Una aplicación de técnicas de *cluster* utilizando variables del sector telecomunicaciones de países de América Latina

José Alberto Candelaria Barrera y Christian James Aguilar Armenta

Resumen

El propósito de las técnicas de agrupamiento conocidas como análisis de *cluster* consiste en integrar elementos en grupos homogéneos de acuerdo a sus similitudes. En este trabajo agrupamos a una selección de países de América Latina, incluyendo a México, utilizando variables del sector de las telecomunicaciones.

I. Introducción.

La motivación para realizar agrupamientos de naciones que pertenecen a la región de América Latina se debe al interés que tenemos por conocer qué tanto se asemeja, o no, el estado actual de las telecomunicaciones en México con el de países pertenecientes a dicha región. Es decir, en virtud de que no se trata de naciones homogéneas es más que lógico pensar que se pueden crear grupos diferentes de naciones de acuerdo a sus características particulares. Las variables del sector de telecomunicaciones que elegimos corresponden al segundo trimestre del año 2018 y son las siguientes:

Conexiones por tecnología de red. Porcentaje de conexiones 2G, 3G y 4G del total de conexiones en la red al final del periodo.

Espectro. Cantidad de espectro asignado medido en MHz para servicios móviles.

Conexiones de IoT. Número de conexiones de IoT en bandas móviles concesionadas, ponderada mediante la variable de suscriptores a servicios móviles.

Suscriptores únicos. Número total de usuarios únicos que se encuentran suscritos a servicios móviles al final del periodo, excluyendo M2M (en inglés, machine to machine).[1]

Penetración de mercado de suscriptores únicos a servicios móviles. Número "real" de personas que utilizan servicios móviles (incluyendo aquellos que posean múltiples tarjetas SIM y/o dispositivos conectados a la red móvil), expresado como porcentaje de la población total de mercado.[2]

Penetración de mercado de suscriptores únicos a Internet móvil. Número "real" de personas que utilizan el Internet móvil (incluyendo aquellos que posean múltiples tarjetas SIM y/o dispositivos conectados a la red móvil), expresado como porcentaje de la población total de mercado.[3]

Conexiones por tipo de dispositivo. Porcentaje de conexiones con smartphones, teléfono básico o dispositivo sólo datos del total de las conexiones al final del periodo.

[1] Definición de GSMA Intelligence para la variable de Unique Susbcribers. https://www.gsmaintelligence.com/

[2] De acuerdo a GSMA el término población se refiere a la población total (incluyendo niños), cuya fuente es UN World Population Prospects 2017 edition. https://www.gsmaintelligence.com/

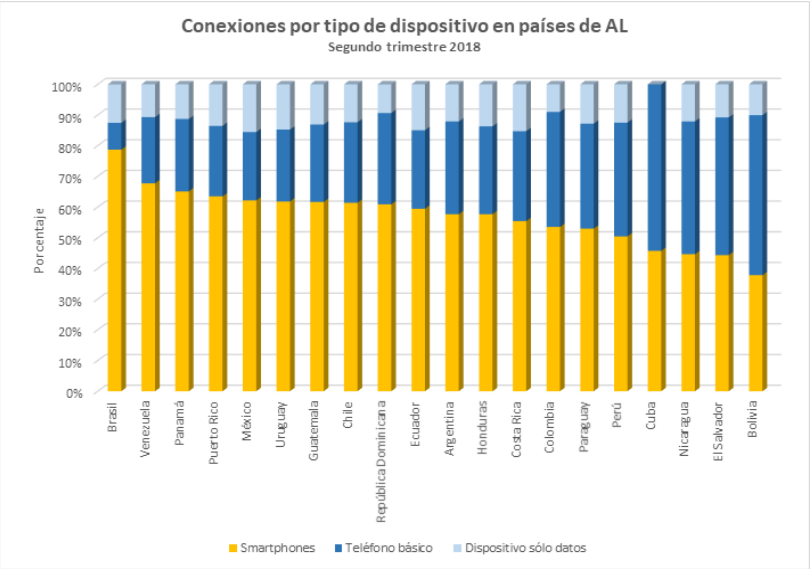
[3] De acuerdo a GSMA el término población se refiere a la población total (incluyendo niños), cuya fuente es UN World Population Prospects 2017 edition. https://www.gsmaintelligence.com/

Conexiones por tipo de contratación. Porcentaje de conexiones prepago y pospago del total de las conexiones al final del periodo.

