METODOLOGÍAS PARA EL PRONÓSTICO DE SERIES DE TIEMPO

ANDREA DEL PILAR GUAVITA CUTA
MARÍA EVANGELINA VERGEL ARESSI

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERÍA
MAESTRÍA EN ANÁLITICA DE DATOS PARA LA INTELIGENCIA DE NEGOCIOS
BOGOTÁ, D.C.
2018

CONTENIDO

RESUMEN

Los proyectos que incluyen series de tiempo son cada vez más comunes de encontrar en el entorno de la consultoría de analítica, los datos que se pueden obtener tienen cada día un mayor volumen y mayor detalle, lo que hace importante el estudio de estas y las oportunidades de mejorar tanto en tiempos de procesamiento como en la precisión del resultados. El presente proyecto busca comparar dos metodologías, la primera generando clústeres de series de tiempo y la segunda a través de un benchmark para diferentes algoritmos de pronóstico, con el fin de determinar cuál es la mejor alternativa de pronóstico en precisión y tiempo de procesamiento, para los datos entregados por Everis correspondientes a un cliente del sector Retail, con la necesidad de realizar pronósticos de las series desagregadas de las ventas por tienda para el mes de mayo; los resultados se medirán mediante el error absoluto medio (MAD), error porcentual absoluto medio (MAPE) y el error cuadrático medio (MSE). Se utilizará la metodología CRISP-DM como guía para el desarrollo de los objetivos de negocio y de minería de datos identificados.

Palabras clave: Analítica, series de tiempo, pronósticos, clusterización.

ABTRACT

Projects that include time series are increasingly common to find in the analytical consulting environment, the data that can be obtained have a greater volume and greater detail every day, what makes it important to study these and the opportunities to improve both in processing times and in the accuracy of the results. The present project seeks to compare two methodologies, the first generating clusters of time series and the second through a benchmark for different forecasting algorithms, in order to determine which is the best alternative forecast in precision and processing time, for the data delivered by Everis corresponding to a client of the Retail sector, with the need to make forecasts of the disaggregated series of sales per store for the month of May; the results will be measured by the mean absolute error (MAD), mean absolute percentage error (MAPE) and the mean square error (MSE). The CRISP-DM methodology will be used as a guide for the development of business and data mning objectives identified.

Key words: Analytics, time series, forecast, clustering

INTRODUCCIÓN

El presente trabajo es desarrollado con el propósito de aplicar y obtener conocimientos adicionales en la maestría de analítica para la inteligencia de negocios de la Pontificia Universidad Javeriana, dentro de su objetivo de formar consultores en análisis de datos para las organizaciones que apoyen la toma de decisiones estratégicas. El cual fue aplicado en la empresa de consultoría Everis.

El trabajo de grado se basó en una problemática actual de negocio de Everis, la cual se quiere solucionar mediante el análisis de datos y la aplicación de modelos de analítica. Para el proyecto fue asignada una base de datos de una empresa del sector retail la cual contiene las ventas diarias por tiendas y por categorías de productos, con el objetivo de realizar un pronóstico de las ventas futuras al horizonte de la data.

Para realizar una comparación con el modelo actual, se realizó la replicación de este, el cual es un modelo de suavización exponencial triple por tienda, con una posterior segregación por categoría de producto teniendo en cuenta la participación de cada una en la totalidad de la tienda.

La primera metodología propuesta consiste en realizar una clusterización de las series de tiempo, para lo cual se aplican tres métodos de clusterización denominados ARIMA, Correlación y Dinamic Time Warping; para el pronóstico se realiza una agregación de las series de tiempo que componen los clústeres resultantes y se utiliza un modelo de suavización exponencial triple, finalmente se realiza la segregación teniendo en cuenta la participación de cada una de las series en la totalidad del clúster al que pertenece.

La segunda metodología consiste en realizar un pronóstico con cinco algoritmos diferentes, iterando en cada serie a través de un foreach, para generar el benchmark y la agrupación final de las series por el menor error generado en los algoritmos utilizados.

Los resultados obtenidos por cada modelo, teniendo presente la medición del error cuadrático medio y Error porcentual absoluto medio, serán comparados junto con los tiempos de preparación y ejecución del modelo para obtener las conclusiones de forma clara y objetiva.

Finalmente, este documento se construye como muestra final de la realización del proyecto donde se cumple con el objetivo de satisfacer la necesidad expresada por la empresa y siguiendo sus directrices, así como también con los requerimientos enmarcados dentro de la formación académica.

METODOLOGÍA

Para poder desarrollar la problemática de negocio de la empresa Everis, utilizaremos la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) la cual es la guía más utilizada cuando se trata de desarrollo de proyectos de Data Mining.

Esta metodología se divide en 4 niveles de abstracción organizados jerárquicamente que van desde los más generales a las tareas más específicas, y se organiza en una serie de seis fases, las cuales se pueden observar en la ilustración 1.

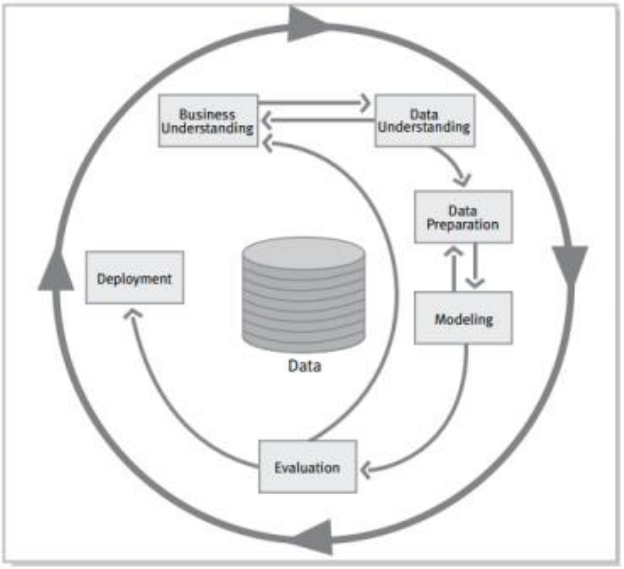


Ilustración 1 Fases de la metodología CRISP-DM. Fuente: CRISP-DM 1.0 SPSS

Fase 1 - Comprensión del negocio: En esta etapa se busca entender el negocio y su problemática entorno a una tarea de Data Mining.
Fase 2 - Compresión de los datos: Comprende la recolección inicial de los datos, con el objetivo de tener un primer acercamiento con el problema, familiarizándose con la información.
Fase 3 - Preparación de los datos: Preparación de los datos para las técnicas que se utilicen en el modelamiento.
Fase 4 - Modelado: Selección de técnicas más apropiadas para cumplir con los objetivos propuestos.
Fase 5 - Evaluación: Se evalúa el modelo teniendo en cuenta el cumplimiento de los criterios de éxito del problema.
Fase 6 – Implementación: Transformación del conocimiento del modelo evaluado en acciones dentro del proceso de negocio.

FASE 1: ENTENDIMIENTO DEL NEGOCIO

1. Definición de objetivos del negocio

1.1. Problemática y antecedentes

Es esencial para las empresas poder generar estrategias acertadas que permitan disminuir costos y generar más ingresos, puntualmente ser más rentables, por lo que el análisis de datos se ha vuelto fundamental en las organizaciones.

Según el libro pronósticos en los negocios de (Hanke, 2006), "Casi todas las organizaciones grandes y pequeñas, privadas y públicas utilizan los pronósticos de manera explícita o implícita, puesto que deben planear para satisfacer las condiciones del futuro sobre las que tienen un conocimiento imperfecto. Además, la necesidad de tener pronósticos está en todas las líneas funcionales, así como en todo tipo de organizaciones. Se requieren pronósticos en las áreas de finanzas, marketing, personal y producción."

Para cualquier empresa es importante poder predecir lo que puede suceder tanto internamente como en el mercado al que pertenecen; internamente es necesario poder predecir sus ventas, pronosticar sus costos para realizar un presupuesto, predecir la capacidad a usar para producir sus bienes y/o servicios, hasta predecir la rotación de su personal y poder realizar oportunamente los procesos de contratación para no tener vacantes; externamente, es frecuentemente usada para predecir la demanda y la oferta de bienes y servicios, enfocando a las empresas en entender como alcanzar el mercado objetivo y generar su estrategia alrededor de estos pronósticos.

1.1.1. Contexto del negocio – Sector Consultoría

Antes de entrar a hablar del negocio de Everis, queremos abordar el contexto del sector para entender en donde se está concentrando, cuáles son sus ventajas y desventajas, cuál es su proyección, entre otras. El sector dentro del cual se clasifica Everis es el de consultoría, denominado Servicios Profesionales, Científicos y Técnicos.

El ranking para este sector está liderado en buena parte por empresas encargadas de la gestión de los recursos humanos para otras empresas, solo tres empresas de soluciones de tecnología o informática se encuentran en el ranking dentro de las 10 primeras posiciones las cuales son: IBM DE COLOMBIA & CIA S.C.A., ORACLE COLOMBIA LIMITADA y UNUSUAL MINDS S.A.S. Además de empresas de especializadas en la gestión del recurso humano y tecnología quienes abarcan cerca de un 65% del sector, se encuentran empresas especializadas en marketing, gestión comercial, auditoria, ingeniería y otras enfocadas en sectores específicos como infraestructura, minería y energía, entre otros.

El sector ha presentado un crecimiento en ventas entre el 14 y 16%, excepto en el 2017 donde se presentó una disminución del 6%. El sector de consultoría no presenta una tendencia clara en cuanto al margen operacional, como se puede observar, en un año alcanzó hasta el 82% de crecimiento y en el siguiente año decreció un 40%, mientras que el margen neto si presentó un crecimiento constante excepto por el año 2017.

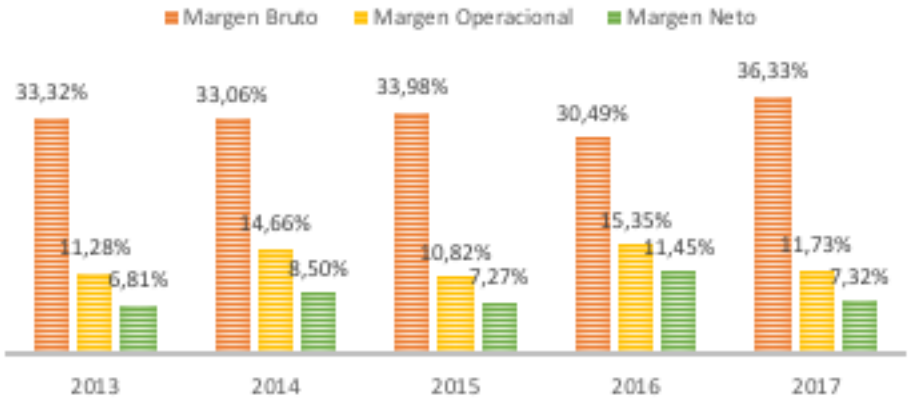


Ilustración 2 Indicadores de Rentabilidad Sector Servicios Profesionales, Científicos y Técnicos 2013 – 2017. Fuente: Emis

De los indicadores más importantes para el sector está el de rotación de cartera el cual ha ido aumentando paulatinamente logrando una estabilización sobre los 100 días, este comportamiento no es favorable para las organizaciones, mientras que la rotación de proveedores ha ido creciendo de forma más significativa logrando estar sobre los 120 días logrando que la diferencia en el tiempo entre las cuentas por pagar y por cobrar sea solo de 20 días.

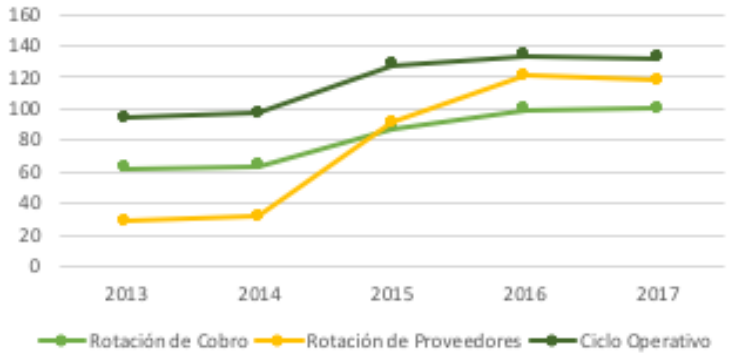


Ilustración 3 Indicadores de Eficiencia Sector Servicios Profesionales, Científicos y Técnicos 2013 – 2017. Fuente: Emis

En conclusión, el sector de consultoría es un sector bastante fluctuante, dadas sus condiciones y entorno hace que cada vez sea más competitivo y sean mayores los retos para poder innovar con servicios y/o productos.

Dentro del ranking por ingresos realizado por Emis, Everis se ubicó en el puesto 96 para el año 2017, luego de estar en 2013 en el puesto 117; para tener una visión más específica de Everis se realizó una subdivisión del sector, encontrando 318 empresas con un objeto social similar en el cual Everis ocupa el puesto 22, mostrando un liderazgo y crecimiento en el sector.

1.1.2. Contexto del negocio – Everis

Everis es una empresa española creada en el año 1996 bajo el nombre DMR Consulting, en el 2006 se convirtió en Everis; en el año 2013 NTT DATA la décima multinacional de servicios IT que presta servicios de consultoría de negocio, estrategia, desarrollo, mantenimiento de aplicaciones tecnológicas y outsourcing, con más de 80.000 profesionales en más de 40 países, adquiere el 100% de Everis, consolidándose aún más en el mercado como una empresa innovadora, enfocada en ser una parte esencial para empresas de múltiples sectores, con soluciones diferenciadoras.

Everis llegó a Colombia en el año 2007, operando sus servicios en diferentes sectores de la economía como el de banca, telecomunicaciones, seguros, energía, sector público e industria, entre sus clientes se destacan Grupo Aval, BBVA, Movistar, Claro, Allianz, AIG, Terpel, Ecopetrol, MinTic, SENA, Bavaria, Postobón entre otros; siendo un aliado estratégico en el desarrollo de cada empresa, generando productos y servicios acorde a sus necesidades.

Al igual que el sector en el que se encuentra, Everis ha tenido un comportamiento variable en cuanto a sus ingresos, aumentando y disminuyendo de un año a otro.

Ilustración 4 Ventas Everis BPO Colombia LTDA 2013-2018. Fuente: Emis

En cuanto a indicadores de rentabilidad se muestra un margen bruto del 36% en 2013 y 2014, con disminución para los años siguientes, y una diferencia de casi 23% con el margen operativo, que ha venido fluctuando entre el 2 y el 5%, es posible ver una tendencia estacionaria en el margen neto, dado que en los últimos años ha sido negativo manteniéndose en –3% o cerca desde el 2014.

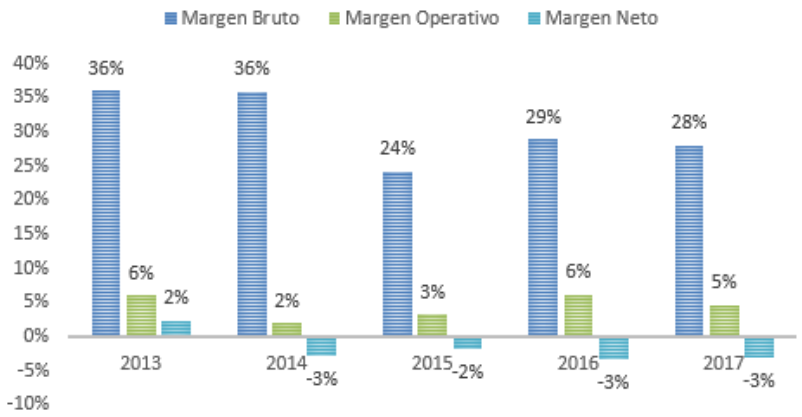


Ilustración 5 Indicadores de Rentabilidad Everis BPO Colombia LTDA 2013 - 2017. Fuente Emis

Everis Colombia se divide en las siguientes unidades de negocio:

- Outsourcing de procesos de negocio - Everis BPO (Business Process Outsourcing): Presta servicios de externalización de procesos de negocio bajo acuerdos de nivel de servicio.
- Outsourcing de sistemas y aplicaciones - Everis ITS&S (IT Solutions and Services): Especializada en la estrategia, asesoramiento, diseño, gestión y prestación de servicios de ingeniería de Software y gestión de infraestructuras.
- Outsourcing de sistemas y aplicaciones - Everis SES (SAP & Enterprise Solutions): ofrece y ejecuta servicios de tecnología bajo plataforma SAP, en las líneas de servicio que abarcan el análisis de información y reporting corporativo, portales y movilidad, arquitectura e integración de procesos, administración de sistemas y desarrollos para mejoras del estándar SAP.
- Tecnología avanzada - T&AS (Technology and Advanced Solutions): Está especializada en la gestión mediante el uso de innovaciones tecnológicas (Digital Architecture, Digital Intelligence, Digital Experience).

T&AS es una unidad de negocio nueva que surge de la necesidad de conocer, desarrollar y poder generar a través de las nuevas tecnologías productos innovadores y disruptivos que puedan ser consolidados, industrializados y comercializados a sus diferentes clientes. Esta unidad de negocio tiene dos enfoques principales:

- User Experience: El objetivo es mejorar la interacción de los usuarios con las aplicaciones mobile y web de sus clientes empresariales.
- Data y Analytics: Tiene dos enfoques centrales el primero es el aprovechamiento de nuevas tecnologías para la arquitectura de BigData, optimizando el almacenamiento, data online, gobierno de la información, tableros de control y monetización del dato; el segundo enfoque es el análisis y creación de modelos analíticos que generen información valiosa, que pueda ser usada en diversas estrategias corporativas y que faciliten la toma de decisiones para las empresas.

En Data y analytics, nace la problemática que desarrollaremos a lo largo de este documento; esta área es relativamente nueva dentro de la organización y tiene como objetivo poder

generar un valor agregado a los clientes actuales a través de consultoría en un tema que aunque no es tan nuevo para las empresas, ahora le están empezando a ver el valor; y es que en la sociedad actual donde casi mundialmente todos estamos conectados, se ve como la tecnología invade cada espacio y momento de nuestras vidas generando un gran volumen de datos, según un informe publicado por la empresa Seagate y la consultora IDC, para el 2025 se habrán creado más de 163ZB de datos en el mundo, un dato que será 10 veces superior al registrado en el 2016.

