



UNIVERSIDAD DE LA REPÚBLICA
Facultad de Ciencias Económicas y de Administración
Instituto de Estadística

Caracterización del gasto de cruceristas en Uruguay mediante técnicas de datamining

Ramón Álvarez-Vaz; Silvia Altmark; Florencia Santiñaque

Febrero, 2020

Serie Documentos de Trabajo

DT (1/20) - ISSN : 1688-6453

Forma de citación sugerida para este documento:

**Álvarez-Vaz, Ramón ; Altmark, Silvia y Santiñaque, Florencia (2020).
Caracterización del gasto de cruceristas en Uruguay mediante técnicas de
datamining[en línea].
Serie Documentos de Trabajo, DT (1/20). Instituto de Estadística, Facultad
de Ciencias Económicas y de Administración, Universidad de la República,
Uruguay.**

Caracterización del gasto de cruceristas en Uruguay mediante técnicas de datamining.

Ramón Álvarez-Vaz ¹; Silvia Altmark ²; Florencia Santiñaque ³
*Instituto de Estadística, Departamento de Métodos Cuantitativos,
Facultad de Ciencias Económicas y de Administración,
Universidad de la República*

Resumen

Desde la temporada 2005-2006 a la 2017-2018, el turismo de cruceros en Uruguay ha crecido 118 % en cantidad de cruceristas y 81 % en divisas. Dado su aporte a la economía uruguaya, es importante caracterizar el gasto de los cruceristas. El artículo presenta una caracterización del gasto de cruceristas que visitan Uruguay, aplicando metodologías relacionadas con datamining, a los datos de cruceros de la temporada 2010-2011 del Ministerio de Turismo. Las variables consideradas para conformar los grupos son el gasto per cápita en Alimentación, Tours y Shopping. Se aplica, el algoritmo de k-means sobre las variables de gastos en escala original, el método PAM sobre las variables de gasto en proporciones y un algoritmo jerárquico de Ward, considerando si existe o no gasto en cada rubro. Las tipologías se asocian con las características socio-demográficas de los cruceristas. Las tres metodologías aplicadas arrojan resultados similares en cuanto a la caracterización de turistas.

Palabras clave: Clusters, cruceros, datamining, gasto.

CÓDIGOS JEL: C10, C18, C30, C83

Clasificación MSC2010: 62D05, 62H30, 62H86, 62P20

¹ *email:* ramon@iesta.edu.uy, ORCID: 0000-0002-2505-4238

² *email:* salt@iesta.edu.uy, ORCID: 0000-0003-3123-2165

³ *email:* florsanmes@gmail.com, ORCID: 0000-0002-8224-6906

ABSTRACT

From the 2005-2006 season to 2017-2018, cruise tourism in Uruguay has grown 118 % in number of cruise passengers and 81 % in foreign currency. Given its contribution to the Uruguayan economy, it is important to characterize the spending of cruise passengers. The article presents a characterization of the spending of cruise passengers visiting Uruguay, applying methodologies related to data mining, to the cruise data of the 2010-2011 season of the Ministry of Tourism. The variables considered to form the groups are the per capita expenditure in Food, Tours and Shopping. The algorithm of k-means on the variables of expenditure in original scale, the PAM method on the expenditure variables in proportions and a hierarchical algorithm of Ward considering whether there is expenditure in each item is applied. Typologies are associated with the socio-demographic characteristics of cruise passengers. The three applied methodologies show similar results in terms of the characterization of tourists.

Key words: Clusters, cruisers, datamining, expenditure.

JEL CODES: C10, C18, C30, C83

Mathematics Subject Classification MSC2010: 62D05, 62H30, 62H86, 62P20 .

1. Introducción

El turismo es una actividad muy relevante para la economía del Uruguay, en términos de divisas, valor agregado (PIB) y empleo. De acuerdo a información del Ministerio de Turismo, el Uruguay pasó de recibir 494 millones de dólares por concepto de ingresos por turismo en 2004 a más de 2300 millones en 2017; en términos de visitantes, se pasó de 1.8 millones a más de 4 millones en el mismo período.

El sector turístico representó en el año 2017 un 8,6% del total del Producto Interno Bruto y el 6,34% de los puestos de trabajo generados en la economía, según estimaciones realizadas en el marco de la Cuenta Satélite de Turismo. Datos de la Balanza de Pagos del Banco Central del Uruguay indican que el turismo representó en 2017 el 49% de las exportaciones de servicios y el 17% de las exportaciones totales del país.

También se pueden señalar otros impactos de la actividad turística que dinamizan la economía: creación de infraestructuras y servicios, mejora de recursos humanos, aplicación de nuevas tecnologías, surgimiento de nuevas oportunidades de negocios, recuperación y/o preservación del patrimonio, puesta en valor de recursos. Sin embargo, el turismo también puede afectar negativamente un destino, cuando no se toman en consideración los impactos medioambientales y socioculturales de ciertas actividades turísticas. El Ministerio de Turismo de Uruguay ha trabajado en el diseño de políticas de desarrollo turístico sostenible, elaborando el Plan Nacional de Turismo Sostenible 2009-2020 y su reciente actualización, H2030.

Además del turismo receptivo e interno, el Uruguay recibe cruceros, que llegan a dos puertos de Uruguay: Montevideo y Punta del Este, entre octubre de un año y abril del siguiente, lo que constituye la “temporada de cruceros”. En este segmento del turismo, se tiene en cuenta la cantidad de personas arribadas en los buques y la cantidad de personas desembarcadas, ya que muchas no descienden de los buques. Si bien se consideran los buques que llegan a cada uno de los dos puertos y la cantidad de personas en cada uno de ellos (sean pasajeros o tripulantes), es importante señalar que las encuestas a cruceros se aplican a las personas desembarcadas, sean pasajeros o tripulantes, ya que tienen la misma implicancia a efectos del cómputo de número de cruceristas y monto del gasto. Según los datos relevados por el Ministerio de Turismo a partir de sus encuestas, el turismo de cruceros presenta una evolución creciente en Uruguay. En la temporada 2004-2005, llegaron 75 buques y desembarcaron 56.167 pasajeros, pero no se dispone de datos de gasto. En la siguiente temporada ya se dispone de todos los datos. En 2005-2006 fueron 99 arribos de buques, 110.827 pasajeros desembarcados y U\$S 4.241.639 el gasto, mientras que en la temporada 2017-2018, las cifras son 140 arribos, 242.466 pasajeros desembarcados y U\$S 7.692.437 el gasto

El presente documento analiza datos de la temporada 2010-2011 y no fue sino hasta la siguiente temporada que Uruguay comenzó a ser país de inicio y fin de itinerario, por tanto, los cruceristas considerados son pasajeros en tránsito.