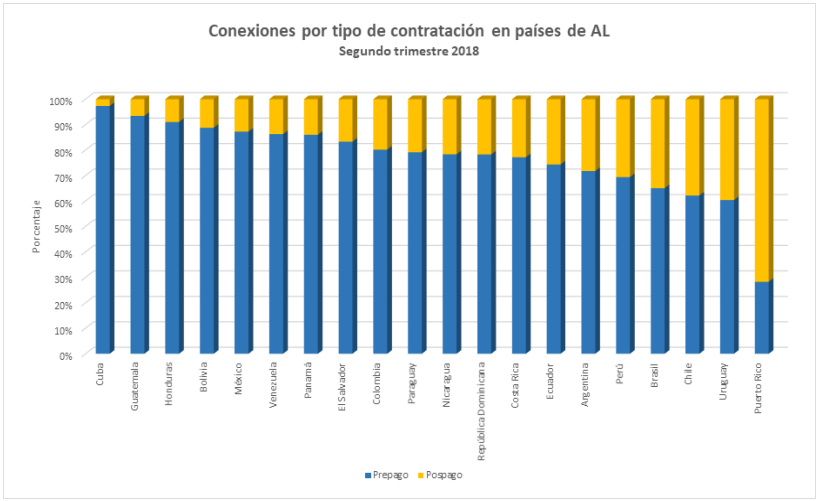
II. Datos

La fuente de nuestros datos es GSMA Intelligence,[4] con excepción de la información del espectro asignado para servicios móviles que es de 5G Américas.[5] Cada una de estas variables se grafica con el fin de comprender el estado actual del sector de las telecomunicaciones en nuestra muestra de países, resaltando con un color diferente, para la mayoría de los casos, los datos de México. Es así que en la primera gráfica se muestra el porcentaje de conexiones por tipo de dispositivo en orden descendente para el caso de los smartphones. Brasil se ubica en el primer lugar de la región con un porcentaje del 78%; en tanto que México se encuentra en el cuarto lugar con el 62%. La segunda gráfica muestra el porcentaje de conexiones por tipo de contratación, mostrándose en orden descendente para el caso de los contratos de prepago. En este sentido, Cuba se encuentra en el primer lugar regional con el 97%, mientras que México ocupa el quinto lugar con el 87%. Por otra parte, en el lado de los contratos de pospago Puerto Rico tiene el porcentaje más elevado con el 71%; en tanto que México se acerca al 13%. La tercera gráfica se refiere al porcentaje de conexiones por tecnología de red, ordenadas de forma descendente para la tecnología 4G. Es así que Brasil muestra el porcentaje más elevado de conexiones con tecnología 4G al alcanzar un porcentaje del 53%. Por su parte México se encuentra muy a la zaga en este tipo de conexiones con un porcentaje de solamente el 22.5%, ubicándose debajo de países como Argentina, Chile, Perú o Ecuador. En lo referente al espectro asignado en cada país de la región tenemos a Brasil y México ocupando las primeras posiciones con 609 MHz y 584 MHz, respectivamente. Las gráficas 5 y 6 muestran la penetración de mercado de suscriptores únicos a servicios móviles y de suscriptores únicos a internet móvil. En el primer caso México se ubica en la posición catorce de veinte con un porcentaje del 63.5% muy detrás del primer lugar que ocupa Chile con un porcentaje del 83.6%. En el segundo caso México ocupa el octavo lugar con el 53.06% por detrás de Brasil (56.3%) y alejado de Puerto Rico que se encuentra en primer lugar con un 63.5%. Finalmente, se observa en la gráfica siete a las conexiones de IoT en bandas móviles concesionadas, dichas conexiones están ponderadas por la variable de número de suscriptores a servicios móviles como una forma de normalización. En este caso nuevamente Brasil y México ocupan los dos primeros lugares.

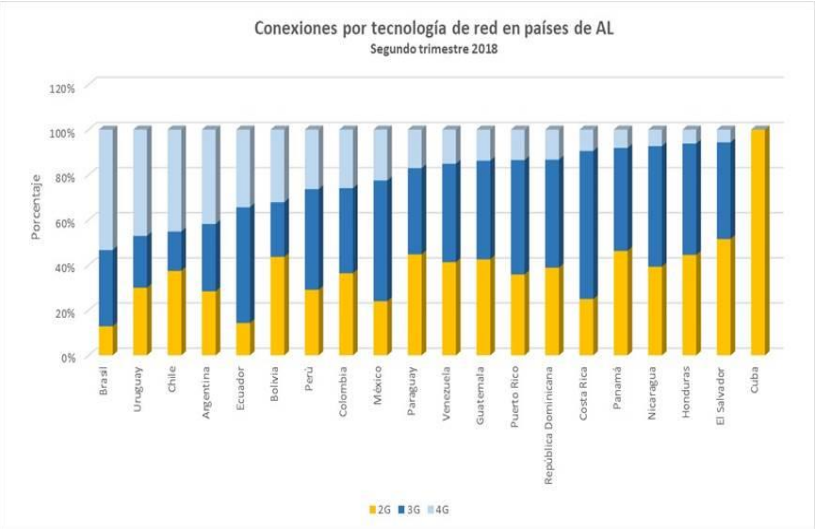
[4] https://www.gsmaintelligence.com/
[5] http://www.5gamericas.org/en/resources/statistics/statistics-latin-america/



