S., Abal, F. J. P., Galibert, M. S. & Aguerri, M. E. (2009), 'Teoría de respuesta al ítem. conceptos básicos y aplicaciones para la medición de constructos psicológicos', *Revista Argentina de Clínica Psicológica* **18**(2), 179–188.

de Kohan, N. C. (2004), 'Teoría de respuesta al ítem: supuestos básicos', *Revista evaluar* **4**(1).

Hernández-Sampieri, R. & Torres, C. P. M. (2018), *Metodología de la investigación*, Vol. 4, McGraw-Hill Interamericana México^ eD. F DF.

Luzardo, M. (2013), 'Consistencia conjunta de las cci y el rasgo en tri multidimensional mediante regresión no paramétrica'.

Martín Guaregua, N., Díaz Torres, C., Córdoba Herrera, G. & Picquart, M. (2011), 'Calibración de una prueba de química por el modelo de rasch', *Revista electrónica de investigación educativa* **13**(2), 132–148.

Moncayo Cabrera, M. A. (2016), 'Determinantes que influyen en el rendimiento académico: un estudio aplicado para colombia a partir de las pruebas icfes-saber 11'.

Montesinos, M. D. H. & French, B. F. (2016), 'Una introducción didáctica a la teoría de respuesta al ítem para comprender la construcción de escalas', *Revista de Psicología Clínica con Niños y Adolescentes* **3**(2), 13–21.

Pérez-Gil, J. (n.d.), 'Modelos de medición: Desarrollos actuales, supuestos, ventajas e inconvenientes: Teoría de respuesta a los items (tri). apuntes de la asignatura: Desarrollos actuales de la medición: Aplicaciones en evaluación psicológica (tema 1). departamento de psicología experimental. universidad de sevilla.[consultado 2 dic 2014]'.

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*https://www.R-project.org/

RStudio Team (2021), *RStudio: Integrated Development Environment for R*, RStudio, PBC, Boston, MA.
*http://www.rstudio.com/

Solórzano Salas, J. & Montero Rojas, E. (2011), 'Construcción y validación de una prueba de comprensión de lectura mediante el modelo de rasch/construction and validation of a reading comprehension test though the rasch model'.

MODELO DE REGRESIÓN LOGÍSTICA EN EL ANÁLISIS DEL RIESGO POR CONSUMO DE SUSTANCIAS PSICOACTIVAS EN UNIVERSITARIOS

Especialización en Estadística

Laura Camila Ramírez Ortiz[1,a]

[1]Escuela de Posgrados, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

Resumen

El uso de sustancias psicoactivas en jóvenes está asociado a comportamientos de riesgo y futura dependencia, conduciendo a problemas emocionales y comportamentales. El presente estudio hizo uso de una regresión logística ordinal para conocer si el sexo y factores familiares están relacionados con el riesgo por consumo de sustancias psicoactivas. Se empleó una muestra de 328 estudiantes universitarios con edades entre los 16 y 36 años quienes contestaron la prueba de detección de consumo de sustancias ASSIST V.3 y dos sub-escalas del instrumento Escala de Clima Familiar (FES): cohesión y adaptabilidad. Los resultados indican que tanto ser hombre como presentar categorías bajas de cohesión y adaptabilidad familiar están asociados a mayores niveles de riesgo por consumo de sustancias psicoactivas. Estos hallazgos sugieren que las intervenciones con jóvenes deberían enfocarse en las relaciones familiares como un factor protector del riesgo de consumo y desarrollar estrategias de prevención selectiva con hombres.

***Palabras clave*:** Regresión logística ordinal, sustancias psicoactivas, factores familiares, sexo, jóvenes.

Abstract

The use of psychoactive substances in young people is associated with risk behaviors and future dependence, leading to emotional and behavioral problems. The present study made use of an ordinal logistic regression to find out if sex and family factors are related with risk by psychoactive substance use. A sample of 328 university students aged between 16 and 36 years was used who answered the ASSIST V.3 substance use detection test and two sub-scales of the Family Climate Scale instrument (FES): Cohesion and adaptability. The results indicate that both being a man and presenting low categories of family cohesion and adaptability are associated with higher levels of risk due to the use of psychoactive substances. These findings suggest that interventions with young people should focus on family relationships as a protective factor of the risk of consumption and develop selective prevention strategies with men.

***Key words*:** Ordinal logistic regression, Psychoactive substances, family factors, sex, youth.

1. Introducción

El intervalo entre la pubertad y la edad adulta temprana es una etapa del desarrollo en la que la experimentación y la toma de riesgos son relativamente frecuentes e incluso normativas. Desde la pubertad, muchos adolescentes comienzan a consumir alcohol, tabaco o cannabis. A lo largo de la adolescencia, el consumo de

[a]Estudiante de Especialización en Estadística. E-mail: laura.ramirez02@uptc.edu.co

sustancias aumenta gradualmente hasta alcanzar su punto máximo durante la adultez emergente, momento en el que comienza a disminuir (Wolf Peuker, Demetrio Caovilla, Batista da Costa & Pereira Mosmann 2020). Ahora bien, el hecho de que el consumo de sustancias esté muy extendido entre los jóvenes no debe llevarnos a ignorar sus consecuencias negativas en la salud. Una amplia evidencia empírica indica que el abuso de sustancias provoca importantes dificultades a corto, mediano y largo plazo, tanto a nivel físico como psicológico (Arellanez-Hernández, Diaz-Negrete, Wagner-Echeagaray & Pérez-Islas 2004). Estudios realizados por el Observatorio de Drogas de Colombia del Ministerio de Justicia y del Derecho, han puesto de relieve que las mayores tasas de consumo de sustancias psicoactivas se presentan en la población joven, que se encuentra entre los 18 y 25 años, donde una gran proporción corresponde a estudiantes universitarios (González Correa, Hernández Ramírez, Velásquez López & Mejía Ocampo 2013).

Entre los principales factores de riesgo que se han identificado para el consumo de alcohol y otras drogas están la vulnerabilidad socioeconómica, las características del funcionamiento familiar, ser hombre y encontrarse dentro del curso de vida de la juventud y adolescencia. Así mismo, estudios longitudinales han permitido detectar factores relacionados tanto con los niveles iniciales de consumo de sustancias como con su trayectoria en el tiempo. Particularmente, el contexto familiar parece ser un factor protector importante contra las conductas problemáticas, como el uso de sustancias psicoactivas (Sánchez-Queija, Oliva, Parra & Camacho 2016), encontrándose que la dimensión afectiva de las relaciones entre padres e hijos es relevante. Específicamente, el vínculo afectivo con los padres, la capacidad de respuesta y apoyo de los padres y la cohesión familiar son variables que se han postulado como factores que previenen el consumo de sustancias en los adolescentes (Goodrum, Smith, Hanson, Moreland, Saunders & Kilpatrick 2020).

La familia es uno de los contextos más relevantes en la vida del ser humano. Los estudios muestran el estrecho vínculo entre las experiencias vividas en la familia y la salud y el desarrollo del individuo (Lardier Jr, Barrios, Garcia-Reid & Reid 2018).A partir de la década de los cincuenta proliferaron los modelos de familia que intentan describir los patrones de interacción que ocurren en el sistema familiar. Existe actualmente un mayor consenso respecto de cuáles son los aspectos que deberían considerarse en la indagación del funcionamiento familiar. Desde el Modelo Circumplejo de Sistemas familiares y Maritales (Olson 2000), el cual ha tenido una gran difusión en los últimos años en el mundo académico y profesional y sostiene que la cohesión, la flexibilidad y la comunicación son las tres dimensiones que principalmente definen el constructo funcionamiento familiar. La cohesión se refiere al grado de unión emocional percibido por los miembros de la familia. La flexibilidad familiar se define como la magnitud de cambio en roles, reglas y liderazgo que experimenta la familia. El grado de cohesión y flexibilidad que presenta cada familia puede constituir un indicador del tipo de funcionamiento que predomina en el sistema: extremo, de rango medio o balanceado. Los sistemas maritales o familiares balanceados tienden a ser más funcionales y facilitadores del funcionamiento (Gau, Lai, Chiu, Liu, Lee & Hwu 2009).

Si bien es evidente la relación entre las características familiares con el uso de alcohol y otras drogas en jóvenes y adolescentes, la influencia mutua o unidireccional entre estas variables y cuánto cada uno contribuye a la ocurrencia del fenómeno necesita ser investigado. Los principales estudios que relacionan estas variables se han llevado a cabo en población adolescente (Wagner, Ritt-Olson, Chou, Pokhrel, Duan, Baezconde-Garbanati, Soto & Unger 2010), haciendo necesario ampliar la comprensión de este fenómeno en la adultez joven, cuya prevalencia es la mayor en Colombia. Dado que actualmente el consumo de sustancias psicoactivas es un problema de salud pública con alta prevalencia en jóvenes, es necesario diseñar programas de prevención y mitigación que impacten sobre los posibles predictores de estos comportamientos.

El presente estudio constituye un fundamento empírico para la práctica de los profesionales que trabajan con este público, ya que al ampliar la comprensión de este fenómeno es posible desarrollar acciones basadas en la evidencia que beneficiarán estas estrategias y acciones.

2. Referente Conceptual

En esta sección se desarrollan algunos conceptos relacionados con el modelamiento estadístico, en particular, la regresión logística con respuesta politómica ordinal, especificando la estimación de sus parámetros, inferencia y medidas de bondad de ajuste.

2.1. Modelos Lineales Generalizados

Cuando se busca explicar o modelar la relación entre una variable Y, denominada respuesta; y una o más variables predictoras, $X1, ..., Xp$, se recurre a análisis por medio de modelos lineales generales. Sin embargo, hay situaciones en las que no se pueden aplicar directamente, como es el caso de variables respuesta no normales como los conteos y las proporciones, donde estos pueden llevar a estimaciones mayores que uno o menores que cero. En respuesta a ésta y otras dificultades surgieron los modelos lineales generalizados como extensión a los modelos lineales clásicos, constituidos por un componente aleatorio, un componente sistemático o predictor lineal y una función de enlace que describe la relación funcional entre los dos componentes anteriores (Agresti 2019). Particularmente, para efectos de este estudio se revisarán los modelos de regresión logística, específicamente con respuesta politómica ordinal.

