LAJED Nº 32 Noviembre 2019 - Abril 2020 95 - 129
ISSN: 2074 - 4706

Clustering, mediterraneidad y comercio internacional: aplicación empírica de los algoritmos *Partitioning Around Medoids y K-means*

Clustering, Landlockedness and International Trade: Empirical Application of the Partitioning Around Medoids and K-means algorithms

Heynz Roberth Gonzáles Argote **
Ulises Amaru Ticona Gonzáles ***

Resumen

El tema de la mediterraneidad ha generado bastante interés en el debate geopolítico, siendo Bolivia uno de los actores principales. Este hecho, junto con las nuevas herramientas de análisis de datos, como la inteligencia artificial y la minería de datos, motivan el presente este estudio, el cual es pionero dentro de la literatura en el marco del análisis de países sin salida marítima mediante algoritmos no supervisados de minería de datos.

* Contacto: heynzg@gmail.com
** Contacto: uticona@gmail.com

En este sentido, se estudia y aplica la teoría de formación de *clusters* a través de los algoritmos *K-means* y PAM (*Partitioning Around Medoids*) con información de indicadores de comercio internacional de 188 países de un periodo de diez años, con el propósito de detectar si la condición de mediterraneidad es un factor limitante en la dinámica comercial de los países.

Los resultados muestran que un subconjunto reducido de los países mediterráneos, entre ellos Bolivia, habrían aliviado, durante la última década, las restricciones que la mediterraneidad implica en los costos y tiempos de exportación e importación.

Palabras clave: *Cluster*, mediterraneidad, litoral, comercio internacional, minería de datos.

Abstract

Landlockedness has generated significative interest in the geopolitical debate, particularly in Bolivia. This fact, along with innovative methodologies such as artificial intelligence and data mining, has motivated this research, which is unprecedented in the literature concerning landlockedness analysis through unsupervised algorithms of data mining.

Consequently, the theory of cluster formation is studied and applied through the K-means and PAM (Partitioning Around Medoids) algorithms using international trade information of one hundred eighty-eight countries over a period of ten years, in order to test whether the landlockedness condition is a limiting factor in the commercial dynamics of countries.

The results show that a reduced subset of the landlocked countries, including Bolivia, would have eased restrictions such as international trade costs and times, during the last decade.

Key words: Cluster, landlocked countries, littoral, international trade, data mining.

Clasificación/Classification JEL: C82, F43, F55, O11, O57

1. Introducción

El comercio que realiza un país no puede ser significativo si se tiene que recurrir a distintos canales de transporte poco eficientes, además de tener que transitar necesariamente por otro país antes de llegar a la costa. De esta forma, siempre estará en poder del país vecino con salida

al mar facilitar u obstruir la comunicación entre el país enclaustrado y las costas. Así, Smith (1796) reconocía los desafíos que enfrentaban los países sin litoral, también denominados mediterráneos, en términos de distancia y también de dependencia de un país de tránsito soberano para el comercio internacional.

¿Por qué los países mediterráneos en desarrollo (LLDC)[1] enfrentan tales retos? Una corriente sugiere que la principal razón es la dificultad en el comercio: los sitios alejados geográficamente muestran rezagos a la hora de percibir las ganancias de la especialización y los beneficios asociados. Otra línea de pensamiento se deriva de la anterior hipótesis: la institucionalidad y tecnología que se genera en toda la dinámica del comercio internacional, logística, infraestructura y procesos productivos es menos avanzada en los países mediterráneos en desarrollo, debido a su menor escala y mayores costos administrativos para su ejecución.

De esta manera, se examinarán los patrones de asociación de variables de comercio internacional que distinguen a los LLDC del resto de países. Se considera inédita la presente aplicación de herramientas de minería de datos no supervisadas, como los *clusters*, sobre el análisis de indicadores mundiales, relacionado a la mediterraneidad. Consecuentemente, en las dos secciones posteriores se hará una revisión de la literatura existente sobre las características de los LLDC y se presentarán hechos estilizados con énfasis en variables de comercio internacional sujetas a estudio. La sección posterior describe la metodología de técnicas de minería de datos de *machine learning* (método de aprendizaje supervisado), junto con los resultados. Finalmente, se presentan las conclusiones del documento.

