Trabajo Fin de Máster

# Ciclos de vida de productos

Paula Zas Castro

**Máster en Técnicas Estadísticas**

**Curso 2019-2020**

# RESUMEN:

Entre las múltiples tareas que se realizan desde el departamento de Tecnología de la Información de FINSA destacan la planificación de la demanda, la evaluación del presupuesto o el estudio de la evolución de las ventas, entre otras. Relacionada con esta última, surge la idea de categorizar ciertos productos en función de las similitudes de sus ciclos de vida.

En el presente trabajo se realiza una clasificación de las series de tiempo de los metros cúbicos de diseños de melamina entregados desde mayo de 2014 hasta agosto de 2019, ambos incluidos. Detectar similitudes entre los diseños de los productos que comercializa es de gran interés para la empresa por diferentes razones. Para empezar, una clasificación apropiada permitiría identificar diseños con ciclos de vida similares. Si, además, se consiguen obtener las características que de finen los grupos, es decir los clusters, podrían predecir a qué clúster pertenecería un diseño que deseasen poner a la venta, y por tanto predecir cómo sería su ciclo de vida. Por otro lado, el hecho de observar en qué punto del ciclo de vida se encuentra un diseño permitiría tomar decisiones, como por ejemplo introducir novedades para potenciar sus ventas si está en un momento de decaída, o dejar de venderlo si detectan que dicho diseño es insalvable. El estudio de la dispersión de los diseños contribuye a esta toma de decisiones ya que le permite a la empresa descubrir cuáles son los diseños que se están vendiendo en más países, cuáles son los que se venden a más clientes diferentes, cuáles son los más versátiles etc. En definitiva, qué diseños se caracterizan por presentar ventas más extendidas en función de diferentes variables consideradas. Por estas razones, se ha considerado interesante incluirlo en el proyecto. En última instancia, la construcción de una aplicación Shiny ha contribuido a la divulgación del trabajo realizado entre el resto de departamentos de la empresa. A través de esta aplicación, se consigue proporcionar a los demás departamentos una herramienta que pueden utilizar para sacar sus conclusiones acerca de los ciclos de vida y de la dispersión de los diseños y tomar decisiones sobre el futuro de los mismos.

El clustering de series temporales consiste en la fragmentación de la base de datos en grupos de tal forma que las series de un mismo grupo sean similares mientras que las series de distintos grupos sean diferentes. La determinación de clusters de series temporales es extremadamente compleja. Los principales problemas a los que nos enfrentamos al realizar un clustering de series de tiempo son: la determinación el criterio de disimilaridad que se desea en el análisis clúster, la habitual elevada dimensión del problema como consecuencia de la longitud de las series, tratar con series potencialmente de diferente longitud, combatir el efecto de outliers y la elección de elementos representativos de los clusters.

A la hora de realizar agrupaciones de series temporales dos vías muy comunes consisten en trabajar directamente con las series en bruto o reemplazar estas series de modo que se traslade el problema a un contexto de datos estáticos. En el primer caso, es común considerar medidas de similitud o distancias que tengan en consideración el carácter dinámico de las series. La otra vía consiste en representar las series temporales mediante datos estáticos y entonces usar directamente los algoritmos estándar existentes. En este proyecto se ha optado por el enfoque basado en características: primero se convierten los datos de una serie temporal en un vector

de características de menor dimensión y posteriormente se aplica un algoritmo de agrupamiento convencional a los vectores de características.

Tras la extracción de un gran número de características (22), se lleva a cabo en una etapa preliminar un análisis de componentes principales con el objetivo de examinar si hay características redundantes y en tal caso eliminarlas del proceso de clasificación. Se pretende evitar el uso de características que no aporten información al clustering por lo que, sobre la base del PCA resultante, se procedió a descartar aquellas con poco peso en todas las componentes (en concreto, sólo una de ellas). Posteriormente se procede a la realización del análisis clúster, para el cual se ha utilizado un método jerárquico aglomerativo. Previamente se ha procedido a la selección apropiada de medidas de vinculación entre grupos y de número de clusters subyacentes. Para ello, se han utilizado diversas técnicas, entre las que destacan la validación de agrupamiento relativo (coeficiente de aglomeración y coeficiente cophenetic) y la validación de agrupamiento interno.

Una vez realizado el análisis clúster, se construyeron los diferentes árboles de decisión que caracterizan a los clusters obtenidos. Los árboles de decisión son un método de aprendizaje supervisado no paramétrico utilizado para la clasificación y la regresión. Estos clasifican las muestras buscando en un subconjunto de características relevantes en lugar del conjunto completo. Por este motivo, en el trabajo hemos utilizado los árboles de decisión para reducir el número de características que hemos obtenido previamente en el clúster de series temporales. De este modo, nos hemos quedado con aquellas características que aparecían en los nodos del árbol de decisión. Los árboles se han utilizado también para caracterizar a los clusters obtenidos tras el proceso de clustering.

El uso conjunto e iterativo de análisis clúster y árboles de decisión hasta que ambos caminos de clasificación resulten congruentes, ha permitido perfilar la solución clúster tanto en la selección de características como en encontrar una solución interpretable y razonable, además de incrementar su estabilidad. Este criterio ha supuesto tener que iterar el procedimiento hasta dos veces.

Tras la realización del procedimiento expuesto anteriormente se concluye que:

- Los métodos de vinculación más adecuados son los métodos del promedio entre grupos y el método de Ward. Ambos conducen a valores de los criterios de optimalidad muy semejantes y a soluciones clúster también muy parecidas.
- En ambos casos, las características más importantes para definir los clusters son tres: la suma de los cuadrados de los cinco primeros coeficientes de autocorrelación parcial, el estadístico basado en la prueba del multiplicador de Lagrange para evaluar el nivel de heterocedasticidad condicional autorregresiva y el coeficiente de Hurst que evalúa el nivel de fluctuación a largo plazo de la serie, es decir si períodos de crecimiento de la serie tienden a ser seguidos por otros períodos de crecimiento o de decrecimiento.
- Las reglas que definen los árboles de decisión para cada grupo de la solución clúster permiten caracterizar los mismos como sigue:

- El clúster con mayor número de series, 195, agrupa series que mayoritariamente se caracterizan por presentar bajos coeficientes de autocorrelación parcial y bajo/moderado nivel de heterocedasticidad.
- Otro clúster agruparía a 64 series y la principal característica que las define es un coeficiente de Hurst razonablemente alto que suponen un comportamiento de persistencia de la serie, es decir series poco fluctuantes.
- El clúster más pequeño agrupa a 22 series que se caracterizan por un nivel de heterocedasticidad elevado y mayor nivel de fluctuación (coeficiente de Hurst bajo) que aquellas de los otros dos grupos.