DESARROLLO DE UN MODELO DE COINTEGRACION Y UN ÁRBOL DE DECISIÓN PARA CARACTERIZAR Y DETERMINAR CLIENTES QUE DEMANDAN FACTURACIÓN MANUAL EN GAS NATURAL DE LA CIUDAD DE BOGOTÁ

ROCIO DEL PILAR PINZON IBAÑEZ
CAMILO ALEXANDER RINCON LOPEZ

FUNDACIÓN UNIVERSITARIA LOS LIBERTADORES
ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA
BOGOTÁ, D. C. 2017

DESARROLLO DE UN MODELO DE COINTEGRACIÓN Y UN ÁRBOL DE DECISIÓN PARA CARACTERIZAR Y DETERMINAR CLIENTES QUE DEMANDAN FACTURACIÓN MANUAL EN GAS NATURAL DE LA CIUDAD DE BOGOTÁ

ROCIO DEL PILAR PINZON IBAÑEZ
CAMILO ALEXANDER RINCON LOPEZ

Trabajo de grado para optar el título de Especialista en Estadística Aplicada

Asesor Juan Camilo Santana Msc. Estadística

FUNDACIÓN UNIVERSITARIA LOS LIBERTADORES
ESPECIALIZACIÓN EN ESTADÍSTICA APLICADA
BOGOTÁ, D. C. 2017

Nota de aceptación

_____
_____
_____
_____
_____
_____

_____
Firma del presidente del jurado

_____
Firma del jurado

_____
Firma del jurado

Salvedad

Las directivas de la fundación universitaria los libertadores, los jurados calificadores y el cuerpo docente no son responsables por los criterios e ideas expuestas en el presente documento. Estos corresponden únicamente a los autores

1. TABLA DE CONTENIDO

LISTA DE GRAFICAS Y ECUACIONES

GLOSARIO

SERIE DE TIEMPO: Consiste en datos que se recopilan, registran u observan a lo largo de incrementos sucesivos de tiempo.

SERIE ESTACIONARIA: Aquellas cuyas particularidades estadísticas fundamentales, como la media y la varianza, permanecen constantes a lo largo del tiempo.

COINTEGRACIÓN: es una estadística característica de las variables en las series de tiempo donde dos o más series de tiempo están cointegradas si comparten una tendencia estocástica común.

ÁRBOLES DE DECISIÓN: Técnica que permite analizar decisiones secuenciales basada en el uso de resultados y probabilidades asociadas. Los árboles de decisión se pueden usar para generar sistemas expertos, búsquedas binarias y árboles de juegos, los cuales serán explicados posteriormente
.

NODO DE DECISIÓN: Indica que una decisión necesita tomarse en ese punto del proceso. Está representado por un cuadrado.

NODO DE PROBABILIDAD: Indica que en ese punto del proceso ocurre un evento aleatorio. Está representado por un círculo.

RAMA: Nos muestra los distintos caminos que se pueden emprender cuando tomamos una decisión o bien ocurre algún evento aleatorio

RESUMEN

Este proyecto consiste en desarrollar un modelo de cointegración y árbol de decisión para la cantidad de clientes que entran en análisis de facturación manual y las variables más influyentes que originan este comportamiento, determinando las correlaciones principales de los clientes con un histórico de cinco años (2012-2017) y un árbol de decisión con datos de los últimos dos años (2016-2017) en Gas Natural de la ciudad de Bogotá.

ABSTRACT

This project consists of developing a cointegration model and decision tree for the number of clients that enter into manual invoicing analysis and the most influential variables that originate this behavior, determining the main customer correlations with a five year historical (2012 -2017) and a decision tree with data from the last two years (2016-2017) in Gas Natural of the city of Bogotá.

INTRODUCCIÓN

El presente trabajo busca desarrollar un modelo de cointegración y árbol de decisión para la cantidad de clientes que entran en análisis de facturación manual y las variables más influyentes que originan este comportamiento en una empresa de gas natural. Los criterios de empresa que generan un proceso de facturación manual en los clientes son la alta utilización del servicio, anomalías en el contador y lecturas ausentes.

La facturación manual es realizada por un analista que verifica por medio de imágenes el flujo de consumo actual y revisa en el sistema el consumo de meses anteriores, a su vez se revisa un histórico de intervenciones operativas en terreno las cuales se mencionan a continuación; cambios de contador, revisiones de fugas, revisiones periódicas y revisiones por reclamo, a partir de lo anterior se define el método a utilizar para facturar el consumo.

1.2. PLANTEAMIENTO DEL PROBLEMA

En la empresa Gas Natural existe un proceso de facturación manual que se debe realizar por algunas de las siguientes razones: alto consumo por fuera del promedio propio del cliente, imposibilidad técnica al tomar la lectura, medidor averiado o encerrado. Por ello el alto nivel de clientes que demanda la revisión manual genera un costo significativo en la operatividad, la cual involucra alrededor de 15 analistas ejecutando el proceso.

Con base en lo anterior se busca desarrollar un modelo de cointegración donde se demuestre si existe una relación a largo plazo entre las variables proceso de Facturación Manual y Anomalías de Contador, el histórico que se tiene disponible para este análisis es de cinco años (Enero de 2012 a Enero de 2017).

De igual forma se generara un modelo de árbol de decisión que permita anticipar la probabilidad de análisis de clientes con Facturación Manual basados en los históricos de los últimos años, este modelo aportara valor predictivo y permitirá saber que clientes tienen mayor propensión a tener un proceso de facturación manual, se tendrá mayor conocimiento sobre los elementos que se correlacionan con el proceso de facturación manual, hecho que permitirá a su vez, diseñar estrategias proactivas.

1.3. FORMULACIÓN DE LA PREGUNTA

Los criterios que se emplean actualmente para reducir la cantidad de clientes que requieren facturación manual no son sistemáticos o basados en fundamentos científicos sólidos, con base en lo anterior surge la siguiente pregunta. ¿Cómo, a través de un modelo de cointegración y árbol de decisión, es posible identificar las variables más influyentes y determinar clientes que demandan facturación manual en Gas Natural de la ciudad de Bogotá?

1.4. JUSTIFICACIÓN

La empresa Gas Natural es una de las compañías multinacionales líderes en el sector de gas y electricidad, con una creciente y diversificada presencia internacional, Gas Natural en Colombia es un grupo de cuatro empresas nacionales independientes, cuya actividad principal es la distribución y comercialización de gas natural por red de tubería. Gas Natural S.A., ESP es la matriz en Colombia que a su vez es filial de la empresa española Gas Natural Internacional SDG S.A, atiende a 2.9 millones de clientes a través de una red de distribución de 21.000 kilómetros, igualmente, operan las compañías Gas Natural Cundiboyacense S.A. ESP; Gas Natural del Oriente S.A., ESP; GasNacer S.A., ESP; y , Gas Natural Servicios S.A.S., empresas de servicios públicos vigilados por la Superintendencia de Servicios Públicos.

Gas Natural siempre se ha destacado por sus altos índices de satisfacción con el cliente en todo nivel, esto hace pensar que la compañía está centrada en prestar un servicio de calidad y responsabilidad en sus procesos de distribución, comercialización, mantenimiento de redes y nuevas tecnologías, también se enfocada en la seguridad de los clientes promoviendo campañas como “Despierta el monóxido de carbono mata” donde se brindan todas las indicaciones para evitar la inhalación de este peligroso gas, que se promueve a través de estribillos y mensajes en diferentes medios de comunicación.
A nivel general, la empresa brinda bienestar a los clientes y crecimiento en todo el país llegando a zonas de difícil acceso y brindando las mejores alternativas para el mercado doméstico, comercial, industrial y vehicular.
La importancia de tener un adecuado proceso de facturación de consumo nos permite mantener nuestros índices de calidad en el

servicio, por ello si ejecutamos este sin errores no existirán reclamos por parte de los clientes.