Tabla 1: Número de cruceristas y monto de gasto por temporada

Cruceristas y gasto en U\$S corrientes			
TEMPORADA	GASTO	PERSONAS	GASTO/PERSONA
2004-2005	S/D	56.167	-----
2005-2006	4.241.639	110.827	38
2006-2007	11.235.466	149.062	75
2007-2008	16.818.273	256.593	66
2008-2009	14.384.413	247.120	58
2009-2010	17.830.909	292.048	61
2010-2011	13.291.304	278.627	48
2011-2012	20.884.091	353.727	59
2012-2013	18.612.467	411.937	45
2013-2014	18.855.505	409.371	46
2014-2015	10.943.470	332.118	33
2015-2016	11.141.587	317.205	35
2016-2017	9.798.264	260.704	38
2017-2018	7.692.437	242.466	32
Fuente: Área de Investigación y Estadística, Ministerio de Turismo			

Además del interés económico del producto turístico cruceros, se señala que es muy habitual que los cruceristas regresen al destino visitado, ya no en esta modalidad, sino como turistas, es decir, alojándose y permaneciendo algunos días en él, con el consiguiente impacto positivo en la exportación de servicios turísticos.

Existen varios antecedentes de análisis en la temática de cruceros: a nivel internacional se pueden citar antecedentes desde 1985 hasta los más recientes en el 2017. Mescon & Vozikis, (1985) estudiaron el impacto económico del turismo de cruceros en el puerto de Miami; un estudio general respecto del transporte de pasaje en este marco se estudia en Herrando et al. (1998); Río & Cruz (2008) realizan un análisis de los perfiles y el gasto realizado por los cruceristas que desembarcan en Bahías de Huatulco (México); Han, W. (2017) estudia el impacto de las actitudes de los consumidores hacia los cruceros en su intención de compra de cruceros; Ma et al. (2018) analizan los factores que motivaron a las compañías de cruceros a seleccionar un puerto específico como puerto base de cruceros en China.

Aunque el turismo de cruceros en Uruguay ha mostrado un importante crecimiento, existen relativamente pocos estudios al respecto: estudios sobre el

gasto de los cruceristas como Risso (2012) y Brida et al. (2014) que en los dos casos analizaron el gasto desde un punto de vista microeconómico, utilizando modelos Heckit. Por otro lado, estudios recientes como Bellani et al. (2017) estudian los determinantes socioeconómicos y comportamentales del gasto en los puertos de desembarco de Uruguay para las temporadas de cruceros 2010 al 2014, mientras que en Brida et al. (2017) introducen el uso de modelos basados en grafos para el estudio de los determinantes del gasto turístico de cruceristas de la temporada 2014-2015.

Por esta razón, el presente trabajo, además de lo metodológico en sí mismo, tiene interés también desde la perspectiva del diseño de políticas públicas orientadas a la promoción del turismo de cruceros, a partir de contar con una mayor información respecto de la caracterización de los cruceristas. En particular, conocer los componentes del gasto y sus determinantes, puede contribuir al mejor diseño de campañas promocionales, así como a definir, junto con los operadores de este producto turístico, cambios en la oferta, de acuerdo al perfil de gasto encontrado.

El artículo está organizado de la siguiente forma: en la segunda sección se explican las técnicas aplicadas y los datos que se utilizan, la tercera sección presenta los resultados, la cuarta sección presenta las conclusiones y finalmente se incluyen bibliografía y anexos.

2. Metodología

Existen muchas definiciones de *datamining* que también se conoce como minería de datos y para eso tomamos la que propone Han et al. (2011): “La minería de datos es el proceso de descubrir patrones interesantes y conocimiento a partir de grandes cantidades de datos. Las fuentes de datos pueden incluir bases de datos, los datos almacenados, la web, otros repositorios de información o de datos que se transmiten en el sistema dinámicamente”. Por lo tanto, a partir de esta definición, lo que interesa es proponer las diferentes técnicas con las que se pueden descubrir patrones en las grandes masas de datos. La tarea que hay que efectuar en la minería de datos es el análisis automático de grandes cantidades de datos donde, para extraer patrones interesantes desconocidos, se pueden agrupar registros de datos, identificar registros poco usuales y, lo más importante, dependencias entre registros para un mismo atributo o para atributos entre sí. Para eso surgen desde el campo de la estadística lo que se conoce como aprendizaje estadístico (*statistical learning*), conceptos que autores como Tibshirani et al. (2013) Hastie et al. (2001) han instaurado, permitiendo clasificar los problemas aprendizaje estadístico como: supervisado (donde en general se dispone de información correspondiente a varios atributos para un conjunto de datos, en los cuales se conoce además un atributo en particular que se toma como variable de respuesta y lo que interesa es poder desarrollar modelos que permitan encontrar relaciones entre la variable de respuesta y sus predictores con fines esencialmente predictivos. Dentro de esta clase de técnicas podemos encontrar la regresión, el análisis discriminante y los métodos CART (*Classification and Regression Trees*)) y no supervisado (donde se dispone de muchos atributos, pero donde no existe una variable de respuesta que

puede usarse para supervisar el análisis. Dentro de estos encontramos todos los métodos de *clustering* o análisis de conglomerados).

Análisis de clusters

Las técnicas de conglomerados o *clusters* tienen como objetivo particionar el conjunto de datos en un número especificado de grupos k , minimizando algún criterio o función objetivo. Los distintos algoritmos pueden clasificarse como jerárquicos (generan una serie de particiones encajadas, de forma que los grupos que se forman en cada paso, comprenden grupos obtenidos a un nivel inferior) o no jerárquicos (necesitan del número de grupos a priori, y a partir de baricentros aleatorios igual a la cantidad de grupos definidos de antemano, en cada paso el algoritmo genera grupos que no se solapan y reasigna las observaciones a cada uno de ellos optimizando la variabilidad dentro de los mismos). Los métodos jerárquicos se caracterizan porque los *clusters* en cada nivel de jerarquía son creados uniendo los *clusters* del nivel inferior. En el nivel inferior cada grupo contiene una única observación y en el nivel más alto existe un solo grupo con el total de observaciones. De esta forma, los *clusters* que se van creando en cada etapa contienen *clusters* del nivel inferior correspondiente (formaciones encajadas). Las estrategias para *clusters* jerárquicos pueden resultar en *clusters* aglomerativos o divisivos. A diferencia de los métodos no jerárquicos, solo se requiere de la definición de una distancia; inicialmente, cada objeto se le asigna a su propio grupo, y entonces los algoritmos proceden iterativamente; en cada etapa unen los dos grupos más similares, continuando hasta que sólo quede un grupo. En cada etapa las distancias definidas entre las agrupaciones se recalculan por la fórmula disimilitud de Lance-Williams, actualizándose de acuerdo con el método de agrupación particular que se utilice. Dentro de los métodos jerárquicos se considera el de *WARD*, que consiste en descomponer la variación total (T) en variación en los grupos W (*within*) y variación entre los grupos B (*between*) y, al estar frente a una partición dada, el método unirá aquellos grupos que produzcan el efecto de hacer mínima la variación *within* en la nueva partición. Dentro de los métodos no jerárquicos se encuentra el de *k-means*, que es uno de los más utilizados, en función de su simplicidad y velocidad de convergencia, que es de orden $(n.p)$, donde n es la cantidad de observaciones y p el número de variables. A partir de un conjunto de n observaciones (x_1, x_2, \dots, x_n) , el método de *k-means* busca encontrar una partición de los n individuos en k subconjuntos con $k \leq n$ (S_1, \dots, S_k) , de manera de minimizar la suma de cuadrados intra clase (SCIC):