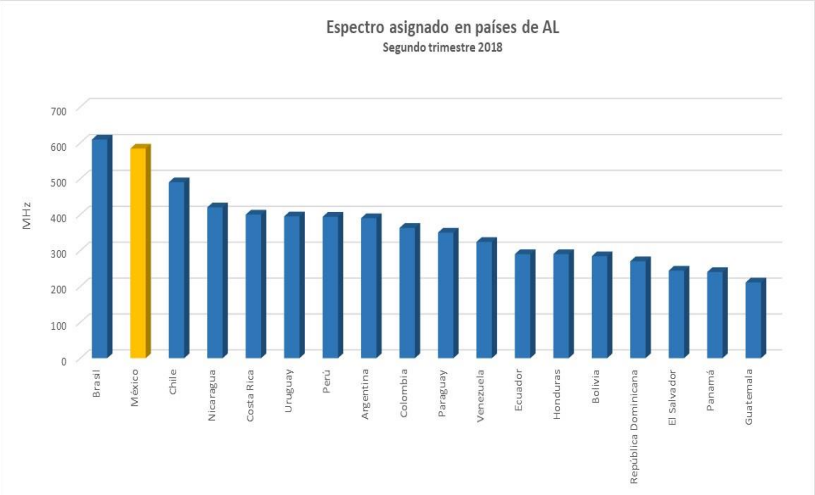
Gráfica 1



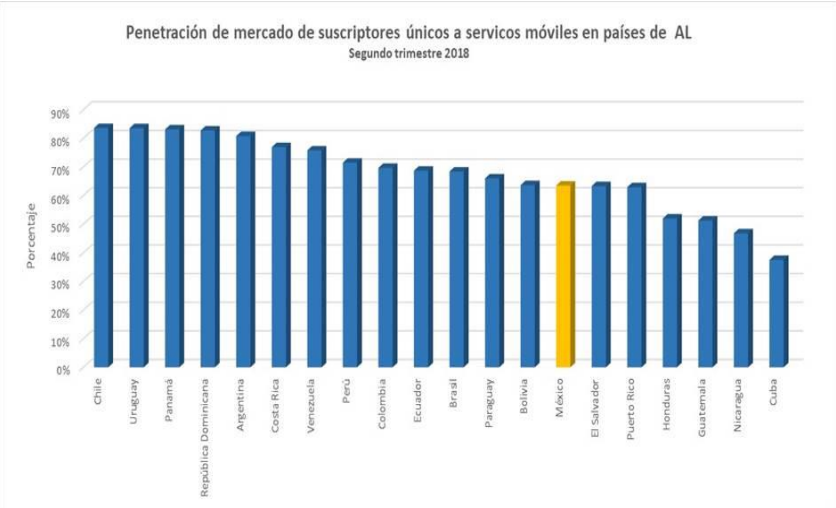
Gráfica 2



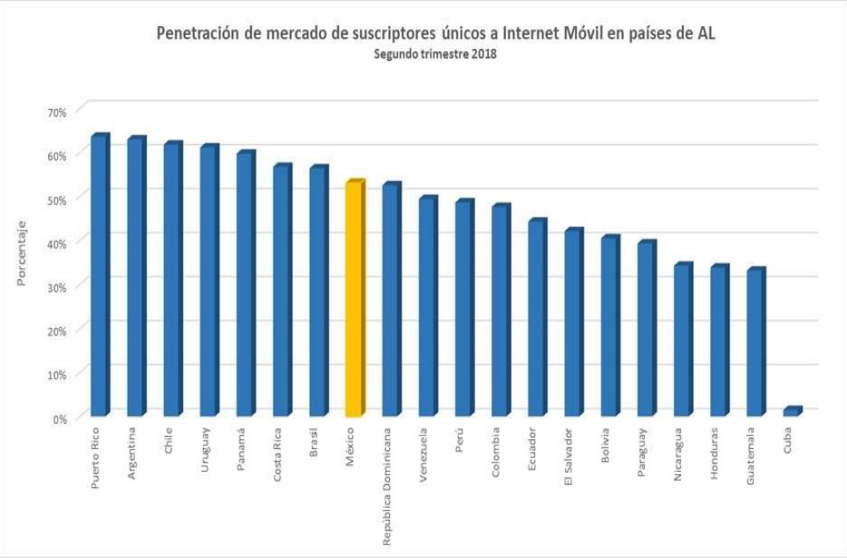
Gráfica 3



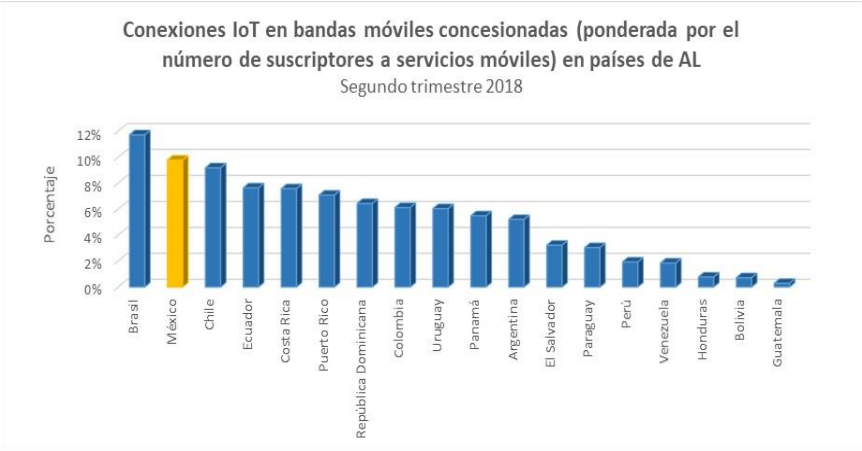
Gráfica 4



Gráfica 5



Gráfica 6



Gráfica 7

III. Análisis de *clusters*

Una vez que hemos graficado nuestros datos procedemos a realizar nuestro análisis de *cluster* y generamos dos diseños. En el primero de ellos utilizamos las siguientes variables: conexiones por tecnología de red (2G, 3G, 4G), espectro asignado medido en MHz para servicios móviles por país, penetración de mercado de suscriptores únicos a servicios móviles, penetración de mercado de suscriptores únicos a Internet móvil y conexiones de IoT ponderadas por la variable de número de suscriptores a servicios móviles. El objetivo de este primer *cluster* es vincular a la variable de espectro asignado para servicios móviles que tiene cada país de nuestra muestra, con las variables de accesos a tecnología. Esto nos servirá de referencia para observar si existe cierta endogeneidad entre estas variables y qué tanto puede ayudar a diferenciar grupos de países a partir de este diseño.

El segundo diseño de *cluster* es similar al primero, pero se añaden las variables de conexiones por tipo de contratación (prepago y pospago) y conexiones por tipo de dispositivo (smartphone, teléfono básico y dispositivo de sólo datos). En el caso de la variable de conexiones por tipo de contratación sobresale en la mayoría de los países, con excepción de Puerto Rico, la de prepago; en tanto que el tipo de dispositivo predominante en los países de la región es el smartphone, con las excepciones de Cuba, Nicaragua, El Salvador y Bolivia. Nuevamente, la intención es construir un diseño de *cluster* que nos permita identificar qué tanto se asemejan o diferencian los países a partir de este diseño.

Una vez que hemos decidido qué variables nos servirán para el diseño de cada *cluster* tenemos que tomar en consideración que dichas variables no se encuentran en la misma unidad de medida. Es por esta razón que nuestros datos tienen que estar estandarizados. El método que se utiliza consiste en calcular la media y la desviación estándar de cada serie de datos; posteriormente, se estandarizan restando a cada observación la media