Con base en la definición anterior se busca que a partir de la implementación de este trabajo se logre una reducción de por lo menos un 25% en la operatividad de clientes que demandan un análisis de facturación manual en la empresa Gas Natural, el impacto ocasionara reducción de los costos en el personal operativo, puestos de trabajo, equipos tecnológicos, formación de personal e índices de error.

1.5. OBJETIVO GENERAL

Desarrollar un modelo de cointegración (ENGLE – GRANGER) con un histórico de cinco años (2012-2017) y un árbol de decisión con datos de los últimos dos años (2016-2017) en Gas Natural de la ciudad de Bogotá.

1.5.1. OBJETIVOS ESPECÍFICOS

Ejecutar un modelo de cointegración (ENGLE – GRANGER) de la serie de tiempo clientes con facturación manual de los últimos cinco años (enero-2012 – enero-2017), incluyendo las variables cantidad de clientes con facturación manual y anomalías de contador.

Realizar un árbol de decisión con los clientes que han pasado por un proceso de facturación manual, de los últimos dos años (mayo-2015 – abril-2017).

Determinar cuáles son las correlaciones principales de los clientes que han pasado por un proceso de facturación manual, por medio de un modelo de árbol de decisión donde se identifique que potencial de clientes tienen mayor probabilidad de tener un proceso de facturación manual.

2. MARCO TEÓRICO

2.1. COINTEGRACIÓN

Dos o más series de tiempo que son no estacionarias de orden I(1), están cointegradas si hay combinación lineal de esas series que sea estacionaria o de orden I(0). El vector de coeficientes que crea esta serie estacionaria es el vector cointegrante.

2.1.1. CARACTERISTICAS DE SERIES DE TIEMPO TEMPORALES

- La mayoría de series tienen una tendencia. Su valor medio cambia con el tiempo son las llamadas series no estacionarias.
- Algunas series describen meandros, es decir, suben y bajan sin ninguna tendencia o linealialidad obvia o de revertir hacia algún punto.
- Algunas series presentan shocks persistentes. Los cambios repentinos en estas tardan mucho tiempo en desaparecer.
- Algunas series se mueven conjuntamente, es decir tienen movimientos positivos.
- La Volatilidad de algunas series cambia en el tiempo o estas series pueden ser más inconstantes en un año que en otro.

2.1.2. PRUEBA DE DICKEY FULLER

Antes de procesar los datos es necesario identificar si las series son estacionarias. Dickey Fuller sugiere las siguientes ecuaciones para determinar la presencia o no de raíces unitarias.

$$\Delta Y_1 = \delta Y_{t-1} + v_1$$
$$\Delta Y_1 = \alpha + \delta Y_{t-1} + v_1$$
$$\Delta Y_1 = \alpha + \beta T + \delta Y_{t-1} + v_1$$

Ecuación 1. Estacionariedad

La diferencia entre estas tres regresiones envuelve la presencia de los componentes determinísticos: intercepto (drift) y tendencia (T). La primera es un modelo puramente aleatorio. La segunda añade un

intercepto y un término de derivada, drift, y la tercera influye intercepto y un término de tendencia.

El parámetro de interés en las tres regresiones es $\boldsymbol{\delta}$

2.1.3. PASOS PARA LA PRUEBA DE HIPOTESIS

1. Planteamiento de hipótesis:

$H_0: \Phi^* = 0$ La serie no estacionaria: tiene raíz unitaria
$H_1: \Phi^* \neq 0$ La serie es estacionaria

Ecuación 2. Pruebas de Hipótesis

2. Estadísticos para la prueba:

$$H_0: \beta = 0, \alpha = 0, ADF \text{ y los valores críticos de Mackinnon}$$

Ecuación 3. Pruebas de Hipótesis

3. Regla de la decisión

Comparan el valor tau con los valores críticos Mackinnon

$SI\ |t^*| \leq |valor\ critico\ DF|$ Rechace H_0 serie estacionaria
$SI\ |t^*| > |valor\ critico\ DF|$ Acepte H_0 serie No estacionaria

Ecuación 4. Pruebas de Hipótesis

4. Conclusión:

PCE Como $|t^*| > |valor\ critico\ DF|$ serie No estacionaria
PDI Como $|t^*| > |valor\ critico\ DF|$ serie No estacionaria

Ecuación 5. Pruebas de Hipótesis

2.2. ÁRBOL DE DECISIÓN

Una de las técnicas más utilizadas dentro del análisis predictivo son los árboles de decisión. Esta técnica tiene múltiples aplicaciones en el campo de la estadística, pero nos vamos a centrar en su uso para

realizar predicciones, concretamente obtener probabilidades de eventos.
Un árbol de decisión es una forma gráfica y analítica de representar todos los eventos (sucesos) que pueden surgir a partir de una decisión asumida en cierto momento.

Nos ayudan a tomar la decisión “más acertada”, desde un punto de vista probabilístico, ante un abanico de posibles decisiones.
Permite desplegar visualmente un problema y organizar el trabajo de cálculos que deben realizarse.

2.2.1. MUESTRA

Hay que distinguir entre una muestra con la que se construye un modelo, sea este de árbol de decisión o de otra naturaleza, de muestras de testeo. Sólo los resultados medidos en muestras de testeo son válidas para estimar lo que pasará con otros clientes.

Una muestra de testeo es una muestra que es independiente de la que se usa para construcción. Puede obtenerse usando funciones pseudoaleatorias que típicamente están disponibles en planillas o bases de datos. Una vez obtenida la muestra es importante revisar que no contenga registros comunes con la muestra de construcción. También es importante verificar que para algunas variables bien conocidas, tales como género, edad, ciudad, etc., sus histogramas para los datos de la muestra se vean con una distribución similar a la de los histogramas con los datos de toda la población.

Durante la construcción del modelo o al final del proceso es importante usar la muestra de testeo para hacer unos ajustes finales. Por ejemplo, ajustes en puntos de corte, ya sea del modelo o de algunas variables. Si así se hizo, entonces esa muestra no puede usarse para estimar la capacidad de discriminación del modelo construido ni de su error. Habrá que usar otra muestra independiente.

2.2.2. SELECCIÓN DE MUESTRA PARA CONSTRUCCIÓN Y VERIFICACIÓN DE REPRESENTATIVIDAD

Existen varias maneras de obtener muestras. Una forma sencilla y directa es crear en una base de datos una columna adicional, y llenarla con la función aleatoria con números entre Cero y un millón. Luego se ordena la base de datos por ese campo y se toma la mitad superior, es

decir, los que están primeros. Si se requiere una muestra de 30% de la población, se toman todos los primeros hasta completar el 30%.

Es muy importante verificar la representatividad de la muestra. Así se descartan posibles errores en los procedimientos de creación de muestras aleatorias.

Se recomienda comparar distribuciones de variables conocidas en la población con las mismas variables en la muestra. Por ejemplo, género, edad, ciudad de residencia, ingresos, etc.

2.2.3. ALMACENAMIENTO Y USO DE MUESTRAS DE TESTEO

Es crítico asegurar la independencia de las muestras de testeo de la muestra utilizada para la construcción. Los errores más frecuentes son:

- Procedimiento mal realizado para sacar muestras de testeo.
- Utilizar toda la información para construir el modelo.
- No almacenar las muestras de testeo debidamente identificadas, dificultando así la creación posterior de muestras de testeo.
- Mezclar algunos registros de la muestra de construcción en la muestras de testeo.
- Uso, aparentemente menor y tangencial, de información en las muestras de testeo para ajustes de modelo.
- Un uso inadecuado de las muestras de testeo puede hacer creer que un modelo tiene cerca de 100% de discriminación (KS = 100) cuando en realidad tiene cerca de cero.