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^k \sum_{x_j \in S_i} \|x_j - \mu_i\|^2 \quad (1)$$

siendo μ_i el centroide de los puntos en el grupo $S_i \forall 1 \leq i \leq k$.

Al inicio, todos los centros de los grupos están en la media de las celdas de Voronoi, que se puede interpretar como el conjunto de puntos de los datos que están más cerca del centro de ese grupo que de cualquier otro grupo. El algoritmo funciona de la siguiente manera: en un primer paso se eligen en

forma aleatoria los centros iniciales. Queda entonces la siguiente secuencia m_1, m_2, \dots, m_k de k centros, luego se asigna cada observación al grupo con la media más próxima (es decir que la partición queda determinada con las medias iniciales), se calculan los S_i de la siguiente manera:

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\| \leq \left\| x_p - m_j^{(t)} \right\| \forall 1 \leq j \leq k \right\} \quad (2)$$

donde cada uno de los x_p queda asignado a uno de los $S_i^{(t)}$. Una vez culminado este paso el algoritmo se actualiza calculando las nuevas medias del grupo y finalmente el algoritmo se detiene luego que al reasignar alguna observación a otro grupo no hay cambios menores a una tolerancia prefijada en la SCIC. Los algoritmos habitualmente usados en los paquetes estadísticos están basados en los que plantearon MacQueen (1967), Lloyd (1957) y Forgy (1965). El algoritmo de Hartigan-Wong es el que se usa habitualmente y el que está implementado en R, al trabajar con centros iniciales que se eligen en forma aleatoria, se recomienda ejecutar el algoritmo varias veces ($n=10$) de manera de ver la estabilidad de los resultados. El algoritmo *Partition Around Medoids* (PAM), al igual que el algoritmo *k-medias*, necesita de una partición inicial dada, a partir de un número de grupos definidos previamente. Una diferencia respecto del algoritmo anterior, es que PAM se basa en la búsqueda de k “individuos representativos” (o medoides) entre el conjunto de observaciones, de manera que representen adecuadamente la estructura de los datos en cada partición. Un medoide se podría definir como el objeto perteneciente a un *cluster* o conglomerado, cuyo promedio de disimilaridad a todos los objetos en el conglomerado es mínima, es decir, que se puede considerar como el punto más céntrico de la agrupación considerada. En general PAM es más robusto que *k-medias* y requiere como argumento de entrada solamente la matriz de disimilitudes entre observaciones y no los datos originales. Como desventaja, dicho método es menos eficiente computacionalmente, debido principalmente a la búsqueda de los medoides (Izenman, 2008). La implementación de dicho algoritmo se describe a continuación: se inicializa con la selección al azar de k de los n puntos de datos como los candidatos a medoides en la fase de construcción, se asigna cada observación al grupo con el medoide más próximo, dependiendo de la distancia elegida (euclidiana, Manhattan o Minkowski). Luego se encuentra un mínimo local para la función objetivo, es decir, una solución de tal manera que el cambio de observación con un medoide haga que la función objetivo decrezca (esto se denomina la fase de intercambio). Se repiten los pasos anteriores hasta que los medoides queden estables (es decir que haya cambios en los medoides). En este caso, a diferencia de lo que ocurre con los *k-means*, los centros son los medoides, los que pueden resultar más robustos. Del mismo modo que para el método anterior, se evalúa cuál es la mejor estructura de grupos en base a la Silueta promedio que se obtiene con cada configuración. El método Silueta en realidad no es método de *clustering* en sí mismo, sino un método de interpretación y validación del número de conglomerados o clusters utilizado. Esta técnica permite obtener una representación gráfica para cada grupo, en la cual se aprecia tanto la cercanía como la separación de cada observación respecto del grupo al que fue asignado. Permite obtener una apreciación de la calidad

relativa de los grupos y una visión general de la configuración de los datos. Tiene la ventaja que puede ser utilizada para datos que hayan sido clasificados a través de cualquier método, como por ejemplo k-medias o k-medoides. Detalles de su construcción se puede encontrar en (Izenman, 2008).

Descripción de datos

Se utilizan los datos de la temporada 2010-2011 del Ministerio de Turismo. Se obtiene para cada uno de los 2 únicos puertos de arribo de cruceros de Uruguay (Montevideo y Punta del Este), la cantidad de arribos (escalas de buques), la cantidad de cruceros (buques), las personas que llegan en cada uno (arribadas) y la cantidad de personas desembarcadas, sean pasajeros o tripulantes. Los datos del puerto de Punta del Este, son proporcionados por la Dirección Nacional de Hidrografía y la Dirección de Transporte Fluvial y Marítimo, mientras que para Montevideo, los datos son proporcionados por Operaciones Portuarias de la Administración Nacional de Puertos (ANP).

Los datos recogidos sobre niveles de gasto, características socio-demográficas y de satisfacción de los turistas surgen de encuestas realizadas por el Área de Investigación y Estadística del Ministerio de Turismo. Las unidades de análisis (grupos de cruceristas) provienen de un diseño muestral probabilístico bietápico estratificado por conglomerados, donde las unidades primarias de muestreo (UPM), son los cruceros, que se clasifican en dos estratos (cruceros que desembarcan en Montevideo (estrato 1) y en Punta del Este (estrato 2). Éstos fueron seleccionados mediante muestreo π -ps (con probabilidad proporcional), en este caso al número de cruceristas. En una segunda etapa se selecciona una cantidad fija de 40 grupos de cruceristas, unidades secundarias de muestreo (USM), que tienen, por lo tanto, un número variable de personas. De esta manera, al ser una cantidad fija de grupos de cruceristas, se obtiene un diseño auto-ponderado. Los expansores que tienen los datos fueron calculados y calibrados por el Área de Investigación y Estadística del Ministerio de Turismo, teniendo, para la temporada estudiada, un total de 3176 grupos.