2. Revisión de la literatura

Hasta donde se pudo evidenciar en la búsqueda del estado del arte de la temática, la metodología empleada es nueva en el análisis de la mediterraneidad, por lo que los estudios revisados serán útiles sobre todo para explicar los resultados de *clusters* de países. En general la literatura se ha enfocado en emplear variables de comercio internacional para medir los costos de la mediterraneidad. También se encontraron múltiples estudios que miden el costo en el crecimiento económico y las interdependencias entre países con y sin litoral. En general, esos estudios encuentran impactos negativos de la mediterraneidad de distintos tipos (Cuadro 1):

1 *Landlocked developing countries*, en inglés, es su denominación conocida en la literatura.

Cuadro 1
Algunos estudios que relacionan mediterraneidad y variables económicas

Autores	Año	Datos	Efecto	Resumen de los hallazgos
Radelet y Sachs	1998	97 países en desarrollo, datos CIF y FOB	Negativo	Los seguros y costos de transporte son dos veces mayores para los mediterráneos que para los países con costa
MacKellar *et al.*	2000	92 países de ingresos medios y bajos (1960-1992)	Negativo	Los países sin litoral muestran una menor tasa de crecimiento, del orden de 1.5%
Raballand	2003	Información de comercio internacional	Negativo	La condición de mediterraneidad (mediante distintas medidas) reduce el flujo comercial en 80%, principalmente por los costos de transporte adicionales
Shrestha y Heffley	2003	Estudio teórico	Negativo	Los países que tienen acceso a los puertos de exportación tienen niveles de bienestar mayores que los países enclaustrados
Faye *et al.*	2004	Descriptivo	Negativo	Los países mediterráneos muestran rezagos en infraestructura y prácticas administrativas, principalmente. Citan el caso de Bolivia, que desaprovechó su ubicación central geográfica
De	2006	Grupo de economías de Asia seleccionadas	Negativo	Mediante un modelo estructural, hallan que los costos de transporte de un país mediterráneo son 55% más altos que los de un país con costa
Grigoriou	2007	Datos de infraestructura de países seleccionados	-	Mejoras en la infraestructura en el país de tránsito incrementaría el comercio internacional del país mediterráneo en 52%
Arvis *et al.*	2010	Información de costos y tiempos de transporte	Negativo	Los países mediterráneos enfrentan una brecha de costos estimada entre 8 a 250%, y una brecha de demoras en tiempo entre 9y 130% por transitar por los países con costa
Lahiri y Masjidi	2012	Modelo de juegos infinitamente repetidos	-	Sostienen que la política de obstrucción unilateral al acceso de los océanos impuesta a los países sin litoral por las economías costeras puede no ser óptima si se considera desde la perspectiva más amplia de costos y ganancias que implican otros mercados

Autores	Año	Datos	Efecto	Resumen de los hallazgos
Wamboye	2012	40 países menos desarrollados (1975-2010)	Negativo	Mediante Método Generalizado de Momentos en Sistema, encuentran que la condición de mediterraneidad genera una brecha en la deuda pública de 22% respecto a los países con litoral
Driffield y Jones	2013	Panel de países (1984-2007)	Negativo	La mediterraneidad implica un impacto en el crecimiento de -0,0682; no tiene impacto negativo sobre la capacidad de atraer IED
Paudel	2014	Panel de países mediterráneos y no mediterráneos	Negativo	El enclaustramiento obstaculiza el crecimiento económico, pero la magnitud del impacto negativo es mayor que en la literatura, controlando por regiones
Mendoza *et al.*	2018	95 países emergentes y en desarrollo (1993-2017)	Negativo	Los efectos directos e indirectos del enclaustramiento marítimo de un país de ingresos medio bajos son significativos con relación a los países con litoral, lo cual repercute en el crecimiento económico (1,0pp) y los niveles de pobreza (1.9%)