3. MARCO DE REFERENCIA

3.1. GENERALIDADES

La unidad de planeación minero energética entrega un balance de gas natural en Colombia, donde relaciona los escenarios de oferta, demanda de gas natural y balance nacional, para este trabajo se muestran los escenarios de oferta:

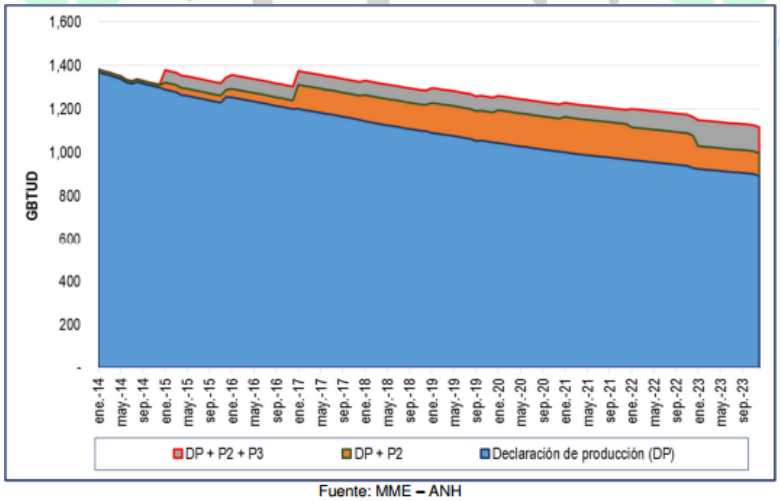
3.1.1. ESCENARIOS DE OFERTA DE GAS NATURAL EN COLOMBIA.

Para el análisis se incluyeron tres escenarios de oferta, que estiman la situación de corto, mediano y largo plazo (10 años es el periodo de análisis). El primer escenario es normativo y los otros dos consideran la

perspectiva sobre reservas de gas natural y disponibilidad complementaria de gas natural, mediante un esquema de suministro proveniente del mercado externo.

Conforme con lo definido por el Decreto 2100 de 2011, y con los lineamientos generales para la realización del Plan Indicativo de Abastecimiento de Gas Natural, el escenario base de oferta de gas natural corresponde a la última declaración de producción nacional e importación por parte de agentes. Sobre éste se considerarán otros escenarios, resultado de la incorporación de reservas probables, reservas posibles y de la construcción de una planta de regasificación en la Costa Atlántica.

En la gráfica 1 se presenta un escenario conformado por la oferta nacional de gas natural, (declaración de producción (DP), las reservas probables y las reservas posibles). Bajo éste escenario, se esperaría una máxima producción en los meses enero de 2015 y enero de 2017 con 1.380 GBTUD y 1.375 GBTUD respectivamente, y posteriormente se espera un comportamiento conforme a la declinación normal de los campos productores alcanzando los 1.116 GBTUD al final del periodo de análisis.



Fuente: MME – ANH

Gráfica 1. Declaración de producción, reservas probables y posibles

Adicionalmente a la oferta nacional, en el año 2013, el país tomó la decisión de disponer de una nueva fuente de suministro, debido al déficit en el balance oferta demanda estimado con las declaraciones de producción y la demanda del escenario medio determinado por UPME. Esta fuente supletoria corresponde a la instalación de una planta de regasificación ubicada en inmediaciones de la ciudad de Cartagena con

una capacidad de 400 MPCD, volumen que hará parte de la oferta en los escenarios planteados e ingresará a partir de enero de 2017.

3.1.2. ESCENARIO BAJO DE OFERTA

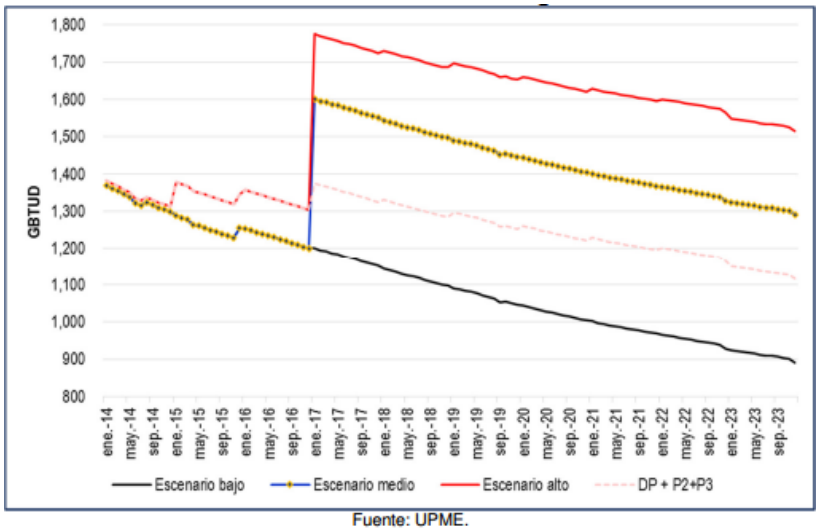
Éste escenario corresponde exclusivamente al volumen informado por los productores en la declaración de producción de gas natural en el año 2014, Resolución Ministerio Minas 72206 de 2014, descrito en el numeral 3.2.

3.1.3. ESCENARIO MEDIO DE OFERTA

En el escenario medio se adicionan a la declaración de producción (esc. bajo) los 400 MPCD que se estima aportará la planta de regasificación, volumen que para efectos de éste ejercicio ingresan a partir de enero de 2017, precisando que para este ejercicio la planta suministrará gas a toda la demanda y no exclusivamente al sector térmico. Este escenario tiene la mayor probabilidad de ocurrencia toda vez que la declaración de producción corresponde con las reservas probadas, volúmenes que tienen una probabilidad de ser producidos del 90%, más la entrada de la planta de regasificación, situación que se incluye en el escenario de manera determinística; sumadas éstas dos consideraciones se puede afirmar que el escenario medio es el que posee menor incertidumbre.

3.1.4. ESCENARIO ALTO DE OFERTA

El escenario alto es el resultado de la combinatoria entre el escenario medio y el aporte esperado por la extracción de reservas probables y posibles, cuyos volúmenes tienen una probabilidad de ocurrencia del 50% y 10% respectivamente, lo que hace que disminuya la probabilidad total de producir esta categoría de reservas. Adicional a lo anterior, debe tenerse en cuenta la situación actual de los precios del petróleo, donde el efecto puede ser un retraso en inversiones, que se traduce en pérdida de volúmenes de producción adicionales. Sin embargo para efectos del ejercicio se incluyeron en el escenario alto de oferta.
En la gráfica 2, se presentan los tres escenarios evaluados y adicionalmente se incluye una combinación de los perfiles de producción de los tres tipos de reservas, situación intermedia entre los escenarios bajo y medio.



Fuente: UPME.

Gráfica 2. Escenarios de oferta de gas natural.

3.2. ESTUDIOS REALIZADOS

A continuación se relacionan los siguientes tres estudios realizados en series de tiempo para la industria de gas natural

- En un estudio realizado en la universidad complutense de Madrid en la cual diseñaron un modelo de predicción de la demanda convencional de gas, con el objetivo de obtener pronósticos de la Demanda Convencional de Gas con frecuencia diaria, obteniendo un modelo SARIMA (1,0,1)x(0,1,1)7, demostrando el efecto sobre la Demanda del dato del día interior, así como un comportamiento estacional diario.

- El autor Gastón Carrazán realiza un análisis de la evolución de la producción mensual de Gas Natural extraído en los yacimientos de la provincia de Salta, desde Enero de 1992 a Junio de 2008 bajo un modelo de series de tiempo obteniendo una relación causal entre la producción de Gas y la de Petróleo y el Ingreso Nacional. El autor resalta que, cuando en el largo plazo aumenta el producto/ingreso, por ello aumentan los requerimientos y la producción de Gas Natural.