Dado que el interés del presente trabajo es el gasto de los cruceristas, se descartaron los grupos que no presentaban gasto o que tenían un monto de gasto imputado, reduciéndose el estudio a 2311 casos (filas). Esto significa que 2311 grupos de cruceristas realizaron algún tipo de gasto. Sin embargo, se han detectado valores atípicos extremos (*outliers*) de gastos totales per cápita (valores por encima de 500 U\$S per cápita) los cuales no son tomados en cuenta para los análisis para evitar sesgos en los resultados a causa de su influencia en la formación de *clusters*, producto de la asimetría que provocan en las variables. Luego de eliminar dichas observaciones, la base a utilizar contara con un total de 2225 grupos de cruceristas.

En la temporada 2010-2011 llegaron a Uruguay 171 cruceros, 76 a Montevideo y 95 a Punta del Este, desembarcando en total 278.627 personas, de las cuales 99.851 lo hicieron en Montevideo y 178.776 en Punta del Este. La nacionalidad mayoritaria de quienes desembarcaron es brasileña (97.272),

seguida de argentina (95.547); norteamericanos y europeos siguen en importancia, con 36.608 y 30.394 personas respectivamente.

En la misma temporada, el gasto total de las personas desembarcadas fue U\$S 13.291.304, discriminados en U\$S 5.232.921 en Montevideo y U\$S 8.058.383 en Punta del Este. Desagregando el gasto total según la nacionalidad de los cruceristas, los brasileños gastaron el 52% del total (U\$S 6.907.191), los argentinos el 22% (U\$S 2.899.730), los norteamericanos el 11% (U\$S 1.514.731) y los europeos el 8% (U\$S 1.116.230). Si se atiende el gasto per cápita en la temporada de referencia, en dólares, en promedio fue de 48, los brasileños 71, los argentinos 30, los norteamericanos 41 y los europeos 37. De acuerdo con la metodología aplicada por el área de Investigación y Estadística del Ministerio de Turismo, los rubros que se utilizan para desagregación del gasto de los cruceristas son: Shopping (Compras), Alimentación, *Tours*, Transporte y otros gastos.

3. Resultados

Para el análisis de conglomerados o análisis de *clusters* se usa el sistema R (Rcoreteam, 2015) se consideran varios escenarios, en función de cómo se considere las variables de gasto: en un primer escenario se considera las variables de gasto en escala original (estandarizados). Para este tipo de variables en este caso se aplicará el método de *k-means*. Un segundo escenario considera las variables de gasto en escala de proporciones, es decir que los gastos por rubro están relativizados contra el gasto total de cada observación. Para este tipo de variables, se aplicará el método PAM. Un último escenario considera las variables de gasto como variables binarias (ausencia o presencia de gasto en cada rubro).

Este tipo de análisis deja de lado el considerar el gradiente del gasto y, a su vez, por la naturaleza de las variables considerada, exige el uso de algoritmos específicos, así como de determinadas distancias. Es así que para este último escenario se aplicará el método de *cluster* jerárquico Ward utilizando una distancia de tipo binaria.

Se decide considerar solamente los gastos en Tour, Alimentación y Shopping, los cuales se toman como elementos centrales en la caracterización del gasto, dada la forma de la distribución de cada gastos donde Transporte y Resto son marginales .

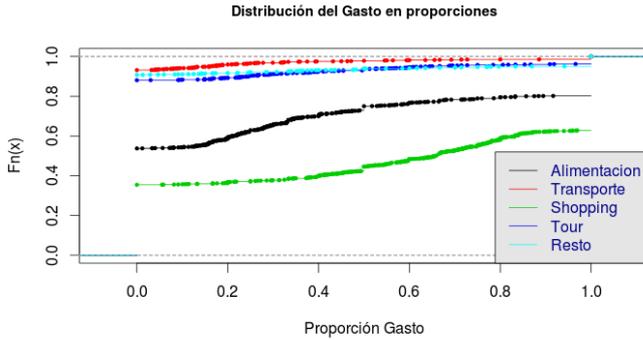


Figura 1: Distribución de los componentes del Gasto en proporciones. Fuente: Elaboración propia

A través de gráficos de dispersión donde la nube de puntos muestra una forma triangular se presenta a continuación las relaciones entre esos tres componentes. Se estudia utilizando las variables de gasto en proporciones tomando los valores originales incluyendo los valores con gasto nulo. Se obtiene entonces una cantidad importante de cruceristas que concentran el gasto en valor nulo (no realiza gasto) o cantidades positivas estrictamente que se repiten, con lo cual al establecerlo en proporciones se produce gran “granularidad”, que es un efecto no deseado al ser variables continuas. Se entiende por “granularidad” la distribución concentrada en pocos puntos, en lugar de una distribución más continua al interior del triángulo. Una posible explicación a esta distribución es la forma como los cruceristas responden a las preguntas de gasto discriminado por rubro que, a pesar de ser variables cuantitativas continuas, las respuestas son en pocos valores que se reiteran frecuentemente. A su vez, resulta difícil captar visualmente el patrón, sobre todos en las aristas de triángulo que aparece en la Figura 3.

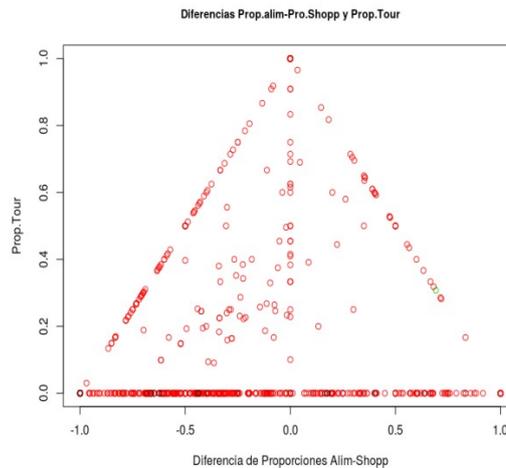


Figura 2: Relación entre proporciones de Gasto, Shopping y Tour. Fuente: Elaboración propia

Para superar esta dificultad se modifican los valores de las proporciones de gasto, incluyendo una perturbación aleatoria con distribución uniforme, lo que

permite ver exactamente cómo es la masa de cruceristas que están en los vértices del triángulo (donde vale cero el gasto per cápita). La Figura 2 muestra que hay una cantidad importante de cruceristas que no gastan en Tour y que no gastan en Alimentación: es el vértice del triángulo que tiene coordenadas (-1,0); lo mismo sucede para los cruceristas que sólo gastan en Alimentación, al tener coordenadas (1,0). Los cruceristas que aparecen proyectados en el punto (0,0) son aquellos que no gastan en Tour y que puede ser que gasten sólo en Alimentación y en Shopping, repartiendo en mitades el gasto, o que sean cruceristas que sólo gastan los rubros del estrato 2 de gasto (Transporte u otro gasto). El punto con coordenadas (0,1), que resulta interesante y que aparece con una superposición de puntos importantes, es el de cruceristas que gastan sólo en Tour. El resto de los cruceristas son los que reparten el gasto entre los tres componentes.