Es en este momento, en el cuál las empresas ven una necesidad clave en poder desarrollar capacidades para almacenar, procesar y más importante aún entender el valor que los datos pueden aportar en su cadena de valor, la gran pregunta aquí es ¿Por qué requieren contratar consultoras, en vez de desarrollar estas capacidades por sí mismos? En una artículo de la revista dinero en el que entrevistaron a varios directores de diferentes consultoras, muchos concluyeron que la respuesta a esta pregunta es el rápido cambio, Rodolfo Guzmán, socio de Arthur D. Little, Colombia respondió que "*Permite a las empresas prepararse mejor para anticipar las tendencias y cambios en sus industrias y mercados, innovar para desarrollar nuevos productos y servicios, y transformarse para competir y triunfar en un entorno de negocios cada vez más complejo y retador.*" Es claro que las empresas pueden realizar los cambios que requieran para acomodarse y seguir siendo competitivos en su entorno, pero definitivamente existe una gran posibilidad que no sea en la velocidad que su entorno lo requiere y en los negocios tiempo significa dinero, una decisión basada en el hoy posiblemente no les permitirá ser competitivos en el mañana.

Pero no sólo el volumen de datos es la razón por la cual las empresas requieren realizar data analytics, al contar con información de rápido acceso, variada, veraz y viable se necesita visualizar el hoy y poder predecir el mañana, sin estas dos últimas variables todo lo que las empresas hagan a nivel de información carecen de sentido, si no logran sacarle el provecho necesario, entender que dicen mis datos y que puedo hacer con ellos.

Cassio Pantaleoni, Country Manager de SAS para Colombia y Ecuador resalta que cada 4 de 10 empresas en Colombia están o han incorporado la analítica en sus procesos como una nueva oportunidad para crecer, mientras que IDC calculó que el crecimiento mundial de inversión de las empresas en analítica será del 25% en el 2016, estimando un valor de inversión en estas soluciones para el 2019 de $48 mil millones de dólares. Estas cifras nos ayudan a entender porque las empresas de desarrollo de software y las consultoras han creado estas áreas, para lograr una participación y posicionamiento en este mercado que tiene una clara tendencia de crecimiento.

Conociendo la necesidad que surge en las empresas por el volumen de los datos y la necesidad de entenderlos, contextualizando porque las empresas acuden a las consultoras y cuál es la estimación en ingresos para este negocio, nos hace falta preguntarnos ¿Qué empresas requieren consultoría en Data Analytics? Aunque la pregunta suena muy abierta y parecería una respuesta obvia, si es importante entender que, aunque todas las empresas de una u otra manera necesitan y realizan analítica, hay empresas que tienen esta necesidad

más latente y es en estas en donde es más importante enfocar los esfuerzos para las consultoras.

En el mismo artículo citado anteriormente también les realizaron esta pregunta a los mismos directivos, y en muchos la respuesta fue unánime y es a las empresas de productos y servicios masivos (B2C), dando unos ejemplos citados por los directores en el artículo, estas empresas son de los sectores como la banca, seguros, consumo masivo, retail, transporte, telecomunicaciones, petróleo, minería, utilities y materiales de construcción. De la misma manera que para las consultoras encuestadas, el segmento de clientes de Everis está en las empresas B2C, el 33% de los ingresos provienen de la banca, el 18% del sector público, el 16% en industria, el 15% en telcos, el 10% en utilities y el 8% en seguros.

Para Everis la estrategia corporativa es acompañar a sus clientes en crear propuestas de valor que les permitan perdurar en el tiempo y crecer al ritmo que el mercado y la tecnología lo hacen; es por esto que su enfoque de crecimiento es ser el aliado en cada proceso de la cadena de valor de su cliente, teniendo como foco sus ejes fundamentales los cuales son la consultoría, la tecnología y la operación, abarcando así todos los procesos de una empresa con visión *end2end,* con equipos de trabajo flexibles, multidisciplinares e híbridos (negocio & digital & tecnología).

Los problemas de negocio que llegan al área de analítica en Everis provienen de sus clientes actuales (los que ya tienen uno o más productos con la consultora) y esto es verdaderamente importante, en primer lugar porque se integra con su estrategia de ser un *partner* para su cliente y no un proveedor fácilmente reemplazable, y en segundo lugar porque genera un crecimiento natural con el cual podrán desarrollar, probar e "industrializar" metodologías de analítica que les permitan competir en el mercado actual con una propuesta de valor.

En data y Analytics el modelo de negocio es costo fijo por proyecto, en este modelo de negocio la clave es el tiempo, cada proyecto tiene una fecha de inicio, unas fechas de hitos y un entregable final; para Everis lograr terminar sus proyectos de analítica en un menor tiempo es clave en dos sentidos, el primero es interno ya que al disminuir el tiempo se logra liberar la capacidad de su "operación" para generar otros proyectos que generen más ingresos con el mismo costo, y el segundo es un valor agregado y diferencial que percibe el cliente a su servicio al recibir el entregable del proyecto en un menor tiempo.

De toda la información que una empresa puede crear, almacenar y procesar, se podría decir que en casi todas, la variable tiempo puede incluirse, si pensamos en los datos generados de ventas, costo de insumos, inventarios, cantidad de clientes, producción, operación, etc. cada una podría ser analizada a través del tiempo y es muy útil lograr entender como ha sido mi comportamiento y lograr predecir cómo será en el futuro desde diferentes perspectivas dentro del mismo negocio; de los proyectos que recibe el área de Data y Anaitycs de Everis, el 65% son problemas relacionados con la predicción y en su mayoría de series de tiempo, han recibido problemas de negocio relacionados con pronosticar el

consumo en clientes empresariales de gasolina, para una estrategia corporativa de tarjetas pre-pagadas de una empresa de hidrocarburos, o el problema de negocio que utilizaremos para desarrollar el proyecto, el cual requiere el pronóstico de categorías por tienda, para la toma de decisiones entorno al enfoque del presupuesto de un retail en publicidad y entorno al inventario.

Si nos enfocamos en el cliente de Everis sobre el que basaremos el proyecto, encontramos que existen unos KPI´S claves que se deben medir en los retail para asegurar sus procesos, al realizar el pronóstico de sus ventas por tienda y por categoría, ayuda a asegurar uno de los procesos fundamentales el cual es la gestión de ventas, los pronósticos se vuelven un insumo clave para las áreas de marketing para generar un funcionamiento eficaz al reducir inventario en categorías que no producen ventas, y potenciar a través de medios publicitarios las mejores categorías.

En estos dos ejemplos existe un factor común y es la gran cantidad de series de tiempo, si pensamos en que los clientes actuales de Everis son de las empresas líderes en Colombia esto se puede volver exponencial, por ejemplo, si realizamos el análisis del proyecto al retail más grande de Colombia podemos deducir lo siguiente: Éxito entre las tiendas del grupo Carulla, Surtimax, Super Inter y sus modelos exprés tenía a 2015 en Colombia 574 almacenes, suponiendo y sólo por ponerle un número, que en promedio por tienda se tengan 50 categorías de productos, esto se convierte en 28.700 series de tiempo que se deben entender, analizar y generar el modelo óptimo en el menor tiempo posible.

Este tipo de análisis para Everis requieren de un esfuerzo bastante grande, en primer lugar porque pueden existir muchas series de tiempo que sean intermitentes o volátiles que por el volumen de datos no sean fáciles de detectar, y que al final representan un desgaste en tiempo y esfuerzo, en segundo lugar porque ejecutar ese volumen de series se vuelve computacionalmente denso y más si se prueban varios modelos, en tercer lugar se vuelve una tarea casi imposible realizar la optimización del modelo escogido para cada serie con el fin de obtener el modelo más preciso y por último el cual es muy específico de problema del retail, es que si eligen pronosticar únicamente por tienda y extrapolar los porcentajes de participación por categoría dentro de cada una de estas, es posible que se pierda precisión.

Encontrar la forma de mejorar estas problemáticas, puede generar un conocimiento que no sólo puede ser aplicable al proyecto del retail, si no a los demás proyectos de pronósticos de series de tiempo que tengan con sus clientes actuales, y quizá cuando cuenten con una metodología estable y aplicable a diferentes sectores, puedan entrar a otro tipo de mercados y generar estrategias diferenciales en analítica.

1.2. Objetivo del negocio

Para la determinación de los objetivos del negocio asumimos algunos supuestos en base a información del mercado de consultoría como lo es el tiempo por proyectos, Kaggle realizó una encuesta en 2017 titulada: Kaggle ML and Data Science Survey, con el objetivo de establecer una visión integral del estado del data science y machine learning. De esta encuesta se obtuvo la distribución en promedio del tiempo dedicado a cada proyecto:

Actividad	Tiempo
Recopilación y limpieza de Datos	36,26%
Selección y construcción del modelo	21,25%
Producción del proyecto	11,01%
Visualización de los datos	13,90%
Encontrar información sobre los datos y comunicarlos	13,08%
Otros	2,30%

Igualmente se logró establecer que un 66% de las personas que trabajan con datos utiliza con mucha frecuencia o la mayoría de veces métodos de análisis de series de tiempo, por lo cual asumiremos este porcentaje para los proyectos de series de tiempo trabajados por Everis con respecto a la totalidad de los proyectos

1.2.1. Objetivo General

Aumentar la utilidad en por lo menos un 20% mediante una metodología que reduzca el tiempo de modelamiento en los proyectos de series de tiempo, permitiendo mayor capacidad para atender nuevos proyectos en la unidad T&AS de everis, en un año.

1.2.2. Objetivos Específicos

1. Disminuir a través de la metodología de análisis de series temporales el tiempo utilizado por cada proyecto en la fase de construcción del modelo en un 5%.
2. Aumentar el nivel de satisfacción de los clientes de Everis mediante una entrega anticipada de los proyectos en un 2% sobre el resultado del indicador actual.

1.3. Criterios de Evaluación de Objetivos de negocio

Entregar la metodología propuesta (manuales, marco teórico y códigos utilizados) en donde se especifique los modelos usados para el análisis de series de tiempo, el procedimiento y los resultados obtenidos, dentro de los cuales se debe especificar un tiempo promedio en la etapa de construcción del modelo, medido como porcentaje frente al tiempo total del proyecto y la precisión alcanzada en base a la data entregada por Everis del retail.

1.3.1. Inventario de recursos

Para el desarrollo del problema de negocio planteado anteriormente contamos con los siguientes recursos:

Bases de datos:

Para este proyecto se entregó una base de datos por Everis Colombia, la cual contiene la información histórica de ventas de los últimos 5 meses del 2018 y está compuesta por las siguientes variables:

IdLocal: Esta variable entrega la codificación de cada tienda
Local: Esta variable es el nombre de la tienda
FechaTransaccion: Esta variable es la fecha de venta
SubCategoria: Esta variable es la una categoría de agrupación de los productos
Categoria: Esta variable es una categoría más detallada de los productos
Venta Neta: Esta variable es la cantidad de venta neta de la fecha

Software:

Se utilizarán las siguientes herramientas para el desarrollo del proyecto:

- Excel
- Rapid miner
- Python
- R

Talento Humano:

Para la realización del proyecto se contará con el apoyo de las siguientes personas:

- Estudiantes de Maestría en Analítica de datos para la inteligencia de negocios: Andrea del Pilar Guavita Cuta y María Evangelina Vergel Aressi, estudiantes de último semestre de la maestría.
- Profesores asesores de trabajo de grado: Jairo Andrés Rendón Gamboa asesor en negocio y Stevenson Bolívar Atuesta asesor en modelos analíticos.
- Asesores empresa: Miguel Ángel Díaz Rodríguez Data Science, Jesús Manuel Ruiz Guzmán Líder Data & Analytics e Iván Herrero Bartolomé Head of Data & Analytics.

1.3.2. Requerimientos y restricciones

Requerimientos

- Base de datos: Se requiere una base de datos con las fechas de consumo de los diferentes productos por las diferentes tiendas, las fechas deben permitir generar variables dummies como día de la semana, quincena, festivos, etc.…

- Alcance del proyecto: El alcance del proyecto está enmarcado en las fases descritas anteriormente en la metodología CRISP-DM, dentro de este alcance no se tendrán en cuenta la última que corresponde a la implementación.

- Privacidad: La privacidad de este proyecto está respaldada por los acuerdos de confidencialidad entre la universidad y Everis, adicional por los acuerdos firmados por los estudiantes y la empresa; adicional como los datos son de una empresa cliente actual de Everis no se dará la información de nombres ni de la empresa, tiendas y productos y se entregarán los datos anonimizados.

- Presentación de informes: Acorde al plan académico entregado por la facultad para el desarrollo del trabajo de grado, se dividirá el trabajo en 3 entregas las cuales son las siguientes:
  - Primera entrega: En esta entrega se incluirá la primera y segunda fase de la metodología CRISP-DM y se realizará el 31 de agosto de 2018.
  - Segunda entrega: En esta entrega se incluirá la tercera y cuarta fase de la metodología CRISP-DM y se realizará el 12 de octubre.
  - Tercera entrega: En esta entrega se incluirá la quinta fase de la metodología CRISP-DM y un capítulo de manejo responsable de la información y se realizará el 23 de noviembre.

- Códigos: Se entregará a la empresa de ser requerido los códigos utilizados para realizar los modelos analíticos.

Restricciones

- Acceso a bases de datos complementarias: Como la información no proviene directamente de Everis, no podremos tener acceso a información adicional o más actualizada de la ya definida.

- Supuestos: Al no conocer cuál es la empresa dueña de los datos sobre la cual realizaremos el desarrollo del trabajo, muchos de los drives utilizados para medir los costos/beneficios del proyecto están sustentados en los indicadores financieros del sector.

- Supuestos: Al no conocer el tiempo y la cantidad de proyectos realizados por Everis en series de tiempo; el tiempo y costo utilizados para medir los costos/beneficios del proyecto son dados por cifras del sector y por una medición del tiempo utilizado en el proyecto por las estudiantes.

- Supuestos: Al no conocer cuáles son los productos reales, los supuestos de las razones de la frecuencia de los consumos son propuestos por las estudiantes al desconocer la información verdadera.

- Supuestos: Al no tener información de por lo menos 1 año atrás, no se podrán determinar estacionalidades, se realizarán más supuestos en el análisis, sin la certeza que en realidad son estacionalidades que se puedan verificar con datos de años anteriores.

- Tiempo: El periodo de tiempo de desarrollo del proyecto será aproximadamente de 4 meses, los compromisos de cada uno de los participantes son los siguientes:
  - Entre las dos estudiantes se dedicará un tiempo semanal de 8 horas
  - Por parte de la empresa se realizará una reunión semanal de 2 horas
  - Por parte de los tutores de la universidad se dispone de un espacio de 1 hora cada 15 días y dos posibles tutorías adicionales a solicitud

1.3.3. Riesgos y contingencias

Riesgos

- Acceso a los datos: El riesgo más grande que podemos tener es no contar con la información requerida para poder desarrollar el proyecto en el tiempo establecido.

- Datos de mala calidad: Contar con una información de mala calidad o no congruente que no permita realizar un modelo analítico eficiente.

- Resultados inconclusos: No lograr los objetivos propuestos referentes a la clasificación y predicción de ventas de productos y tiendas. Este riesgo se ve en dos vías, la primera es no lograr los indicadores esperados para la predicción y el segundo es no lograr realizar una clasificación por tienda y por producto.

Contingencias

- Acceso a los datos y datos de mala calidad: Para lograr solventar estos dos riesgos se realizarán seguimientos semanales con Everis para contar con los datos a tiempo y lograr validar con ellos la calidad de los datos para realizar la limpieza respectiva acorde al entendimiento de la información; si con las reuniones no se logra solventar el riesgo como segunda medida se aumentará la intensidad horaria para el desarrollo del proyecto, como última medida en caso que con las dos anteriores no se logré solventar el riesgo se trabajará con la base de datos entregada lo que afectará directamente al modelo y los resultados que se obtendrán.

- Resultados inconclusos: Para minimizar este riesgo tanto en el modelo no supervisado como en el modelo supervisado, se realizarán dos modelos por cada uno para tener dos opciones y minimizar la probabilidad de no lograr los objetivos.

1.3.4. Terminología técnica y de negocios – Glosario

Retail: es un término de la lengua inglesa que se emplea para nombrar a la venta minorista.

Aprendizaje no supervisado: es un método de aprendizaje donde no existe una variable función, si no que el modelo es ajustado a las observaciones en el proyecto lo utilizaremos para los modelos de clusterización.

Aprendizaje supervisado: es un método de aprendizaje donde se tienen variables de entrada y una variable de salida la cual puede ser un valor numérico o una etiqueta de clase, en el proyecto los utilizáremos para el pronóstico de series de tiempo.

Margen Bruto: es el beneficio directo que obtiene una empresa por un bien o servicio, es decir, la diferencia entre el precio de venta de un producto y su coste de producción.

Margen Operacional: es el beneficio que obtiene una empresa por un bien o servicio, es decir, la diferencia entre el precio de venta de un producto y su coste de producción más los costos indirectos antes de impuestos.

Margen Directo: es igual a las ventas netas menos el costo de ventas, menos los gastos operacionales, menos la provisión para impuesto de Renta, más otros ingresos menos otros gastos.

Clúster: es un término que se emplea para hacer referencia a una concentración de series de tiempo relacionadas entre sí por unas características específicas.

Pronóstico: es la predicción de lo que sucederá con las series de tiempo a analizar dentro del marco de un conjunto dado de condiciones.

Tendencia: es la componente de largo plazo que constituye la base del crecimiento o declinación de una serie histórica.

Variación cíclica: es un conjunto de fluctuaciones en forma de onda o ciclos, en una frecuencia de tiempo determinada.

Variación estacional: el componente de la serie de tiempo que representa la variabilidad en los datos debida a influencias de las estaciones, se llama componente estacional.

Variación irregular: este componente explica la variabilidad aleatoria de la serie, es impredecible, es decir, no se puede esperar predecir su impacto sobre la serie de tiempo.

Error medio absoluto (MAD): es una medida de dispersión del error del pronóstico en valor.

Error Medio Cuadrático (MSE): el MSE es una medida de dispersión del error de pronóstico, sin embargo, esta medida maximiza el error al elevar al cuadrado.

Error porcentual medio absoluto (MAPE): es una medida de dispersión del error en términos porcentuales y no en unidades como la anterior medida.

1.3.5. Costo-beneficio

Para determinar el costo-beneficio de realizar el proyecto nos basaremos en los siguientes supuestos:

1. Para determinar el costo directo de un proyecto de analítica para Everis, tomaremos como único driver el costo salarial de un científico de datos en Colombia, basándonos en los salarios de este tipo de cargos en EEUU según una encuesta realizada por KdNuggets y llevaremos este valor a salario a pesos colombianos.