y dividiendo con| la desviación estándar[6]. Una vez hecho lo anterior, el primer paso consiste en calcular la *Suma de Cuadrados al interior de los Grupos* (Sum of Squares Within Groups o *SSW* por sus siglas en ingles) de nuestras variables, y graficar dicha suma *versus* un número designado de 15 grupos. Es decir, nuestro objetivo es minimizar la *SSW*. A este tipo de gráfica se le conoce como *Screeplot*, y nos permite determinar de una forma visual el número de *clusters* óptimo a utilizar. La regla consiste en observar en qué parte se rompe la estructura de la curva para volverse cada vez más plana (ver gráficas 8 y 11).

Una vez que se tiene el número de *clusters* para cada caso procedemos al agrupamiento de las diferentes naciones, lo cual nos ayuda a determinar qué países son los más similares entre ellos. La primera técnica de agrupamiento que utilizamos es la *jerárquica* mediante el método Ward,[7] la cual nos arroja un Dendograma. Posteriormente, cambiamos a una técnica de *partición* mediante la utilización del algoritmo de K-means,[8] la cual nos arroja los *clusters* para cada caso.

En las siguientes subsecciones presentamos nuestros dos diseños de *cluster.*

1. Primer cluster

El Screeplot se muestra en la gráfica 8 y se puede observar un quiebre en la curva de la gráfica a partir del séptimo *cluster,* por lo que dejamos ese número como nuestra selección grupos. Una vez determinado lo anterior mediante el método Ward obtenemos nuestro primer dendograma (Gráfica 9) en el cual observamos que México se agrupa junto a Ecuador y Costa Rica; en tanto que países como Brasil y Cuba no se encuentran dentro de ningún grupo. Por otra parte, Argentina, Chile y Uruguay forman un *cluster*; mientras que un trío de naciones centroamericanas (Guatemala, Honduras y Nicaragua) forman un grupo por su cuenta. El último *cluster* es el más grande de todos y se encuentra formado por seis naciones, entre ellas Bolivia, Colombia y Perú.