- El análisis de series de tiempo para la predicción de los precios de la energía en la bolsa de Colombia utiliza la serie histórica de datos entre enero de 1996 y junio de 2007. Los resultados identifican algunos períodos de intervención por parte del

regulador, sólo hasta el momento en que fueron incorporadas dichas intervenciones a las técnicas de estimación. Se logró obtener un modelo que cumpliera todos los estadísticos de prueba tanto en la significancia de los parámetros como en el análisis de residuales. La predicción de los Precios de la Energía en Bolsa en el largo plazo (años) requiere modelos más elaborados que incluyan variables como la operación del sistema de transmisión nacional, la estructura del mercado, los mecanismos de contratación, la simulación del despacho económico, como es el caso de los modelos de Análisis de Equilibrio (Guang 2005).

4. MARCO METODOLÓGICO

4.1. ANÁLISIS DE COINTEGRACIÓN

A continuación se relacionan los pasos para el desarrollo del modelo:

Etapa 1: Descripción de las variables:

- Cantidad de clientes con facturación manual: Esta variable es el conteo de clientes únicos que requieren un análisis de facturación manual. (La variable en la base es nombrada como DF)

- Anomalías de contador: Esta variable es el conteo de clientes únicos que tienen anomalía de contador identificada en terreno. (La variable en la base es nombrada como DA)

Etapa 2: Recolección de Datos:

- Se obtienen la serie de tiempo clientes con facturación manual que inicia en (enero-2012 – enero-2017), para el ejercicio se nombra con las siglas FMGN (Facturación Manual Gas Natural)

Etapa 3: Desarrollo del modelo:

- Para realizar este análisis se utiliza el software estadístico R
- Prueba de Dickey Fuller para la variable DF
- Prueba de Dickey Fuller para la variable DA

- Desarrollo de ecuación de largo plazo
- Validación de residuales
- Ecuación del error
- Residuales del modelo a corto plazo

4.2. ÁRBOL DE DECISIÓN EN FUNCIÓN DE PROBABILIDADES

A continuación se relacionan los pasos para el desarrollo del modelo:

Etapa 1: Descripción de las variables:

- Variable: Nombre de la variable
- Descripción: Criterio cálculo de la variables
- Rol: El rol define si es una variable dependiente o independiente
- Tipo: Define si es texto, fecha, numero etc.
- Medida**:** Es el valor que toma la variable, se puede identificar de manera cualitativa o cuantitativa

Variable	Descripción	Rol	Tipo	Medida
Indicador Proceso Facturación Manual	Indica si la observación tiene facturación manual	Dependiente	Texto	SI/NO
Consumo Promedio Efectivo Últimos 6 Meses	Es el promedio de los últimos seis meses excluyendo Consumos en ceros	Independiente	Numérica	Mts^3
Consumo Último Mes	Consumo registrado en la última Factura	Independiente	Numérica	Mts^3
Consumo Promedio efectivo Últimos 9 Meses	Es el promedio de los últimos nueve meses excluyendo Consumos en ceros	Independiente	Numérica	Mts^3

Gráfica 3. Descripción y definición de variables

Etapa 2: Recolección de Datos:

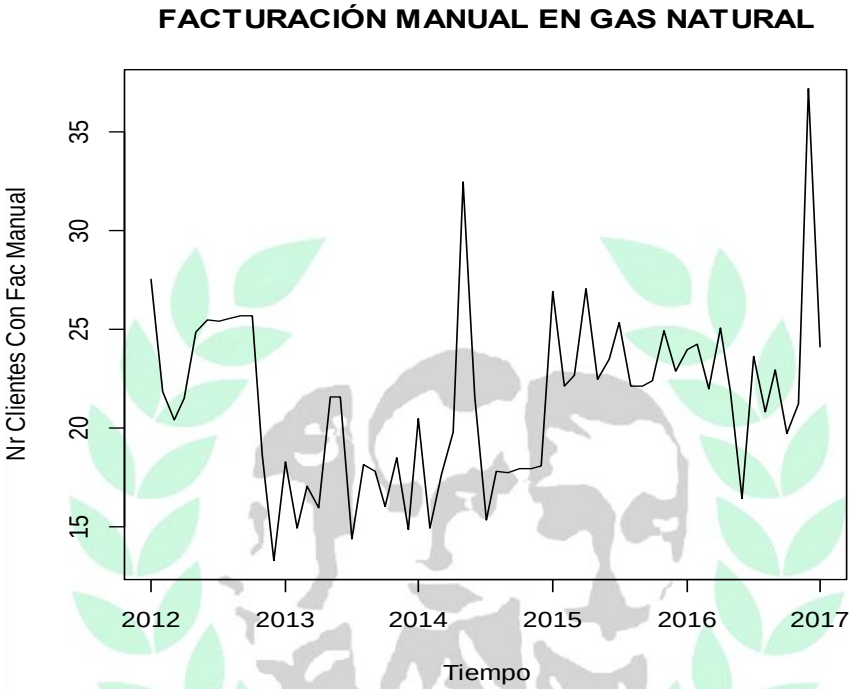
- Se obtiene una base de datos con 21.114 observaciones, esta es una muestra del 1% del total de los clientes de gas natural, se relacionan los clientes que sí y no han pasado por un proceso de facturación manual con el fin de realizar un balance en la muestra de datos.

Etapa 3: Desarrollo del modelo:

- Para realizar este análisis se utiliza el software estadístico SAS ENTERPRISE MINER.
- Se crea es un nodo de exploración de estadísticos básicos, solo se analizan las variables más relevantes que son: Consumo efectivo últimos 6 meses, Consumo efectivo últimos 9 meses y Consumo último mes.
- Se crea un nodo de partición de los datos, este lo que hace es dividir las observaciones en una base de entrenamiento y una base real, para el pronóstico.
- Se crea un nodo de árbol decisión. Este nodo tiene dos funcionalidades la primera es que podemos diseñar el árbol manualmente y la segunda es dejar que el sistema lo realice, en este caso se dejara que el sistema diseñe el árbol, cabe agregar que el SAS MINER postula el mejor modelo, sin embargo se revisa la lógica del modelo.
- Análisis discriminante.

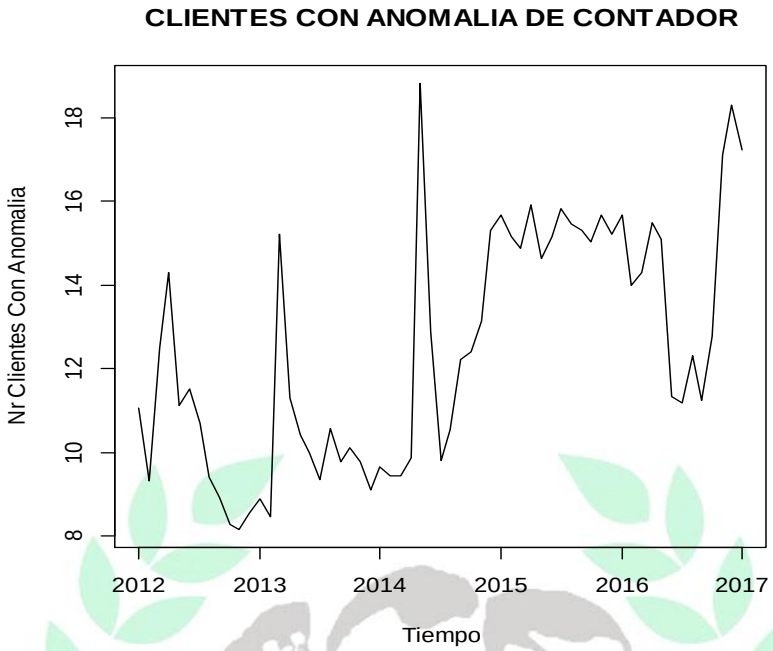
5. ANÁLISIS DE RESULTADOS

5.1. ANÁLISIS DE COINTEGRACIÓN



Gráfica 4. Serie de tiempo Variable – DF

En la Grafica 4, se visualiza la serie de tiempo variable facturación manual, en donde se evidencian: en primera instancia cambios abruptos en las observaciones (outliers) en mayo de 2014 con 32.445 clientes y en febrero de 2016 con 64.258 clientes, esta situación se da por campañas que se utilizan para encontrar clientes auto reconectados es decir un cliente que estuvo suspendido y el mismo se conecta sin autorización. En segunda instancia un outliers de cambio de nivel entre las observaciones de mayo de 2015 hasta octubre de 2015, este cambio se da por campañas de normalización de medidores realizadas en 2014, estas hacen referencia a acciones que se realizan en terreno y en el sistema de la empresa como actualizaciones o reparaciones del medidor.