2.1.1. Regresión logística ordinal

Para variables respuesta que tienen más de dos categorías se utilizan generalizaciones de la regresión logística, donde c denota el número de categorías de la variable respuesta Y, cuyas probabilidades $(\pi_1, \ldots, \pi_c)$ satisfacen que $\sum_j \pi_j = 1$, siempre que las observaciones sean independientes (Agresti 2019). Ahora bien, cuando las categorías de la variable respuesta tienen un orden natural, la forma más óptima de estudiar los datos es especificando esto en el modelo, para lo cual se cuenta con la regresión logística ordinal (RLO), que permite relacionar la variable respuesta con variables predictoras métricas y no métricas (Díaz Monroy, Morales Rivera & León Dávila 2018).

Según Dobson (2002) existen varios modelos diferentes que pueden ser utilizados para estos análisis, es de nuestro interés el modelo de odds proporcionales o modelo logit acumulado, el cual se basa en el supuesto que el efecto de las covariables $X1, ..., Xp$ es igual para todas las categorías en la escala logarítmica. Si el predictor lineal tiene un término en el intercepto β_{0j} que dependen de la categoría j, pero las otras variables explicativas no dependen de j, dicho modelo se expresa de la siguiente manera:

$$\log \frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_J} = \beta_{0j} + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1} \tag{1}$$

Generalmente, si la variable respuesta D tienen G categorías $(D = 0, 1, 2, \ldots, G-1)$, entonces hay $G-1$ formas de dicotomizar la respuesta: $(D \geq 1$ ó $D < 1; D \geq 2$ ó $D < 2, \ldots, D \geq G-1$ ó $D < G-1)$. Con la categorización de D, se puede definir la "odds" o "ventaja" de que $D \geq g$ dividida por la probabilidad de que $D < g$ así:

$$\text{odds}(D \geq g) = \frac{\mathbf{P}(D \geq g)}{\mathbf{P}(D < g)} \quad \text{donde } g = 1, 2, 3, \ldots, G-1 \tag{2}$$

Bajo este modelo, el odds ratio que evalúa el efecto de una variable explicativa para cualquiera de las divisiones o categorizaciones anteriores será el mismo independientemente de donde se realice el punto de corte sobre las categorías, es invariante al punto utilizado para la dicotomización, lo cual implica que si hay G categorías en la respuesta, solo hay un parámetro (β) para cada una de las variables predictoras o explicativas. Sin embargo sigue habiendo constantes separadas (α_g) para cada una de las $G-1$ comparaciones. Esto contrasta con la regresión logística politómica, donde hay $G-1$ parámetros para cada variable predictora, así como constantes separadas para cada una de las $G-1$ comparaciones (Dobson 2002).

2.1.2. Extensión del modelo ordinal a k variables

Para añadir más de una variable explicativa a este modelo basta con expandir el predictor lineal, representado por $\underline{X}$ (Arias Benítez 2018). El modelo se puede expresar por:

$$P(D \geq g \mid \underline{X}) = \frac{1}{1 + \exp\left[-\left(\alpha_g + \sum_{i=1}^{k} \beta_i X_i\right)\right]}, \quad g = 1, 2, 3, \ldots, G-1 \tag{3}$$

El odds para la respuesta mayor o igual al nivel g sería el siguiente:

$$\text{odds}(D \geq g \mid \underline{X}) = \frac{P(D \geq g \mid \underline{X})}{P(D < g \mid \underline{X})} = \exp\left(\alpha_g + \sum_{i=1}^{k} \beta_i X_j\right) \tag{4}$$

2.1.3. Estimación de parámetros

Se puede estimar los parámetros del modelo por máxima verosimilitud (MV), maximizando la función de verosimilitud por el método iterativo de Newton-Raphson (Díaz Monroy, Morales Rivera & León Dávila 2018):

$$L(\alpha, \beta \mid Y, X) = \cdots = \prod_{i=1}^{n} \prod_{j=2}^{g-1} \left[\frac{1}{1 + e^{-(\alpha_1 + \beta' X_j)}}\right]^{\delta_{j1}} \left[\frac{1}{1 + e^{-(\alpha_j + \beta' X_j)}} - \frac{1}{1 + e^{-(\alpha_{j-a} + \beta' X_j)}}\right]^{\delta_{ij}} \tag{5}$$

donde:

$$\delta_{ij} = \begin{cases} 1 & \text{si el i-ésimo individuo muestra } Y = y_j \\ 0 & \text{en caso contrario} \end{cases} \tag{6}$$

De aquí por las propiedades de los estimadores de MV[1],

$$\hat{\theta}_{k,MV} \overset{\text{asint.}}{\sim} N\left(\theta_k, \sqrt{\hat{F}_{kk}^{-1}}\right) \tag{7}$$

Se puede realizar la prueba de Wald para juzgar la hipótesis:

$$\begin{aligned} H_0 &: \beta_k = 0 \\ H_1 &: \beta_k \neq 0 \end{aligned} \tag{8}$$

con el estadístico de contraste

$$\frac{\hat{\beta}_k}{\sqrt{\hat{F}_{kk}^{-1}}} \overset{H_0}{\sim} N(0,1), \quad \text{ó equivalentemente} \quad \frac{\hat{\beta}_k^2}{\hat{F}_{kk}^{-1}} \overset{H_0}{\sim} \chi_1^2 \tag{9}$$

2.1.4. Interpretación de los coeficientes de regresión

Como en otros modelos logísticos una interpretación adecuada de los coeficientes β nos remite a los conceptos de riesgo relativo, *odds* y razón *odds*(Díaz Monroy, Morales Rivera & León Dávila 2018). En ese sentido, si se considera que las p-variables conforman un vector X, es decir, que $X = (X1, X2, ..., Xp)$, se puede probar, por las propiedades de la función exponencial, que los *odds* del evento $Y = 1$ se pueden escribir como

$$\begin{aligned} O(X) &= \frac{P(Y=1)}{P(Y \neq 1)} = \frac{P(Y=1)}{1 - P(Y=1)} \\ &= \exp\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right) \end{aligned} \tag{10}$$

[1]F es la matriz de información de Fisher.

Suponiendo que se tienen dos perfiles específicos, es decir, dos individuos k y l determinados por los valores que asuman en cada una de las p-variables; estos son:

$$\begin{aligned} &\text{individuo } k: \quad X_{k1}, X_{k2}, \ldots, X_{kp} \\ &\text{individuo } l: \quad X_{l1}, X_{l2}, \ldots, X_{lp} \end{aligned} \tag{11}$$

el valor de los *odds* en cada uno de ellos, de acuerdo con (10) son $O(Xk)$ y $O(Xl)$, respectivamente. Así, $O(Xk)$ representa los *odds* correspondientes al primer perfil y $O(Xl)$ los relacionados con el segundo. Mediante manipulación algebraica sencilla se obtiene la siguiente expresión:

$$RR = \frac{O\left(X^k\right)}{O\left(X^l\right)} = \exp\left[\sum_{i=1}^{p} \beta_i \left(X_{ki} - X_{li}\right)\right] \tag{12}$$

donde X^k y X^l denotan el vector de observaciones para los individuos k y l,respectivamente. La expresión (12) corresponde a una medida relativa del riesgo relacionada con un perfil respecto de otro en términos de los parámetros de la regresión logística.

2.1.5. Ajuste del modelo

Cuando se tiene un modelo con p variables y otro con $k < p$ variables, el problema es decidir cuál de los dos modelos se ajusta mejor a los datos.

Al primer modelo se le nota por M y al más simple por M^*. La estadística de razón de verosimilitud es

$$\begin{aligned} G^2 &= -2\ln\frac{L\left(M^*\right)}{L(M)} \\ &= -2\ln L\left(M^*\right) - 2\ln L(M) \\ &= G^2\left(M^*\right) - G^2(M) \end{aligned} \tag{13}$$

Esta estadística G^2 mide los desvíos entre los datos (valores observados) y los valores ajustados (pronosticados) por el modelo logístico, y se define

$$G^2 = 2\sum(\text{ observ. })\ln\left(\frac{\text{observ.}}{\text{ajuste}}\right) \tag{14}$$

Para muestras de tamaño grande, la estadística G^2 tiene distribución ji-cuadrado, con un número de grados de libertad igual a la diferencia entre los grados de libertad de los respectivos errores en los dos modelos, es decir, $gl = p - k$.

El cociente (o razón) de verosimilitud G^2 es útil para determinar si hay diferencia significativa entre incluir en el modelo todas las variables (modelo saturado) o incluir tan solo algunas de ellas. G^2 sirve para evaluar si las variables $X_1, X_2, \ldots, X_p$, integradas en conjunto al modelo, contribuyen más a .explicar"las modificaciones que se producen en $P(Y = 1)$ que con k de estas variables ($k < p$). La mayoría de los paquetes estadísticos muestran a G^2 descompuesto en la forma $-2\ln L\left(M^*\right)$ y $-2\ln L(M)$, donde $-2\ln L\left(M^*\right)$ corresponde a la razón de verosimilitud del modelo ajustado únicamente por el intercepto(Agresti 2019).

3. Metodología

Estudio cuantitativo, de alcance descriptivo-correlacional y corte transversal.

3.1. Participantes

La muestra está conformada por 328 estudiantes de pregrado pertenecientes a una institución de educación superior de la ciudad de Tunja, cuya edad oscila entre los 16 y 36 años ($M = 20.81$).

3.2. Instrumentos

Prueba de detección de consumo de alcohol, tabaco y sustancias (ASSIST)(OMS 2011): Es un instrumento desarrollado por la Organización Mundial de la Salud (OMS) como herramienta técnica para ayudar a la identificación temprana de riesgos para la salud y trastornos debido al uso de sustancias en la atención primaria de salud, la atención médica general y otros entornos. La consistencia interna de la prueba es alta (Alcohol $\alpha = .86$. marihuana $\alpha = .84$ y cocaína $\alpha = .90$).