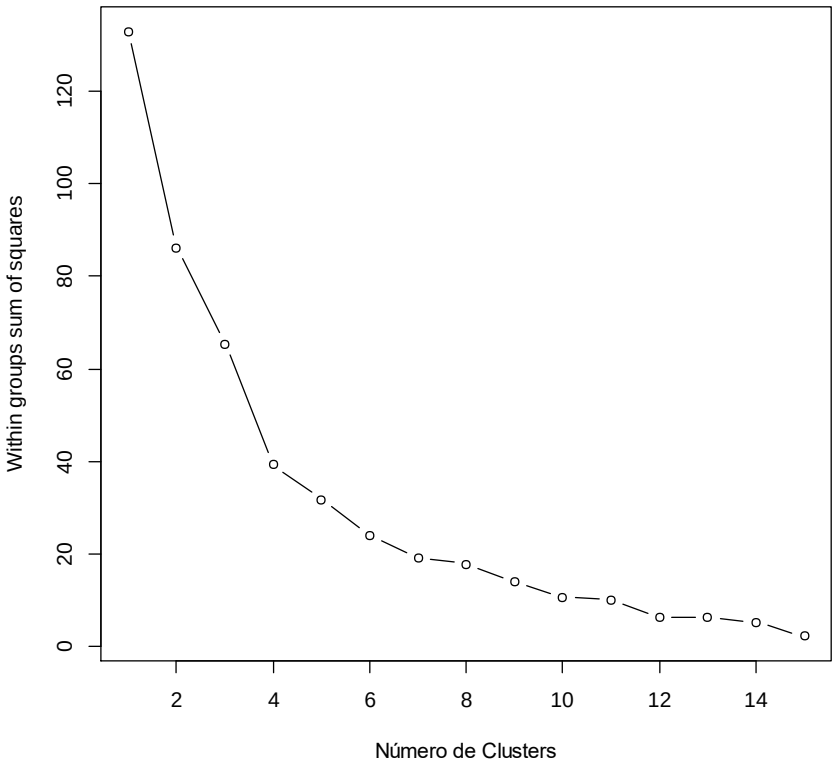
Por otra parte, cuando cambiamos de una técnica *jerárquica* a una técnica de *partición* mediante el algoritmo de K-means (Gráfica 10) obtenemos que México queda agrupado nuevamente con Costa Rica y Ecuador. Mediante esta técnica Brasil queda agrupado junto a Argentina, Chile y Uruguay; en tanto que Bolivia es ahora la queda fuera de todo *cluster*, lo mismo que Cuba. La República Dominicana, Panamá y Puerto Rico vuelven a formar un *cluster,* igual que Guatemala, Honduras y Nicaragua.