	Colombia USD	EEUU USD	% Sobre EEUU
Salario mínimo 2018	$262,9	$1.256,70	21%

Ítems	EEUU	Colombia
Científico de datos Anual USD	$141.000	$29.497
Pesos Colombianos Anual	$423.000.000	$88.491.048
Científico de datos Mensual USD	$10.444	$2.185
Pesos Colombianos Mensual	$31.333.333	$6.554.892
Pesos Colombianos Diario	$1.692.000	$353.964

*Cifras tomadas de DatosMacro.com, calculadas con TRM $3.000 y 13,5 salarios mensuales

2. Según la encuesta realizada por Kaggle en 2017 titulada: Kaggle ML and Data Science Survey, con el objetivo de establecer una visión integral del estado del data science y machine learning, en la cual se da la siguiente distribución en las actividades dentro de un proyecto, en promedio, donde se relaciona dentro de la selección y construcción del modelo

Actividad	Tiempo
Recopilación y limpieza de Datos	36,26%
Selección y construcción del modelo	21,25%
Producción del proyecto	11,01%
Visualización de los datos	13,90%
Encontrar información sobre los datos y comunicarlos	13,08%
Otros	2,30%

3. Tomaremos para los proyectos en series de tiempo, en los cuales según la encuesta de kaggle el 66% de las personas que trabajan con datos utiliza con mucha

frecuencia o la mayoría de las veces métodos de análisis de series de tiempo, por lo cual tomaremos este valor como referencia para la participación de ingresos por este tipo de proyectos.

4. El 65% del costo el cuál es destinado en este tipo de modelos, y calcularemos el tiempo por proyecto en 3 meses hábiles según el promedio de tiempo que tardan por proyecto actual en Everis.

5. Como el objetivo no es disminuir en costo (personal) calcularemos el ingreso estimado de un proyecto con respecto al indicador financiero del Margen Directo de Everis, obtenidos a través de la empresa Emis, para el año 2017:

Margen	2017
Margen Directo	28,2%

Con estos supuestos, determinaremos si al disminuir en un 20% el tiempo por proyecto de series de tiempo al incluir en sus procesos la metodología implementada, esta liberación de capacidad cuánto ingreso representaría para Everis.

Por científico de datos	Sin Metodología	Con Metodología	Diferencia
Costo Anual ST	$55.218.414	$55.218.414	$0
# Proyectos ST Anuales	4	5	1
Días hábiles x Proyecto ST	39	31	8
Costo X proyecto ST	$13.804.603	$11.043.683	$2.760.921
Ingreso Por Proyecto ST	$19.226.467	$19.226.467	$0
Ingreso Anual	$76.905.869	$96.132.336	$19.226.467
Utilidad Directa	$21.687.455	$40.913.922	$19.226.467

Se disminuirá el costo en $2,7 millones al disminuir el tiempo en 8 días hábiles menos por proyecto, si se mantiene el mismo ingreso por proyecto de $19,2 millones, se generaría un ingreso de $40,9 millones en vez de $21,6 millones.

2. Definición de Objetivo de minería de datos

2.1. Objetivo de minería de datos

2.1.1. Objetivo General

Generar dos metodologías para el análisis de las series de tiempo que permita encontrar un balance entre los resultados obtenidos y el tiempo dedicado, identificando series de tiempo que son difíciles de pronosticar las cuales generan desgaste y ruido en los modelos, con el fin de facilitar la labor de análisis descriptivo de datos para Everis y que ayude a realizar pronósticos más precisos.

2.1.2. Objetivos específicos

1. Identificar las series de tiempo intermitentes o volátiles, las cuales no amerita generar un esfuerzo en modelamiento, y permita enfocar los esfuerzos en las series de tiempo que pueden generar valor.

2. Generar agrupaciones de series de tiempo que permitan reducir el tiempo de modelado, generando pronósticos en series agregadas para posterior pronóstico por serie.

3. Generar un modelo de pronóstico de serie una a una, que permita medir la perdida de información que se da con la agregación de las series de tiempo.

4. Medir el desempeño de cada modelo comparando el error generado y el tiempo utilizado, para poder dar alternativas en el pronóstico y análisis de tiempo.

2.2. Criterios de Evaluación del objetivo de minería de datos

1. Entregar la caracterización de las similitudes encontradas entre las series de tiempo especificando las series de tiempo que no generan valor al realizar modelamientos, para disminuir el tiempo en cada proyecto.

2. Entregar un Benchmark basado en la precisión de la metodología implementada con los diferentes modelos utilizados, y validar si en realidad logra generar modelos más precisos, respecto al utilizado por Everis.

3. Entregar el pronóstico diario del último mes de las ventas por categoría por tienda del modelo que generó los mejores resultados que mejore la precisión del modelo actual.

3. Definición del Plan del proyecto

Referente a los objetivos previamente determinados la siguiente matriz describe, los objetivos, roles, herramientas y responsables para la entrega del proyecto:

DEFINICION PLAN DE PROYECTO					
OBJETIVO DE NEGOCIO	Realizar un análisis descriptivo y predictivo de la información de 40 tiendas y 52 productos, con el fin de identificar patrones de consumo por tienda y producto que permitan tomar decisiones oportunas para mejorar las ventas de las tiendas y productos.		OBJETIVO DE ANALITICA	Desarrollar dos modelos analíticos que permitan tener un análisis descriptivo y predictivo del retail, de las tiendas y de las categorías de los productos, que genere valor al analizar el comportamiento de las series de tiempo y que permita controlar y generar alertas del comportamiento de las ventas de cada uno.	
Objetivos especificos de negocio	**Objetivos especificos de analitica**	**Responsables**	**Evaluación**	**Actividades**	**Herramientas**
Generar un análisis descriptivo por tienda y por producto que permita tomar decisiones para mejorar los ingresos.		Andrea Guavita Maria Vergel	Entrega de indicadiores de medición	Socialización del objetivo del proyecto	Emis, información bibliografica, informacion de contexto Everis
				Contextualización de la problemática	
				Contextualización de la minería de datos	
				Definición de objetivos de negocio	
				Entendimiento de los datos	Base de datos entregada Excel, Rapid Miner, R
				Preparación de los datos	
	Generar clusters de las tiendas y las categorías de productos acorde a su consumo de ventas que generen información de valor e insigths de negocio que puedan ser utilizados por Everis.	Andrea Guavita Maria Vergel	Entrega de Clusters	Desarrollo y evaluacion del modelo	
Detectar los productos y tiendas que tengan un comportamiento similar para generar estrategias por clúster y no uno a uno.	Generar un pronostico de las ventas en diferentes frecuencias (quincenal, mensual, semanal diaria) para todo el retail, por tienda y por categoría de producto, que permita controlar y generar alertas del comportamiento de cada uno, y entender cual pronostico mensual, quincenal, semanal o diario genera los mejores resultados.	Andrea Guavita Maria Vergel	Señal de rastreo tiendas Señal de rastreo categoria de productos		
Generar un análisis predictivo del consumo por tienda y por categoría de producto que permita tomar decisiones acertadas sobre el indicador de ventas reales Vs ventas pronosticadas.			CFE MAD MSE MAPE	Comparación de modelos predictivos	
				Selección de modelo predictivo	
				Resultados y conclusiones	Word

3.1. Cronograma del Plan de trabajo

El cronograma en tiempo es el siguiente:

Actividad	Responsable	Duración (horas)	Junio	Julio				Agosto					Septiembre				Octubre				Noviembre				
			S4	S1	S2	S3	S4	S1	S2	S3	S4	S5	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S5
Fase 0 Kick-off																									
Socialización del objetivo del proyecto	Everis	Reunión																							
Contextualización de la problemática	María / Andrea	16																							
Contextualización de la minería de datos	María / Andrea	16																							
Fase I Entendimiento del Negocio																									
Definición de objetivos de negocio	María / Andrea/Everis	8																							
Valoración de la situación	María / Andrea	8																							
Determinación de los objetivos de la minería de datos	María / Andrea	8																							
Plan del proyecto	María / Andrea	8																							
Fase II Entendimiento de Datos																									
Descriptiva de las series de tiempo	María / Andrea	16																							
Exploración de datos	María / Andrea	16																							
Verificación de la calidad de los datos	María / Andrea	8																							
Fase III Preparación de Datos																									
Limpieza de Datos	María / Andrea	16																							
Definición de las estructura de las series de tiempo	María / Andrea	16																							
Vista minable de las series de tiempo	María / Andrea	8																							
Fase IV Modelado																									
Selección de técnica de segmentación	María / Andrea	8																							
Prueba de modelos analiticos de predicción	María / Andrea	8																							
Generación del diseño de prueba	María / Andrea	8																							
Construcción de los modelos	María / Andrea	16																							
Evaluación de los modelos	María / Andrea	16																							
Fase V Evaluación																									
Evaluación de resultados	María / Andrea/Everis	24																							
Discusión de resultados	María / Andrea/Everis	16																							
Documentación	María / Andrea	22																							
Presentaciones																									
Presentación 1	María / Andrea	1																							
Presentación 2	María / Andrea	1																							
Presentación Final	María / Andrea	1																							

FASE 2: ENTENDIMIENTO DE LOS DATOS

1. Recolección inicial de datos

Se cuenta con una base de ventas de una empresa de retail que comprende los primeros 5 meses del año, contiene las siguientes variables: ID y nombre de la tienda, fecha de la transacción, los productos se identifican con una categoría y una subcategoría y por último el valor de la venta neta. Cada registro representa una única venta por subcategoría de producto en una tienda.

No conocemos el método de recolección de esta información ya que fue suministrada por Everis en excel, de forma anonimizada, los nombres de las tiendas no corresponden a los nombres reales y se desconoce la ubicación geográfica.

2. Descripción de los Datos

La base cuenta con 5 variables las cuales se describen a continuación:

Variable	Tipo	Descripción
IdLocal	Numérica	Toma valores de 1 a 40, de tipo identificador que se asigna a cada tienda. Cada tienda tiene un único ID.
Local	Categórica	Nombre de la tienda, anonimizada.
FechaTransaccion	Fecha	Corresponde al momento, día de la venta, va desde el 3 de enero de 2018 hasta el 30 de mayo de 2018
Categoria	Categórica	Clasifica los productos en grupos, determinada por la empresa de retail.
SubCategoria	Categórica	Clasifica a un grado de mayor detalle el producto, determinada por la empresa de retail.
VentaNeta	Numérica	Representa el valor de la venta en pesos colombianos, ajustados por un factor para asegurar la confidencialidad de los datos

La base original cuenta con 216.545 registros de ventas donde se da la información desagregada de las ventas, los productos están divididos en 22 categorías y 81 subcategorías como se muestra en el Anexo 1, las ventas corresponden a 40 tiendas, en el Anexo 2 se muestran los ID y nombres de las mismas.

3. Exploración de los Datos

En el histórico de las ventas se puede identificar un patrón, con picos ascendentes y descendentes que se mantiene casi de forma regular a lo largo de la serie, se presentan diferencias en esta normalidad desde la mitad del mes de marzo hasta comienzos de abril, con picos más altos, a mitades del mes de mayo se presenta otro pico importante. Las ventas presentan una distribución normal con unos datos atípicos en los costados inferiores.

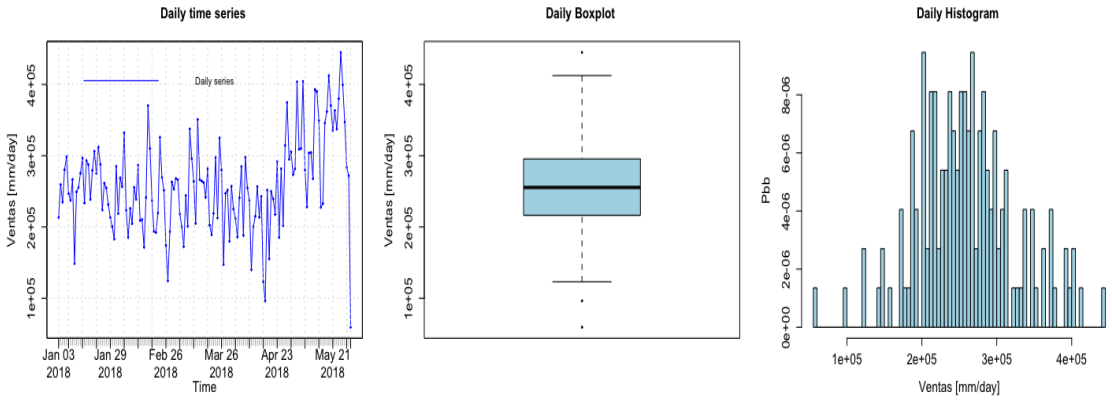


Ilustración 6 Serie de Tiempo Venta Neta Total Empresa Retail Enero - Mayo 2018. Fuente: Autores

Las ventas netas mensuales se muestran en la ilustración 12, en la que se puede evidenciar que los meses con mayores ventas fueron Enero y Marzo, quienes tienen la mayor cantidad de días, mientras que febrero que solo tiene 28 días muestra el valor de ventas más bajo.

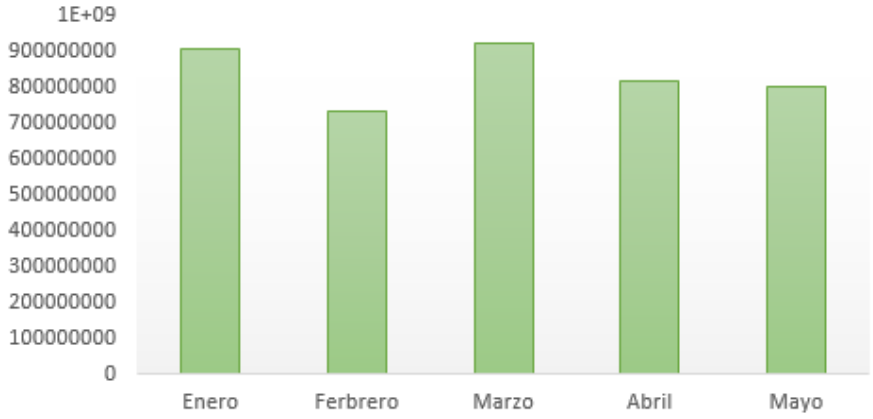


Ilustración 7 Venta Neta por Mes Empresa Retail Enero - Mayo 2018. Fuente: Autores

Analizando la serie por día, se puede observar que el día 1, 30 y 31 son los de ventas más bajas, el 31 está muy por debajo en el valor de las ventas, esto debido a que este día solo está en los meses de enero y marzo, del día 3 al 6 se presentan las ventas más altas, el día 16 también tiene un pico alto, se presentan bajas ventas en los días 17 y 24 con una reactivación a los dos días siguientes la cual se mantiene por tres días.

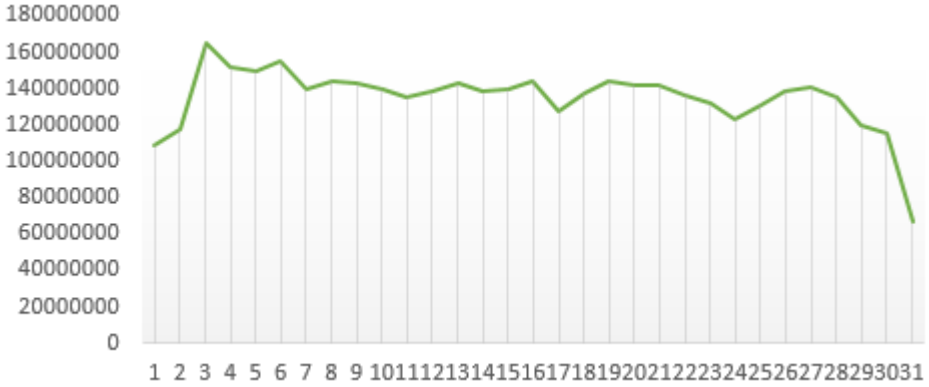


Ilustración 8 Venta Neta por Día Empresa Retail Enero- Mayo 2018. Fuente: Autores

3.1. Análisis por tiendas

Las tiendas en las cuales se perciben mayores ventas son El País de Nunca Jamás, Basin City y Atlántida con ventas de $328.136.279, $315.836.232 y $267.703.761 seguidas por Macondo, Metrópolis y Tierra Media con valores cercanos a $200.000.000. Los valores de ventas más bajos los tiene las tiendas Isla Glubbdubdrib, Arcadia, Zenda, Camelot, Narnia e Icaria con valores por debajo de los 40.000.000.

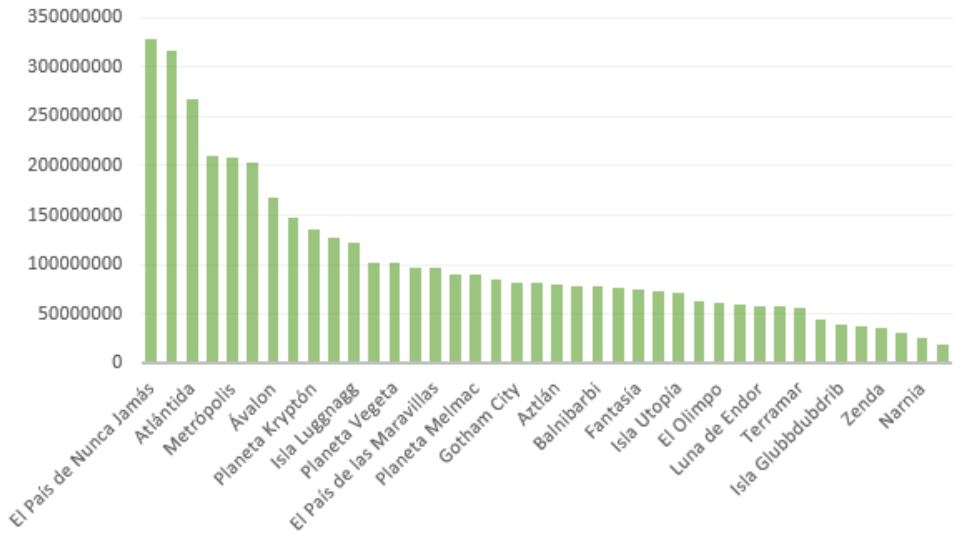


Ilustración 9 Venta Neta por Tienda Empresa Retail Enero - Mayo 2018. Fuente: Autores

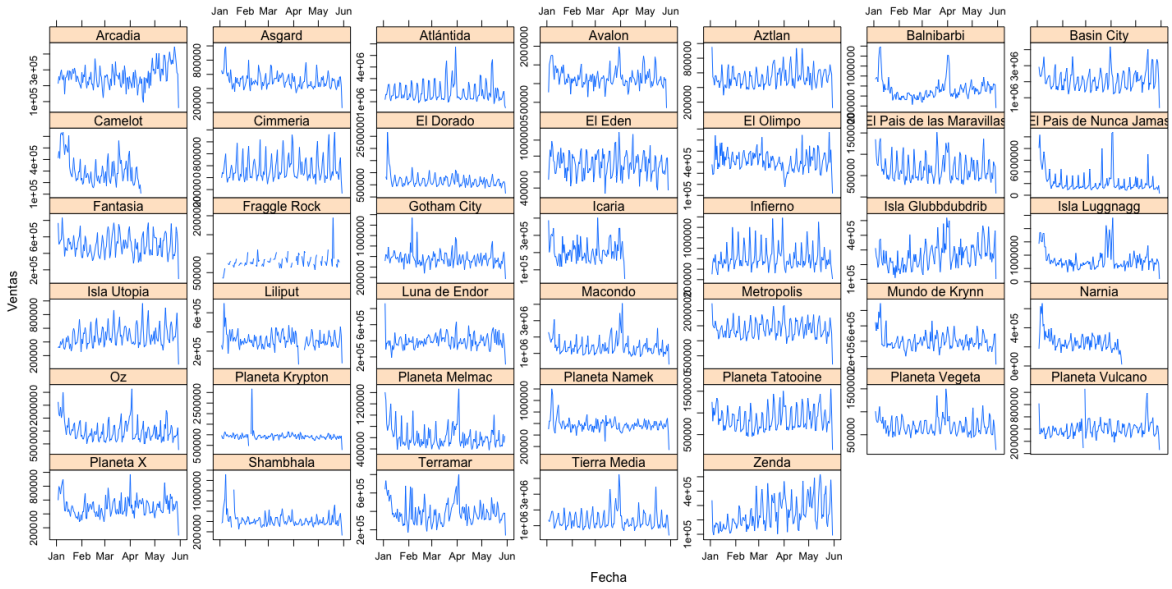


Ilustración 10 Serie de Tiempo por Tienda Empresa Retail Enero - Mayo 2018. Fuente: Autores

Si analizamos las series de tiempo por tienda, se puede observar que a simple vista se podría realizar un agrupamiento por su comportamiento, adicional se observa que tiendas como Icaria y Narnia no tienen la misma cantidad de datos u otras como Fraggie Rock que tienen

intermitencia en la fecha de ventas puede ser por falta de información o que tiene horarios diferentes a las otras tiendas y hay fechas en las que no abre.