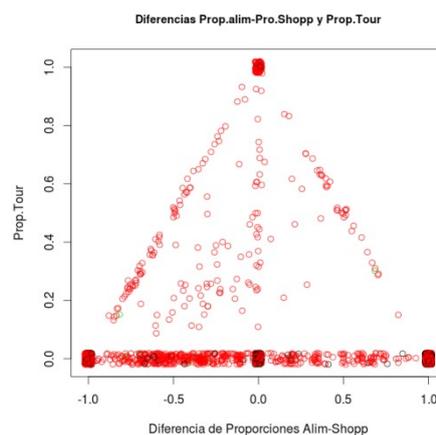


Figura 3: Relación entre proporciones de Gasto, Shopping y Tour (perturbación aleatoria). Fuente: Elaboración propia

3.1 Análisis de clusters con los gastos en escala original

Para el análisis de las variables en la escala original se procede en una primera instancia a estandarizar los datos. En este caso se utiliza la metodología *k-means* con cinco grupos, que a partir del estudio de la varianza intra clase (VIC) que se visualiza en la Figura 4, presenta una inflexión o descenso en el ritmo de crecimiento, lo cual indica que el número de grupos predeterminado es apropiado:

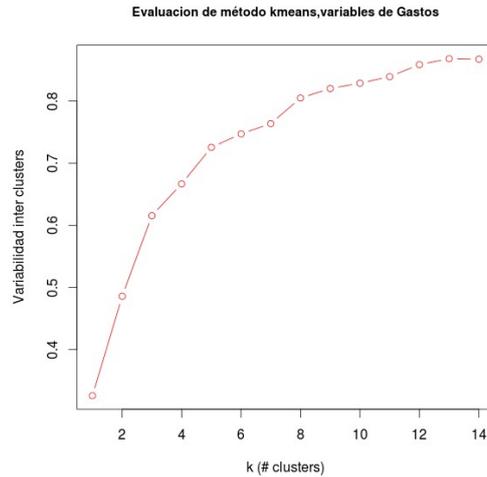


Figura 4: Variabilidad intra clase con método k-means. Fuente: Elaboración propia

En Tabla 2 se aprecian los promedios dentro de cada grupo, de las variables gasto por rubro y del gasto total per cápita. Al estar estandarizados los datos, los valores negativos representan gastos que se encuentran por debajo de la media global por rubro.

	Gasto Tour	Gasto Alim	Gasto Shopping	N
1	-,0179	-0,267	-0,454	1417
2	-0,212	-0,114	2,908	134
3	3,810	0,0823	-0,079	102
4	-0,142	-0,158	0,815	384
5	-0,274	2,368	-0,269	188
Total	0	0	0	2225

Tabla 2: Promedios de cada tipo de Gasto según grupos por método k-means. Fuente: Elaboración propia

A partir de la Tabla 2 y las Tablas 6, 7, 8 y 9 del Anexo se puede caracterizar los grupos obtenidos de la siguiente manera: el grupo 1 se caracteriza por ser el *cluster* de mayor tamaño (contiene el 63,7% de los grupos de cruceristas encuestados), los niveles de gastos promedios se encuentran por debajo de la media de cada rubro respectivamente. Además este grupo está conformado mayoritariamente por parejas de turistas y el 64% aprox. de los turistas menciona que es su visita por primera vez al país; el grupo 2 se caracteriza por poseer el segmento de cruceristas que realizan mayor gasto en el rubro *Shopping* (los valores se encuentran a casi tres desvíos respecto del gasto medio de este rubro). Su tamaño es casi diez veces menor que el grupo 1. Está conformado por grupos de cruceristas de 2, 3 y 4 personas. El porcentaje de turistas que vistan por primera vez es de 69% aproximadamente; el grupo 3 concentra los turistas que se destacan por un nivel de gasto significativo por

encima de la media en el rubro *Tour*. El mismo incluye a 102 grupos de cruceristas que en su mayoría se componen de 2 y 4 personas. Es el grupo con mayor porcentaje de turistas que visitan por primera vez el país (85% aprox.); el grupo 4 es otro segmento de turistas que poseen un nivel de gasto significativo en el rubro *Shopping* aunque menor al grupo 2. Su tamaño de 384 grupos de turistas. Se compone mayoritariamente por grupos de 2 personas, el 66% declara que es su visita por primera vez al país y posee el mayor índice de retorno por segunda vez (17% aprox.); el último grupo se compone de los grupos de cruceristas que realizaron el mayor nivel de gasto en el rubro Alimentación. Está conformado por grupos de 2 y 4 personas. El 72% aprox. de los turistas visitaron por primera vez al país. En cuanto a la composición por sexo y conformación etaria de los grupos, se puede apreciar que los resultados inter clases arrojan resultados similares: predominan turistas del sexo femenino (en promedio más del 80% dentro de cada cluster), y predominan las personas mayores a 65 años (en promedio representan más del 60% inter cluster).

3.2 Análisis de clusters con los gastos en escala de proporciones

En una segunda instancia y utilizando las variables de gasto por rubros per cápita en proporciones respecto del gasto total per cápita, se aplica la metodología PAM. Para dicho algoritmo y en base al análisis del gráfico silueta (Figura 5), también se trabaja con cinco grupos.

Para poder visualizar mejor cómo es la calidad de los grupos generados se evalúa la misma a través de la generación de varios gráficos de silueta con submuestras en cada grupo del 20 % las que se seleccionan mediante muestreo aleatorio simple. Se ensayaron varias muestras donde siempre se ve que hay un grupo mayoritario muy homogéneo, mientras que hay un grupo muy heterogéneo que, con valores negativos del índice en silueta, son un indicio que hay un grupo cuyos integrantes (grupos de cruceristas) se parecen más a los grupos de cruceristas de los otros clusters.