---

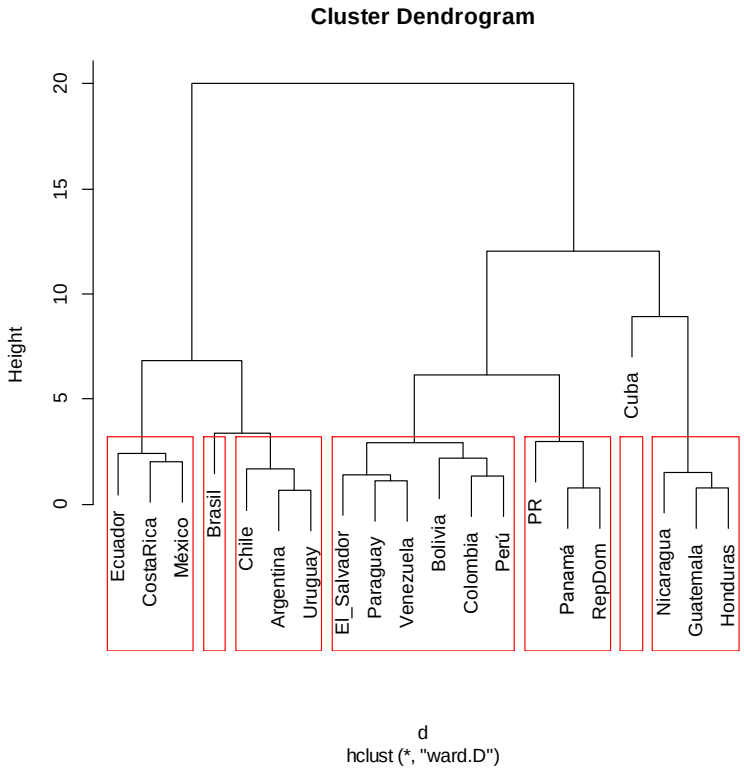
6 $(x_i - \bar{x})/_{Std.dev.(x)}$

7 Peña, D., 2002. Análisis de Datos Multivariantes, McGraw Hill.

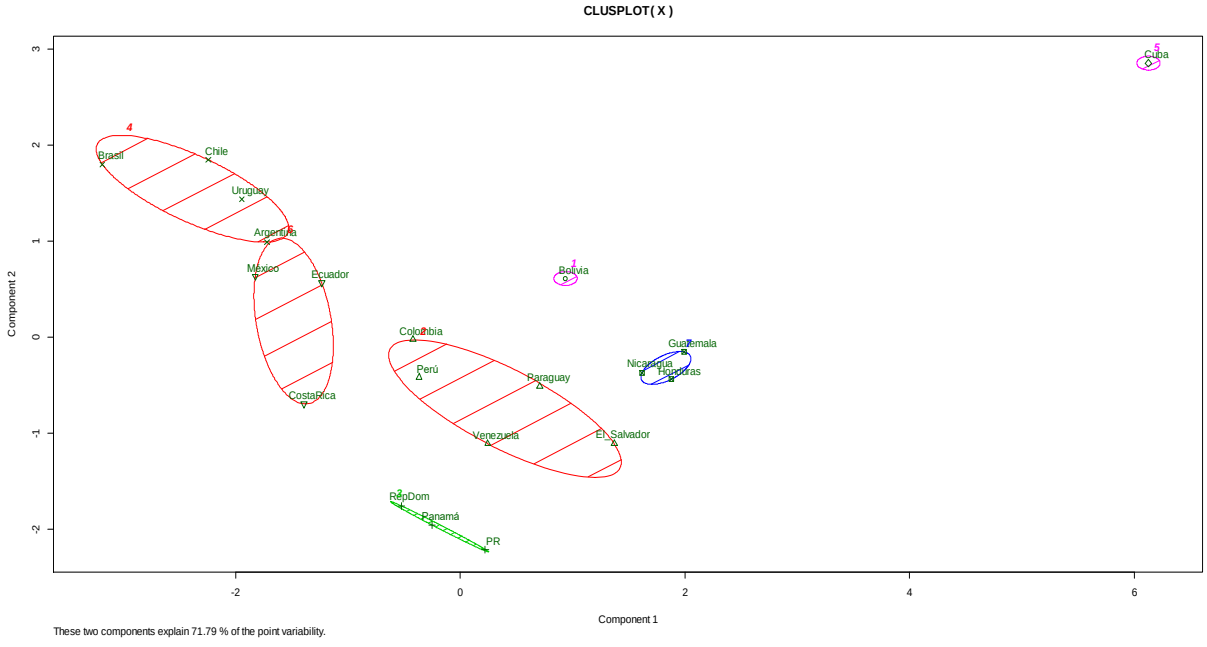
8 Idem


Within groups sum of squares
0
20
40
60
80
100
120
2
4
6
8
10
12
14
Número de Clusters

