

Caracterización de los componentes del gasto de los turistas de cruceros mediante técnicas de data mining

Silvia Altmark, Ramón Alvarez^a, Florencia Santiñaque

Universidad de la República, IESTA, Departamento de Métodos Matemáticos Cuantitativos



^aramon@iesta.edu.uy

Resumen

Existen muchas situaciones donde es necesario hacer predicciones en variables económicas, como el gasto por ejemplo, que puede descomponerse en varios rubros. Tal es el caso del gasto de los turistas de cruceros que llegan a Uruguay en cada temporada, entre los meses de octubre y abril. Este tipo de turismo ha venido creciendo en Uruguay, desde la temporada 2006-2007 con 130 cruceros arribados, a 225 cruceros arribados en la temporada 2011-2012, lo que significa un importante aporte de divisas al país.

De acuerdo a este planteo, este trabajo propone en convertir los componentes del gasto en proporciones, y caracterizar las mismas mediante una herramienta gráfica, como los gráficos ternarios o triangulares. Por otro lado se comparan los resultados, con la tipología de cruceristas que surge de aplicar métodos de clusters jerárquicos y no jerárquicos. Se utilizan los datos correspondientes a las temporadas de cruceros 2010-2011 y 2011-2012, cuya fuente es el Ministerio de Turismo y Deporte. Los mismos surgen de una muestra de pasajeros a través de una encuesta cara a cara con diseño muestral complejo.

Objetivos Principales

- Transformación en proporciones** En este trabajo se propone convertir los componentes del gasto en proporciones, y caracterizar las mismas mediante herramientas gráficas, como son los gráficos ternarios o triangulares.
- Creación de tipologías de cruceristas** Por otro lado se comparan los resultados, con la tipología que surge de aplicar métodos de clusters jerárquicos y no jerárquicos.

Metodología

Gráficos Triangulares

Los gráficos triangulares son un tipo de gráfico baricéntrico que permiten trabajar a la vez con 3 variables que tienen la característica de tener una suma constante por observación; son un caso particular (para 3 variables) de lo que se denomina datos composicionales. [Hamilton, 2015], [David Meyer y Hornik, 2015]