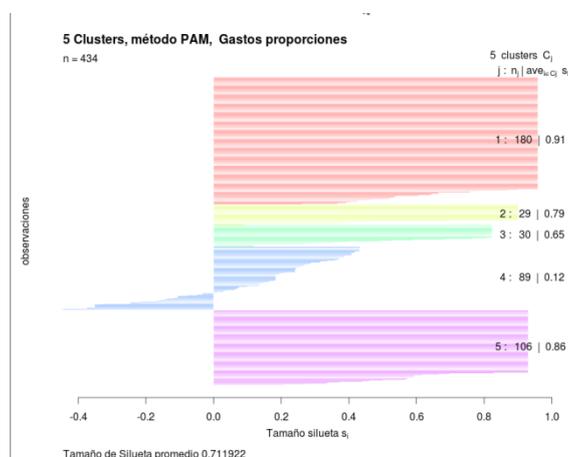


Figura 5: Gráfico Silueta utilizando método PAM. Fuente: Elaboración propia

En la Tabla 3 y Tablas 10, 11, 12 y 13 del Anexo la segmentación de viajeros es en base a gasto en proporciones (que ignora el nivel de gasto total), utilizando

la metodología PAM con cinco grupos, muestra las siguientes características: el grupo 1 es el que presenta mayor proporción de gasto en el rubro Tour en promedio (65.4%). Está conformado mayoritariamente por grupos de dos, tres y cuatro personas. Es el grupo con mayor porcentaje de turistas que visitan por primera vez (77% aprox.); El grupo 2 se caracteriza por poseer mayor proporción promedio de gasto en *Shopping* (57.3%) y Alimentación (34.1%). Predominan los grupos de dos personas (51% aprox.). El 66% aproximadamente visita por primera vez el país; en el grupo 3, en promedio los individuos destinan el 96.2% de su gasto total al rubro Alimentación. De similares características en cuanto a conformación de grupos de personas. El 63% aproximadamente visita por primera vez al país mientras que alrededor de un 12% de turistas de este grupo afirma que ha visitado más de cinco veces nuestro país; el grupo 4 se destaca por poseer las proporciones de gastos promedios más bajos. En este grupo se encuentra el porcentaje más alto de grupos de cruceristas individuales, de dos personas y más de cinco personas. Casi el 75% declararon estar visitando por primera vez el país; por último, el grupo 5 se caracteriza por poseer un gasto predominante en el rubro *Shopping* (97.1%). El 66% ha declarado que visita por primera vez el país mientras que el 15% declaró estar visitando por segunda vez al país.

	Gasto Tour	Gasto Alim	Gasto Shopping	N
1	0,654	0,074	0,066	212
2	0,035	0,341	0,573	378
3	0,004	0,962	0,009	529
4	0,016	0,047	0,005	144
5	0,008	0,15	0,971	188
Total	0	0	0	2225

Tabla 3: Promedios de cada tipo de Gasto según los grupos por método PAM.
Fuente: Elaboración propia

3.3 Análisis de cluster con los gastos en escala dicotómica

Para este análisis se modifica la perspectiva del análisis, tal como se mencionó antes, al considerar las variables de gasto como variables binarias de ausencia/presencia. Se construye tipologías de turistas, teniendo que recurrir a algoritmos adecuados, en este caso el método de *cluster* jerárquico de Ward, usando para ello distancia de tipo binaria. Analizando el dendrograma (Figura 5) que surge al aplicar el algoritmo anteriormente mencionado, se propone un corte en el árbol de agregación, de manera de tener cinco grupos con los que se obtienen los siguientes resultados.

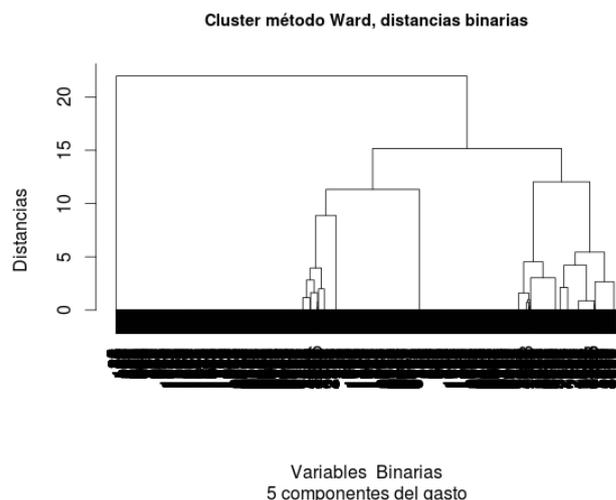


Figura 6: Dendrograma utilizando método Ward. Fuente: Elaboración propia

De la Tabla 4 se observa que los tamaños de los *clusters* son más homogéneos. Además se puede describir los cinco grupos de la siguiente forma: el grupo 1 concentra los turistas que destinan únicamente su gasto al rubro Shopping. Se caracteriza por poseer grupos de dos, tres y cuatro personas. El 65.60% de los turistas de dicho grupo visita por primera vez el país y es el grupo con mayor grado de retorno por segunda vez (15% aprox.); el grupo 2 está conformado por aquellos turistas que destinan en promedio el 87.3% del gasto al rubro Alimentación y el 87% al rubro Shopping. Mayoritariamente conformado por grupos de dos y cuatro personas. El 67% visita por primera vez el país; el grupo 3 concentra los turistas que solo gastan en el rubro Alimentación. Está conformado mayoritariamente por grupos de dos, tres y cuatro personas. El 59% del *cluster* está visitando por primera vez; el grupo 4 se caracteriza por estar integrado de turistas que gastan poco en los tres rubros. Mayoritariamente conformado por grupos de dos personas y posee mayor cantidad de grupos unitarios. El 71% de los turistas de este *cluster* visitan por primera vez; el grupo 5 concentra los turistas que destinan su gasto al rubro Tour y en segundo lugar al rubro Shopping. Mayoritariamente concentrado por grupos de dos, tres y cuatro personas por grupo. Posee el mayor porcentaje de visitas por primera vez (77%). Al igual que en las metodologías anteriores las variables socio-demográficas como edad y sexo no tienen mayor disimilaridad entre grupos.

	Gasto Tour	Gasto Alim	Gasto Shopping	N
1	0,00	0,00	1,00	860
2	0,014	0,873	0,870	539
3	0,00	1,00	0,00	412
4	0,114	0,261	0,195	210
5	1,00	0,348	0,500	204
Total	0,119	0,462	0,645	2225

Tabla 4 Promedios de cada tipo de Gasto según grupo por método jerárquico.
Fuente: Elaboración propia

Conclusiones

El objetivo del presente trabajo es obtener una caracterización del gasto de turistas cruceristas que visitan Uruguay. Para llevar a cabo dicho objetivo se aplicaron metodologías relacionadas con la minería de datos, en particular se aplicaron metodologías de análisis de *clusters* jerárquicos y no jerárquicos. Las variables que se toman en cuenta para la conformación de los grupos son las variables de gasto per cápita en los rubros de Alimentación, Tour y Shopping. Se aplica por un lado el algoritmo de k-means sobre las variables de gastos en escala original; en un segundo escenario se aplica el método PAM sobre las variables de gasto en proporciones y un tercer escenario se construye aplicando un algoritmo jerárquico de Ward con distancia binaria al considerar las variables de gasto como presencia o ausencia de los mismos. Las diferentes tipologías se analizan y asocian con las características socio-demográficas de los turistas, como ser: número de personas de los grupos encuestados, composición por sexo de dichos grupos, composición etaria de los grupos, número de visitas realizadas a nuestro país. Se utilizan datos de la temporada 2010-2011 del Ministerio de Turismo y Deporte.