Gráfica 8



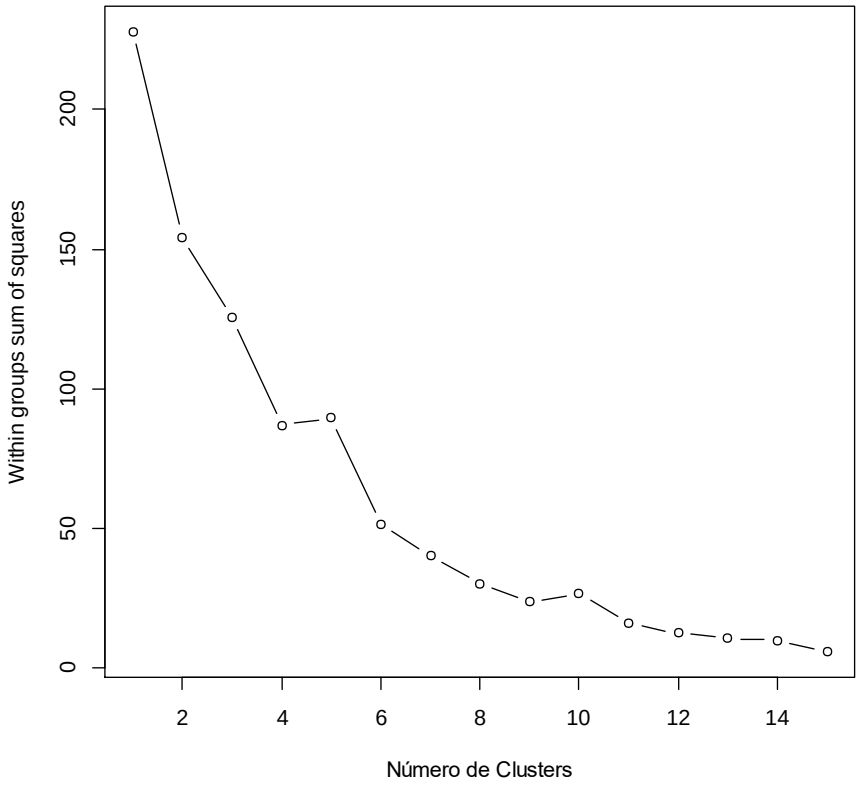
Gráfica 9



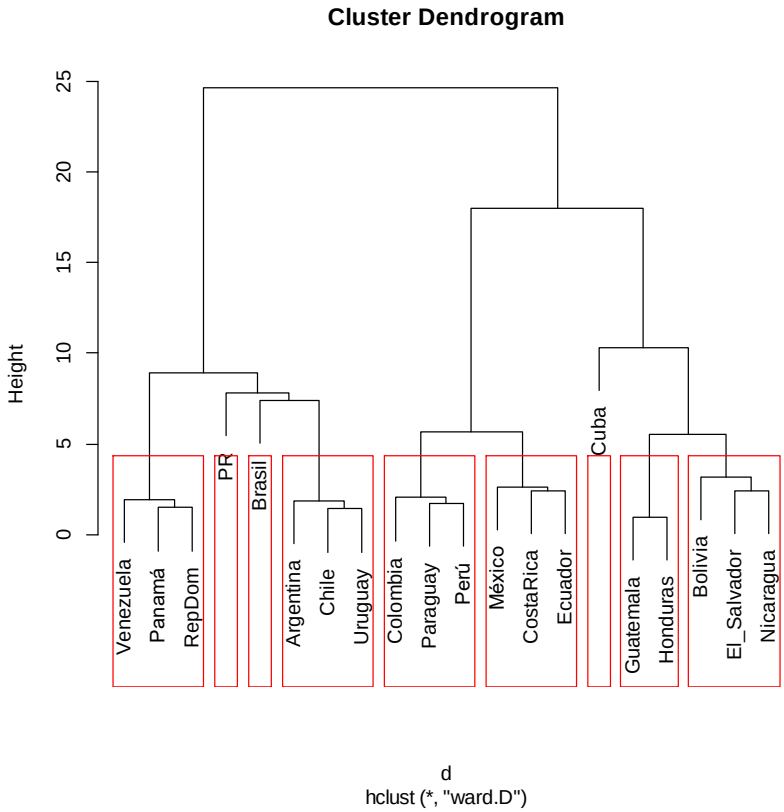
Gráfica 10

2. Segundo cluster

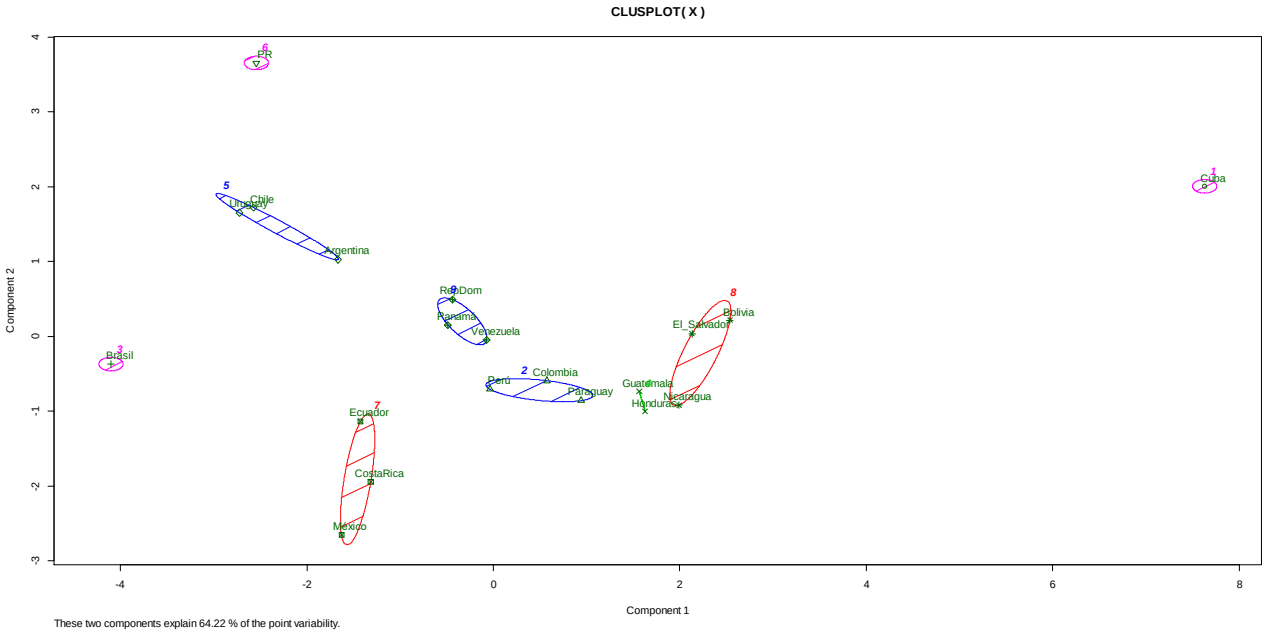
Como se mencionó anteriormente, el segundo diseño de *cluster* involucra a las mismas variables que el primero más las variables de conexiones por tipo de dispositivo y por tipo de contratación. Si observamos la gráfica del Screeplot (ver Gráfica 11) observamos un pico en la misma a partir del quinto *cluster*; sin embargo, nos decidimos por el décimo *cluster* ya que es a partir de ahí que la curva pierde casi completamente su pendiente. Es así que al trabajar con diez *clusters* tenemos, mediante la aplicación del método Ward, nuestro nuevo dendograma en el cual México vuelve a agruparse junto a Costa Rica y Ecuador. Brasil, Cuba y Puerto Rico quedan aislados; en tanto que Argentina, Chile y Uruguay demuestran que las similitudes entre ellos son muy fuertes quedando en un solo *cluster* como en el caso anterior. Cuando cambiamos al método de K-means la estructura de los clusters no cambia en lo más mínimo, conservándose la misma que en el caso jerárquico.