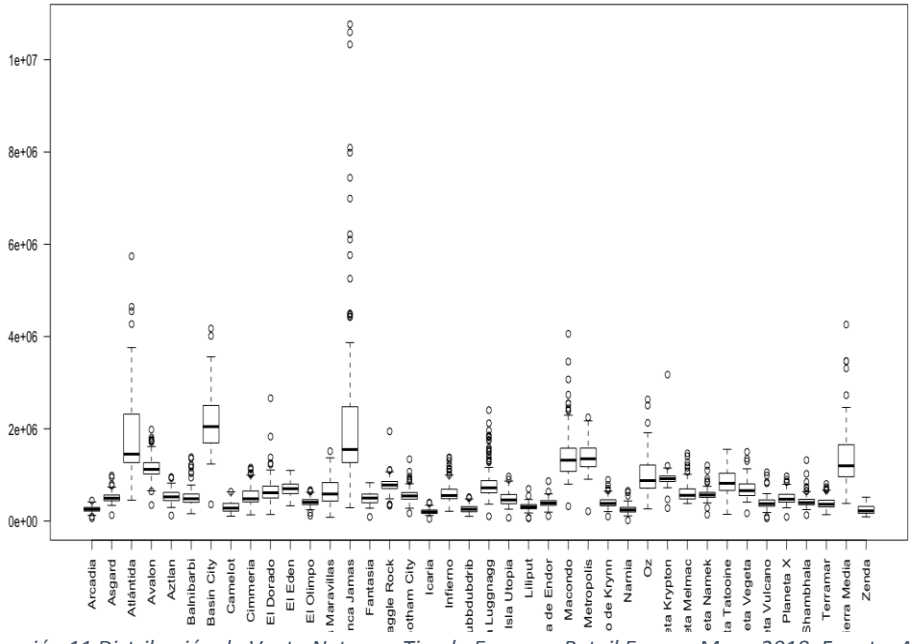


Ilustración 11 Distribución de Venta Neta por Tienda Empresa Retail Enero - Mayo 2018. Fuente: Autores

Referente a la distribución de los valores por tienda, se puede observar que la gran mayoría presenta una distribución similar, donde la mediana es muy cercana en todas las tiendas, y los rangos intercuartil no superan $875.212, hay unas tiendas con un comportamiento atípico, por ejemplo, el país de nunca jamás la cual tiene muchos datos outliers y tiene el volumen de ventas más alto en un día.

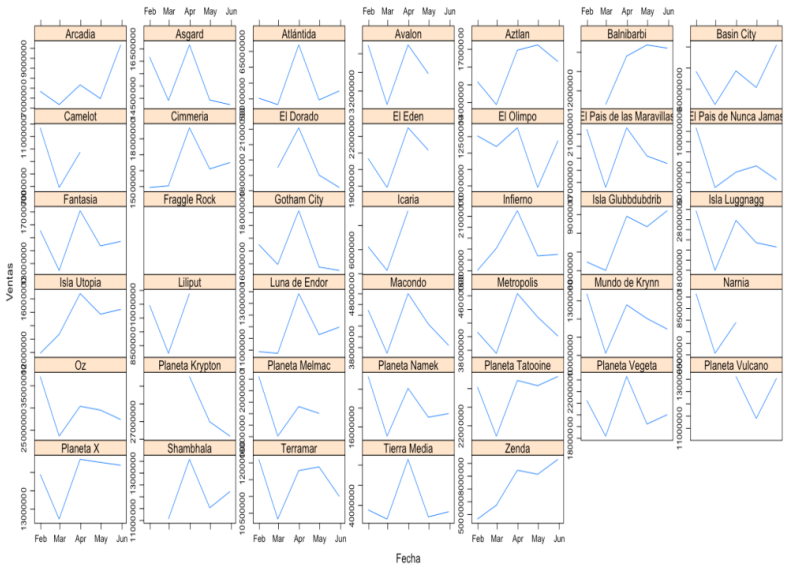


Ilustración 12 Venta Neta Mensual por Tienda Empresa Retail Enero - Mayo 2018. Fuente: Autores

Si vemos la serie de tiempo resumida por mes, se puede observar con más detalle las similitudes de ciertas series de tiempo por el comportamiento de ventas mensual, y se ve nuevamente un error en la serie de tiempo de la tienda Fraggie Rock, probablemente por la cantidad de datos faltantes que en cada mes no logra graficarla.

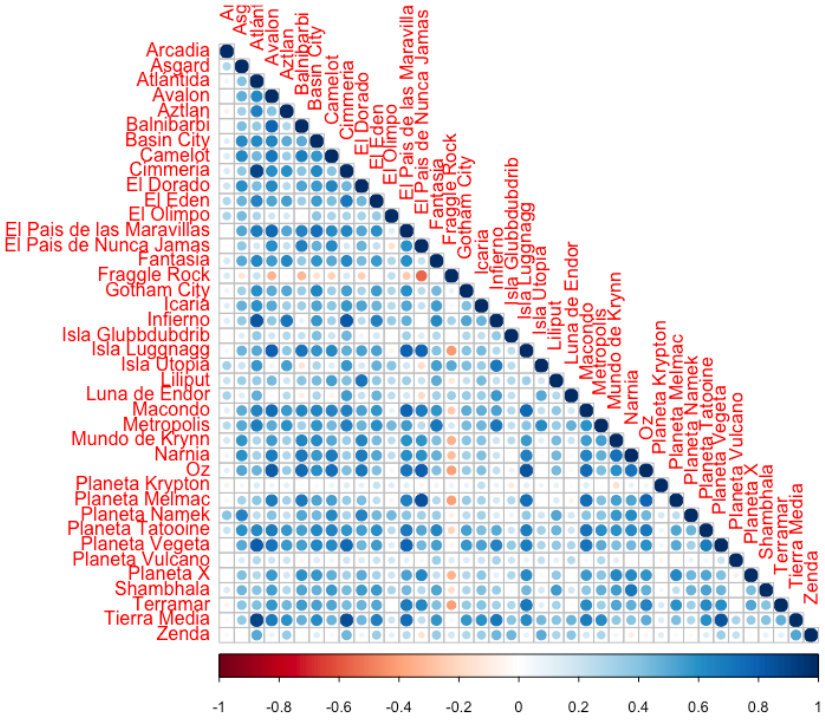


Ilustración 13 Correlación de Venta Neta entre Tiendas Empresa Retail Enero - Mayo 2018. Fuente: Autores

Al realizar el test de Pearson la correlación estimada en total de todas las tiendas es de 0,17 la cual no es muy alta, pero si realizamos el análisis de correlaciones entre las ventas de las tiendas, podemos encontrar que existen varias correlaciones negativas y positivas entre las tiendas, las negativas podría indicar sí las tiendas están cerca geográficamente mientras en una tienda se incrementan las ventas en la otra decaen, mientras que positivo significa una relación lineal entre estas tiendas.

La relación positiva es más fuerte que la negativa, en realidad, las negativas se ubican en la tienda Fraggle Rock por la cantidad de missing values, las tiendas con correlación positiva más alta de 0.909 son Cimmeria - Avalon y Zenda - Atlántida, seguida por Atlántida – El dorado con 0,907.

3.2. Análisis por categoría de producto

La categoría con el valor más alto es Comidas Preparadas con un valor de $1.006.711.256, representado el 24% del total de las ventas netas, seguido por otras categorías de comidas y bebidas como bebidas No Alcohólicas, Comidas Rápidas y Bebidas Calientes, la categoría tabaco ocupa el cuarto lugar con $252.153.657, la categoría con menor valor de ventas es

la de recargas que tiene un 60% menos que la categoría de insumos, penúltima en la gráfica. Se puede evidenciar que en estos meses el impuesto por bolsa está representando un 0,05% del total de las ventas netas, antes del 1 de julio las bolsas se daban al final de las compras de forma gratuita, pero desde esa fecha los establecimientos la cobran con el ánimo de disminuir el uso de las bolsas plásticas en pro del medio ambiente, objetivo que dados los datos analizados no se cumple, ya que se tiene un incremento desde el mes de enero a marzo y una disminución en los meses de abril y mayo.

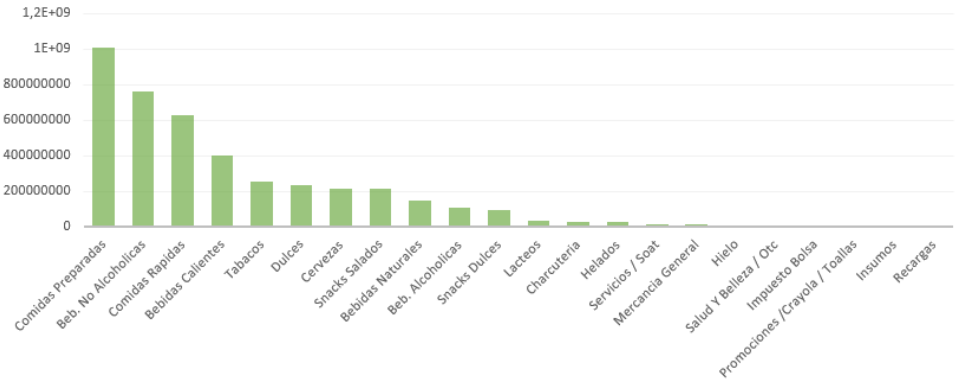


Ilustración 14 Venta Neta por Categoría Empresa Retail Enero - Mayo 2018. Fuente: Autores

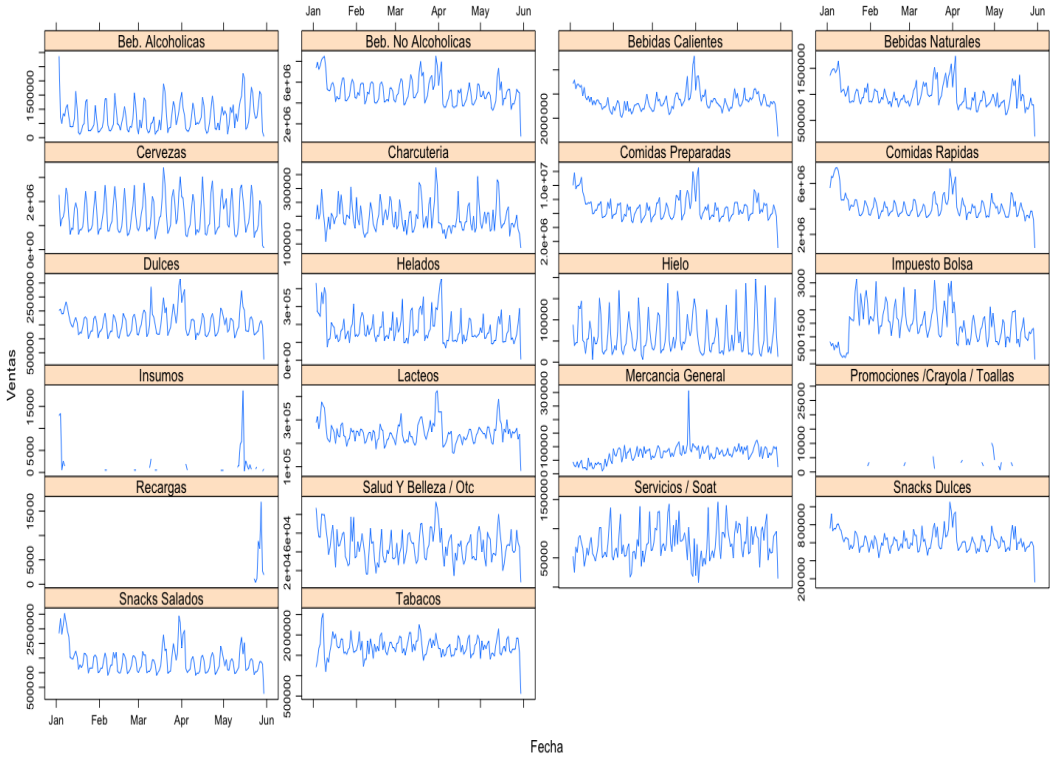


Ilustración 15 Serie de Tiempo por Categoría de Producto Empresa Retail Enero - Mayo 2018. Fuente: Autores

Si analizamos las series de tiempo por categoría, podemos observar que al igual que en las tiendas, a simple vista se pueden agrupar unas categorías por su consumo, por ejemplo, las

de bebidas calientes con comidas preparadas o el hielo con las cervezas; también podemos ver categorías de las que carecemos de información como recargas, Promociones e insumos en donde se debería evaluar no tomarlas en cuenta para el proyecto.

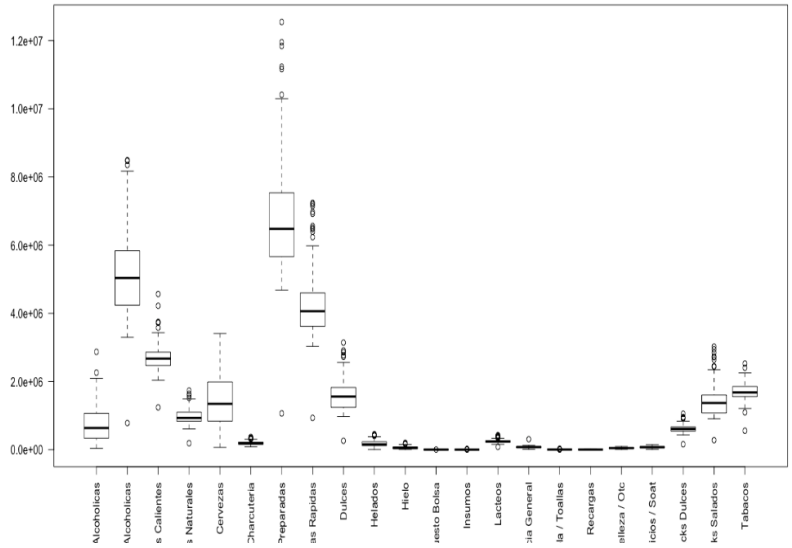


Ilustración 16 Distribución de Venta Neta por Categoría de Producto Empresa Retail Enero - Mayo 2018. Fuente: Autores

En cuanto a la distribución al haber categorías de tan alto consumo se dificulta a través de esta grafica visualizar la distribución de las categorías con menores ventas, en cuanto a datos atípicos las categorías que más las presentan son las comidas rápidas y los snacks salados.

Si analizamos las series de manera mensual, encontramos consumo de patrones mensuales mucho más parecidos entre las categorías de los productos.

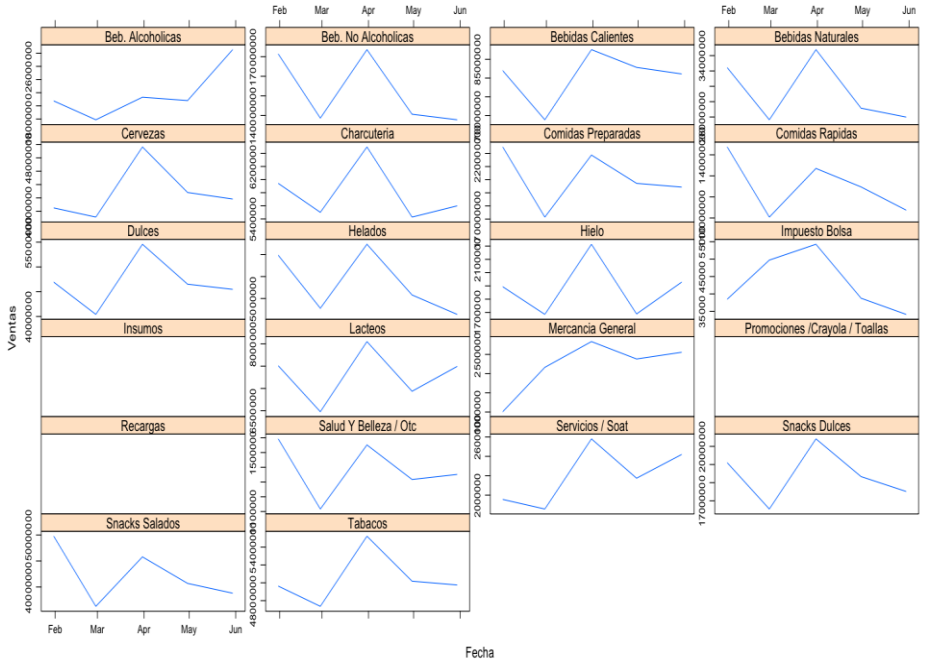


Ilustración 17 Venta Neta Mensual por Categoría de Producto Empresa Retail Enero - Mayo 2018. Fuente: Autores

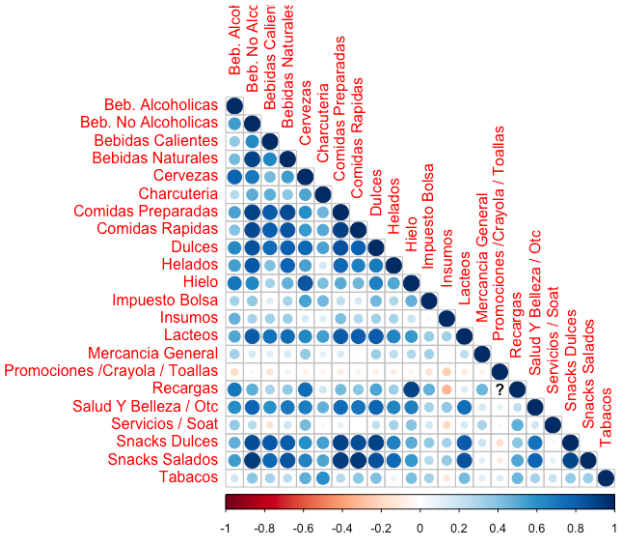


Ilustración 18 Correlación de Venta Neta entre Categoría de Producto Empresa Retail Enero - Mayo 2018. Fuente: Autores

Al realizar el test de Pearson la correlación estimada en total de todas las categorías es de 0,55 que si bien no demuestra una alta relación si es mucho más alta que la que encontramos por tiendas, lo que significa que los patrones de consumo son más relacionados por categoría que por tienda, pero si realizamos el análisis de correlaciones entre las ventas de las categorías, podemos encontrar que existen varias correlaciones positivas y muy pocas negativas entre las categorías.