Escala de Clima Familiar (FES)(Moos, Moos & Trickett 1995): Esta escala aprecia las características socio-ambientales de todo tipo de familia, evalúa y describe las relaciones interpersonales entre los miembros de la familia, los aspectos de desarrollo que tienen mayor importancia para ella y su estructura básica. Cuenta con una confiabilidad $\alpha = 0.79$. Para el estudio se emplearon únicamente las subescalas de cohesion y de adaptabilidad.

3.3. Análisis

Se realizaron análisis descriptivos y se propuso un modelo de regresión ordinal utilizan como variable respuesta el riesgo de consumo y las otras variables como predictoras. Los análisis y procesamiento de los datos se realizó con el sofware libre R Core Team (2021).

4. Resultados y conclusiones

Como se puede observar en la Tabla 1, la muestra está conformada en un mayor porcentaje por hombres, el nivel de cohesión y adaptabilidad corresponde en su mayoría a la categoría "buena" y el riesgo por consumo es en su mayoría de nivel bajo.

Variable		**N**	(%)
Sexo	Hombre	171	(52.13)
	Mujer	157	(47.87)
Cohesión	Buena	202	(61)
	Promedio	85	(26)
	Mala	41	(13)
Adaptabilidad	Buena	191	(58.2)
	Promedio	89	(27.1)
	Mala	48	(14.7)
Riesgo Consumo	Alto	11	(3.4)
	Medio	132	(40.2)
	Bajo	185	(56.4)

Tabla 1: Estadísticos descriptivos de las variables de estudio (N=328).

Al analizar la proporción de hombres y mujeres según el nivel de riesgo (Figura 1) se observa que para los niveles medio y alto dicha proporción es mayor en hombres (48 % y 5.2 % respectivamente) con respecto a las mujeres (32 % y 1.3 % respectivamente).

FIGURA 1: Diagrama de mosaico: Riesgo - Sexo

Con respecto a las variables familiares, se observa (ver Figura 2) que para el nivel de riesgo alto la mayor proporción corresponde a aquellos estudiantes que presentan una cohesión (17.1 %) y adaptabilidad (12.5 %) "mala" con respecto a las categorías "promedio" y "buena". Así mismo, el nivel de riesgo bajo presenta mayores proporciones en quienes refieren una cohesión (71.8 %) y adaptabilidad (75.4 %) "buena".

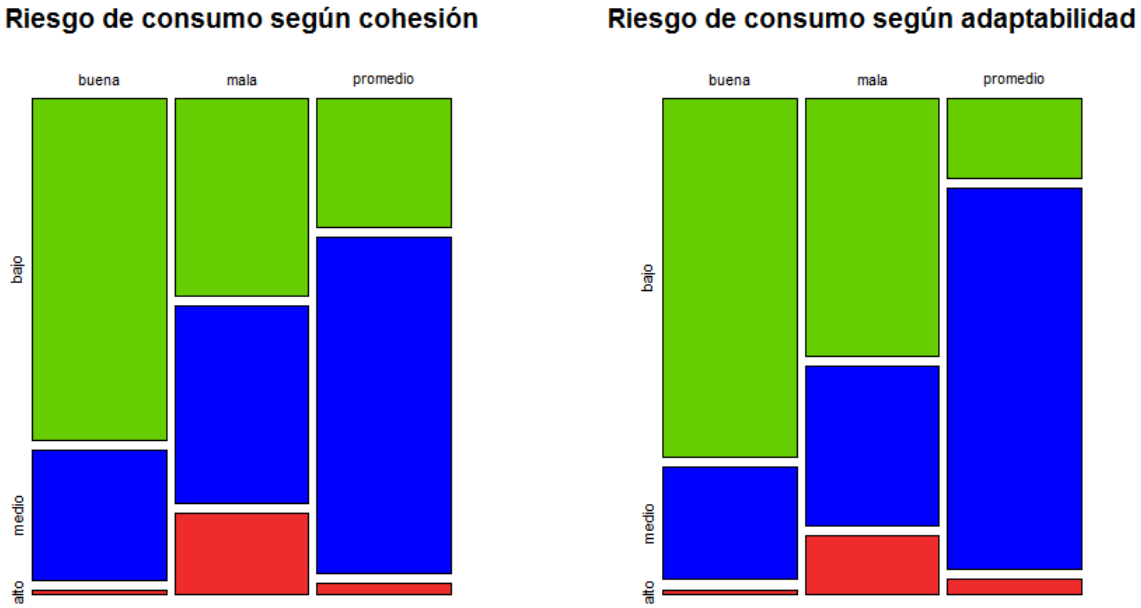

FIGURA 2: Diagrama de mosaico: Riesgo - Factores Familiares

Factores asociados al riesgo por consumo de sustancias psicoactivas

En la Tabla 2 se presentan los resultados obtenidos de la regresión logística ordinal para el riesgo por consumo de sustancias psicoactivas en el que se emplearon como variables regresoras el sexo, la cohesión y la adaptabilidad familiar. La significancia y coeficientes de regresión se verificaron mediante el contraste de razón de verosimilitud y contraste de Wald, los cuales indicaron que las variables tienen un efecto estadísticamente significativo para un nivel de significancia del 5 %.

Tabla 2: Parámetros del modelo de riesgo por uso de sustancias.

Variable	**Estimación**	**Error**	**valor z**	**Pr(>|z|)**	**OddsRatio**	**Intervalo de confianza**	
bajo|medio	2.07	0.269	7.70	1.37e-14	7.93	NA	NA
medio|alto	6.10	0.496	12.3	9.36e-35	448.	NA	NA
SexoHombre	1.10	0.276	3.97	7.23e-5	2.99	1.76	5.21
CohesionMala	1.63	0.513	3.18	1.48e-3	5.11	1.90	14.2
CohesionPromedio	1.33	0.313	4.25	2.14e-5	3.78	2.06	7.03
AdaptabilidadMala	0.467	0.489	0.953	3.40e-1	1.59	0.602	4.13
AdaptabilidadPromedio	2.10	0.320	6.57	4.99e-11	8.18	4.43	15.6

De acuerdo con los coeficientes estimados en la variable sexo, los hombres tienen aproximadamente 3 veces más probabilidades de presentar categorías altas de riesgo de consumo que categorías bajas con respecto a las mujeres.

Así mismo, quienes refieren una cohesión familiar mala y promedio tienen 5.11 y 3.78 más probabilidades respectivamente de presentar categorías altas de riesgo de consumo que categorías bajas con respecto a quienes refieren una buena cohesión familiar.

Finalmente, aquellas personas que presentan una adaptabilidad familiar mala y promedio tienen 1.59 y 8.18 más probabilidades respectivamente de presentar categorías altas de riesgo por consumo que categorías baja, esto con respecto a quienes refieren una buena adaptabilidad familiar.

Para verificar el supuesto de regresión paralela (odds proporcionales) se utilizó el test de Brant, con el cual se determinó con un nivel de significancia del 5 % que el supuesto de regresión paralela se sostiene (Tabla 3).

Tabla 3: Test de Brant.

	X2	**df**	**Pr(>Chisq)**
SexoHombre	1.13	1	0.29
CohesionMala	0.99	1	0.32
CohesionPromedio	1.32	1	0.25
AdaptabilidadMala	0.88	1	0.35
AdaptabilidadPromedio	2.69	1	0.1

Considerando que la categoría “mala” correspondiente a la variable de adaptabilidad familiar no resultó ser estadísticamente significativa, se realizó la prueba de razón de verosimilitud para conocer si toda la variable era significativa, encontrando que el modelo que la contiene presenta un ajuste estadísticamente mejor que el modelo anidado que sólo incluye el sexo y la cohesión familiar (Tabla 4).

Tabla 4: Prueba de razón de verosimilitud.

	AIC	**logLik**	**LR.stat**	df	Pr(>Chisq)
Modelo 2	465.75	-227.88			
Modelo 1	419.13	-202.57	50.62	2	1.019e-11 ***

A partir de este estudio se puede concluir que los hombres tienen una mayor probabilidad de presentar consumos de interés clínico, dado que las categorías de riesgo mayores indican problemas de salud, sociales, económicos, legales, relacionales y posible dependencia, los cuales requieren intervención profesional, por lo cual es importante desarrollar estrategias de prevención selectiva e indicada con ellos.

Con respecto a los factores familiares, tanto la cohesión como la adaptabilidad juegan un papel importante en la presentación de consumos de riesgo. Particularmente, cuando los miembros de la familia no están compenetrados, evitan ayudarse y apoyarse entre sí (cohesión), o bien no se les permite actuar libremente ni expresar sus sentimientos (adaptabilidad) la probabilidad de presentar consumos de mayor riesgo es mayor con respecto a quienes sí perciben esa unión y posibilidad de expresión familiar. Por lo anterior, las acciones de prevención frente al consumo problemático de sustancias psicoactivas deben enfocarse fortalecer estas características familiares como factores protectores.

Referencias Bibliográficas

Agresti, A. (2019), *An introduction to categorical data analysis*, John Wiley & Sons.

Álvarez-López, Á. M., Carmona-Valencia, N. J., Pérez-Rendón, Á. L. & Jaramillo-Roa, A. (2020), 'Factores psicosociales asociados al consumo de sustancias psicoactivas en adolescentes de Pereira, Colombia', *Universidad y Salud* **22**(3), 213–222.

Arellanez-Hernández, J. L., Diaz-Negrete, D. B., Wagner-Echeagaray, F. & Pérez-Islas, V. (2004), 'Factores psicosociales asociados con el abuso y la dependencia de drogas entre adolescentes: análisis bivariados de un estudio de casos y controles', *Salud mental* **27**(3), 54–64.

Arias Benítez, M. (2018), 'Regresión ordinal y sus aplicaciones'.

Botzet, A. M., Dittel, C., Birkeland, R., Lee, S., Grabowski, J. & Winters, K. C. (2019), 'Parents as interventionists: Addressing adolescent substance use', *Journal of substance abuse treatment* **99**, 124–133.

Díaz Monroy, L. G., Morales Rivera, M. A. & León Dávila, L. R. (2018), *Análisis estadístico de datos categóricos*, Universidad Nacional de Colombia.

Dobson, A. J. (2002), *An introduction to generalized linear models*, CRC press.