Fuente: Elaboración propia

Asimismo, el documento de UN-OHRLLS[2] (2013) analiza a detalle el impacto de la mediterraneidad sobre las perspectivas de desarrollo de los LLDC, mediante indicadores económicos, institucionales y sociales. Este estudio desarrolla un modelo econométrico estructural para estimar empíricamente el costo de la mediterraneidad en el desarrollo. El modelo pone énfasis en los múltiples canales de vínculo entre la mediterraneidad y el desarrollo, además del comercio internacional. Al mismo tiempo, a través de la construcción de umbrales específicos de los países, la metodología provee una medida de costo en el desarrollo para cada LLDC que se investiga. Se consideran las siguientes variables: ingreso *per cápita*, calidad institucional, integración económica, latitud, *dummy* de mediterraneidad, población económicamente activa, superficie terrestre y un indicador de recursos naturales.

Este trabajo recurre a cuatro métodos para estimar los coeficientes del modelo: Mínimos Cuadrados Ordinarios, Mínimos Cuadrados en 2 Etapas, Regresiones Aparentemente no Relacionadas y Método Generalizado de Momentos. La distancia al Ecuador (latitud) y la

2 United Nations of the High Representative for the Least Developed Countries, Landlocked Developing Countries and Small Island Developing States.

condición de mediterraneidad presentan coeficiente negativo, indicando la existencia de una brecha entre los LLDC y los países con litoral, además de cierta influencia de la condición geográfica aproximada por la latitud. El nivel de ingresos, la institucionalidad y la integración económica inciden positivamente en el desarrollo, en concordancia con la literatura. Los resultados son robustos a la metodología de estimación.

3. Hechos estilizados

En línea con la literatura revisada, se han seleccionado un conjunto de indicadores de comercio a ser utilizados como datos de ingreso a los algoritmos *K-means* y PAM (*Partitioning Around Medoids*). Estos indicadores, extraídos del *World Development Indicators* del Banco Mundial, son descritos a continuación:

- Costo de exportación (US$ por contenedor) - IC.EXP.COST.CD. El costo mide los honorarios cobrados para un contenedor de 20 pies en dólares de EE. UU. Se incluyen todas las tarifas asociadas para completar los procedimientos para exportar los bienes. Estos incluyen los costos de los documentos, los aranceles administrativos para el despacho de aduanas y el control técnico, los honorarios de los agentes de aduanas, los cargos por manipulación de terminales y el transporte terrestre. La medida del costo no incluye aranceles ni impuestos comerciales. Solo se registran los costos oficiales[3].
- Costo de importación (US$ por contenedor) - IC.IMP.COST.CD. El costo mide los honorarios cobrados para un contenedor de 20 pies en dólares de EE. UU. Se incluyen todas las tarifas asociadas para completar los procedimientos para importar los bienes. Estos incluyen los costos de los documentos, los aranceles administrativos para el despacho de aduanas y el control técnico, los honorarios de los agentes de aduanas, los cargos por manipulación de terminales y el transporte terrestre. La medida del costo no incluye aranceles ni impuestos comerciales. Solo se registran los costos oficiales.

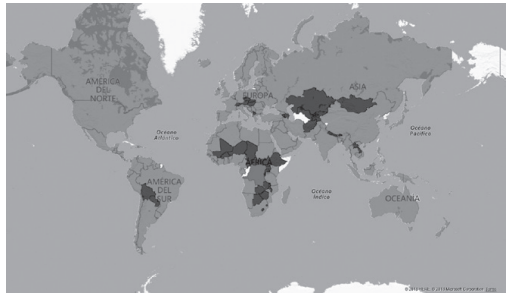
3 Se hacen varias suposiciones para el negocio encuestado: tiene 60 o más empleados; está ubicado en la ciudad más poblada del país; es una compañía privada de responsabilidad limitada; no opera dentro de una zona de procesamiento de exportaciones o un polígono industrial con privilegios especiales de exportación o importación; es de propiedad nacional sin propiedad extranjera; y exporta más del 10% de sus ventas. También se realizan suposiciones sobre los bienes comercializados: el producto comercializado viaja en una carga seca, son cargas de contenedores completos de 20 pies. Finalmente, las suposiciones sobre el producto: no es peligroso ni incluye artículos militares; no requiere refrigeración ni ningún otro entorno especial; no requiere ninguna norma de seguridad ambiental o fitosanitaria especial que no sea una norma internacional aceptada.