Las tres estrategias utilizadas tienen como elemento común que parten de la base de que la segmentación de los cruceristas se realiza independientemente de características socio-demográficas de dichos turistas, es decir únicamente se utilizan las variables de gasto per cápita de los rubros señalados anteriormente para la conformación de los *clusters*. También es importante señalar que en las tres metodologías se valida que el número de adecuados son cinco.

Las caracterizaciones encontradas bajo las tres metodologías si bien en detalle pueden encontrarse diferencias, coinciden en que se detecta al menos un grupo de cruceristas en que su nivel de gasto en todos los rubros son relativamente bajos (por debajo de la media de cada uno). Mayoritariamente

conformado por grupos de turistas individuales o de dos turistas. También se ha encontrado coincidencia en la existencia de al menos un grupo en que los turistas concentran su gasto en el rubro *Shopping* y/o Alimentación. Estos grupos se conforman en general por grupos de turistas de dos, tres y hasta cuatro personas. Por otro lado se detecta también un grupo de turistas que destinan casi exclusivamente su gasto en el rubro *Tour* siendo en general éste grupo que el posee el máximo porcentaje de turistas que visitan por primera vez al país. Por otra parte, en las tres metodologías, las variables referentes a la composición por sexo y conformación etaria de los grupos arrojan resultados inter clases similares: predominan turistas del sexo femenino (en promedio más del 80% dentro de cada *cluster*), y predominan las personas mayores a 65 años (en promedio representan más del 60% inter *cluster*).

Bibliografía

Bellani, A., Brida, J. G., & Lanzilotta, B. (2017). El turismo de cruceros en Uruguay: determinantes socioeconómicos y comportamentales del gasto en los puertos de desembarco. *Revista de Economía del Rosario*, 20(1), 26.

Brida, J. G., Fasone, V., Scuderi, R., & Zapata-Aguirre, S. (2014). Research note: Exploring the determinants of cruise passengers' expenditure at ports of call in Uruguay. *Tourism Economics*, 20(5), 1133-1143.

Brida, J. G., Santiñaque, F., & Lanzilotta, B. (2017). Modelos basados en grafos: una aplicación al estudio del gasto de cruceristas en Uruguay//Graph-Based Models: An Application to the Study of Cruise Passengers' Expenditure in Uruguay. *Revista de Métodos Cuantitativos para la Economía y la Empresa*, 24, 270-291.

Forgy, E.W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometric Society Meeting, Riverside, California, 1965. Abstract in *Biometrics* 21 (1965) 768.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Han, W. (2017). Consumer Attitudes and Purchase Intentions of Cruises in China.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). Elements of statistical learning. Springer Verlag.

Herrando, J. A., Chapapría, V. E., & Piqueras, V. Y (1998). El Turismo de Cruceros como Modalidad del Transporte de Pasaje.

Izenman, A. J. (2008). Modern multivariate statistical techniques (Vol. 1). New York: Springer.

Lloyd, S. (1957). Least squares quantization in pcm. Bell Telephone Laboratories Paper, Marray Hill.

Ma, M. Z., Fan, H. M., & Zhang, E. Y. (2018). Cruise homeport location selection evaluation based on grey-cloud clustering model. *Current Issues in Tourism*, 21(3), 328-354.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Symposium on Math,

Statistics, and Probability (pp. 281–297). Berkeley, CA: University of California Press.

Mescon, T. S., & Vozikis, G. S. (1985). The economic impact of tourism at the port of Miami. *Annals of Tourism Research*, 12(4), 515-528.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

Río, M. C., & Cruz, M. T. K. (2008). Perfil y análisis del gasto del crucerista: El caso de Bahías de Huatulco (México). *Cuadernos de turismo*, (22), 47-78.

Risso, W.A. (2012). “El gasto de los cruceristas en Uruguay 2008–2010”. *Revista de Turismo y Patrimonio*, 10 (3), 393–406.

Tibshirani, T. ; Gareth, J.; Hastie, T. y Witten, D. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer. Recuperado a partir de <http://link.springer.com/content/pdf/10.1007/978-3-319-00327-6.pdf>

Anexo

Tablas de resultados de los grupos creados considerando las variables de gasto en escala original:

<i>Cluster por k-means con variables gasto en escala original</i>						
Nro. de Personas	1	2	3	4	5	Total
1	8,78%	6,41%	0,52%	5,05%	2,26%	6,87%
2	54,84%	35,17%	30,35%	46,44%	37,60%	49,10%
3	13,78%	20,78%	15,23%	19,01%	12,50%	15,07%
4	14,00%	25,23%	29,33%	14,72%	24,34%	16,68%
5	2,55%	3,36%	16,64%	2,99%	3,71%	3,59%
Más de cinco	6,07%	9,04%	7,93%	11,80%	19,58%	8,69%
Total	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

Tabla 5: Porcentaje de cruceristas por método k-means según número de personas. Fuente: Elaboración propia

<i>Cluster por k-means con variables gasto en escala original</i>						
% Hombres	1	2	3	4	5	Total
0%	20,05%	18,98%	13,28%	15,06%	9,10%	17,64%
(0- 50]%	68,01%	66,74%	68,70%	76,83%	80,37%	70,75%
[50-100%)	6,89%	9,67%	16,80%	5,93%	7,63%	7,52%
100 %	5,05%	4,62%	1,21%	2,18%	2,90%	4,09%
Total	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

Tabla 6: Porcentaje de cruceristas por método k-means según porcentaje de hombres. Fuente: Elaboración propia

	<i>Cluster por k-means con variables gasto en escala original</i>					
% adultos	1	2	3	4	5	Total
0%	14,76%	6,34%	18,39%	8,66%	11,18%	13,04%
(0- 50]%	9,12%	11,57%	18,31%	10,78%	15,53%	10,71%
[50-100%)	9,18%	12,07%	15,53%	16,26%	20,25%	12,05%
100 %	66,94%	70,02%	47,78%	64,31%	53,03%	64,19%
Total	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

Tabla 7: Porcentaje de cruceristas por método k-means según porcentaje de adultos. Fuente: Elaboración propia