Gráfica 5 Serie de tiempo Variable – DA

En la Grafica 5, se visualiza la serie de tiempo variable clientes con anomalía de contador, en donde se evidencian cambios abruptos en las observaciones (outliers) en mayo de 2014 con 18.920 y diciembre de 2016 con 18.298 clientes con anomalía de contador, este comportamiento está dado por campañas de detección de contadores encerrados, es decir clientes que aún tienen el medidor dentro del lugar de su vivienda, por ende en el momento de tomar la lectura una persona debe darle acceso al técnico de lecturas para que pueda leer el contador, por esta razón se incrementó las anomalías de contador.

5.1.1. PRUEBA DE DICKEY FULLER PARA LA VARIABLE DF

Al implementar esta prueba se consideró inicialmente la existencia de tendencia determinística, en este caso para probar la existencia de una raíz unitaria, se deben contrastar las hipótesis:

$H_0: \Phi^* = 0$ Implica que la serie tiene raíz unitaria, por tanto la serie NO es estacionaria

$H_1: \Phi^* < 0$ Implica que y la serie NO tiene raíz unitaria, por tanto la serie es estacionaria

Estadísticos de Prueba	Valor Obtenido	5pct
Raíz Unitaria	-3,8273	-3,45
Tendencia	5,0613	4,88
Intercepto	7,5721	6,49

Gráfica 6. Prueba DICKEY FULLER para la variable DF

En la gráfica 6, se rechaza H_0 del estadístico de prueba tau3 indicando que no hay raíz unitaria. De igual manera en la tendencia al 5pct (phi2), se rechaza H_0 indicando que hay tendencia. A su vez Intercepto al 5pct (phi3): Se rechaza H_0 indicando que hay intercepto.

5.1.2. PRUEBA DE DICKEY FULLER PARA LA VARIABLE DA

Al implementar esta prueba se consideró inicialmente la existencia de tendencia determinística, En este caso para probar la existencia de una raíz unitaria, se deben contrastar las hipótesis:

$H_0: \Phi^* = 0$ Implica que la serie tiene raíz unitaria, por tanto la serie NO es estacionaria

$H_1: \Phi^* < 0$ Implica que y la serie NO tiene raíz unitaria, por tanto la serie es estacionaria

Estadísticos de Prueba	Valor Obtenido	5pct
Raíz Unitaria	-3,7076	-3,45
Tendencia	4,7419	4,88
Intercepto	6,9168	6,49

Gráfica 7. Prueba DICKEY FULLER para la variable DA

En la gráfica 7, se rechaza H_0 del estadístico de prueba tau3 indicando que no hay raíz unitaria. De igual manera en la tendencia al 5pct (phi2), no se rechaza H_0 indicando que no hay tendencia. A su vez Intercepto al 5pct (phi3): Se rechaza H_0 indicando que hay intercepto.

5.1.3. ETAPA 1 DESARROLLO DE ECUACIÓN DE LARGO PLAZO.

De acuerdo a las pruebas de DICKEY FULLER procedemos a generar la ecuación de relación de largo plazo:

Coefficients:

	Estimate	std.Error	t.value	Pr(>|t|)	
(Intercept)	1,82999	0.23697	7.722	1.79e-10	***
bdd2w	0,4835	0.09446	5.119	3.65e-10	***

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

Residual standard error: **0.1692 on 58 degrees of freedom**

Multiple R-squared: **0.3112,** Adjusted R-squared: **0,2993**

F-statistic: **26.2 on 1 and 58 DF,** p-value: **3,65e-06**

Gráfica 8. Análisis de Ecuación a largo Plazo

*Ecuación en logaritmos

A partir de la gráfica 8, se genera la ecuación de largo plazo la cual se menciona a continuación:

$$\log(DF_t) = \alpha_0 + \alpha_1 \log(DA_t) + e_t$$

$$Yt = \alpha_0 + \alpha_1 + t + e_t$$

Ecuación 6. Largo Plazo

$$\widehat{\boldsymbol{\alpha}}_{\boldsymbol{o}} = \mathbf{1.82}$$
$$\widehat{\boldsymbol{\alpha}}_{\mathbf{1}} = \mathbf{0.48}$$

$\widehat{\boldsymbol{\alpha}}_{\boldsymbol{o}}$ = Facturación Manual

$\widehat{\boldsymbol{\alpha}}_{\mathbf{1}}$ = Anomalía de Contador

A partir de la Ecuación 6, se demuestra que existe una relación significativa a largo plazo entre las variables de *Facturación Manual* y

Anomalía de Contador, siendo 1.82 la tasa de crecimiento mensual de clientes con facturación manual y por cada unidad porcentual de anomalías de contador se incrementa las facturaciones manuales en 48.35%.

5.1.4. ETAPA 2 VALIDACIÓN DE RESIDUALES.

El siguiente paso es validar que los residuales son *I(0).*

$H_0: \Phi^* = 0$ Implica que los residuales NO son *I(0)*
$H_1: \Phi^* < 0$ Implica que los residuales son *I(0)*

Coefficients:

	Estimate	std.Error	t.value	Pr(>|t|)	
z.lag.1	-0,67101	0,15797	-4.248	0,000081	***
z.diff.lag	-0,08116	0,14029	-0,579	0,000232	

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

Residual standard error: **3,469 on 57 degrees of freedom**

Multiple R-squared: **0.3693,** Adjusted R-squared: **0,3472**

F-statistic: 16,69 **on 2 and 57 DF,** p-value: **-1,972e-06**

Value of test-statistic is: -4.2477

Critical values for test statistics:

	1pct	5pct	10pt
tau1	-2,6	-1,95	-1,61

Gráfica 9. Análisis de Residuales

A partir de la Gráfica 9, se confirma al calcular los residuales indican que son I(0) al nivel de significancia del 5%.