Gau, S. S.-F., Lai, M.-C., Chiu, Y.-N., Liu, C.-T., Lee, M.-B. & Hwu, H.-G. (2009), 'Individual and family correlates for cigarette smoking among Taiwanese college students', *Comprehensive psychiatry* **50**(3), 276–285.

González Correa, A., Hernández Ramírez, E. M., Velásquez López, C. A. & Mejía Ocampo, J. A. (2013), 'II Estudio epidemiológico andino sobre consumo de drogas en la población universitaria, Comunidad Andina de Naciones (CAN): informe Universidad de Antioquia, 2003. Proyecto PRADICAN (Programa Antidrogas Ilícitas de la Comunidad Andina)'.

Goodrum, N. M., Smith, D. W., Hanson, R. F., Moreland, A. D., Saunders, B. E. & Kilpatrick, D. G. (2020), 'Longitudinal relations among adolescent risk behavior, family cohesion, violence exposure, and mental health in a national sample', *Journal of abnormal child psychology* **48**(11), 1455–1469.

Kliewer, W., Murrelle, L., Prom, E., Ramirez, M., Obando, P., Sandi, L. & Karenkeris, M. d. C. (2006), 'Violence exposure and drug use in Central American youth: Family cohesion and parental monitoring as protective factors', *Journal of Research on Adolescence* **16**(3), 455–478.

Lardier Jr, D. T., Barrios, V. R., Garcia-Reid, P. & Reid, R. J. (2018), 'Preventing substance use among Hispanic urban youth: Valuing the role of family, social support networks, school importance, and community engagement', *Journal of Child & Adolescent Substance Abuse* **27**(5-6), 251–263.

Moos, R. H., Moos, B. S. & Trickett, E. J. (1995), *Manual de Escalas de clima social*, 4 edn, Madrid: TEA.

Olson, D. H. (2000), 'Circumplex model of marital and family systems', *Journal of family therapy* **22**(2), 144–167.

OMS (2011), *La prueba de detección de consumo de alcohol, tabaco y sustancias (ASSIST).*

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*https://www.R-project.org/

Sánchez-Queija, I., Oliva, A., Parra, Á. & Camacho, C. (2016), 'Longitudinal analysis of the role of family functioning in substance use', *Journal of Child and Family Studies* **25**(1), 232–240.

Wagner, K. D., Ritt-Olson, A., Chou, C.-P., Pokhrel, P., Duan, L., Baezconde-Garbanati, L., Soto, D. W. & Unger, J. B. (2010), 'Associations between family structure, family functioning, and substance use among Hispanic/Latino adolescents.', *Psychology of Addictive Behaviors* **24**(1), 98.

Wolf Peuker, A. C., Demetrio Caovilla, J., Batista da Costa, C. & Pereira Mosmann, C. (2020), 'Uso de alcohol y otras drogas por adolescentes: asociaciones con problemas emocionales y comportamentales y el funcionamiento familiar', *Psicologia Clínica* **32**(2), 315–334.

APLICACIÓN DE LOS MODELOS ARIMA EN LOS PRECIOS DE LA CEBOLLA JUNCA (ALLIUM FISTULOSUM) EN EL MUNICIPIO DE AQUITANIA

Especialización en Estadística

YESID FERNANDO MONTAÑA MONTAÑA[1,a]

[1]Escuela de Posgrados, Seccional Duitama, Universidad Pedagógica y Tecnológica de Colombia, Duitama, Colombia

Resumen

El presente artículo tiene como objetivo realizar pronósticos de los precios de la cebolla larga (Allium fistulosum) del municipio de Aquitania, departamento de Boyacá. Para ello se construye una serie temporal con los precios obtenidos de la página de Corabastos BOGOTÁ, D.C. El estudio cumple con las características de una investigación del tipo cuantitativo y la técnica utilizada es la aplicación de modelos ARIMA para la predicción de los precios de la cebolla larga. El estudio fue realizado con los precios desde enero del 2010, hasta octubre del 2021, estos permitieron evidenciar que, en los próximos seis meses a partir de octubre, el precio de la cebolla va en aumento, luego es prudente que el agricultor realice el siembro en el mes de noviembre del presente año.

***Palabras clave*:** Autocorrelación, Modelos ARIMA, Predicción, Precios cebolla larga.

Abstract

The objective of this article is to forecast the prices of Welsh onion (Allium Fistulosum) in the municipality of Aquitania, department of Boyacá. For this, a time series chart is built with the prices obtained from the "Corabastos" page BOGOTÁ, D.C.

The study fulfills with the characteristics of a quantitative type investigation and the technique used is the application of ARIMA models for the prediction of welsh onion prices.

This study was carried out with prices from January 2010 to October 2021, these data allowed to show that, in the following six months from October, the price of welsh onion is increasing, therefore, it is prudent for the farmer to sow in the month of November of this year.

***Key words*:** Autocorrelation, ARIMA models, Prediction, Welsh onion prices.

[a]Estudiante de Especialización en Estadística. E-mail: yesid.montana@uptc.edu.co

1. Introducción

El ser humano siempre ha buscado como predecir la ocurrencia de diferentes eventos en el futuro, desde pronosticar el clima hasta saber el comportamiento de las variables económicas que sustentan la existencia de las empresas o inclusive de un país; ya sea para buscar nuevas estrategias o resistir el impacto que producen las bajas demandas de un producto o la inflación en los costos de artículos necesarios para la producción de una materia prima.

La utilidad de las predicciones económicas es un hecho constatado en los países desarrollados para anticiparse a la toma de decisiones o a la aplicación concreta de políticas económicas. Una muestra de lo anterior lo presenta (Lopez, Flores & Sanchez 2017) en su artículo sobre el uso de modelos SARIMAX los cuales permiten predecir del tráfico total de pasajeros aéreos a nivel nacional, con el fin de evitar congestiones y brindar un mejor servicio a los pasajeros.

Un estudio sobre la predicción de la precipitación en el Departamento de Cundinamarca y la ciudad de Bogotá D.C, Colombia(Ramírez & Malagón 2018). Para el estudio se tomaron 133 registros de estaciones meteorológicas del IDEAM (Instituto de Hidrología, Meteorología y Estudios Ambientales), comprendidas entre enero del 2010 y diciembre del 2016 con una frecuencia mensual, utilizando interpoladores determinísticos espaciotemporales, series de tiempo y análisis de datos funcionales. Se obtienen entonces predicciones espacio-tiempo para los años 2017, 2018, 2019 y 2020 (48 meses) a escala local y/o regional, con un buen nivel de detalle y de baja complejidad.

Los estudios anteriores muestran la aplicación de modelos estadísticos que brindan pronósticos con alto grado de confiabilidad en las predicciones, las cuales permiten tomar decisiones acertadas respecto a un estudio que sea de interés.

El ministro de Agricultura y Desarrollo Rural, Rodolfo Zea Navarro, destacó que *"en equipo, con los productores y con las estrategias de financiamiento, emprendimiento y comercialización que hemos implemementado, logramos un crecimiento en el sector agropecuario de* $6,8\,\%$*en los tres primeros meses de 2020"(Minagricultura 2020)*. Lo anterior muestra la importancia del agro en la economía Colombiana, y por ese motivo es de gran importancia para el agricultor que su trabajo presente utilidades significativas como muchas otras profesiones del país. Para lograr que el trabajo del campo sea rentable es prudente tener conocimiento de las fechas de siembro con el fin de generar ganancias a la hora de la producción.

Según la encuesta realizada por el (DANE 2017) , en Colombia se cosecharon 14533 hectárea de Cebolla junca, con una producción de 289975 toneladas y rendimiento promedio de 39.9 toneladas por hectárea al año. El departamento de Boyacá es el principal productor con 195358 toneladas que corresponden al $67,4\,\%$ de la producción total, seguido por los departamentos de Nariño, Risaralda y Santander. Cabe resaltar que los mayores rendimientos en la producción se presentan en el departamento de Boyacá con 55,2 toneladas por hectárea al año, superando el promedio nacional.

El cultivo de la cebolla es un producto preferido por un gran número de agricultores debido a su rendimiento, ya que en promedio una planta de cebolla de rama puede producir entre $1,8$ y $2,2$ kilogramos de seudotallos en el periodo de cosecha. Otras de las grandes ventajas con la que cuenta el producto es que permite hacer varios cortes, o cosechas, normalmente 3, realizando el primer corte entre los 5 y 6 meses después de la siembra para alcanzar rendimientos desde mínimos de 30 toneladas a máximos de hasta 75 toneladas por hectárea, siendo el más común, alrededor de 40 toneladas por hectárea, el segundo corte a los 3 meses después y tercer corte a los 3 meses. Las exigencias del mercado en Colombia están relacionadas con la apariencia y la calidad, especialmente en lo que tiene que ver con el color del seudotallo, su textura y la pungencia (DANE 2017).

En Colombia la zona de cultivo con mayor área sembrada, producción, rendimiento y buena calidad de la cebolla se localiza en el municipio de Aquitania (departamento de Boyacá). Este municipio se encuentra a una altura entre los 3000 y los 3400 metros sobre el nivel del mar. La cebolla de rama pertenece a una especie vegetal que para su desarrollo, producción constante y buena calidad demanda buenos volúmenes de agua, siendo necesaria la aplicación de riego complementario, especialmente durante las temporadas de menor precipitación o en épocas de verano. Las condiciones agroecológicas más adecuadas para el desarrollo del cultivo de la cebolla de rama o cebolla junca son: Temperatura $11-20$ grados centígrados, altura sobre el

nivel del mar de 1500 a 3000 metros. Suelos de textura media franca (F) a franco-arcillosa (FAr), profundos, con buena retención de humedad y medio a alto contenido de materia orgánica.

El beneficio que le brinda el cultivo de la cebolla larga al agricultor aquitanence es: el área de siembra está en el rango de la mayoría ($51,82\,\%$) de predios de la zona, es decir, entre los 500 y los 5000 metros cuadrados; la siembra es cerca de la laguna; posee buenas condiciones técnicas; produce este alimento como monocultivo; no realiza rotación de cultivo; tiene buena capacidad financiera; paga a un intermediario para la negociación del producto en Corabastos (Bogotá); se informa de los precios; cuenta con contactos y lleva más de 15 años en la labor de producir cebolla de rama en la zona(DANE 2017).