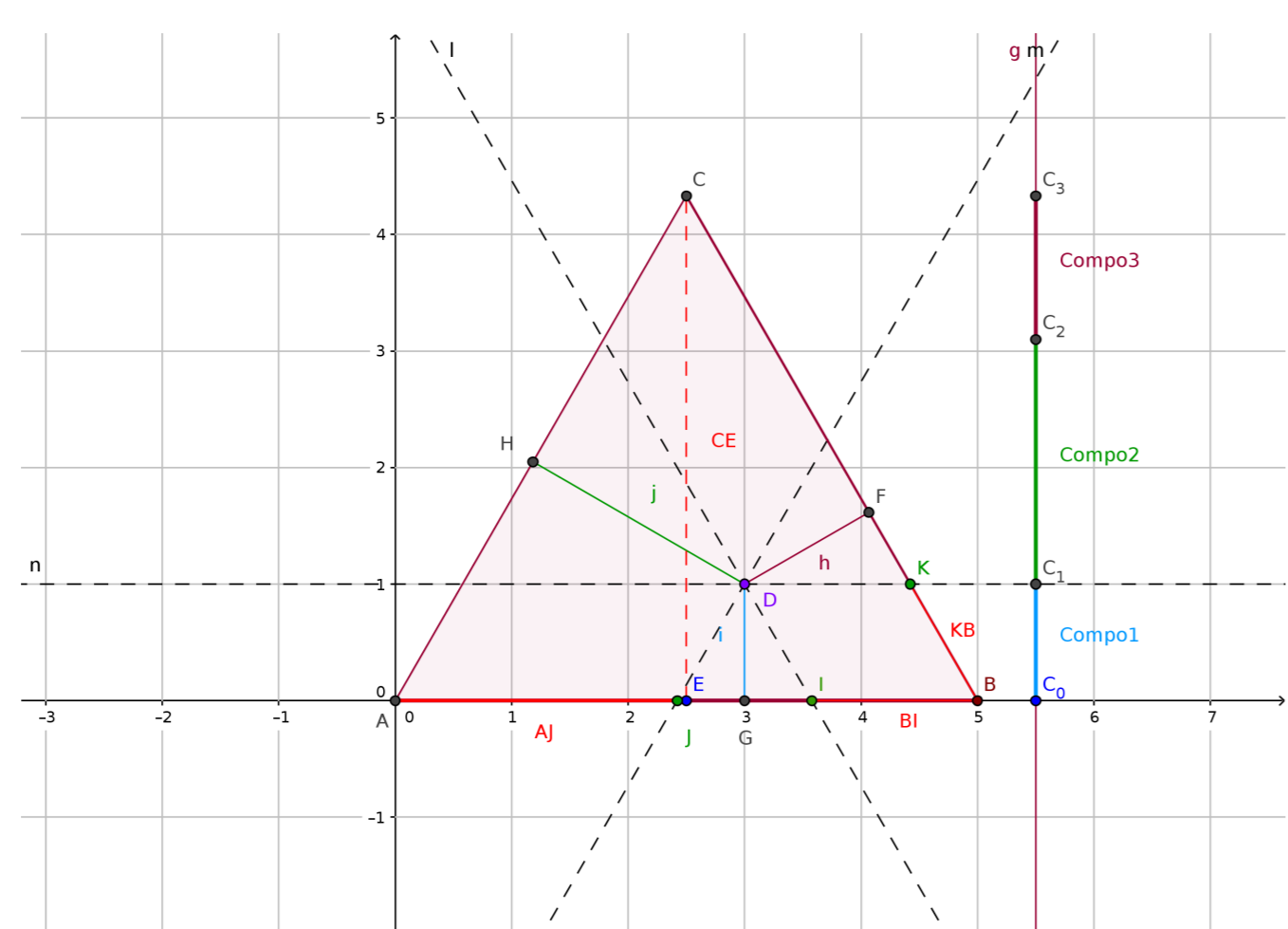


Figura 1: Gráficos triangulares

Análisis de Cluster

- Tipología (a)**, al aplicar el algoritmo de *kmeans* sobre los gastos por rubro en niveles
- Tipología (b)**, al aplicar el método PAM (Partition around medoids) sobre los gastos en proporciones. [Kaufman y Rousseeuw, 1990]
- Tipología (c)**, a través de algoritmos de cluster jerárquicos para datos binarios al considerar presencia o ausencia de gastos por rubro. [Maechler et al., 2015]