- Tiempo para exportar (días) - IC.EXP.DURS. Es el tiempo necesario para cumplir con todos los procedimientos requeridos para exportar bienes, y se registra en días naturales. El cálculo del tiempo para un procedimiento comienza desde el momento en que se inicia y se ejecuta hasta que se completa. Si se puede acelerar un procedimiento por un costo adicional, se elige el procedimiento legal más rápido. Se supone que el exportador no pierde el tiempo y se compromete a completar cada procedimiento restante sin demora. Los procedimientos que se pueden completar en paralelo se miden como simultáneos. El tiempo de espera entre los procedimientos, por ejemplo, durante la descarga de la carga, se incluye en la medida.
- Tiempo para importar (días) - IC.IMP.DURS. Es el tiempo necesario para cumplir con todos los procedimientos requeridos para importar bienes, y se registra en días naturales. El cálculo del tiempo para un procedimiento comienza desde el momento en que se inicia y se ejecuta hasta que se completa. Si se puede acelerar un procedimiento por un costo adicional, se elige el procedimiento legal más rápido. Se supone que el importador no pierde el tiempo y se compromete a completar cada procedimiento restante sin demora. Los procedimientos que se pueden completar en paralelo se miden como simultáneos. El tiempo de espera entre los procedimientos, por ejemplo, durante la descarga de la carga, se incluye en la medida.

Para el proceso de formación del set de datos, se consideró la generación de dimensiones de información relacionadas a país-continente, cualidad marítima y georreferenciación de los países (Gráfico 1). Se recurre a herramientas de inteligencia de negocios como *Power BI* para realizar cálculos auxiliares y sobre todo para la visualización de los datos y presentación de los resultados del estudio.

Gráfico 1: Países mediterráneos por continente

Salida marítima			
Continente	**Sí**	**No**	**Total**
África	37	16	53
América	33	2	35
Asia	35	11	46
Europa	34	8	42
Oceanía	12	0	12
Total	**151**	**37**	**188**



Nota: En el siguiente link se pone al servicio de los lectores un Dashboard en Power BI para poder ampliar la revisión de los datos del presente documento: https://app.powerbi.com/view?r=eyJrIjoiNGI0MTI0ZDctYTJhOC00MzJhLTkxNWMtYjhkNTQxYjFhY2JmIiwidCI6IjYxZGU5YzBjLTRkNWEtNGE4NS05ODIjLTg4OWIwOGZmZmU5YSIsImMiOjZ9view?r=eyJrIjoiNGI0MTI0ZDctYTJhOC00MzJhLTkxNWMtYjhkNTQxYjFhY2JmIiwidCI6IjYxZGU5YzBjLTRkNWEtNGE4NS05ODIjLTg4OWIwOGZmZmU5YSIsImMiOjZ9

Fuente: Elaboración propia

Es importante acudir al cálculo de las medidas de tendencia central para un mejor entendimiento de la distribución de los datos de las variables de estudio. Una comparación del promedio de los costos de exportación entre los países con salida al mar y los LLDC de cinco continentes entre dos periodos 2005 y 2014 permite observar que la relación entre los promedios de los costos de exportación en Asia y África supera el 200% entre los países con salida marítima y los LLDC. Esta diferencia no es tan evidente para los países de América, y en el caso de los países europeos es prácticamente inexistente (Cuadro 2).