El factor costo para producir la cebolla es una de las variables con mayor valor, debido a que el cultivo de cebolla de rama pastusa en Aquitania (Boyacá) arrojó unos costos totales por hectárea de $72960374 en mayo de 2017. A una fase inicial de establecimiento correspondieron $38614521; a un segundo corte, $18013388; y a la fase del tercer corte, $16332466, con unos rendimientos de 1500 rollos de 30 kilos por hectárea en la primera la fase, 1300 rollos en la segunda y en la tercera 1200, es decir, rendimientos por hectárea de 45000, 39000 y 3.000 kg, respectivamente. En esta forma, se obtuvo un $80\,\%$ de cebolla extra y un $20\,\%$ de cebolla de primera, conservando la misma proporción durante los próximos ciclos (DANE 2017).

De los datos anteriores se estima que en promedio los costos de producción por rollo de $18200. No hace falta ser economista para saber que si un agricultor vende el cultivo a un valor inferior a $18200 ya entra en perdidas. Debido a que la cantidad de agricultores de la región son pequeños productores, es decir, personas que no cuentan con los ingresos suficientes para sostener un cultivo durante todo el año, es de gran importancia conocer las fechas en las que no se debe sembrar cebolla de manera que al vender el producto, los precios no sean inferiores al costo de producción.

El municipio de Aquitania conserva un clima, altura, humedad y demás variables que influyen en la producción de un cultivo aptas para sembrar: Papa, Papa Criolla, Arveja, Zanahoria, Nabo, Haba entre otros; de modo que el pequeño agricultor puede optar por la siembra de otros cultivos diferentes a la cebolla que generen mayor utilidad.

De la anterior problemática se deduce la pregunta, ¿ Cómo predecir las fechas de sembradío del cultivo de cebolla junca en el municipio de Aquitania con el fin de evitar bajos precios a la hora de la producción?

El propósito del artículo está encaminado a proporcionar intervalos de tiempos críticos para el precio del cultivo de cebolla larga, y así alternar los diferentes cultivos en la región, de manera que aumente los ingresos de los agricultores.

La presente investigación tiene un enfoque cuantitativo. Se inicia con los conceptos introductorios del enfoque moderno, o estocástico, en análisis de series temporales, desarrollado por Box y Jenkins (1970) con el objetivo de estimar y diagnosticar modelos dinámicos de series temporales en los que la variable tiempo juega un papel fundamental (De Arce & Mahía 2003).

2. Referente Conceptual

El análisis de series de tiempo ayuda a detectar *regularidades* en las observaciones de una variable y derivar *leyes* a partir de ellas, o bien para explorar toda la información incluida en la variable para **predecir** mejor el futuro.

A inicios de del siglo XX, Slutzky y Yule mostraron que las series de tiempo con propiedades similares pueden generarse como sumas o restas (simples o ponderadas) de procesos aleatorios; desarrollaron los procesos de media móvil y autorregresivos como modelos para representar series de tiempo.

En 1970 Box y Jenkins publican un texto de análisis de series de tiempo. Introducen los modelos univariados para series de tiempo, que usan sistemáticamente la información contenida en los valores de la serie. Estos modelos son parsimoniosos en el sentido de producir pronósticos. Existe hasta los días presentes una metodología para la aplicación de los modelos univariantes, llamada *metodología de Box y Jenkins* , la cual consiste en:

1. Postular clase general de modelos.

2. Identificar un modelo provisional
3. Estimar parámetros del modelo.
4. Ejecutar pruebas de diagnóstico.
5. Usar el modelo para pronóstico.

(De Arce & Mahía 2003).

En el presente artículo se estudia la teoría necesaria para llevar a cabo la metodología anterior, para esto se inicia con los cimientos de las series temporales, y esto es los procesos estocásticos.

2.1. Proceso estocástico

Un proceso estocástico es una colección $\{Y_t\}_{t=1}^{\infty}$ de variables aleatorias iid(independientes e idénticamente distribuidas) , para $t = 1, 2, 3, \cdots, \infty$, ordenada de acuerdo con el parámetro discreto t, que en nuestro contexto es el tiempo (Otero 1993).

Una serie de tiempo es una realización de un proceso estocástico, es decir, una muestra aleatoria $\{y_t\}_{t=1}^{T}$.

Para tener una mayor comprensión de los conceptos de proceso estocástico y serie temporal, se analiza la Figura 1:

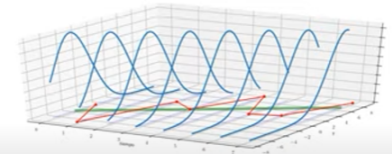

FIGURA 1: Relación proc. estocas y serie temporal

De la figura anterior podemos decir que la colección de curvas de densidad azules, es el proceso estocástico, es decir, para cada instante de tiempo t la variable aleatoria Y_t tiene una función de distribución. La trayectoria roja es la serie de tiempo, en decir, una muestra de realizaciones $Y_t = y_t$ que toma la variable aleatoria en el tiempo t. Generalmente los procesos estocásticos no son observables, luego lo que se busca describir el proceso estocástico que genere la serie conformada por las observaciones obtenidas.

Una serie de tiempo es una parte de un proceso estocástico para el cual ya se tienen las realizaciones del proceso (un valor por periodo)

$$\underbrace{y_1, y_2, y_3, \cdots, y_{T-1}, y_T}_{datos\ \ observados}, \underbrace{y_{T+1}, y_{T+2}, \cdots}_{datos\ \ futuros}$$

Lo que se busca es que de los datos observados proporcionen información útil para predecir los datos futuros.

Las leyes probabilistas que rigen cualquier proceso estocástico se describen mediante las funciones de probabilidad (*f.d.p*) conjunta de todos y cada uno de los vectores de variables aleatorias que construyen el proceso. Sin embargo, para fines prácticos los procesos se suelen describir, mediante sus momentos, entre los cuales se destacan:

- **Media de un proceso estocástico**

$$\mu_t = E(Y_t) \tag{1}$$

  y generalmente es una función del tiempo(no constante).

- **Función de autocovarianza**

$$\gamma_{t,t+k} = cov(Y_t, Y_{t+k}) = E((Y_t - E(Y_t))(Y_{t+k} - E(Y_{t+k}))) \quad k = 0, \pm1, \pm2, \pm3, \cdots \tag{2}$$

En particular de la ecuación anterior para $k = 0$ se obtiene la función de varianza del proceso:

$$\gamma_{t,t} = cov(Y_t, Y_t) = E((Y_t - E(Y_t))(Y_t - E(Y_t))) = E((Y_t - E(Y_t))^2) = Var(Y_t) \tag{3}$$

(Casimiro 2009)

- **Función de autocorrelación**

$$\rho_{t,t+k} = \frac{\gamma_{t,t+k}}{\sqrt{\gamma_{t,t}}\sqrt{\gamma_{t+k,t+k}}} \tag{4}$$

Los procesos estocásticos se dividen en dos clases:

- Estacionarios
- Evolutivos

Es importante aclarar que para propósitos de este artículo se estudian los procesos estacionarios.

2.2. Estacionariedad

De estacionariedad puede hablarse en un sentido amplio, y en otro estricto. Un proceso estacionario es *estricto* si los vectores:
$(Y_{t1}, Y_{t2}, \cdots, Y_{tn})$ y $(Y_{t1+s}, Y_{t2+s}, \cdots, Y_{tn+s})$ poseen la misma *f.d.p*, independientemente de s para un n dado. Lo anterior se interpreta como que el proceso estocástico no sufre ningunas alteración al considerar tiempos históricos diferentes. Esta restricción aplicada a la clasificación de un proceso es demasiado fuerte y en la práctica no sucede; por lo tanto, se concibe otro tipo de estacionariedad en sentido *amplio* (o estacionariedad de segundo orden o de covarianza estacionaria, o débilmente estacionario) cuando se verifique que:

$$\mu_t = \mu < \infty$$

Lo anterior significa que la media del proceso es constante.

$$\gamma_{t,t+k} = \gamma_k < \infty$$

Lo anterior significa que la autocovarianza es solo función del lapso temporal considerado, y no del tiempo histórico.

Para comprender estos conceptos que son la base de los modelos ARIMA, se muestra la siguiente figura tomada de (Otero 1993, p.204).

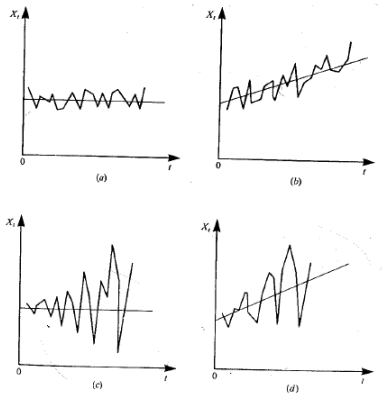

FIGURA 2: Criterios de estacionariedad

(a) representa un proceso estacionario en sentido amplio.
(b) proceso no estacionario a causa de la evolución de las medias (tendencias).
(c) la falta de estacionariedad se debe al crecimiento de la varianza en el tiempo.
(d) no hay constancia ni en la media ni el la varianza.

Como se mencionó anteriormente la mayoría de procesos que representan sistemas económicos no se ajustan a las condiciones de estacionariedad, pero es posible eliminar la tendencia y estabilizar sus varainzas para transformarlos en otros aproximadamente estacionarios.

En el presente artículo se enfoca al análisis y predicción de series de tiempo estacionarias o transformables a estacionarias bajo la metodología de Box-Jenkins. La metodología involucra **modelos autoregresivo (AR)**, **Modelo de medias móviles (MA)**, y **el modelo mixto (ARMA)**, y de ser necesario un modelo ARMA integrado, es decir un ARIMA.

Antes de definir los modelos anteriores es prescindible definir los conceptos de *Operador de rezagos* L y *Ruido blanco*.

Operador de rezagos L: Un operador L es una función tal que:

$$L(y_t) = y_{t-1}$$

El operador transforma una serie en otra serie atrasada en un periodo respecto de la original. Se puede observar que:

$$L(y_{t-1}) = L(L(y_t)) = L^2(y_t)$$

En general: $L(y_{t-k}) = L^k(y_t)$.