5.1.5. ETAPA 3 ECUACIÓN DEL ERROR

Coefficients:

	Estimate	std.Error	t.value	Pr(>|t|)	
(Intercept)	0,006151	0,024464	0,251	0,802418	
bdd1w1	-0,616922	0,156578	-3,940	0,000232	***
bdd1w2	0,126345	0,175472	0,720	0,474557	
error.ecm	0,023766	0,008599	2,764	0,007755	**

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

Residual standard error: **0,1875 on 55 degrees of freedom**

Multiple R-squared: **0.2542,** Adjusted R-squared: **0,2136**

F-statistic: 6,25 **on 3 and 55 DF,** p-value: **0,0009956**

Gráfica 10. Análisis de Residuales

$$\Delta\log(DF)_t = \Phi_0 + \Phi_1\Delta\log(DF)_{t-1} + \alpha\hat{e}_{t-1} + \theta_1\Delta\log(DA)_{t-1} + \sqrt{t}\ vt{\sim}(0,\sigma^2)$$

Ecuación 7. Ecuación modelo del error

Modelo ajustado

$$\widehat{\boldsymbol{\Phi}}_{\boldsymbol{0}} = \mathbf{0.0061}$$
$$\widehat{\boldsymbol{\Phi}}_{\boldsymbol{1}} = \mathbf{-0.6116}$$
$$\widehat{\boldsymbol{\theta}}_{\boldsymbol{1}} = \mathbf{0.1263}$$
$$\widehat{\boldsymbol{\alpha}} = \mathbf{0.0237}$$

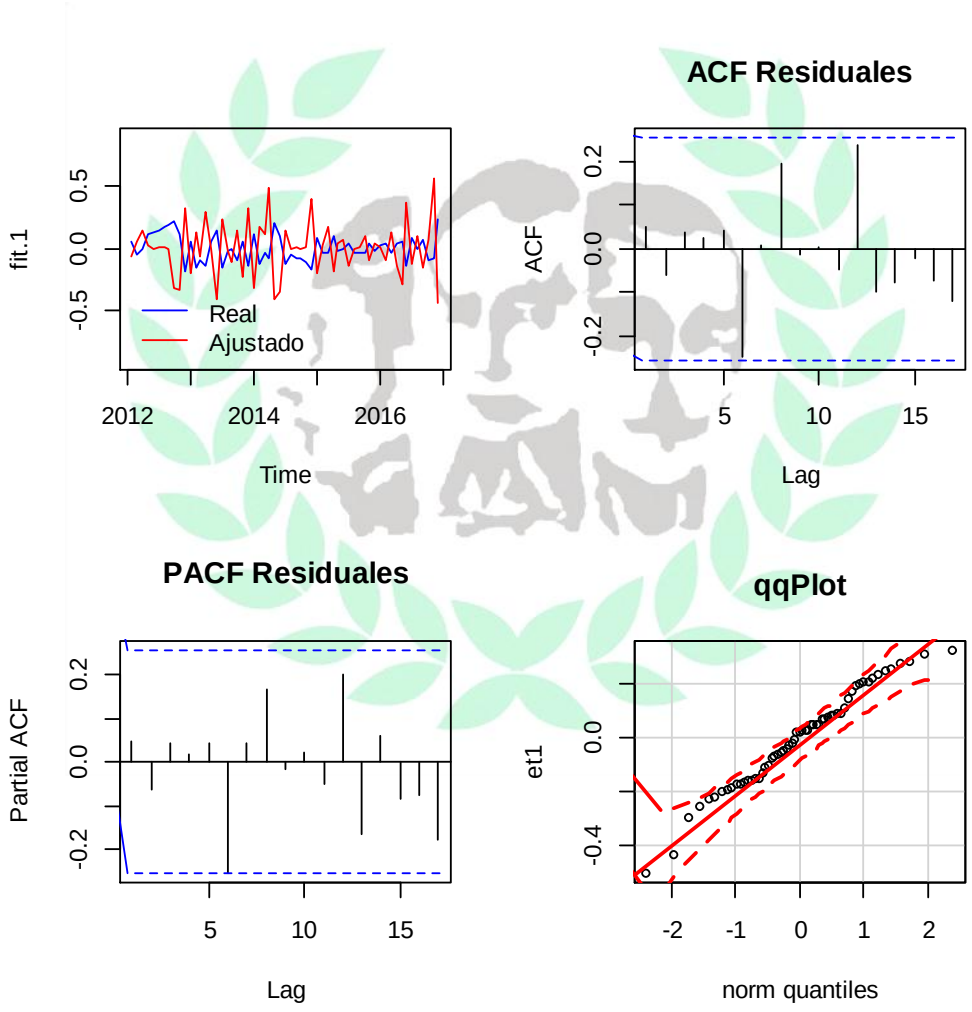
A partir de la ecuación 7 modelo del error, se define ante un incremento porcentual de la demanda de usuarios que tienen facturación manual de hace un mes se incrementa el 12.63% en anomalías de contador.

Ante incrementos porcentuales de anomalías de contador de hace un mes se disminuye la tasa de crecimiento en un -61.7%, se revisan las anomalías de contador del mes pasado encontrando que hay una caída

representativa en clientes con anomalías de contador, afectando la tasa de crecimiento.

Los desequilibrios que se presentan en el tiempo cuando la variable clientes con anomalías de contador no es capaz de pronosticar la variable clientes con facturación manual son corregidos al 2,37% en la ecuación de largo plazo.

5.1.6. ETAPA 4 RESIDUALES DEL MODELO A CORTO PLAZO.



Gráfica 11 Residuales

En la Grafica 11 residuales, se visualiza la siguiente información

- **ACF Residuales:** El grafico de la función de auto correlación simple está dada por $pk = \frac{yk}{y_0}(k = 1,2, \dots)$; $p_0 = 1$, en este caso de indica que no hay auto correlación en los componentes de los residuales.

- **PACF Residuales:** El grafico de la función de auto correlación simple parcial está dada por $\phi_{kk} = \frac{A_k}{B_k}(k = 1,2, \dots)$: $\phi_{11} = p_{11}$ n este caso de indica que no hay auto correlación parcial en los componentes de los residuales.

- **QQPLOT:** El resultado de la gráfica cumple el supuesto básico de normalidad teórica para los residuales del modelo, como se puede observar nos hay puntos fuera de la banda.

5.1.7. ETAPA 5 PRUEBAS DE LJUNG BOX Y JARQUE BERA.

- **Prueba de Ljung Box.**

Box-Ljung test

Data: et1

x-squared = 22.775, df = 24, p-value = 0.5331

Gráfica 12. Prueba de Ljung Box

A partir de la Gráfica 12 de la prueba de Ljung Box que está dada por $Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{p}^2{}_k}{m-k}$ se obtiene el siguiente análisis.

H_0 La correlación entre sus residuales es cero.
H_a Existe una correlación diferente de cero entre dos de sus residuales.

La prueba de Ljung-Box para los residuales es satisfactoria el p-valor es mayor de 0.05 con 0.5331% indicando que no hay auto correlación entre los componentes de los residuales.

- **Prueba de Jarque bera.**

Jarque Bera Test

Data: et1

x-squared = 1.0372, df = 2, p-value = 0.5953

Gráfica 13. Prueba de Jarque bera

A partir de la Gráfica 13 de la prueba de Jarque Bera que está dada por $JB = \frac{n-k+1}{6}(S^2 + \frac{1}{4}(c-3)^{2)}$ se obtiene el siguiente análisis.

H_0 Los residuales siguen una distribución normal.
H_a Los residuales NO siguen una distribución normal.

La prueba de Jarque Bera para los residuales es satisfactoria el p-valor es mayor de 0.05 con 0.5953% cumpliendo el supuesto de normalidad.

5.2. ANÁLISIS ÁRBOL DE DECISIÓN

5.2.1. ETAPA 1 NODO DE EXPLORACIÓN.

Se crea es un nodo de exploración de estadísticos básicos, solo se analizan las variables más relevantes que son:

- Consumo efectivo últimos 6 meses
- Consumo efectivo últimos 9 meses
- Consumo último mes

Análisis variable independiente consumo promedio efectivo últimos seis meses agrupada por la variable dependiente indicador proceso de facturación manual.