Las categorías con correlación positiva más altas son, en primer lugar con 0.94 Snack Salados – Dulces, en segundo lugar Snacks Salados – Comidas rápidas y Snacks Salados – Bebidas calientes con 0,93; como se puede observar la que mayor correlaciones tiene con las otras categorías, son los Snacks Salados y Snacks Dulces lo que ya puede dar indicio de realizar estrategias de mercadeo de ventar cruzada en esta categoría, ya que se incrementa con el consumo de cualquiera de las otras, se podría decir que existe una alta probabilidad que estas se consuman en el mismo momento.

Por el consumo de las categorías también podemos inferir que no pertenecen a un hipermercado ya que no tienen productos de canasta familiar, verduras, etc., sino a un modelo de negocio “express” en los que están entrando los hipermercados a competir con tiendas de barrio, en las que se ofrece solo un pequeño número de productos de forma especializada a una escala más reducida y en un espacio pequeño.

4. Calidad de datos

4.1. Coincidencia del significado y valores contenidos

Se encuentra un registro con categoría – y subcategoría -, la venta registra un valor negativo de 97,75, en la tienda Atlántida el día 15/01/2018.

Se registran valores negativos para la variable venta neta en las subcategorías galletas y postres, los cuales podrían darse por el ajuste realizado a la base, pero para el análisis a realizar puede ser contraproducente tener estos valores negativos ya que una venta no se da negativa al no ser por devoluciones.

4.2. Redundancias y correlaciones

Dado que cada tienda un único ID y el nombre de la tienda no nos proporciona información relevante para el problema, es redundante tener dentro de la base el nombre de las tiendas.

4.3. Consistencia

Para los días 1 y 2 de enero no se presenta información de Venta Neta por lo cual se asume que en estos días no se realizó apertura de las tiendas.

Para el día 30 de mayo se presenta una venta neta un 80% por debajo del promedio de las ventas de los días del resto de la serie de tiempo. No se conocen los motivos de este atípico.

El 32.5% de las tiendas no tienen registros del total de días de la serie de tiempo, 9 de las tiendas tienen entre 1 y 6 días sin registro de venta, mientras que las tiendas Camelot, Fraggle Rock, Icaria y Narnia solo cuentan con información de un 69%, 67%, 64% y 69%, respectivamente, del total de registros.

Se desconoce si los productos tienen el mismo precio y si se registran dentro de la misma categoría en todas las tiendas.

FASE 3: PREPARACIÓN DE LOS DATOS

1. Selección de los Datos

Dadas las condiciones de la base, las cuales se expusieron en la fase anterior, se realizarán algunas exclusiones de datos para el cumplimiento del objetivo propuesto, a continuación, se relacionan:

Eliminación del día 30 de mayo

Como el día 30 de mayo presenta una atipicidad que expuesta ante el cliente no pareciese tener sentido se eliminara dentro de la predicción, este día.

Locales

Las tiendas Narnia, El Dorado, Icaria y Fraggle Rock, solo cuentan con el 69%, 69%, 64% y 67% correspondiente, de los datos en la serie, estas tiendas no serán tenidas en cuenta y por lo tanto no se realizará pronóstico sobre las mismas.

Categorías

Dentro de las categorías se encuentran series de tiempo intermitentes, como las de los productos: -, Insumos, Promociones / Crayola / Toallas y Recargas, que presentan por encima del 70% registros faltantes, por lo que no serán tenidas en cuenta y por lo tanto no se realizará pronóstico sobre las mismas.

2. Limpieza de Datos

La base de datos se encuentra en condiciones adecuadas para el análisis y modelamiento a realizar, solo es necesario las siguientes acciones:

Tipificación de series

Se le asigna un id a cada series tienda_categoria, que se compone de la palabra serie y un número que permite su identificación, para posteriormente realizar el análisis teniendo presente el nombre de tienda y categoría de producto. En total se identificaron 732 series.

Eliminación Registros negativos

Dentro de la base de datos se encuentran 35 registros, los cuales corresponden a comidas rápidas y comidas preparadas, en los cuales la venta neta tiene un valor negativo, dado que no es coherente con la naturaleza de la variable, estos registros no serán tenidos en cuenta.

Missing Values

Para los registros que se encuentran en blanco en las series de tiempo de las tiendas por categoría, se realiza una imputación con el promedio de la serie.

3. Construcción de los Datos

Para los modelos que se realizarán se utilizarán dos vistas minables:

1. Series de tiempo por tienda, excluyendo las anteriormente nombradas
2. Series de tiempo por categoría de producto por tienda, exceptuando las series de tiempo con un volumen de missing values superior a 15%, lo que representa más de 22 días.

FASE 4: MODELAMIENTO

1. Selección de la técnica de modelado

Para poder llevar a cabo el objetivo propuesto en el presente proyecto realizaremos tres modelos diferentes con el fin de realizar la comparación en el tiempo de modelado y precisión del modelo:

Detección de series intermitentes:

Las series intermitentes son muy frecuentes en el estudio de demanda, en este caso particular el volumen de ventas de algunos productos, como el hielo, bebidas alcohólicas u

otros tienen demandas intermitentes que no permiten realizar el pronóstico, el primer paso antes de ejecutar los modelos será la detección de estas series con el fin de darles un tratamiento diferente, y no incluirlas dentro de la vista minable.

Modelo HoltWinters desagregado:

Para poder tener una comparación en cuanto a tiempo y desempeño del modelado vamos a realizar un modelo similar al utilizado por Everis, el modelo es el de suavización exponencial triple aditivo, realizado por tienda, con el porcentaje de valores de participación por categoría por tienda se desagregaron los pronósticos, fue elegido por su ventaja de fácil adaptabilidad frente al ingreso de nueva información, ya que el cliente tiene la facilidad de generar y poder alimentar constantemente el modelo con grandes volúmenes de información, sin embargo la desagregación implica un desgaste operativo que puede conducir a errores.

Modelo Clusterización – pronóstico HoltWinters:

La agrupación de series de tiempo y otras secuencias de datos ha cobrado un sentido importante con el crecimiento de desafíos de investigación que nacen a partir de la búsqueda de optimización de tiempos y mejora de resultados en el análisis de datos.

Cuando se tiene un volumen importante de series de tiempo, el agrupamiento de estas aporta un conocimiento a nivel exploratorio que no sería fácil o alcanzable sin la ayuda de modelos no supervisados como lo es la clusterización. Por medio de los conglomerados se pueden observar patrones entre las diferentes observaciones (series de tiempo) que conducen a interesantes aportes a nivel descriptivo y facilitan la tarea de análisis para el proceso de modelamiento.

Algunas de las aplicaciones que se han encontrado en este tipo de metodologías y que han tenido un impacto significativo se destacan a continuación:

- Para no ir tan lejos, en Cali, Colombia se propuso una clusterización como técnica de análisis exploratorio de registros múltiples en datos meteorológicos con el objetivo de identificar estaciones atípicas, es decir las que no se agrupan con otras y estaciones que no son representativas de la muestra, estas estaciones tienen un gran valor dentro del estudio ya que contienen elementos de especial interés permitiendo reducir una compleja cantidad de información.
- Otra aplicación de la agrupación de series de tiempo se ha dado dentro de la industria de videojuegos, a través del análisis de datos temporales se pudo obtener la información del perfil de los jugadores, no solo basado en el comportamiento si no relacionado con sus habilidades de aprendizaje, de esta forma fue posible segregar a los jugadores en diferentes grupos y poder desarrollar productos interactivos con extensiones personalizadas y de uso sencillo para satisfacer las preferencias de cada uno de los segmentos.

En los casos expuestos anteriormente es posible observar la aplicación y el valor que se genera en cuanto a información, al analizar las series de tiempo en conjunto y no de forma unitaria.

La agrupación de series de tiempo es la división en conjuntos, la partición se realiza de tal manera que los objetos en el mismo grupo son más similares entre sí que los objetos de otros grupos de acuerdo con un criterio definido.

Dada la importancia del tema se han realizado varios estudios que han permitido el desarrollo de metodologías basadas en diferentes conceptos como lo es la complejidad, características o propósito de la agrupación, varias de estas metodologías se han implementado en aplicaciones de analítica. Para llevar a cabo el proyecto se van a utilizar algunos de las distancias ofrecidas en el paquete TSclust en R, las cuales se dividen en los siguientes grupos:

a. Basadas en modelos: En este tipo de clusterización, se asume para todas las series de tiempos analizadas un modelo o familia de modelo de series de tiempo, se comparan los parámetros del modelo y con estos genera los clústeres; para el proyecto utilizaremos clusterización a través del modelo ARIMA.
b. Basadas en características: En este tipo de clusterización, se compara una característica para todas las series de tiempo (autocorrelación, periodograma, etc..) en base a estas se agrupan los clústeres; para el proyecto utilizaremos el modelo de correlación.
c. Basadas en datos brutos: En este tipo de clusterización, no se asume que las series de tiempo tienen características de series de tiempo, solo compara cada dato en las diferentes series a través de una función de distancia. Estos enfoques basados en características tienen como objetivo representar la estructura dinámica de cada serie mediante un vector de características, disminuyendo la dimensionalidad y el tiempo de cálculo dando como resultado procedimientos de agrupamiento más eficientes; para el proyecto utilizaremos el modelo de Dinamic Time Warping.
d. Basadas en predicción: La clusterización basada en la predicción busca agrupar series de tiempo las cuales serán cercanas en un horizonte futuro, se comparan las densidades de las predicciones de cada serie.

Para la selección de la medida de disimilitud seleccionada para el modelo se tiene en cuenta en primera medida el propósito del agrupamiento, el agrupamiento puede estar basado en la estructura o en la forma, la disimilitud basada en la estructura está medida por una distancia dCOR o dCORT en la cual se agrupa series con patrones de aumento / disminución más cercanos entre sí, por lo cual frecuentemente se emplea para el análisis de indicadores económicos o financieros, por otro lado el agrupamiento dado por el concepto “basado en la forma”, como la distancia euclidiana, se centra en comparar los perfiles geométricos de

la serie, generalmente es aplicada a series cortas, ya que con series largas puede presentar fallas, igual que cuando se presentan registros atípicos.

Las medidas basadas en modelos incluidas en TSclust solo son aplicables en series de tiempo con procesos subyacentes estacionarios y lineales, mientras que las medidas basadas en predicción están libres del requisito de linealidad, pero suponen estructuras autorregresivas.

Para el pronóstico a los diferentes clústeres resultantes se les aplicará el modelo que ya ha sido utilizado en el problema por parte de Everis, dado que, si se varia se van a obtener resultados diferentes, los cuales no va a ser posible atribuirlos al algoritmo de predicción o a la agrupación de las series de tiempo.

Modelo por resultados de pronóstico:

Para este modelo se realizará un modelo iterativo que recorra cada una de las series de tiempo ejecutando diferentes tipos de algoritmos, al final realizaremos la agrupación de las series en las que mejor pronóstico generaron y realizaremos el benchmark de los diferentes modelos y sus resultados.

Los algoritmos que utilizaremos son los siguientes:

1. ETS con tendencia
2. ETS sin tendencia
3. HoltWinters
4. Arima
5. Sarima

2. Diseño de prueba:

Para todos los modelos se realizó la división de base de datos de entrenamiento desde el 3 de enero hasta el 30 de abril y se realizó el pronóstico diario de los 29 días que tenemos de mayo.

3. Construcción de los modelos:

Detección de series intermitentes: Para hacer este paso previo al modelado se realizó una sencilla formula en Excel que contara la cantidad de missing values y el porcentaje sobre el total de observaciones, estás se descartaron para la construcción de los modelos, el resultado de este paso fue un total de 442 series para analizar.

Modelo HoltWinters desagregado: Para este modelo se utilizaron los parámetros que optimiza el paquete de TSeries, al ser 36 series de tiendas se volvía una tarea innecesaria realizar la parametrización de cada una, y se utilizaron como fecha de inicio el 3 de enero

de 2018 y como frecuencia 30 días; para realizar la desagregación diaria se utilizó el porcentaje total de las ventas de cada categoría en cada tienda.

Modelo de Clusterización: Tanto para los modelos de ARIMA, Correlación y DTW no se requirió realizar ninguna parametrización adicional, estos paquetes calculan los datos internamente y generan una clusterización con las distancias calculadas de los resultados obtenidos, para hallar el número de k en cada uno de los clústeres se realizaron varias pruebas en las que se revisaba en que momento la desagregación de los clústeres no generaba valor ya que la cantidad de series en cada una no era significativa, por lo cual se llegó a 4 clúster en el modelo ARIMA, 3 de Correlación y 3 de DTW.

Se realizó un apply para sumar los resultados de todas las series que se encontraban en cada clúster y esta serie grande fue la que se modelo y pronóstico, con los mismos parámetros que el modelo HoltWinters desagregado.

Finalmente, al igual que en modelo desagregado con los porcentajes de participación en ventas de cada serie sobre la serie agregada, se desagregaron los valores de manera diaria.

Modelo por resultados de pronóstico: Para realizar la iteración de cada una de las series de tiempo se utilizó el paquete Foreach, en la construcción de este modelo en ninguno de los algoritmos usados excepto en el de SARIMA se realizó cambios en los parámetros, para todos los modelos se tomó la optimización de los parámetros de los paquetes utilizados (forecast, tseries). Para poder parametrizar los campos del SARIMA se utilizaron como muestra las 10 primeras series y con los resultados del MAD y MSE se fueron cambiando los parámetros para obtener el mejor resultado promedio en estas series de muestra.

Finalmente, para todos los modelos se generó el pronóstico de los 29 días de mayo y se obtuvieron como indicadores de error MSE ponderado, por el peso de cada una de las series dentro del total de las mismas y MAD.

4. Evaluación del modelo

Modelo HoltWinters desagregado: Al ser un volumen de series de datos muy grande se dificultó analizar los parámetros, se realizaron 2 pruebas con los siguientes parámetros y en los dos casos los valores promedio de error de las series fueron mayores que utilizando los parámetros optimizados del modelo tseries:

1. Alpha = 0,5, Beta = 0,6
2. Alpha = 0,4, Beta = 0,6

Modelo de Clusterización: En este modelo lo más complicado fue encontrar el número correcto de k, para realizar los modelos se probaron desde 10 hasta 3 clúster por cada modelo.

Para el pronóstico, con el objetivo de comparar los resultados del método de clusterización, se aplicó el mismo modelo utilizado en el modelo de Everis.

Modelo por resultados de pronóstico: En este modelo con lo que ya se había probado para los demás algoritmos dejamos los parámetros optimizados por el algoritmo, solo se realizó los cambios de la parametrización al modelo SARIMA el cual fue de la siguiente manera:

1. (1,0,0) x (0,0,1) – Con dos componentes estacionarios uno en AR para el primer componente y el componente cíclico se colocó un valor estacionario en el MA
2. (0,0,0) x (1,0,0) – Con un solo componente estacionario en el componente cíclico en el AR
3. (1,1,0) x (0,1,1) - El cual fue el modelo escogido, con dos componentes estacionaros en la tendencia regular en el AR y en I, y dos componentes estacionarios en el componente cíclico en el I y en el MA

Particularmente en el modelo de SARIMA es difícil elegir una parametrización, dado que cuando se cambia un campo mejora los pronósticos en unas series y en otras los empeora, por lo cual se realiza el promedio de los errores de las 10 primeras series y el que genera un menor error es el utilizado, obviar para todas las series si tienen tendencia, ciclicidad, autocorrelaciones, etc. es una mala práctica, sin embargo, para la obtención del pronóstico se aplica a todas las series el mismo modelo.

FASE 5: EVALUACIÓN

1. Evaluación de resultados de minería de datos con respecto a criterios de éxito empresarial

Modelo HoltWinters desagregado: Este modelo generó uno de los resultados más bajos, adicional la dificultad de la implementación para futuros pronósticos, la dificultad de desagregar las series de tiempo en porcentajes tiene más contras que pros, en primer lugar toma un tiempo mayor lograr desagregar los valores pronosticados en las tiendas por categoría, se pierde toda la variabilidad de los datos a lo largo de la serie al aplicarle un único porcentaje para la totalidad de pronósticos por serie, no se tienen en cuenta por ejemplo las series intermitentes y se les asigna un valor a cada día del pronóstico cuando se sabe que estas deben tener un tratamiento diferente.

Partiendo de estos resultados se podría pensar que cualquier modelo podría mejorar el pronóstico obtenido, a continuación, describiremos en tiempos el modelo para validar la cantidad de minutos dedicado y poder en la evaluación de los resultados comparar el tiempo utilizado y la medición de los errores.