Ruido Blanco Es un caso simple de los procesos estocásticos $\{\varepsilon_t\}$, donde los valores son *iid* a lo largo del tiempo de manara que:

- $E(\varepsilon_t) = 0$ (media cero)
- $Var(\varepsilon_t) = \sigma^2$ (No hay heteroscedasticidad)
- $Cov(\varepsilon_t, \varepsilon_{t+k}) = 0$ (No hay autocorrelación)

2.3. Modelos estacionales lineales

2.3.1. Modelo autoregresivo $AR(p)$

El modelo autorregresivo de orden p , $AR(p)$, se define mediante la expresión:

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \tag{5}$$

Donde $\varepsilon_t \sim (0, \sigma_\varepsilon^2)$, es decir, un proceso ruido blanco.
Usando el operador L en la ecuación anterior se tiene:

$$y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p} = \delta + \varepsilon_t$$

$$y_t - \phi_1 L(y_t) - \phi_2 L^2(y_t) - \cdots - \phi_p L^p(y_t) = \delta + \varepsilon_t$$

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p) y_t = \mu + \varepsilon_t$$

$$\Phi(L)y_t = \delta + \varepsilon_t$$

La importancia del polinomio $\Phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p$ es con el fin de que se debe cumplir que las soluciones de la ecuación característica $\Phi(L) = 0$ deben estar fuera del círculo de radio uno en el plano complejo, para que el modelo $AR(p)$ sea estacionario. Otra forma de estudiar la estacionalidad amplia se da siempre y cuando se cumpla que $|\phi_i| < 1$, ver demostración en (Mauricio 2007, p.36).

Propiedades de los modelos AR_p

- $E(y_t) = \mu = \frac{\delta}{1-\phi_1-\phi_2-\cdots,\cdots\phi_p}$
- $Var(y_t) = \gamma_0 = \frac{\delta^2}{1-\phi_1^2-\phi_2^2-\cdots-\phi_p^2}$
- $Cov(y_t, y_{t-k}) = \gamma_k = \phi^k\gamma_0$
- $\rho_k = \phi^k$

Este modelo se caracteriza en la predicción de una variable en determinado tiempo, a partir de sus valores pasados. El valor de p en este tipo de modelo a demás de denotar en orden, también denota el número de rezagos a tener en cuenta, es decir la cantidad de periodos que se necesitan para obtener una predicción con el mínimo error.

Pronósticos de los modelos $AR(p)$ **cuando los parámetros** ϕ_i **son conocidos.**

La previsión de los procesos $AR(p)$ se realiza de forma iterativa. Formamos el pronóstico de 1 paso y, dado esto, formamos el pronóstico de 2 pasos y así sucesivamente. El pronóstico de 1 paso en el tiempo T de un proceso AR (p) se define como:

$$\hat{y}_{T+1} = \hat{\delta} + \hat{\phi}_1 y_T + \hat{\phi}_2 y_{T-1} + \cdots + \hat{\phi}_p y_{T+1-p}$$

(Racine 2019)

2.3.2. Modelo de Medias Móviles ($MA(q)$)

El modelo de medias móviles de orden q, $MA(q)$, se define mediante:

$$y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q} \tag{6}$$

Donde $\varepsilon_t \sim (0, \sigma_\varepsilon^2)$, es decir, un proceso ruido blanco.

Los primeros momentos se obtienen fácilmente:

Aplicando esperanza matemática a ambos lados de la ecuación se llega a:

$$E(Y_t) = \mu$$

Constante, y simplificando:

$$E((Y_y - \mu)^2) = \sigma^2(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)$$

como la cantidad de parámetros es finita, entonces la varianza es contante. Luego cumple las condiciones de estacionariedad.

Pronósticos de los modelos $MA(q)$ Hay dos fuentes de error en nuestras previsiones. La primera es que no conocemos el valor de los errores futuros ε_{T+i}. La segunda es que no conocemos los verdaderos parámetros del proceso, lo que da lugar a errores en las estimaciones de los parámetros y en las estimaciones de perturbaciones dentro de la muestra.

Procesos de previsión de $MA(q)$ cuando ε_{T-i} y θ_i son estimados.

Ahora, va a surgir una incertidumbre adicional porque se desconoce los verdaderos parámetros del proceso que dan lugar a errores en las estimaciones de los parámetros y en las estimaciones de perturbaciones dentro de la muestra. Claramente, lo anterior agregará incertidumbre adicional a nuestras previsiones. Los pronósticos seguirán siendo insesgados (el error de pronóstico seguirá teniendo una media de cero), sin embargo, el error de pronóstico tendrá una variación mayor. Por lo tanto, el pronóstico en el siguiente periodo.

$$\hat{y}_{T+1} = \hat{\mu} + \hat{\theta}_1\hat{\varepsilon}_T + \cdots + \hat{\theta}_q\hat{\varepsilon}_{T+1-q}$$

(Racine 2019)

2.3.3. Modelo autorregresivo de medias móviles $ARMA(p, q))$

El modelo $ARMA(p, q)$ se define mediante la ecuación:

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q} \tag{7}$$

Escrito en términos del operador L tenemos que:
$y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p} = \delta + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q}$, entonces $\Phi(L)y_t = \Theta(L)\varepsilon_t$; para $\delta = 0$, tales que las soluciones de las ecuaciones $\Phi(L), \Theta(L)$ están fuera del círculo unitario, para garantizar la estacionalidad del modelo, y ε_t es un proceso ruido blando.

Otra característica importante de estos modelos es que el término ε_t recibe el nombre de *Innovación*, por tratarse de una parte de la serie que no puede predecirse a partir de su propio pasado.

Para pronosticar valores futuros de una serie aplicando los conceptos matemáticos anteriormente explicados es necesario y suficiente que las series sean estacionarias. La mayoría de las series, y más las económicas no son estacionarias en media ni en varianza, por ese motivo la teoría se direcciona a las estrategias que se deben aplicar para lograr la estacionariedad.

Como los modelos $ARMA$ solo pueden ser aplicados a series que no muestren ningún tipo de tendencia, será necesario recurrir a métodos de eliminación de tendencia (en media y en varianza).

Tendencia en varianza o heteroscedasticidad. Prácticamente todas las series procedentes del mundo socioeconómico presentan heteroscedasticidad, en menos o myor grado.

Dado que tampoco existen contrastes univariantes de homoscedasticidad muy efectivos, en la práctica se opta por eliminar esa tendencia en todas las series, sometiendolas a algún tipo de transformación.

Las transformaciones más habituales son las transformaciones logarítmicas o cualquier otra perteneciente a la familia *Box-Cox*

$$y_t^* = \begin{cases} y_t^{\lambda} & si \quad -1 \leq \lambda \leq 1 \\ \\ ln(y_t) & si \qquad \lambda = 0 \end{cases} \tag{8}$$

Lo más usado para los casos de modelos univariantes es transformación logarítmica.

Para eliminar la **tendencia en media** se hace uso de las *diferencias sucesivas* de la misma serie. Así si la y_t muestra tendencia en media, ya no la tendrá con la siguiente transformación:

$$\Delta y_t = y_t - y_{t-1} \tag{9}$$

Cuando la tendencia es creciente o decreciente lineal, y realizamos una diferencia, entonces la nueva serie Δy_t, no tendría tendencia en media, es decir la línea de la tendencia sería una línea paralela al eje horizontal.

En toda transformación que se le realice a una variable, después de hallar lo que se necesita, se debe regresar a la variable original, es decir, se debe aplicar la transformación inversa. En el estudio que se está realizando, existe la transformación inversa a la **diferenciación** y es la **integración**. De este modo, la serie que necesita de una diferencia para ser estacionaria en media se dice que es **integrada de orden 1**(Racine 2019, p.56)

$$y_t \longrightarrow \underbrace{y_t - y_{t-1}}_{Diferenciación} \longrightarrow \Delta y_t$$

Proceso inverso,

$$\Delta y_t \longrightarrow \underbrace{y_{t-1} + \Delta y_t}_{Integración} \longrightarrow y_t$$

Se debe tener claro que una serie se puede diferenciar más de una vez, hasta lograra la estacionariedad en media. En la práctica por mucho se llega a una diferencia de tercer orden y esto se da cuando la línea de tendencia tiene forma polinómica de tercer grado, y si es integrada de orden 2, es porque la línea de tendencia tiene forma parabólica.

2.3.4. Modelos $ARIMA(p, d, q))$

Un modelo $ARIMA$ técnicamente es lo mismo que aplicar un modelo $ARMA(1, 1)$ a una serie homoscedástica con una diferencia a la parte regular (Δy_t) , que aplicar un modelo $ARIMA(1, 1, 1)$ a la serie original (en el caso de ser homoscedástica). y_t en general es un proceso $ARIMA(p, d, q)$

$$\Delta^d y_t = \phi_1 \Delta^d y_{t-1} + \phi_2 \Delta^d y_{t-2} + \cdots + \phi_p \Delta^d y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \tag{10}$$

(Racine 2019, p.55). En la expresión anterior se encuentra eliminado la media de la serie δ , ya que será prácticamente nulo, al trabajar con una serie en diferencias.

Todos los análisis respectivos se harán usando el software estadístico R y en el se utilizan algunos test para verificar el tipo de modelo que se acopla a la serie que se desea analizar. A continuación se muestra los test que se vana a utilizar.

2.3.5. Test de estacionariedad y Ruido blanco

Una de las pruebas que se va a realizar en el software es la estacionariedad usando la prueba de ***Dickey-Fuller*** para raíz unitaria.

$$\begin{cases} H_0 : \text{No es estacionaria (raíz unitaria)} \\ \qquad\qquad vs \\ H_1 : \text{Es estacionaria} \end{cases}$$

También se hará la prueba de ***Ljung Box*** para verificar la existencia de **Ruido blanco**. Recordemos que **Ruido blanco** significa que el error:

- Media igual a cero.
- Varianza constante.
- No estar serialmente relacionada.

Prueba de hipótesis:

$$\begin{cases} H_0 : \text{ Ruido blanco} \\ \qquad\qquad vs \\ H_1 : \text{No hay ruido blanco} \end{cases}$$

(Otero 1993).