CON_FAC_PRE=NO

Medidas estadísticas básicas			
Ubicación		Variabilidad	
Media	20.57210	Desviación std	55.74234
Mediana	14.00000	Varianza	3107
Moda	0.00000	Rango	4586
		Rango intercuartil	18.00000

Límites de confianza básicos suponiendo normalidad			
Parámetro	Estimación	95% Límites de confianza	
Media	20.57210	20.28366	20.86053
Desviación std	55.74234	55.53913	55.94704
Varianza	3107	3085	3130

CON_FAC_PRE=SI

Medidas estadísticas básicas			
Ubicación		Variabilidad	
Media	71.64990	Desviación std	217.67974
Mediana	30.00000	Varianza	47384
Moda	0.00000	Rango	7065
		Rango intercuartil	42.00000

Límites de confianza básicos suponiendo normalidad			
Parámetro	Estimación	95% Límites de confianza	
Media	71.64990	67.49706	75.80274
Desviación std	217.67974	214.78282	220.65643
Varianza	47384	46132	48689

Gráfica 14. Estadísticos Básicos

Análisis variable independiente consumo promedio efectivo últimos nueve meses agrupada por la variable dependiente indicador proceso de facturación manual.

CON_FAC_PRE=NO

Medidas estadísticas básicas			
Ubicación		Variabilidad	
Media	20.61935	Desviación std	55.98097
Mediana	14.00000	Varianza	3134
Moda	0.00000	Rango	4818
		Rango intercuartil	18.00000

Límites de confianza básicos suponiendo normalidad			
Parámetro	Estimación	95% Límites de confianza	
Media	20.61935	20.32968	20.90902
Desviación std	55.98097	55.77690	56.18656
Varianza	3134	3111	3157

CON_FAC_PRE=SI

Medidas estadísticas básicas			
Ubicación		Variabilidad	
Media	70.56863	Desviación std	214.55541
Mediana	29.00000	Varianza	46034
Moda	0.00000	Rango	6900
		Rango intercuartil	41.00000

Límites de confianza básicos suponiendo normalidad			
Parámetro	Estimación	95% Límites de confianza	
Media	70.56863	66.47539	74.66187
Desviación std	214.55541	211.70007	217.48938
Varianza	46034	44817	47302

Gráfica 15. Estadísticos Básicos

Análisis variable independiente consumo último mes agrupada por la variable dependiente indicador proceso de facturación manual.

CON_FAC_PRE=NO

Medidas estadísticas básicas			
Ubicación		Variabilidad	
Media	18.17536	Desviación std	49.06183
Mediana	13.00000	Varianza	2407
Moda	0.00000	Rango	4905
		Rango intercuartil	17.00000

Límites de confianza básicos suponiendo normalidad			
Parámetro	Estimación	95% Límites de confianza	
Media	18.17536	17.92149	18.42923
Desviación std	49.06183	48.88298	49.24200
Varianza	2407	2390	2425

CON_FAC_PRE=SI

Medidas estadísticas básicas			
Ubicación		Variabilidad	
Media	64.02415	Desviación std	210.23233
Mediana	26.00000	Varianza	44198
Moda	0.00000	Rango	6737
		Rango intercuartil	41.00000

Límites de confianza básicos suponiendo normalidad			
Parámetro	Estimación	95% Límites de confianza	
Media	64.02415	60.01339	68.03492
Desviación std	210.23233	207.43453	213.10718
Varianza	44198	43029	45415

Gráfica 16. Estadísticos Básicos

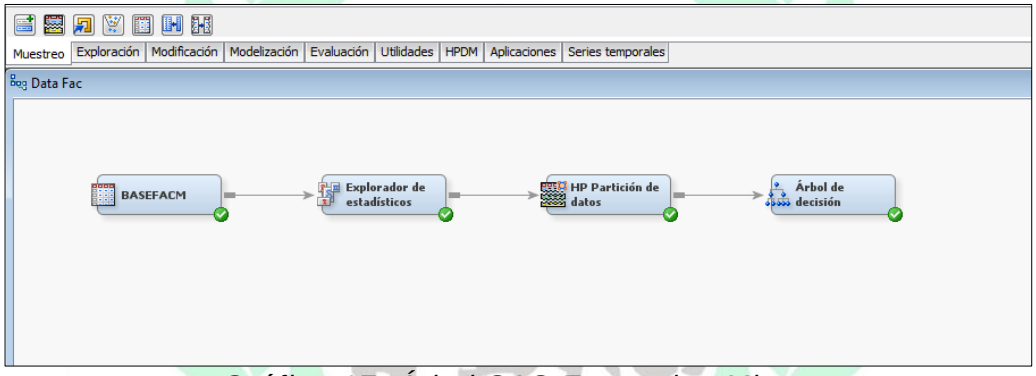
Análisis variable independiente mes ciclo RTR agrupada por la variable dependiente indicador proceso de facturación manual.

5.2.2. ETAPA 2 NODO DE PARTICIÓN

Se crea un nodo de partición de los datos, este lo que hace es dividir las observaciones en una base de entrenamiento y una base real, para el pronóstico.

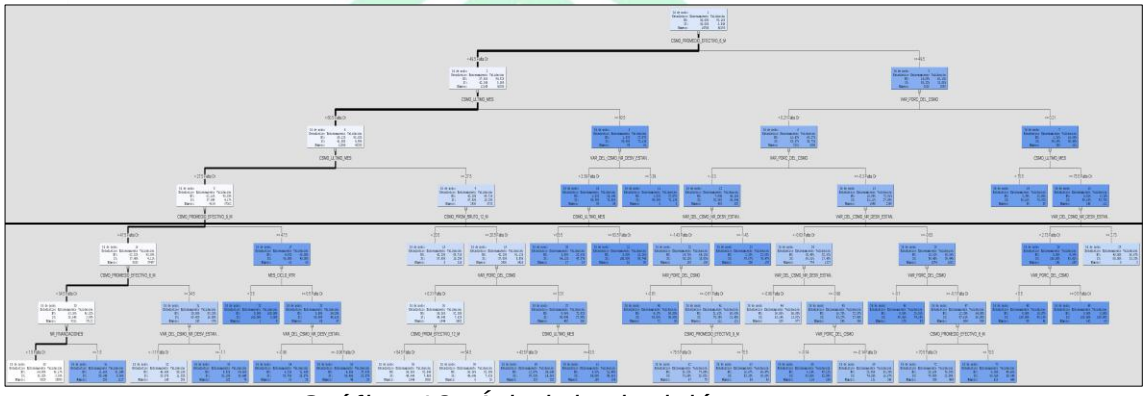
5.2.3. ETAPA 3 NODO DE ÁRBOL DE DECISIÓN

Se crea un nodo de árbol decisión. Este nodo tiene dos funcionalidades la primera es que podemos diseñar el árbol manualmente y la segunda es dejar que el sistema lo realice, en este caso se dejara que el sistema diseñe el árbol, cabe agregar que el SAS MINER postula el mejor modelo, sin embargo se revisa la lógica del modelo.



Gráfica 17. Árbol SAS Enterprise Miner

Modelo sugerido y estadisticos.



Gráfica 18. Árbol de decisión

A partir de la Gráfica 18 árbol de decisión se obtienen el siguiente resultado:

Nota:

Por extensión de las ramas del árbol se explicara solo la parte más relevante que esta delineada con negro, es decir se explicaran 3 ramas principales.

Gráfica 19. Rama Uno

Resultado Rama1	
Id de nodo Estadístico	**Probabilidad Evento**
Proceso Facturación (NO)	57,66%
Proceso Facturación (SI)	42,34%

Gráfica 20. Resultado Rama Uno

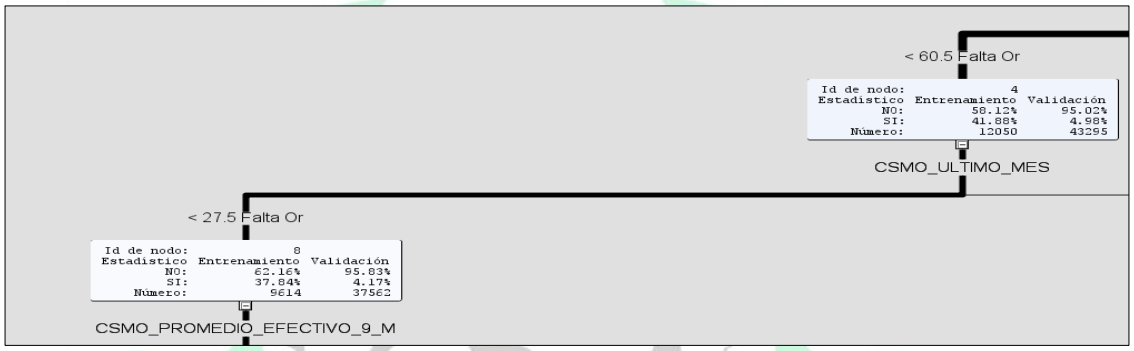
- La rama 1 define que los clientes que tienen un consumo promedio efectivo los últimos seis meses menor de 49.5 mts^3 tienen una probabilidad de tener facturación manual en un 42,34% y de no tener en un 57,66%.