	<i>Cluster por k-means con variables gasto en escala original</i>					
Visitas	1	2	3	4	5	Total
1-Primera vez	64,11%	68,85%	84,61%	65,56%	72,04%	66,59%
2-Dos veces	14,17%	14,01%	7,19%	16,99%	8,15%	13,66%
3-Tres veces	7,42%	8,14%	3,87%	7,65%	7,44%	7,30%
4-Cuatro veces	3,68%	2,14%		1,50%	1,63%	2,80%
5-Cinco veces	0,97%			2,09%	0,47%	1,01%
6-Más de cinco veces	9,65%	6,86%	4,32%	6,21%	10,27%	8,64%
Total	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

Tabla 8: Porcentaje de cruceristas por método k-means según número de visitas. Fuente: Elaboración propia

Tablas de resultados de los grupos creados considerando las variables de gasto como proporciones:

Cluster por PAM con variables gasto en proporciones						
Número de Personas	1	2	3	4	5	Total
1	2,64%	4,25%	7,44%	11,05%	7,97%	6,87%
2	46,30%	51,29%	48,10%	52,00%	49,01%	49,10%
3	17,74%	15,27%	13,48%	9,88%	16,12%	15,07%
4	17,77%	17,16%	18,90%	12,81%	15,49%	16,68%
5	7,88%	4,15%	3,82%	3,42%	2,21%	3,59%
Más de cinco	7,66%	7,87%	8,25%	10,85%	9,21%	8,69%
Total	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

Tabla 9: Porcentaje de cruceristas por método PAM según número de personas. Fuente: Elaboración propia

Cluster por PAM con variables gasto en proporciones						
% de hombres	1	2	3	4	5	Total
0%	19,29%	11,83%	16,01%	13,74%	21,27%	17,64%
(0- 50]%	67,98%	76,39%	69,99%	71,66%	69,38%	70,75%
[50-100%)	10,14%	9,08%	8,47%	5,72%	5,96%	7,52%
100 %	2,60%	2,70%	5,53%	8,88%	3,40%	4,09%
Total	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

Tabla 10: Porcentaje de cruceristas por clusters PAM según porcentaje de hombres. Fuente: Elaboración propia

<i>Cluster por PAM con variables gasto en proporciones</i>						
% de adultos	1	2	3	4	5	Total
0%	14,73%	9,81%	12,93%	6,52%	15,08%	13,04%
(0- 50]%	13,99%	13,63%	10,18%	9,33%	9,25%	10,71%
[50-100%)	13,27%	10,26%	13,07%	13,02%	11,74%	12,05%
100 %	58,01%	66,31%	63,82%	71,13%	63,93%	64,19%
Total	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

Tabla 11: Porcentaje de cruceristas por clusters PAM según porcentaje de adultos. Fuente: Elaboración propia

<i>Cluster por PAM con variables gasto en proporciones</i>						
Visitas	1	2	3	4	5	Total
1-Primera vez	77,13%	65,55%	62,59%	74,85%	65,64%	66,59%
2-Dos veces	10,37%	14,23%	12,71%	11,61%	15,11%	13,66%
3-Tres veces	4,25%	7,36%	9,54%	0,74%	7,68%	7,30%
4-Cuatro veces	1,59%	3,53%	2,67%	1,39%	3,07%	2,80%
5-Cinco veces	0,19%	1,70%	0,75%	2,11%	0,89%	1,01%
6-Más de cinco veces	6,47%	7,62%	11,74%	9,31%	7,61%	8,64%
Total	100,00 %	100,00%	100,00%	100,00%	100,00%	100,00%

Tabla 12: Porcentaje de cruceristas por clusters PAM según número de visitas. Fuente: Elaboración propia

Tablas de resultados de los grupos creados considerando las variables de gasto binarias:

<i>Cluster por Ward con variables gasto binarias</i>						
Número de Personas	1	2	3	4	5	Total
1	8,36%	3,96%	7,60%	10,36%	2,44%	6,71%
2	48,56%	47,92%	47,85%	49,77%	45,83%	48,07%
3	15,71%	17,62%	12,29%	10,28%	16,29%	15,00%
4	14,20%	20,75%	17,59%	15,13%	15,74%	16,69%
5	2,07%	3,00%	3,39%	5,09%	9,09%	3,61%
Más de cinco	11,09%	6,74%	11,29%	9,37%	10,61%	9,92%
Total	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

Tabla 13: Porcentaje de cruceristas por clusters jerárquico según número de personas. Fuente: Elaboración propia

<i>Cluster por Ward con variables gasto binarias</i>						
% de Hombres	1	2	3	4	5	Total
0%	21,94%	13,28%	16,73%	14,60%	15,35%	17,44%
(0- 50]%	67,69%	75,61%	70,23%	70,61%	71,84%	70,78%
[50-100%)	6,86%	8,01%	7,83%	6,18%	11,31%	7,76%
100 %	3,51%	3,10%	5,21%	8,61%	1,50%	4,01%
Total	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

Tabla 14: Porcentaje de cruceristas por clusters jerárquico según porcentaje de hombres. Fuente: Elaboración propia

<i>Cluster por Ward con variables gasto binarias</i>						
% de adultos	1	2	3	4	5	Total
0%	14,64%	9,45%	13,80%	6,41%	14,97%	12,56%
(0- 50]%	9,46%	12,51%	9,25%	9,55%	14,13%	10,64%
[50-100%)	12,61%	12,40%	12,87%	11,58%	15,45%	12,84%
100 %	63,29%	65,64%	64,08%	72,46%	55,45%	63,96%
Total	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

Tabla 15: Porcentaje de cruceristas por clusters jerárquico según porcentaje de adultos. Fuente: Elaboración propia

Cluster por Ward con variables gasto binarias						
Nro de Visitas	1	2	3	4	5	Total
1-Primera vez	65,60%	66,93%	58,94%	71,23%	77,12%	66,28 %
2-Dos veces	15,35%	13,62%	13,16%	10,71%	10,67%	13,55 %
3-Tres veces	6,98%	7,79%	10,09%	3,29%	5,25%	7,30%
4-Cuatro veces	3,16%	3,43%	5,28%	2,87%	0,71%	3,37%
5-Cinco veces	0,90%	1,22%	0,85%	1,45%	0,48%	0,97%
6-Más de cinco veces	8,01%	7,01%	11,69%	10,45%	5,77%	8,53%
Total	100,00%	100,00%	100,00%	100,00%	100,00%	100,00 %

Tabla 16: Porcentaje de cruceristas por clusters jerárquico según número de visitas. Fuente: Elaboración propia

Instituto de Estadística

Documentos de Trabajo



Eduardo Acevedo 1139. CP 11200 Montevideo, Uruguay

Teléfonos y fax: (598) 2410 2564 - 2418 7381

Correo: ddt@iesta.edu.uy

www.iesta.edu.uy

Área Publicaciones

Febrero, 2020

DT (1/20)