Actividad	Tiempo dedicado minutos
Limpieza de datos	30
Construcción de vista minable	60
Construcción de modelo	300
Ejecución del modelo	30
Desagregación de series	240
Medición de errores	60
Total minutos	**720**
Total horas	**12 horas**

Modelo de Clusterización: Este modelo es la primera metodología utilizada para mejorar tanto la precisión como el tiempo de procesamiento, en los ítems de limpieza de datos y construcción de la vista minable el tiempo fue el mismo que en el modelo anterior, lo que cambio fueron las siguientes actividades, el tiempo de la construcción de modelo fue mucho menor que el anterior, lo más complejo del modelo fue elegir el k para realizar la agrupación de las series a una para el pronóstico, referente a los criterios esta metodología efectivamente reduce los tiempos de procesamiento, sin embargo tiene varias desventajas, en primer lugar el paquete utilizado a pesar que calcula los valores resultantes para cada modelo, estos son una caja negra y disminuyen la posibilidad de realizar un análisis descriptivo para cada clúster, porque si bien se pueden extraer las distancias, difícilmente se puede entender el modelo de agrupación, por ende no es posible determinar características singulares de los clústeres como estacionalidad, correlación, etc… en segundo lugar la agregación de las series de tiempo tienen cierta desventaja ya que de alguna manera a nivel de negocio no es consecuente que se estén sumando series que no tienen relación entre ellas, como por ejemplo, sumar bebidas calientes de la tienda uno con hielo de la tienda dos, por último, al igual que el modelo anterior, al realizar el mismo proceso de desagregación con un único valor porcentual se pierde la variabilidad de los datos.

Actividad	Tiempo dedicado minutos
Limpieza de datos	30
Detección series intermitentes	30
Construcción de vista minable	60
Construcción de modelo clusterización	120
Ejecución del modelo	30
Agregación de series	20
Construcción modelo pronóstico	45
Desagregación de series	240
Medición de errores	60
Total minutos	**635**
Total horas	**10 horas y media**

Efectivamente, comparándolo con el modelo anterior se ve una disminución del 11,8% del tiempo, después compararemos los resultados en cuanto a precisión del pronóstico.

Modelo por resultados de pronóstico: Este modelo fue el que en la construcción nos requirió más tiempo, puesto que debimos asegurar que el modelo realizara la iteración por cada una de las series, ejecutara los algoritmos escogidos y guardara una lista de los resultados de cada algoritmo por cada serie, tiempo que no será requerido en el futuro dado que solo se carga la nueva base y se corre el código ya elaborado, por lo cual el tiempo de construcción no se tomara para el total de tiempo utilizado, la mayor ventaja de este modelo es realizar el pronóstico por cada serie, lo cual, al no hacer ninguna agregación o desagregación que condujera a la perdida de información y al probar diferentes tipos de algoritmos se puede realizar una descripción del clúster conociendo las características de las series, como desventaja podemos encontrar que algunos de los algoritmos utilizados no generaron valor más si un mayor tiempo en la ejecución.

Actividad	Tiempo dedicado minutos
Limpieza de datos	30
Detección series intermitentes	30
Construcción de vista minable	60
Construcción de modelo (no se toma en el total)	780
Ejecución del modelo	60
Total minutos	**180**
Total horas	**3 horas**

En comparación en tiempo de ejecución con respecto al modelo realizado por Everis se aumentó en un 33%, cuando revisemos los resultados en precisión compararemos los resultados con los modelos anteriores.

2. Proceso de revisión de resultados

Modelo HoltWinters desagregado: En el modelo HoltWinters se obtuvieron los siguientes resultados en promedio para todas las series:

MSE	MAD
$ 10.836.384.442.198,80	$ 10.836.384.442.198,80

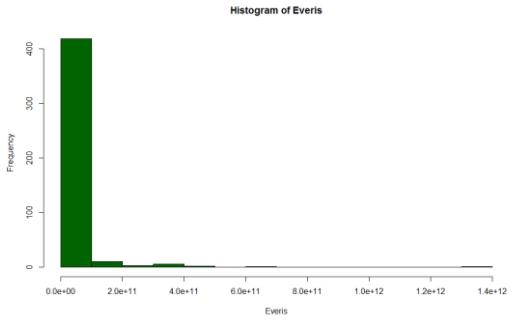


Ilustración 19 Histograma de MAPE Modelo HW Desagregado. Fuente: Autores

Con este modelo se presenta un error de 0 en 7 series y el error mas alto se da en la serie de la tienda Atlántida categoría Comidas Preparadas, el cual corresponde a $1.369.904.262.638,23 en el mes de mayo.

Como dijimos anteriormente este no es el mejor modelo, puesto que este asume que todas las series tienen un componente de tendencia y estacionalidad, por lo cual en todos los casos generó pronósticos con tendencia positiva sin ser necesario, generando un pronóstico lineal de las ventas.

Modelo de Clusterización:

En el modelo de clusterización jerárquico se obtuvieron los siguientes resultados:

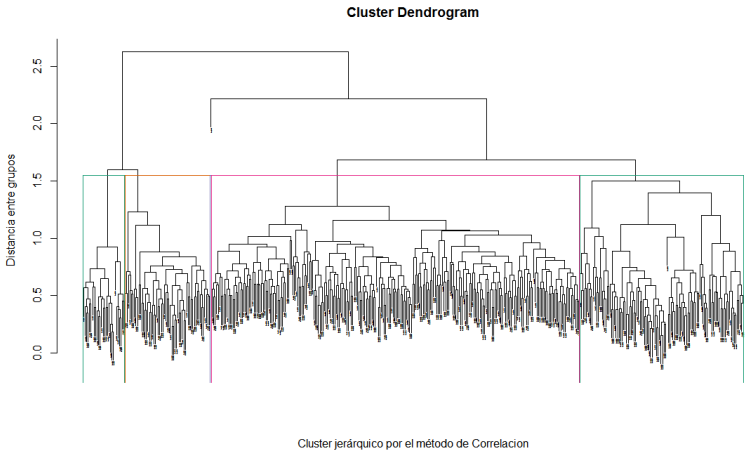


Ilustración 20 Clúster Jerárquico por el método de Correlación. Fuente: Autores

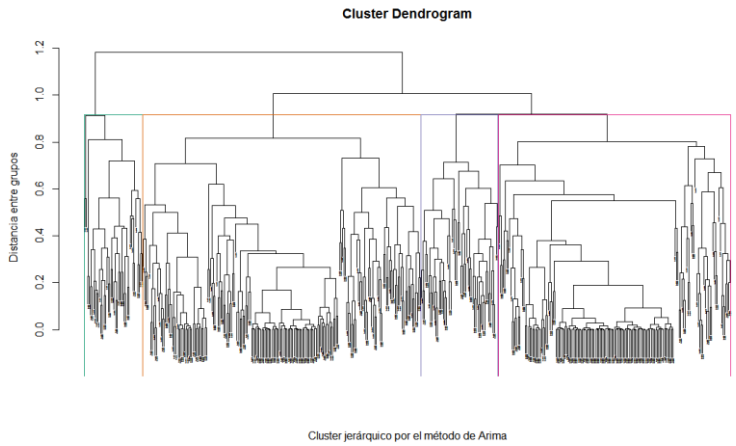


Ilustración 21 Clúster Jerárquico por el método Arima. Fuente: Autores

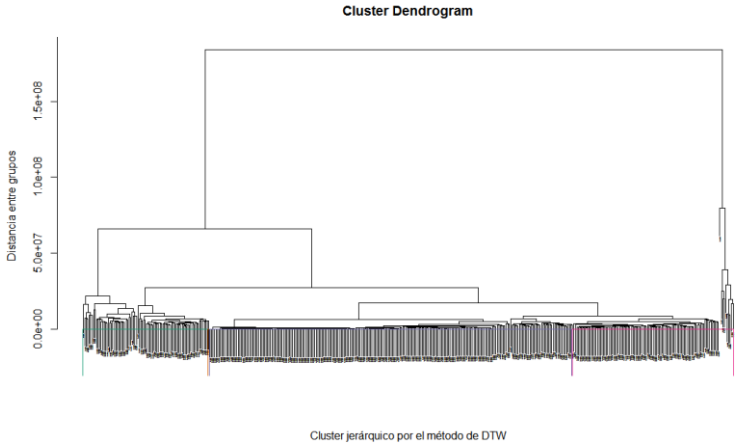


Ilustración 22 Clúster Jerárquico por el método DTW. Fuente: Autores

Se construyeron para el modelo de ARIMA 4 clústeres, y para los de Correlación y Dinamic Time Warping 3, los errores promedios obtenidos en el pronóstico con el modelo HoltWinters en cada uno de los modelos de clústeres es el siguiente:

Clúster	MSE	MAD
Clúster Arima	$ 2.718.128.439.513,73	$ 12.043.170,65
Clúster COR	$ 3.227.957.381.101,20	$ 12.020.342,15
Clúster DTW	$ 1.636.017.885.835.910,00	$ 34.455.896,08

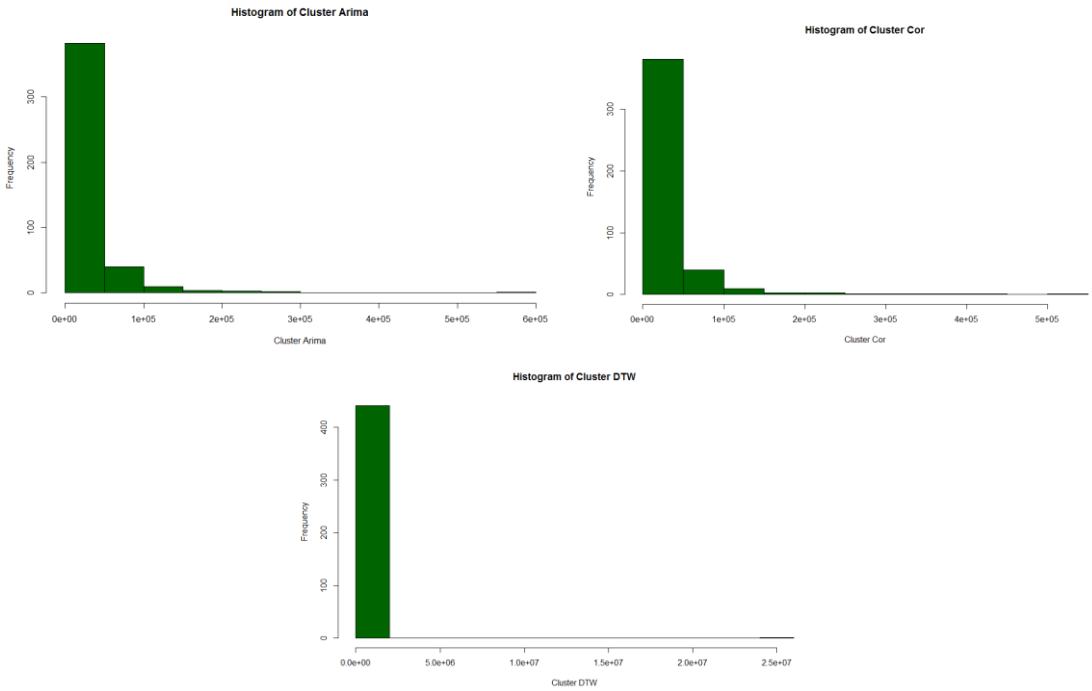


Ilustración 23 MAPE Clústeres ARIMA – COR – DTW. Fuente: Autores

Como se puede observar el error promedio en todos los clústeres con el mismo método de pronóstico disminuyó en 99,999889% en el clúster Arima, 99,999889% en el clúster de correlación y un 99,999682% en el clúster de DTW.

Si analizamos los resultados se puede observar que, realizando el mismo modelo de pronóstico, pero agregando por Clúster más no por tienda se obtienen mejores resultados, esto comprobaría que realizando la clusterización se mejoran los resultados por pronóstico, aunque la disminución en tiempos de modelamiento no tenga la misma proporción de disminución.

Si comparamos los resultados entre los modelos utilizados de cada clúster, podemos concluir que los modelos de agrupamiento de arima y cor generan resultados muy similares lo que no afecta en gran proporción el resultado del pronóstico, por el contrario, el agrupamiento del DTW fue el de más bajos resultados dentro de este modelo.

Modelo por resultados de pronóstico:

Para el modelo por pronósticos se construyeron 5 modelos: ARIMA, SARIMA, holtWinters, ETS con tendencia y ETS sin tendencia obteniendo los siguientes resultados:

Modelo	MSE	MAD
Arima	$ 1.050.528.840.840,82	$ 2.874.641,10
HW	$ 2.669.945.273.835,78	$ 18.675.066,06

ETS	$ 274.210.919.574.284.000,00	$ 18.675.066,06
SINETS	$ 6.816.988.870.987.650.000.000.000.000,00	$ 9.563.458,25
SARIMA	$ 37.776.345,50	$ 9.563.458,25

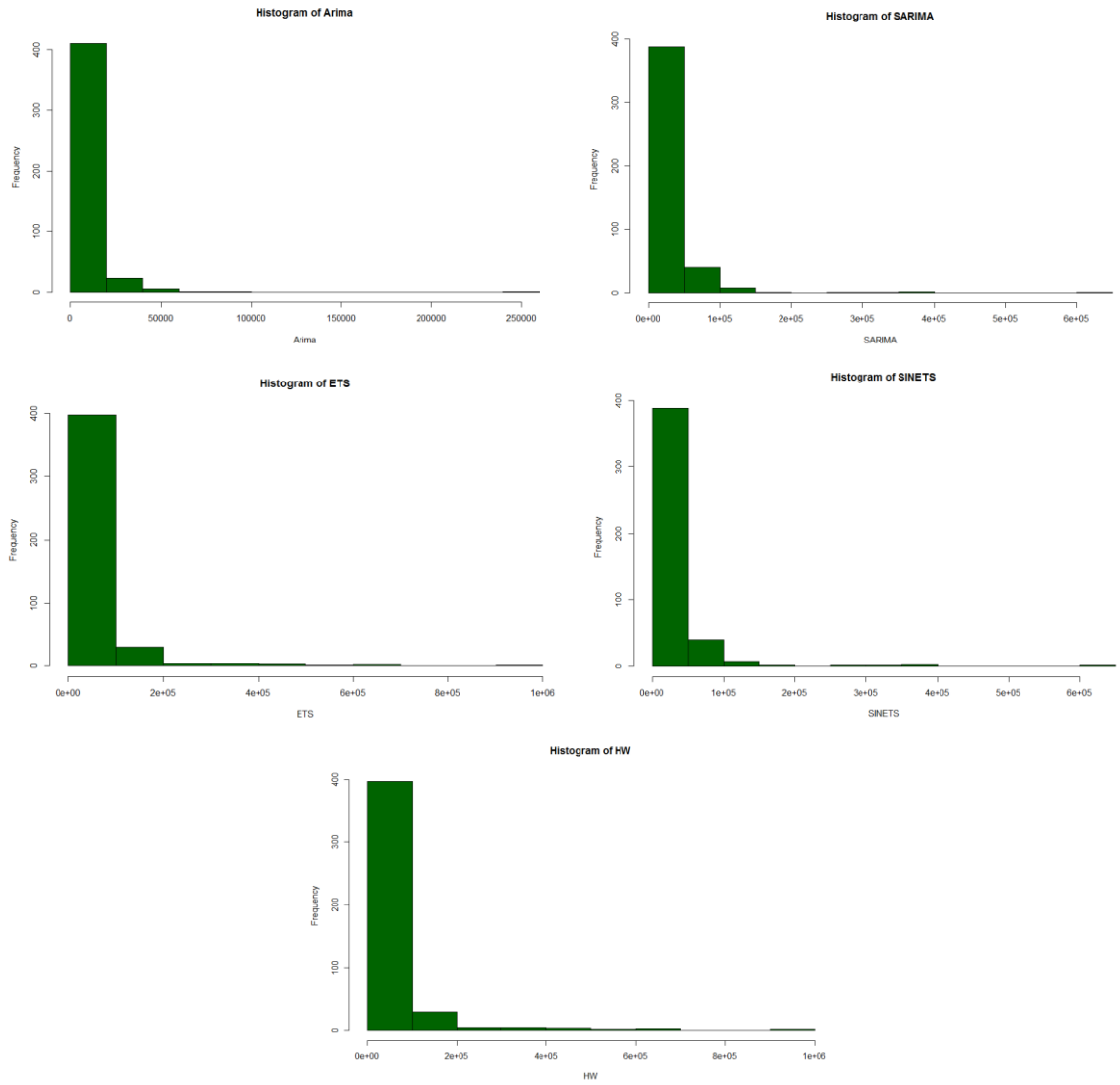


Ilustración 24 MAPE Modelo por serie ARIMA – SARIMA – ETS con y sin tendencia y HW. Fuente: Autores

La mejora del promedio del error en cada uno de los modelos fue la siguiente: en comparación con el modelo planteado por Everis, el modelo de ETS sin tendencia y Sarima fue un 99,999912% mejor, el modelo ETS con tendencia y HoltWinters fue un 99,99983% mejor y el modelo ARIMA un 99,99997% mejor.

El 70% de las series generó el mejor resultado con un ARIMA, ya que este tipo de algoritmo funciona muy bien cuando la frecuencia de los valores a pronosticar es menor a la anual, adicional porque al tener una cantidad reducida de observaciones difícilmente se puede

encontrar una estacionalidad, por lo que estos modelos eliminan este factor y presenta una tendencia de tipo polinominal.

El 3% de las series generó un mejor pronóstico con el modelo HoltWinters esto significa que estas series tienen claramente una tendencia y una estacionalidad, puede ser que en el tiempo analizado tengan consumos recurrentes en días de la semana.

El 10% de las series generó un mejor pronóstico con el modelo SARIMA esto significa al igual que las anteriores que son series con un componente de estacionalidad, pero se comportan mejor al agregarle el componente de promedios móviles.

El 17% de las series de tiempo generaron el mejor de los resultados con el modelo de ETS con tendencia

Por último, solo el 1% de las series generaron el mejor resultado clusterizando las series mediante un ARIMA.

Comparando los modelos utilizados y teniendo como referente el MAPE en cada serie se obtuvieron los resultados mostrados en las siguientes ilustraciones:

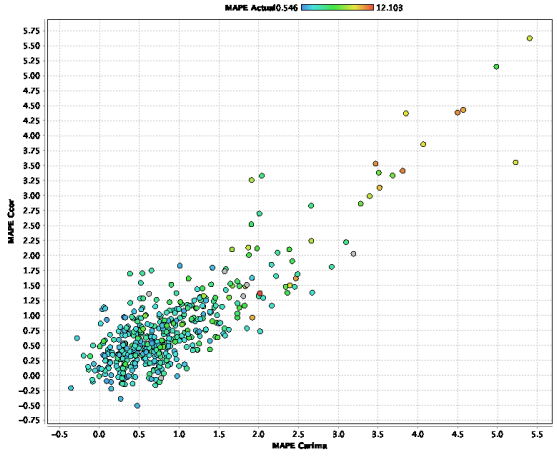


Ilustración 25 MAPE Comparación MAPE modelo Clúster Fuente: Autores.