Resultados

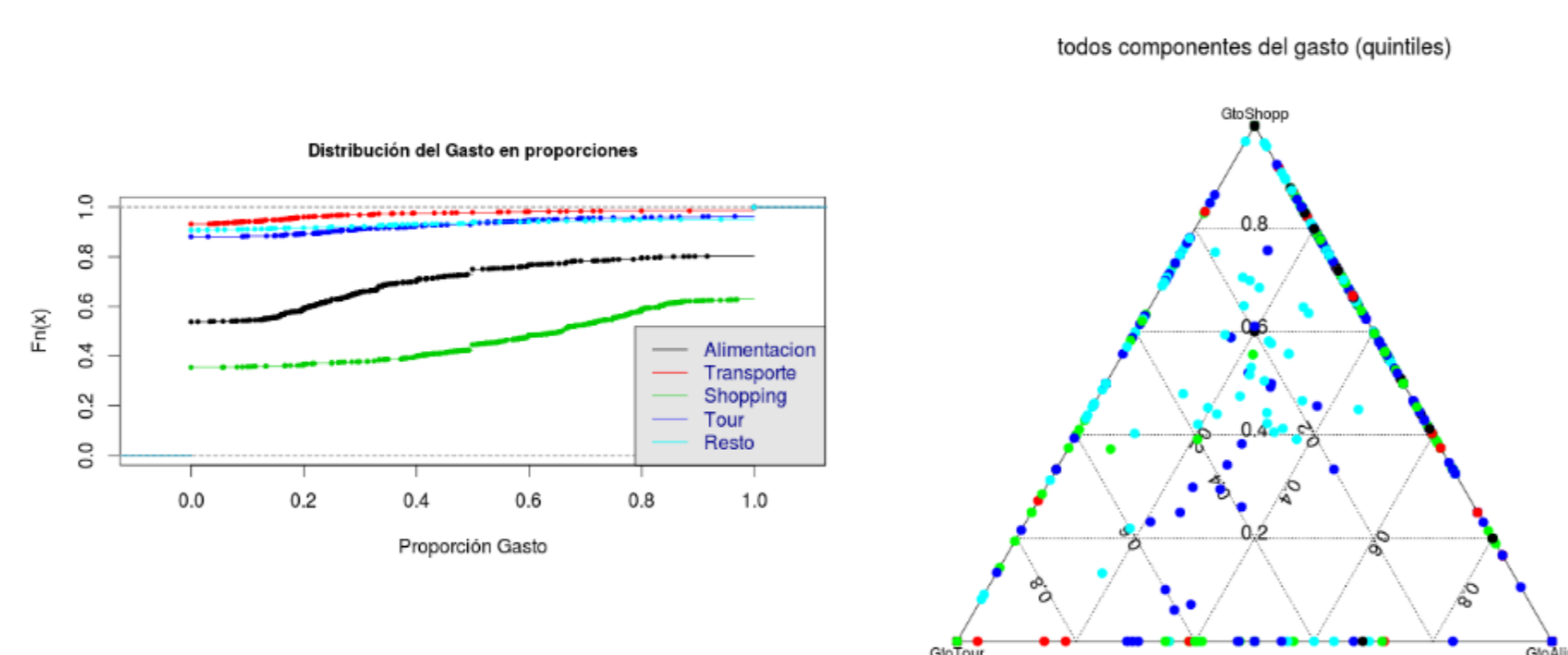


Figura 2: Componentes del Gasto

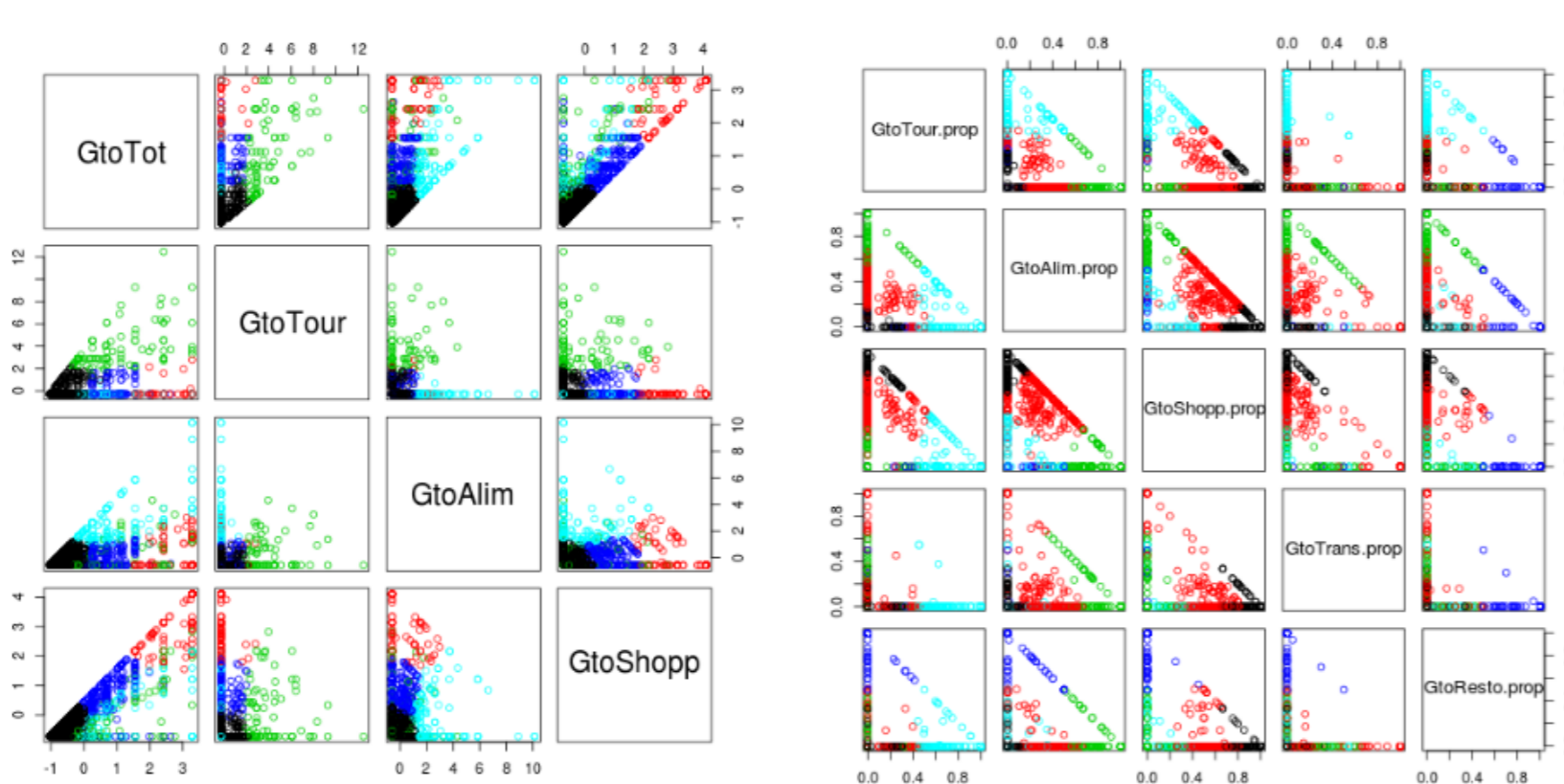


Figura 3: Componentes Gasto en niveles (a) y proporciones (b)

- Dado que el interés del presente trabajo es el gasto de los cruceristas, se descartaron los grupos de cruceristas que no presentaban gasto o que tenían un monto de gasto imputado, reduciéndose el estudio a 2311 casos (filas).
- Los rubros que se utilizan para desagregación del gasto de los cruceristas son: Shopping (Compras), Alimentación, Tours, Transporte y Otro gastos, siendo los de mayor importancia Shopping y Alimentación.

En la figura 3 pueden verse los gastos discriminados por clusters formados sobre los gastos en niveles (figura (a) mediante *k-means* y los clusters formados sobre los gastos en proporciones mediante algoritmo PAM en la (figura (b))

cluster	Gasto Total	Gasto Tour	Gasto Alim	Gasto Shopping	N
1	-0,577	-0,0179	-0,267	-0,454	1417
2	2,495	-0,212	-0,114	2,908	134
3	0,94	3,81	0,0823	-0,079	102
4	0,675	-0,142	-0,158	0,815	384
5	0,982	-0,274	2,368	-0,269	188
Total	0	0	0	0	2225

Cuadro 1: Gasto promedio según Grupos (Gastos en niveles)

cluster	G. Tour	G. Alim	G. Shopping	G. Transporte	G. Resto	N
1	0,654	0,074	0,066	0,187	0,019	212
2	0,035	0,341	0,573	0,029	0,021	378
3	0,004	0,962	0,009	0,017	0,008	529
4	0,016	0,047	0,005	0,006	0,926	144
5	0,008	0,15	0,971	0,004	0,002	188
Total	0,0734	0,303	0,526	0,029	0,0682	2225

Cuadro 2: Gasto promedio según Grupos (Gastos en proporciones)

cluster	G. Tour	G. Alim	G. Shopping	G. Transporte	G. Resto	N
1	0	0	1	0	0	860