Gráfica 11



Gráfica 12



Gráfica 13

IV. Conclusiones

El propósito del presente trabajo consiste en realizar una aplicación de dos de las técnicas de agrupamiento más reconocidas en la literatura de análisis de datos multivariados, siendo la primera de ellas la jerárquica del método Ward, y la segunda la de partición del método K-means. A partir de una serie de variables del sector de las telecomunicaciones de todos los países de América Latina realizamos dos diseños, el primero de los cuales considera el porcentaje de conexiones por tecnología de red (2G, 3G, 4G), el espectro asignado para servicios móviles por país, penetración de mercado de suscriptores únicos a servicios móviles, penetración de mercado de suscriptores únicos a Internet móvil y conexiones de IoT ponderadas mediante la variable de número de suscriptores a servicios móviles. El segundo diseño considera las mismas variables más las conexiones por tipo de contratación (prepago y pospago) y conexiones por tipo de dispositivo (smartphone, teléfono básico y dispositivo de sólo datos). Los resultados de ambas técnicas, una vez estandarizados los datos, identifican que México mantiene similitudes importantes con dos países de la región: Costa Rica y Ecuador. Resalta que, de las variables sin estandarizar en las que estos tres países se asemejan se encuentra la de conexiones por tipo de dispositivos, específicamente en teléfono básico: México (22.2%), Ecuador (25.6%), Costa Rica (29.3%). En lo referente a los dispositivos de sólo datos, la similitud entre México y estos dos países también es importante: México (15.5%), Ecuador (14.9%), Costa Rica (15.2%) (ver Gráfica 1). Por otro lado, la variable ponderada de conexiones IoT en bandas móviles concesionadas (ver Gráfica 7) también demuestra que estas naciones se encuentran relativamente cerca una de la otra: México (9.7%), Ecuador (7.6%), Costa Rica (7.5%). Las conexiones de prepago reflejan un acercamiento entre Ecuador (74.3%) y Costa Rica (77.2%), no tanto así con México que tiene un porcentaje más elevado (87.2%). En tanto que la penetración de mercado de suscriptores únicos a Internet móvil muestra a México (53%) y a Costa Rica (56.7%) relativamente cerca uno del otro, mientras que Ecuador se ubica un poco debajo (44.2%). Es así que los resultados de nuestros *clusters* parecen encontrarse en contraposición con la creencia de que México tendría que agruparse con las naciones de tamaño poblacional similar, como sería el caso de Brasil. Aunque, en el primer diseño de *cluster* (ver Gráfica 10) se observa que el grupo conformado por Argentina, Brasil Chile y Uruguay, se llega a tocar con el grupo formado por Costa Rica, Ecuador y México, no así en el caso del segundo diseño. Esto refleja el hecho de que el resultado obtenido depende de las variables seleccionados. Finalmente se deja abierta la posibilidad de explorar otros diseños de *cluster* o de datos diferentes del mismo sector de las telecomunicaciones que puedan arrojar resultados distintos.

V. Referencias

[1] GSMA Intelligence. https://www.gsmaintelligence.com/

[2] Peña, D., 2002. Análisis de Datos Multivariantes, McGraw Hill.

[3] 5G Américas. http://www.5gamericas.org/en/resources/statistics/statistics-latin-america/