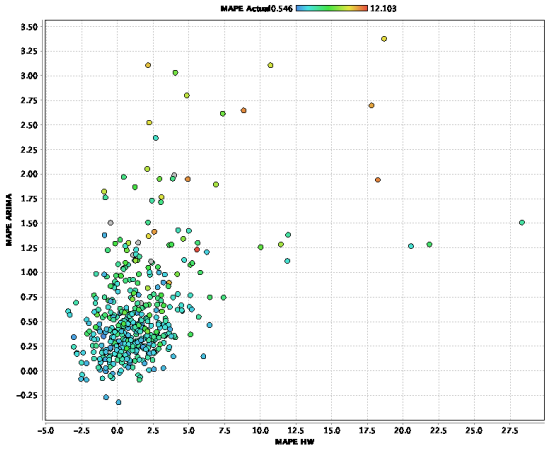


Ilustración 26 MAPE Comparación MAPE modelo por pronóstico Fuente: Autores.

Por un lado, se puede observar que la relación entre los errores de los dos clústeres ARIMA y COR tienen una correlación lineal, lo que significa que, si bien uno pudo generar mejores resultados que el otro la diferencia en sus errores no es significativa, cualquiera de los dos modelos con el pronóstico HoltWinters generaría el mismo resultado y se puede observar cómo las series en rojo con peores pronósticos se mejoraron con estos modelos.

Por otra parte, en la segunda gráfica se observa que al parecer no existe una relación entre los MAPE de los modelos HoltWinters y ARIMA, pero en la mayoría de las series los dos generan buenos resultados, si se realizará la comparación entre los MAPE de los modelos SARIMA y ARIMA son perfectamente relacionados linealmente y se puede observar como las series en rojo con peores pronósticos se mejoraron con estos modelos.

Adicional en las dos gráficas se identifican que hay series en las cuales ninguno de los modelos utilizados logro predecirlas de forma adecuada, teniendo en cuenta que son modelos simples y posiblemente algunas de estas series se les deberían realizar un estudio con más detalle y la aplicación de modelos más robustos, las series con peores resultados en las metodologías se mencionan a continuación:

Serie	MAPE min
Asgard - Cervezas	1,74
Asgard - Helados	1,57
Atlántida - Cervezas	2,87
Atlántida - Helados	1,96
Cimmeria - Cervezas	2,33
El Dorado - Cervezas	2,60
El Edén - Cervezas	2,69
El Edén - Charcuteria	2,63
El País de Nunca Jamás - Cervezas	2,20
Fantasía - Cervezas	2,34
Isla Luggnagg - Impuesto Bolsa	2,22
Isla Utopía - Helados	1,61
Macondo - Cervezas	1,85

Macondo - Mercancia General	1,61
Tierra Media - Helados	1,93
Tierra Media - Mercancia General	1,55

De los cuales categorías como cervezas, helados y mercancía general se repiten, lo que puede ser debido al comportamiento del producto indiferentemente de las tiendas, se podría corroborar lo anterior con análisis posteriores y estudios que permiten identificar los patrones de comportamiento de estas series en específico.

3. Conclusiones

Observados los resultados de las metodologías, tanto utilizada actualmente, como propuestas podemos concluir lo siguiente:

- Se debe continuar con este proyecto analizando las series intermitentes, ya que el porcentaje de series intermitentes fue de del 42,8% del total de series analizadas; se podría validar con modelos de datos por conteo los cuales se utilizan ampliamente para pronosticar la demanda intermitente y pronosticar todo el modelo incluyendo también estas series de datos.

- Viendo los resultados obtenidos en el modelo de clusterización se podría probar con diferentes tipos de algoritmos de pronóstico en los clústeres generados en el ARIMA o COR que generaron mejores resultados, para validar si con estos se logra mejorar el modelo, el modelo de clusterización de DTW en estos casos no es muy útil, puesto que lo que intenta es relacionar los puntos de series de tiempo con longitudes (cantidad de observaciones) diferentes, por lo que prácticamente agrupo todas las series en un solo clúster y algunas pocas en los otros dos, aun si se aumentaba el k, seguía ocurriendo lo mismo, este tipo de clusterización podría ser útil en otro tipo de datos con otro tipo de características.

- Dadas las características y el bajo rendimiento del modelo de clusterización Dinamic Time Warping, podría tomarse otro modelo de agrupación de datos en bruto que permita generar unos resultados más interesantes para su análisis.

- En el modelo de clusterización de COR al aumentar la cantidad k, no genera valor, en realidad con 2 clúster que sería con correlación o sin correlación es más que suficiente para realizar el pronóstico.

- La metodología actual y de clusterización dada la agregación y segregación hace que se pierda información importante para el pronóstico de las series, aumentando los valores de error.

- Los resultados de pronóstico por clusterizacón realizados en este trabajo no son concluyentes o aplicables a otros sectores u otro tipo de datos y tampoco si se

llegase a requerir realizar desagregación por subcategoría o hasta nivel de producto, puede que los resultados desmejoren.

- Si lo que se requiere es precisión en el pronóstico la mejor metodología es la segunda propuesta, sin embargo, podría probarse si mejora el resultado, realizando la metodología de clusterización de forma inversa, es decir primero pronosticando y agrupando en base al desempeño de cada modelo en las diferentes series para posteriormente clusterizar y optimizar los parámetros por serie agregada de los clústeres resultantes. Ya que la mayor dificultad de esta metodología fue la parametrización del modelo, dadas las características que no son extensivas en la totalidad de las series.

- Aunque el tiempo de construcción del modelo en la segunda metodología fue mayor respecto a los demás, esta actividad solo se realizará una vez ya que es escalable a cualquier tipo de base de datos, siempre que se cumpla con las característica de ser series univariadas y se construya la vista minable determinada en el trabajo; si el volumen de series de tiempo aumenta y no se quiere aumentar el tiempo de procesamiento en esta metodología, se puede incluir en el código la optimización a través de un algoritmo paralelo para que ejecute al mismo tiempo varias series; con esta mejora que para este modelo fue innecesaria por el corto tiempo de ejecución se lograrían los mismos resultados en menor tiempo.

- Difícilmente se logrará encontrar los parámetros perfectos para cada serie aún más en los modelos SARIMA o ARIMA, donde es necesario determinar los componentes de la serie para obtener el mejor resultado, por lo cual cualquiera de las dos metodologías propuestas se ve limitada en este sentido y puede que la parametrización escogida funcione bien para unas series, pero no para otras.

MANEJO RESPONSABLE DE LA INFORMACIÓN

Aunque para el proyecto no fueron suministrados ni tratados datos de carácter personal, como cumplimiento de los requerimientos académicos se dan las siguientes generalidades y recomendaciones descritas en la ley 1581 de 2012, asegurando que se respete el derecho constitucional que tienen todas las personas con respecto a su información recopilada, en cuanto a su protección y tratamiento.

Para dar cumplimiento con lo anterior se deben tener en cuenta los siguientes términos:

Autorización: Consentimiento previo, expreso e informado del Titular para llevar a cabo el Tratamiento de datos personales;

Dato personal: Cualquier información vinculada o que pueda asociarse a una o varias personas naturales determinadas o determinables;

Titular: Persona natural cuyos datos personales sean objeto de Tratamiento;

Tratamiento: Cualquier operación o conjunto de operaciones sobre datos personales, tales como la recolección, almacenamiento, uso, circulación o supresión.

Los lineamientos básicos para el tratamiento de los datos personales acerca de la autorización, el tratamiento de los datos y la publicación o difusión de los datos o los resultados del tratamiento son:

Para el tratamiento de los datos personales, y en consecuencia todas sus actuaciones deberán ser enmarcadas en los principios de legalidad, finalidad, libertad, veracidad o calidad, transparencia, acceso y circulación restringida, seguridad y confidencialidad.

Se debe solicitar una autorización, de forma previa e informada, al titular de los datos para poder recopilarlos y tratarlos, esta autorización debe ser explícita y expresada de cualquier manera que se evidencie una aceptación del tratamiento de los datos por parte del titular.

Se debe informar debidamente al Titular sobre la finalidad de la recolección y los derechos que le asisten por virtud de la autorización otorgada.

El tratamiento realizado a los datos personales debe ser correspondiente a la autorización firmada, y con la finalidad expresada al titular, garantizando al titular en todo momento, el pleno y efectivo ejercicio del derecho de hábeas data.

La información debe ser conservada bajo las condiciones de seguridad necesarias para impedir su adulteración, pérdida, consulta, uso o acceso no autorizado o fraudulento.

REFERENCIAS

Montero, Pablo; Vilar, José A. TSclust: An R Package for Time Series Clustering
Montero, Pablo; Vilar, José A. Análisis de Series Temporales usando R: Clúster, Clasificación y Contraste de Hipótesis.
Cáceres, Gustavo; Rodríguez, Jorge E. Agrupamiento de datos de series de tiempo. Estado del arte
Castro, Lina M.; Carvajal, Yesid; Ávila, Álvaro. Análisis clúster como técnica de análisis exploratorio de registros múltiples en datos meteorológicos
Vindel, Rafael; Menéndez, Héctor D.; Camacho, David. Combinando Series Temporales y Clustering para extraer Perfiles Evolutivos de Jugadores

Trabajos citados

Hanke, J. (2006). *Pronósticos en los negocios.*

TABLA DE ILUSTRACIONES

ANEXOS

Anexo 1. Categoría y Subcategoría de Productos

Categoría	Subcategoría	Categoría	Subcategoría
Beb. Alcoholicas	Licores	Comidas Rapidas	Horneados
Beb. No Alcoholicas	Aguas		Perros
	Energizantes		Comidas Rapidas
	Gaseosas		Fritos
	Jugos		Dulces
	Te		Galletas
	Isotonicos		Cerveza
	Otros		Postre
Bebidas Calientes	Cafes	Dulces	Dulces
	Otros		Chocolates
	Te		Masticables
	Chocolates		Nueces Y Mezclas
	Adiciones		Otros
Bebidas Naturales	Jugos		Galletas
	Otros		Ponques
	-	Helados	Helados
Cervezas	Cerveza	Hielo	Hielo
	Cerveza Sin Alcohol	Impuesto Bolsa	-
Charcuteria	Carnes	Insumos	Insumos
	Quesos	Lacteos	Lacteos
	Otros	Mercancia General	Mercancia General
Comidas Preparadas	Fritos		Otros
	Sandwich		Diarios/Revistas
	Salchipapa		Cerveza
	Frutas	Promociones /Crayola / Toallas	Por Promocion
	Adiciones		Otros
	Hamburguesas	Recargas	Recargas
	Insumos	Servicios / Soat	Servicios
	Menu Del Dia	Snacks Dulces	Dulces
	Arepas		Galletas
	Desayuno		Ponques
	Porciones		Otros
	Asopados Minuta		Chocolates
	Arroces Minuta		Nueces Y Mezclas
	Carnes Minuta	Snacks Salados	Galletas
	Otros		Nueces Y Mezclas

	Postre		Papas
	Productos Listos		Otros
Salud Y Belleza / Otc	Medicamentos		Chicharrones
	Aseo Y Otros		Servicios
		Tabacos	Tabacos

Anexo 2. ID y Nombre de Tiendas

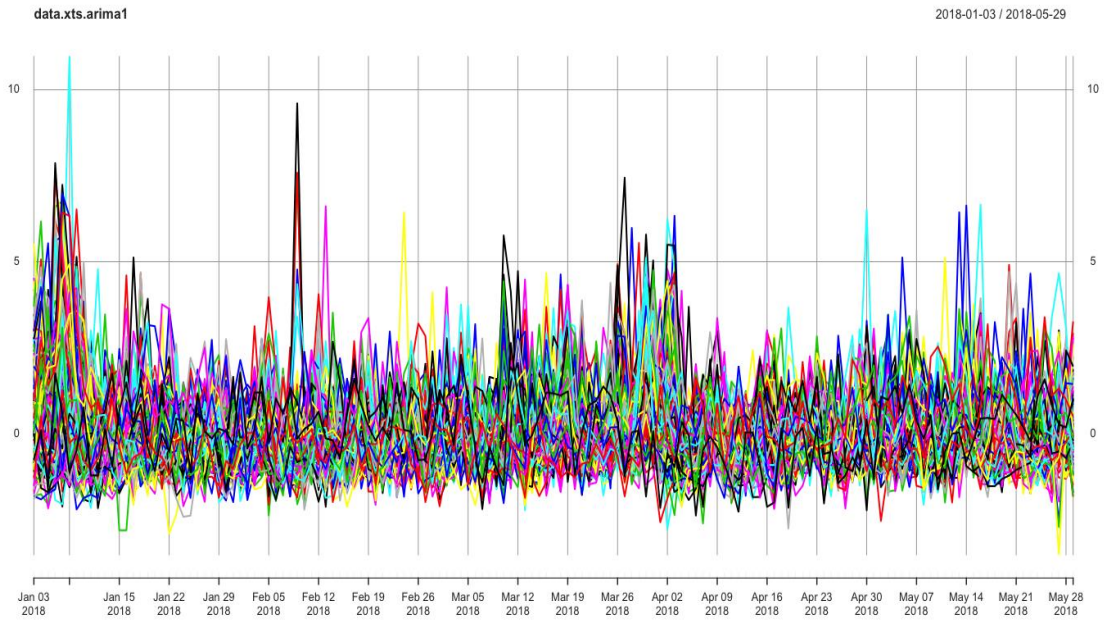
1	El Olimpo	21	Planeta Tatooine
2	Narnia	22	Macondo
3	Asgard	23	Planeta Kryptón
4	Atlántida	24	Arcadia
5	Tierra Media	25	Planeta X
6	El País de las Maravillas	26	Liliput
7	El País de Nunca Jamás	27	Isla Utopía
8	Camelot	28	Aztlán
9	El Dorado	29	Planeta Melmac
10	El Edén	30	Terramar
11	Ávalon	31	Zenda
12	Fantasía	32	Cimmeria
13	Oz	33	Planeta Vulcano
14	Shambhala	34	Icaria
15	Planeta Namek	35	Mundo de Krynn
16	Infierno	36	Basin City
17	Planeta Vegeta	37	Fraggle Rock
18	Gotham City	38	Balnibarbi
19	Luna de Endor	39	Isla Luggnagg
20	Metrópolis	40	Isla Glubbdubdrib

Anexo 3. Medidas de distancia TSClust

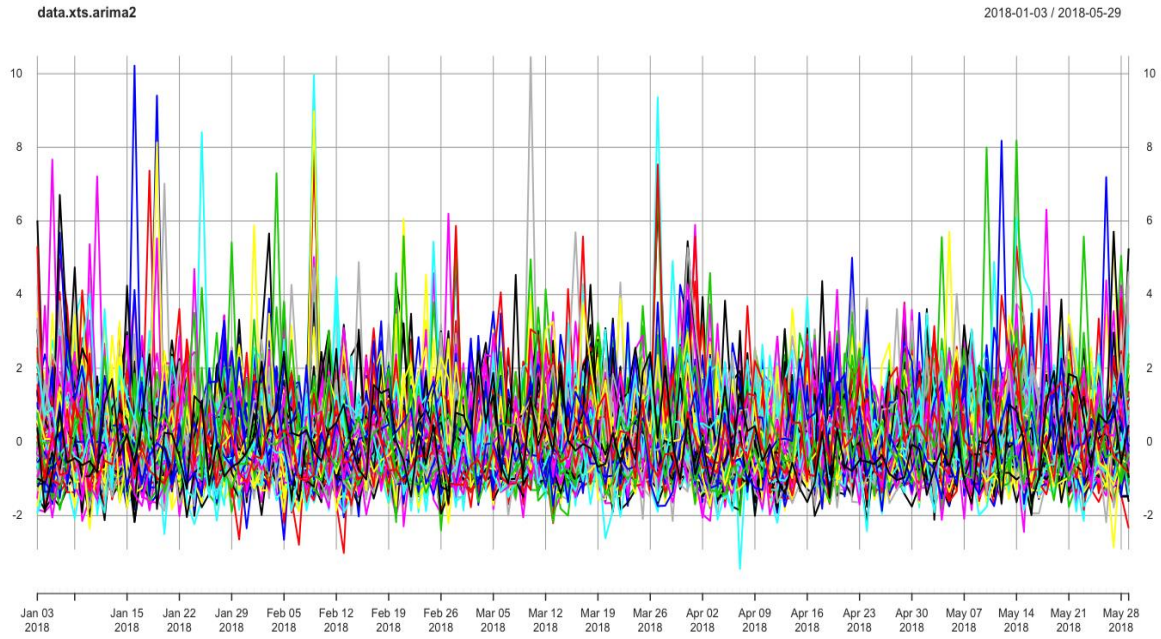
Distancia	Descripción	Ventaja/Desventajas
Minkowski distance/ Euclidean distance/ Manhattan distance	La noción de proximidad se basa en la cercanía de los valores observados en los puntos de tiempo correspondientes para que las observaciones se traten como si fueran independientes.	Invariante a cualquier traslación y rotación solo para n = 2. Las características con grandes valores y variaciones tienden a dominar sobre otras características.
Correlation-based distances / Autocorrelation-based distances	Criterio de disimilitud basado en el factor de correlación de Pearson entre las series de tiempo.	Incapaz de detectar la magnitud de las diferencias de dos variables.