3. Metodología

El tipo de investigación del presente artículo es cuantitativo. Los análisis se realizan con el software R Core Team (2020) y el entorno de desarrollo integrado RStudio Team (2020). Se inicia con la importación de la base de datos que conforman la serie de tiempo, luego se estudia la estacionariedad y el proceso de ruido blanco a través de los test, llegado el caso, que la serie no sea estacionaria, se aplican diferentes transformaciones para lograr la estacionariedad, para saber el números de autorregresivos y medias móviles se analizan los correlogramas de la *función de autocorrelación* y *función de autocorrelación parcial*, por último se realizan los pronósticos para periodos futuros.

Los datos se obtuvieron de la página principal de **Corabastos**, *https* : *//www.corabastos.com.co/*, en la sección de *Información de interés*, histórico de precios, luego se ingresa a tendencias por producto, se selecciona el producto de **Cebolla larga**, se delimita el intervalo de tiempo que se desea, y finalmente se descarga el archivo *Excel* que lleva los precios del producto en los respectivos días de mercado. Por motivos que no todos los días son de mercado en la plaza de corabastos para el cultivo de la cebolla, y además la cantidad de días no es constante en cada mes, fue necesario realizar una nueva base de datos que muestre el promedio de precios mensuales de la cebolla, de manera que los periodo de tiempo para la serie son mensuales.

4. Resultados y conclusiones

Luego de construir la base de datos que lleva los precios mensuales de la cebolla larga en la plaza mayorista de corabastos desde enero del 2010, hasta octubre del 2021, se procedió a realizar un tratamiento estadístico, más conocido como *metodología de Box-Jenkins.*

4.1. Identificación del modelo

4.1.1. Gráfico de la serie

Con la gráfica se determina si la serie es estacionaria, es decir, si la serie de tiempo varía al rededor de un nivel fijo.

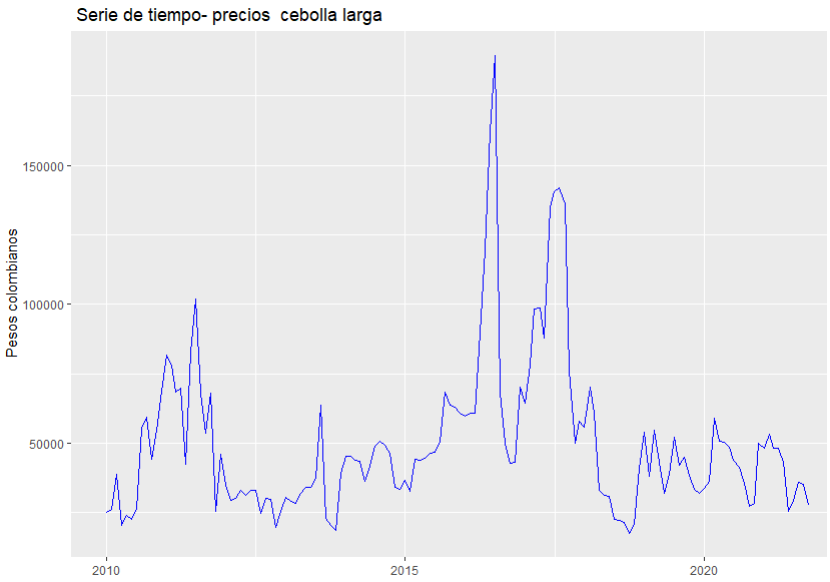

FIGURA 3: Serie de tiempo original

En la figura aparece el gráfico de la serie original, indica cambios en la varianza de la serie, y tendencia en ciertos intervalos de tiempo, por lo tanto, no parece ser estacionaria. La estacionariedad se prueba por el correlograma muestral y la prueba de Dickey- Fuller.

4.1.2. Análisis de Autocorrelación

La unción de autocorrelación **acf** proporciona la autocorrelación en todos los rezagos posibles.

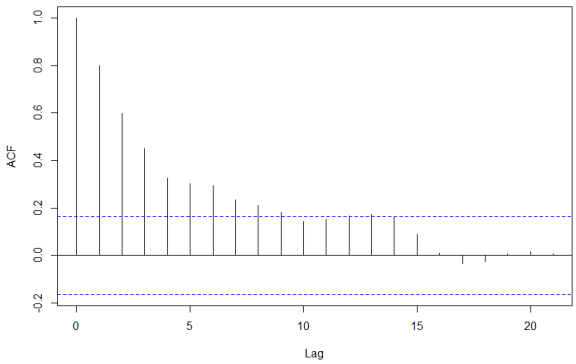

FIGURA 4: Función de autocorrelación

En el gráfico se observa que la autocorrelación continúa disminuyendo a medida que aumenta el número de rezagos, lo que confirma que no existe una asociación lineal entre las observaciones separadas por rezagos más grandes.

Ahora se analiza la gráfica de autocorrelación parcial (**pacf**) la cual muestra la autocorrelación entre todos los datos que están separados k rezagos, después de tener en cuenta la correlación, el gráfico nos ayuda a identificar el número de coeficientes autorregresivos (AR) en un modelo ARIMA.

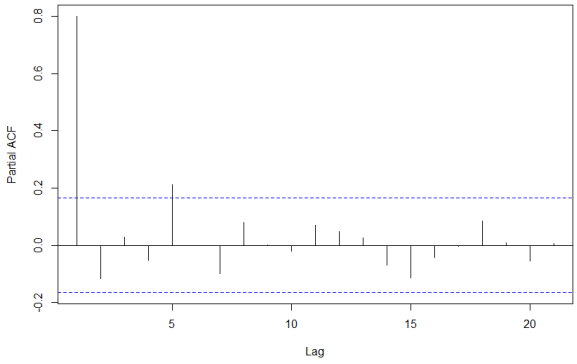

FIGURA 5: Función de Autocorrelación Parcial

Observando las funciones de autorrelación y autocorrelación parcial, se puede llegar a la conjetura que la serie de tiempo se puede modelar mediante un proceso $AR(1)$, ya que en la FIGURA 4 se muestra un descenso rápido, y en la FIGURA 5 se muestra un rezago. Sin embargo, se debe realizar una prueba formal para el estudio de la estacinariedad.

4.1.3. Prueba de Dickey-Fuller

Hipótesis:

$$\begin{cases} H_0 : \text{La serie es No estacionaria: Tiene raíz unitaria.} \\ \qquad\qquad vs \\ H_1 : \text{La serie es estacionaria: no tiene raíz unitaria.} \end{cases}$$

```
Augmented Dickey-Fuller Test

data:  serie_precios
Dickey-Fuller = -2.7321, Lag order = 5, p-value = 0.2717
alternative hypothesis: stationary
```

Como el $p-valor$ es mayor al 0.05 por lo que la hipótesis nula no se rechaza, la serie original no es estacionaria.

Tomando la primera diferencia de la serie

Al aplicarle la primera diferencia a la serie se obtiene la siguiente gráfica.

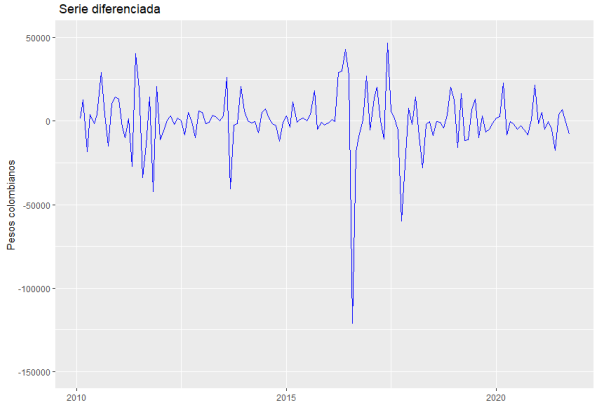

FIGURA 6: Serie diferenciada de primer orden.

Al aplicar la primera diferencia se observa que la serie se estabiliza en torno a un valor medio.

Nuevamente se muestran las funciones de autocorrelación y autocorrelación parcial.

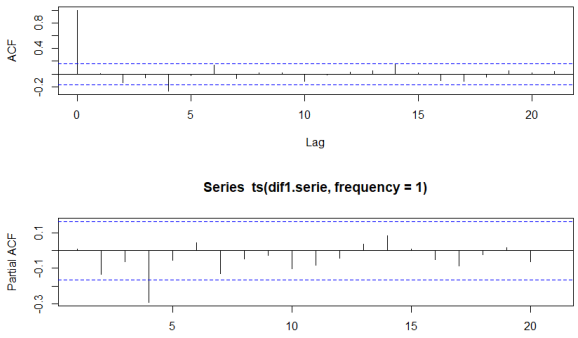

FIGURA 7: ACF y PACF de la serie diferenciada.