Gráfica 21. Rama 1.1.

Resultado Rama1.1	
Id de nodo Estadístico	**Probabilidad Evento**
Promedio < 60.5 Mt3 (NO)	58,12%
Promedio < 60.5 Mt3 (SI)	41,88%

Gráfica 22. Rama 1.1.

- La rama 1.1 define que los clientes que tienen un consumo el último mes de menos de 60.5 mts^3 tienen probabilidad de tener consumo promedio efectivo los últimos seis meses menor de 49.5 mts^3 en un 41,88% y de no tener en un 58,12%.



Gráfica 23. Rama 1.2.

Resultado Rama1.2	
Id de nodo Estadístico	**Probabilidad Evento**
Promedio < 27.5 Mt3 (NO)	62.16%
Promedio < 27.5 Mt3 (SI)	37.84%

Gráfica 24. Rama 1.2.

- La rama 1.2 define que los clientes que tienen consumo promedio los últimos nueve meses menor de 27.5 mts^3 tienen probabilidad de tener un consumo el último mes de 60.5 mts^3 en un 62,16% y de no tener en un 37,84%.

5.2.4. ETAPA 4 RESULTADO DEL MODELO

Se tiene la posibilidad de tener clientes con facturación manual y clientes sin facturación manual, lo que esta nombrado en el modelo como (SI/NO), después de ejecutar el modelo de árbol de decisión se

analizan las salidas de diagnóstico, donde se ha sometido el modelo a una prueba de análisis cuyo resultado puede ser positivo o negativo, si la prueba es positiva el diagnostico de que el cliente puede tener facturación manual será correcto, esto significa que el modelo se ajustó, la misma situación se presenta si la prueba es negativa cuando el cliente tiene facturación manual.

El otro escenario del modelo espera que las probabilidades de error sean pequeñas, es decir el modelo indica que tiene facturación manual cuando no la tiene o indica que no tienen facturación manual cuando si la tiene.

	VERDADERO DIAGNOSTICO		
PRONOSTICO DEL MODELO	EVENTO FACTURACIÓN MANUAL **(SI)**	EVENTO FACTURACIÓN MANUAL **(NO)**	TOTAL
EVENTO FACTURACIÓN MANUAL **(SI)**	**6.845 (93%)**	**547(7%)**	**7.392**
EVENTO FACTURACIÓN MANUAL **(NO)**	**3.265(44%)**	**4.127(56%)**	**7.392**

Gráfica 25. Resultado del modelo

En la gráfica 25, a partir del resultado del modelo se determina:

Verdaderos positivos y verdaderos negativos.

- La probabilidad de que el modelo estime bien los verdaderos positivos es de un 93% (6.845/7.392), esto solo para los que sí tienen facturación manual y el modelo indica que tiene facturación manual.

- La probabilidad de que el modelo estime bien los verdaderos negativos es de un 56% (4.127/7.392), esto solo para los que no tienen facturación manual y el modelo indica que no tiene facturación manual.

Falsos positivos y falsos negativos

- La probabilidad de que el modelo estime mal los falsos positivos es de un 7% (547/7.392), esto solo para los que no tienen facturación manual y el modelo indica que si tienen.

- La probabilidad de que el modelo estime mal los falsos negativos es de un 44% (3.265/7.392), esto solo para los que tienen facturación manual y el modelo indica que no la tiene.

6. CONCLUSIONES

Se construyó un modelo de cointegración (ENGLE – GRANGER) de la serie de tiempo clientes con facturación manual de los últimos cinco años (enero-2012 – enero-2017), encontrando que a partir de la ecuación a largo plazo existe una relación significativa entre las variables mencionadas, se demuestra que siendo 1.82 la tasa de crecimiento mensual de clientes con facturación manual y por cada unidad porcentual de anomalías de contador se incrementa las facturaciones manuales en 48.35%.

Se concluye a partir de la ecuación de corto plazo, que ante un incremento porcentual de la demanda de usuarios que tienen facturación manual de hace un mes se incrementa el 12.63% en anomalías de contador.

Ante incrementos porcentuales de anomalías de contador de hace un mes se disminuye la tasa de crecimiento en un -61.7% de clientes con anomalía de contador, se revisan las anomalías de contador del mes pasado encontrando que hay una caída representativa afectando la tasa de crecimiento a corto plazo.

Los desequilibrios que se presentan en el tiempo cuando la variable clientes con anomalías de contador no es capaz de pronosticar la variable clientes con facturación manual son corregidos al 2,37% en la ecuación de largo plazo.

Se desarrolló un árbol de decisión con los clientes que han pasado por un proceso de facturación manual, de los últimos dos años (mayo-2015 – abril-2017). Se concluye que la variable consumo efectivo de los últimos seis meses, menor a 45mt3 tienen un 42% de probabilidad de tener facturación manual.

Se deduce a partir de las variables del árbol de decisión: consumo último mes y consumo promedio efectivo de los últimos nueve meses, que se mantiene un equilibrio en la probabilidad de los eventos, colocando estas como las más significativas del modelo.

Con base en el modelo del árbol de decisión se encuentra que la predicción para los verdaderos positivos tiene una probabilidad del 93% lo que indica un buen ajuste del modelo para este segmento, sin embargo la probabilidad de los falsos negativos es de un 44%.

7. RECOMENDACIONES

- Se recomienda a la empresa de gas natural de la ciudad de Bogotá generar campañas de normalización de contadores, ya que la relación que existe entre las variables facturación manual y anomalías de contador es altamente significativa a largo plazo, es decir si se logra bajar la cantidad de anomalías de contador se lograra disminuir la cantidad de clientes que requiere un análisis manual de facturación.

- Se sugiere aplicar el árbol de decisión diseñado en este trabajo para el total de los usuarios del servicio de gas natural, con este se podría detectar la cantidad de clientes que son altamente probables de tener facturación manual, una vez conociendo este nominal se deberá generar un plan de choque para evitar que estos lleguen tener un proceso de facturación manual y este replique en la operatividad de la empresa.

8. BIBLIOGRAFIA

BERAN, Jan. Fitting long-memory models by generalized linear regression, 1993. Biometrika 80(4), p. 817–822.

DICKEY, David y FULLER Wayne. Distribution of the estimators for Autoregressive Time Series With a Unit Root. Journal of the American Statistical Association, Vol. 74, Issue 366 (Jun., 1979), p. 427-431.

ARMANDO Aguirre Jaime. Introducción al tratamiento de series temporales: aplicación a las ciencias. Madrid: Diaz de Santos S.A. (1994).

AUTOMIND,(2017). [online] Available at: http://www.automind.cl/articulos/metodos_estadisticos/Metodos_estadisticos_con_arboles_de_decision.htm (referencia estudio 1) [Accessed 2 Jun. 2017].

DATAPRIX (2017). dataprix. [online] Available at: http://www.dataprix.com/blog-it/analisis-datos/analisis-predictivo-sas-arboles-decision (Referencia sas) [Accessed 2 Jun. 2017].