	Medidas basadas en las funciones de autocorrelación estimadas	
Dynamic Time Warping	Busca encontrar patrones midiendo la proximidad entre curvas continuas	No solo trata la serie como dos conjuntos de puntos, sino que tiene en cuenta el orden de las observaciones.
Distancia basada en forecast	Medida de disimilitud basada en la comparación de las densidades de pronóstico para cada serie en un futuro horizonte de interés.	Es compatible con el propósito del proyecto.
Compression-based dissimilarity measures / Permutation distribution clustering	La similitud de dos series de tiempo se basa en la medición del nivel de información compartida por ambas series de tiempo medida como el número de bits al comprimir las dos series concatenadas. La disimilitud entre series se describe en términos de divergencia entre las distribuciones de permutación de los patrones de orden en la incrustación en m de la serie original.	Capta el dinamismo de la información Captar la heterogeneidad existente en los distintos clúster.
Piccolo distance	Define una medida de disimilitud en la clase de procesos ARIMA invertibles como la distancia euclidiana entre los operadores AR (1) que se aproximan a las estructuras ARIMA correspondientes. Para la clase de procesos	Su enfoque nos permite superar el problema de obtener aproximaciones ARMA ad hoc para cada una de las series sometidas a agrupación

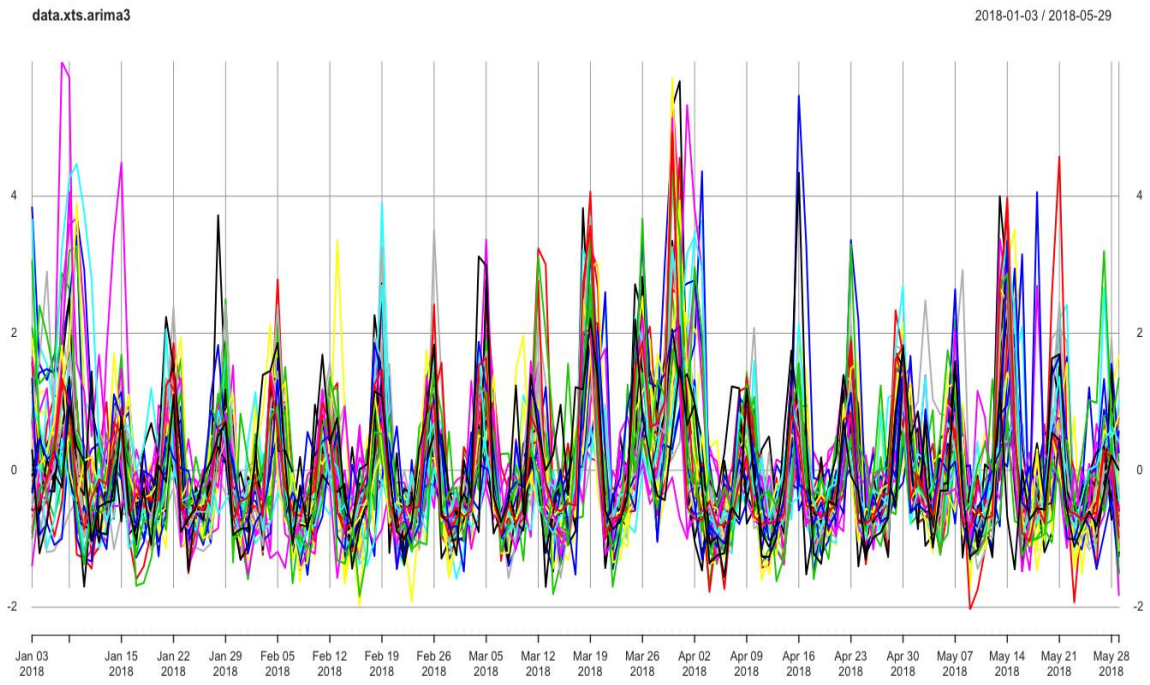
Anexo 4. Grafico Series de tiempo por clúster con datos normalizados – Arima1



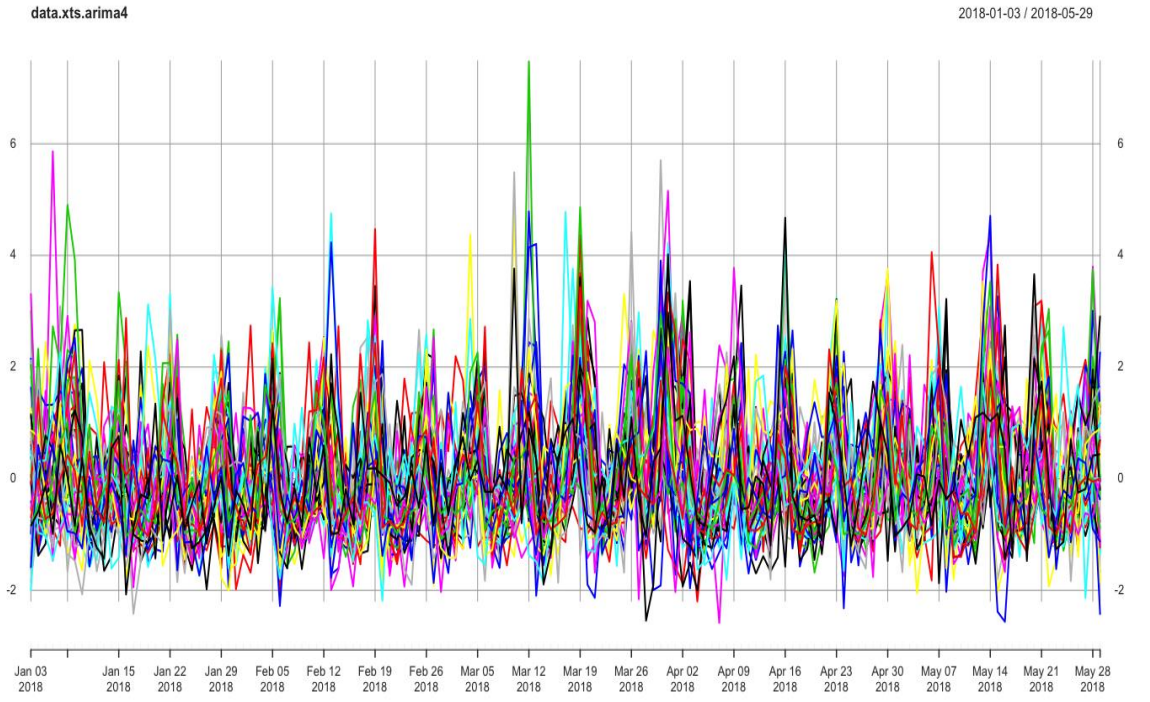
Anexo 5. Grafico Series de tiempo por clúster con datos normalizados – Arima2



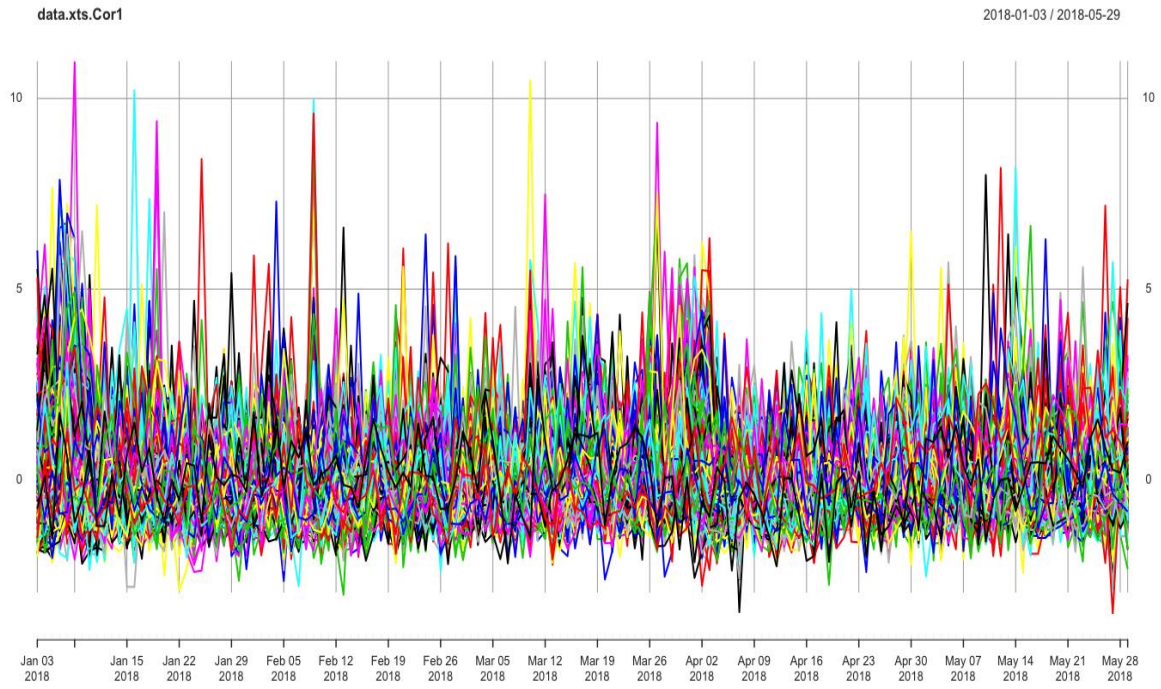
Anexo 6. Grafico Series de tiempo por clúster con datos normalizados – Arima3



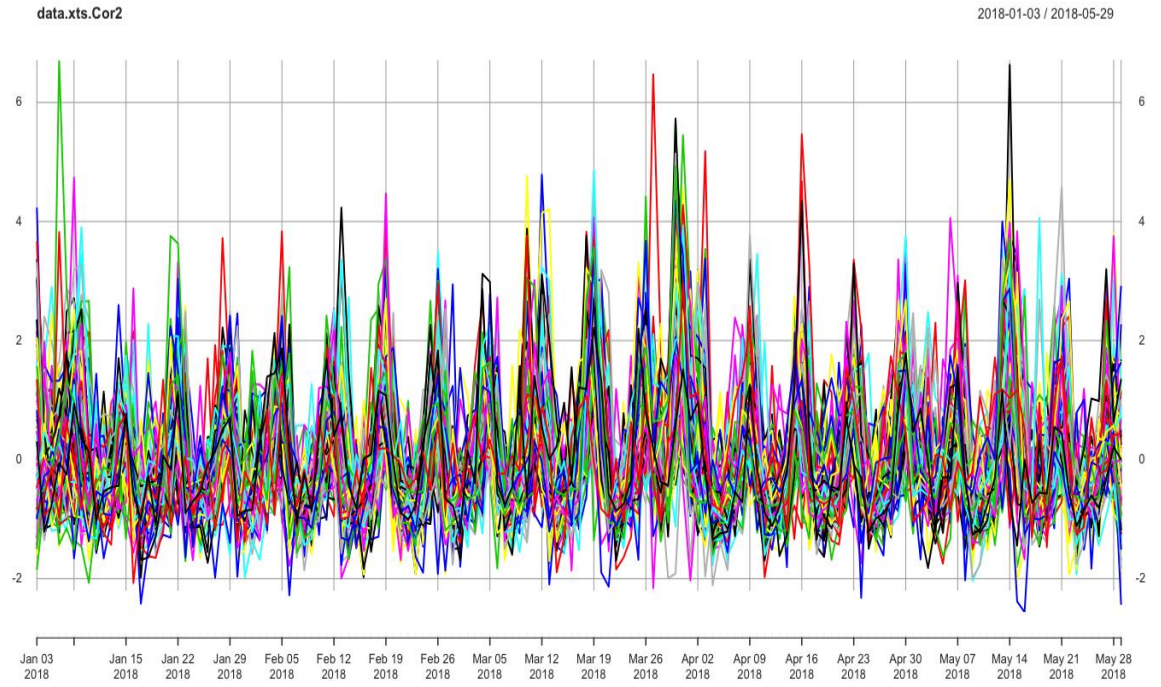
Anexo 7. Grafico Series de tiempo por clúster con datos normalizados – Arima4



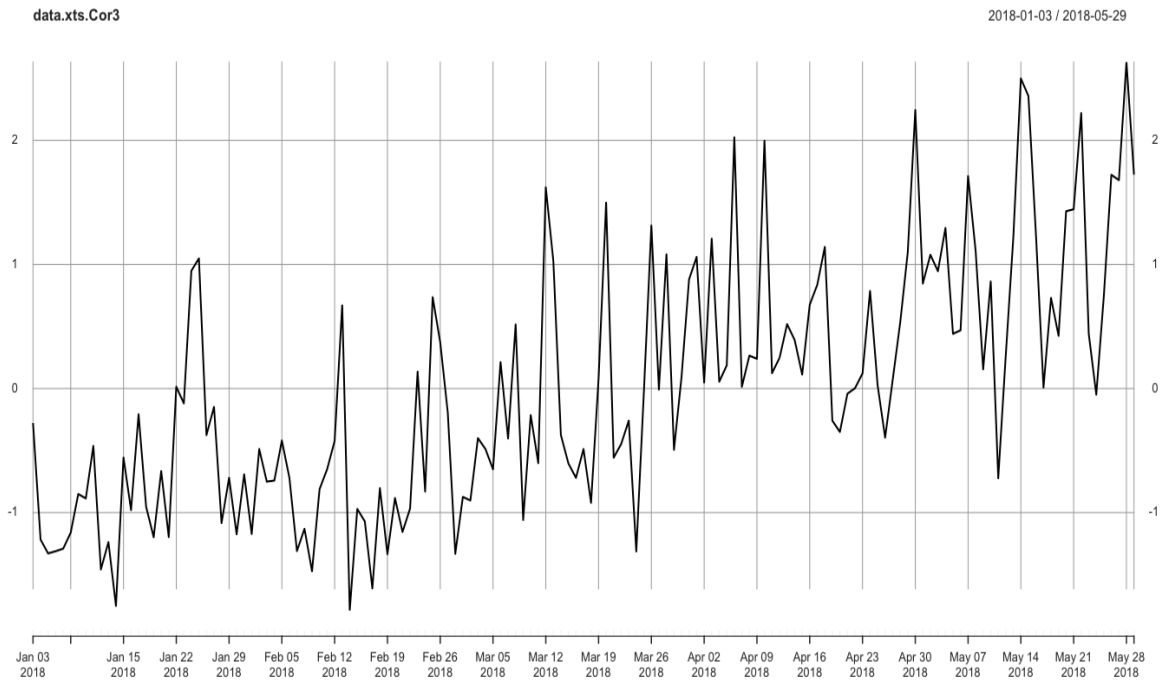
Anexo 8. Grafico Series de tiempo por clúster con datos normalizados – Cor1



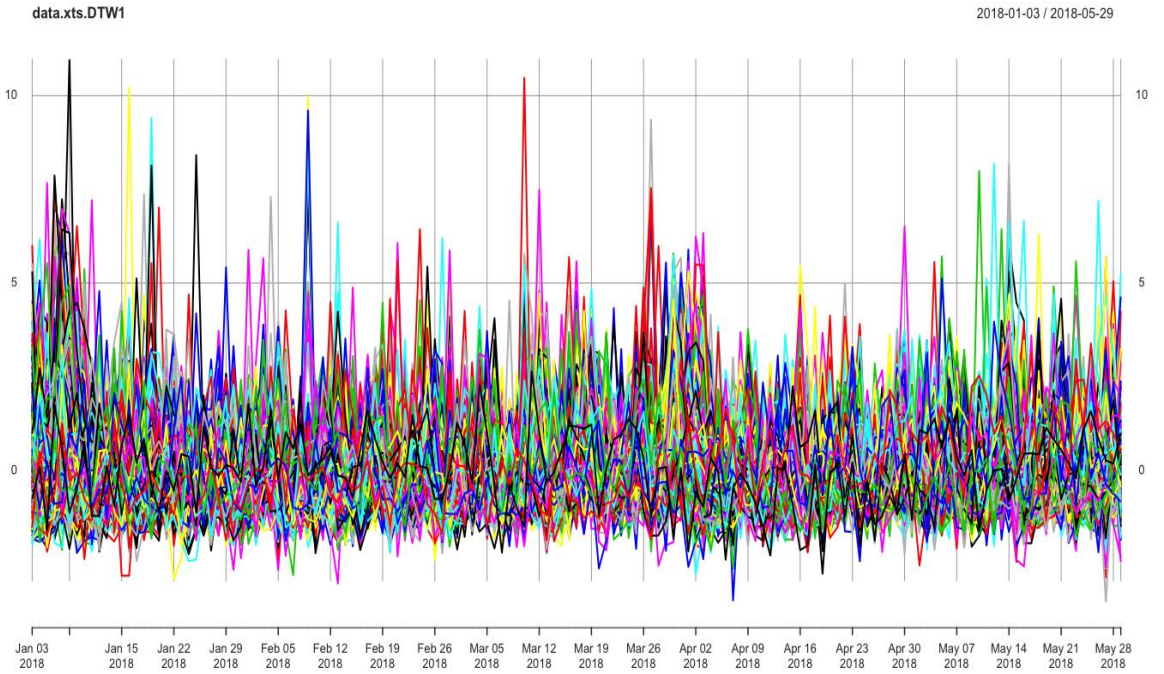
Anexo 9. Grafico Series de tiempo por clúster con datos normalizados – Cor2



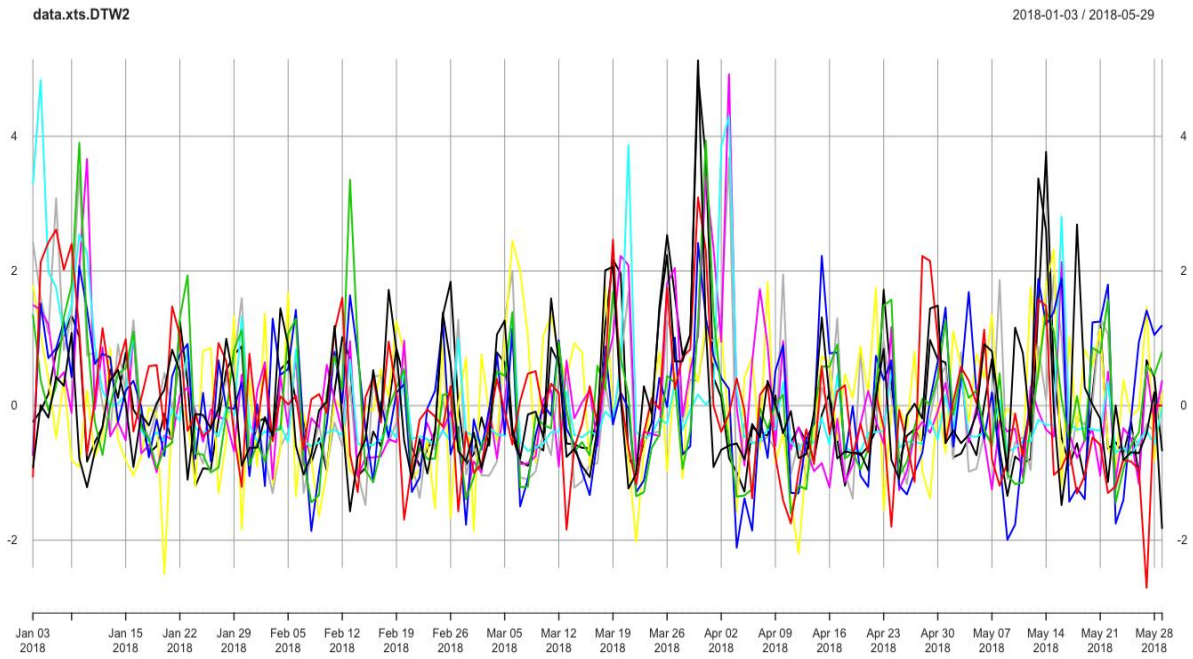
Anexo 10. Grafico Series de tiempo por clúster con datos normalizados – Cor3



Anexo 11. Grafico Series de tiempo por clúster con datos normalizados – DTW1



Anexo 12. Grafico Series de tiempo por clúster con datos normalizados – DTW2



Anexo 13. Grafico Series de tiempo por clúster con datos normalizados – DTW3