Los correlogramas muestran un declinamiento rápido, es una muestra de la estacionariedad en la serie diferenciada. Para confirmar se realiza nuevamente la prueba de **Dickey- Fuller**.

```
Augmented Dickey-Fuller Test

data:  dif1.serie
Dickey-Fuller = -5.7674, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

Como el $p - valor$ es menor a 0.05, se comprueba que la la serie en primera diferencia es estacionaria.

Una vez se ha obtenido estacionariedad en la serie, se debe identificar las partes del modelo. En la identificación de la forma del modelo se lleva a cabo las autocorrelaciones y las autocorrelaciones parciales.

4.2. Ajuste del modelo

El número de medias móviles se extraen del correlograma de autocorrelación, y el número de autorregresivos del correlograma de autocorrelación parcial. El proceso de identificación a partir de los correlogramas muestrales se suguiere un $ARIMA(4,1,1)$, es decir, un proceso autorregresivo integrado de medias móviles, con 4 autorregresivos, 1 diferencia y 1 media móvil; también se puede pensar en un $ARIMA(0,1,1)$.

Para escoger el mejor modelo realizamos las estimaciones de cada modelo por el método de máxima verosimilitud.

Modelo 1

```
arima(x = serie_precios, order = c(4, 1, 1), method = "ML")

Coefficients:
ar1      ar2      ar3      ar4      ma1
0.1216  -0.1722  -0.0402  -0.2802  -0.1510
s.e.  0.2799   0.0811   0.0884   0.0848   0.2937

sigma^2 estimated as 283654297:  log likelihood = -1572.43,  aic = 3156.86

Training set error measures:
ME     RMSE     MAE       MPE     MAPE     MASE          ACF1
Training set 38.08989 16782.63 10481.5 -5.496762 21.89727 1.010268 -0.0006612616
```

Modelo 2

```
arima(x = serie_precios, order = c(0, 1, 1), method = "ML")

Coefficients:
ma1
0.0126
s.e.  0.0978

sigma^2 estimated as 318271845:  log likelihood = -1580.35,  aic = 3164.7

Training set error measures:
ME     RMSE      MAE       MPE     MAPE      MASE        ACF1
Training set 16.88896 17777.25 10288.56 -4.545691 21.31529 0.9916717 -0.00158847
```

Para escoger el mejor modelo se compara el criterio de información (AIC) de Akaike para un conjunto de modelos y se elige los modelos con valores AIC más bajos, por lo tanto decidimos el modelo $ARIMA(4,1,1)$.

4.2.1. Modelo matemático de la serie temporal

$$\Delta y_t = 0.1216\Delta y_{t-1} + -0.1722\Delta y_{t-2} - 0.0402\Delta y_{t-3} - 0.2802\Delta y_{t-4} - 0.1510\varepsilon_{t-1} + \varepsilon_t \tag{11}$$

4.3. Validación del modelo

Ahora se va a estudiar los supuestos sobre la parte aleatoria del modelo. El análisis se inicia con la interpretación en los siguientes gráficos.

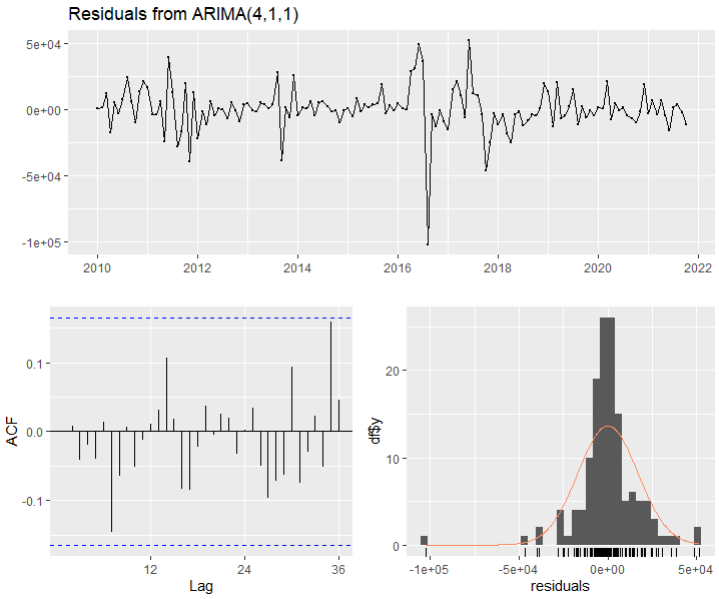

FIGURA 8: Comportamiento de los residuales.

En el gráfico superior se muestra la serie residual arrojada por el modelo, al observar la gráfica pareciera que se cumple el supuestos de **ruido blanco**, es decir, media constante cero; el gráfico inferior izquierdo muestra la función de autocorrelación de la cual se deduce la no autocorrelación en los errores, y el tercer gráfico muestra la normalidad lo cual garantiza la existencia de una sola media y varianza constante.

4.3.1. Ruido blanco

Hipótesis

$$\begin{cases} H_0 : \text{Existe ruido blanco.} \\ \qquad\quad vs \\ H_1 : \text{No existe ruido blanco.} \end{cases}$$

Inicialmente el siguiente gráfico nos da una idea de la existencia de **ruido blanco** en la serie de residuos.

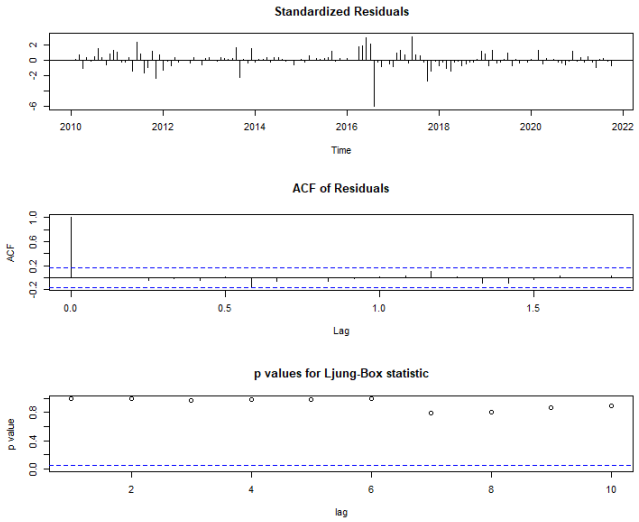

FIGURA 9: Ruido blanco.

Los primero gráficos muestran estacionariedad en los residuos estandarizados, así como la función de autocorrelación, el tercer gráfico muestra lo valores p arrojados por el test estadístico de Ljung-Box, en el cual se observa la existencia de ruido blanco, ya que todos los puntos están por arriba de la línea de referencia que indica el 0.05.

Para asegurar la existencia de ruido blanco hacemos uso directamente del test de Ljung-Box, el cual arroja los siguientes resultados:

```
Ljung-Box test

data:  Residuals from ARIMA(4,1,1)
Q* = 9.978, df = 19, p-value = 0.9535

Model df: 5.   Total lags used: 24
```

Como el $p - valor$ está muy por arriba del nivel de signifancia del 0.05, no hay evidencia suficiente para rechazar el ruido blanco, por lo tanto se puede asegurar que los residuales cumplen el supuesto de ruido blanco.

Como el modelo cumple todo lo necesario para realizar predicciones, ahora se va a calcular los posibles precios de la cebolla larga para los próximos 12 meses, ya que es el tiempo que dura la máxima producción.

4.4. Pronósticos

A continuación se muestran los precios pronosticados para los próximos 12 meses iniciando en noviembre del año 2021:

```
         Point Forecast      Lo 95     Hi 95
Nov 2021       27141.32  -5868.471  60151.11
Dec 2021       26576.01 -19427.347  72579.36
Jan 2022       27098.78 -25862.201  80059.77
Feb 2022       29436.77 -28863.775  87737.32
Mar 2022       29754.40 -30684.345  90193.14
Apr 2022       29527.89 -32925.104  91980.88
May 2022       29205.12 -36034.302  94444.55
Jun 2022       28536.88 -39740.664  96814.42
```

```
Jul 2022       28431.27 -43343.237 100205.77
Aug 2022       28609.93 -46484.887 103704.75
Sep 2022       28767.18 -49156.649 106691.00
Oct 2022       28947.06 -51557.624 109451.74
```

Los anteriores resultados muestran pronostico puntual y el intervalo de confianza al 95 % para los precios de la cebolla larga.

4.4.1. Gráfico del pronóstico

.

FIGURA 10: Pronosticos para los próximos 12 meses.

4.5. Conclusiones

- Los valores de las perdiciones puntuales muestran un aumento en el precio con una tendencia baja, sin embargo, los intervalos de confianza dan mayor información sobre las posibles alzas en los precios.
- Según los pronósticos arrojados por el modelo ARIMA permite a los agricultores tomar la decisión de sembrar el cultivo de la cebolla larga en el próximo mes, con el fin de no obtener pérdidas al cabo de seis meses.
- Se esperaba un comportamiento en la trayectoria de predicción con mayor tendencia o similitud con el pasado de la serie, una razón por la que no se cumple lo anterior es que los modelos ARIMA son regulares para predecir cuando la serie presenta cambios estructurales como se muestra en la serie original.
- Se deja constancia de la aplicación de la metodología Box y Jenkins, así como los diferentes test estadísticos para su aplicación en otras series de tiempo que involucran variables económicas de otros campos.

Referencias Bibliográficas

Casimiro, M. P. G. (2009), 'Análisis de series temporales: Modelos arima', *Universidad del País Basco* **1**(1), 1–169.

DANE (2017), 'El cultivo de la cebolla de rama (allium fistulosum l.) y un estudio de caso de los costos de producción.', *Boletín mensual INSUMOS Y FACTORESASOCIADOS A LA PRODUCCIÓN AGROPECUARIA.* .

De Arce, R. & Mahía, R. (2003), 'Modelos arima', *Programa CITUS: Técnicas de Variables Financieras* .

Díaz, L. G. & Morales, M. (2002), 'Análisis estadístico de datos categóricos', *Notas de Clase del Departamento de Estadística de la Universidad Nacional de Colombia. Bogotá: Universidad Nacional de Colombia* .

Leiva-Valdebenito, S. A., Torres-Avilés, F. J. et al. (2010), 'Una revisión de los algoritmos de partición más comunes en el análisis de conglomerados: un estudio comparativo', *Revista Colombiana de Estadística; Vol. 33, núm. 2 (2010); 321-339 Revista Colombiana de Estadística; Vol. 33, núm. 2 (2010); 321-339 0120-1751* .

Lopez, A. M., Flores, M. A. & Sanchez, J. I. (2017), 'Modelos de series temporales aplicados a la predicción del tráfico aeroportuario español de pasajeros: Un enfoque agregado y desagregado', *Estudios de economía aplicada* **35**(2), 395.

Mauricio, J. A. (2007), 'Análisis de series temporales', *Universidad Complutence de Madrid* .

Minagricultura (2020), 'Aporte del sector agropecuario en la economía colombiana.'.
*https://www.upra.gov.co

Otero, J. M. (1993), 'Econometría series temporales y predicción.', *Facultad de Ciencias Económicas y Empresariales. Universidad de Málaga* .

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*https://www.R-project.org/

Racine, J. S. (2019), *Reproducible econometrics using R.*, Oxford University Press, USA.

Ramírez, D. D. & Malagón, D. A. (2018), 'Interpoladores determinísticos espacio-temporales, series de tiempo y análisis de datos funcionales para el estudio y la predicción de la precipitación en cundinamarca y bogotá dc.'.

RStudio Team (2020), *RStudio: Integrated Development Environment for R*, RStudio, Inc., Boston, MA.
*http://www.rstudio.com/