2	0,014	0,873	0,87	0	0,287	539
3	0	1	0	0	0	458
4	0,114	0,261	0,195	1	0,01	210
5	1	0,348	0,5	0	0	244
Total	0,119	0,462	0,645	0,09	0,068	2311

Cuadro 3: Gasto promedio según Grupos (Gastos variables binarias)

Conclusiones

- Se consideran los 3 rubros más importantes
- Se consideran la totalidad de los gastos (en proporciones)
- Se deja de lado el gradiente de gasto y analizan como atributos de ausencia o presencia (para combatir la gran asimetría de cada rubro)
- Con cada método de clustering surgen tipologías diferenciadas (ver los cuadros 1, 2 y 3)

Futuros Pasos

- Probar métodos de cluster difusos. [Kaufman y Rousseeuw, 1990]
- Construir perfiles de gasto usando toda la información que se dispone complementaria al tipo y nivel de gasto, en particular las socio-demográficas. Para eso se puede probar de aplicar análisis factorial de correspondencias sobre las variables binarias de gastos + el bloque de variables socio-demográficas

Referencias

- [David Meyer y Hornik, 2015] David Meyer, A. Z. y Hornik, K. (2015). *Vcd: Visualizing Categorical Data. R package version 1.4-1.*
- [Hamilton, 2015] Hamilton, N. (2015). *ggtern: An Extension to 'ggplot2', for the Creation of Ternary Diagrams.* R package version 1.0.6.0.
- [Kaufman y Rousseeuw, 1990] Kaufman, L. y Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley, New York.
- [Maechler et al., 2015] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2015). *cluster: Cluster Analysis Basics and Extensions.* R package version 2.0.3 — For new features, see the 'Changelog' file (in the package source).