Cuadro 2
Promedios de los indicadores de comercio internacional
(en dólares y días)

Indicador	África		Asia		América		Europa		Oceanía		Total
	2005	2014	2005	2014	2005	2014	2005	2014	2005	2014	
Costo de exportación	1.559	2.052	1.186	1.711	1.029	1.297	1.011	1.202	1.025	905	1.398
Con litoral	1.198	1.411	797	990	1.010	1.276	1.010	1.163	1.025	905	1.103
Sin litoral	2.404	3.534	2.282	4.004	1.323	1.645	1.013	1.370			2.601
Tiempo para exportar	37	29	32	26	20	16	16	13	22	21	24
Con litoral	32	24	24	18	20	16	16	13	22	21	20
Sin litoral	48	40	56	49	30	26	15	13			40
Costo de importación	1.920	2.715	1.328	1.911	1.350	1.656	1.089	1.272	1.100	904	1.670
Con litoral	1.399	1.785	887	1.097	1.345	1.634	1.099	1.255	1.100	904	1.297
Sin litoral	3.136	4.867	2.570	4.501	1.426	2.010	1.039	1.344			3.191
Tiempo para importar	45	36	36	28	24	18	17	13	24	23	28
Con litoral	38	29	27	20	24	17	17	12	24	23	23
Sin litoral	62	52	61	54	35	29	15	13			47
Total	**890**	**1.208**	**646**	**919**	**606**	**747**	**533**	**625**	**543**	**463**	**780**

Fuente: Elaboración propia

La relación porcentual de las diferencias de las medias entre los LLDC respecto a los países con salida marítima brinda un panorama altamente diferenciado en Asia, seguido por África, y mucho menos diferenciado en América. Sin embargo, en Europa esta diferencia es prácticamente nula para el periodo 2005 (Cuadro 3).

Cuadro 3
Promedios de las brechas entre países mediterráneos y con litoral, 2005
(en porcentajes)

Indicador	África	América	Asia	Europa
Costo de exportación	100,7	31,0	186,1	0,2
Tiempo para exportar	52,2	50,9	139,0	6,3
Costo de importación	124,2	6,0	189,7	5,5
Tiempo para importar	62,1	45,8	125,0	-10,9

Fuente: Elaboración propia

Una década después, la diferencia de las medias entre los países con salida al mar y los LLDC muestra un escenario aún más polarizado, donde los países mediterráneos de Asia revelan costos de exportación e importación por contenedor superiores en 405% y 410%, respectivamente. Sin embargo, en América la diferencia de las medias de los costos de exportación disminuyó de 131% el año 2005 a 129% el año 2014. Los tres indicadores adicionales también mostraron un incremento no tan significativo como el reflejado por los países asiáticos y africanos (Cuadro 4).

Los incrementos en la diferencia de las medias de los indicadores en el año 2014 en Europa también sufrieron un incremento, pero no es significativo, indicando que los países mediterráneos tienen costos y tiempos de exportación similares a los países que no son mediterráneos.

Cuadro 4
Promedios de las brechas y variaciones respecto a 2005, entre países mediterráneos y con litoral, 2014
(en dólares)

Indicador	África	América	Asia	Europa
Costo de exportación	150%	29%	305%	18%
	Δ+	Δ+	Δ+	Δ+
Tiempo para exportar	65%	64%	171%	-0,1%
	Δ+	Δ+	Δ+	Δ-
Costo de importación	173%	23%	310%	7%
	Δ+	Δ+	Δ+	Δ+
Tiempo para importar	80%	69%	170%	2%
	Δ+	Δ+	Δ+	Δ+

Fuente: Elaboración propia
Nota: Δ+ = aumentó; Δ- = disminuyó

Europa también mostró incrementos en la diferencia de las medias para la gestión 2014, **aunque** no tan significativos como en el resto de continentes, demostrando que los países mediterráneos de esta región tienen costos y tiempos de exportación similares a los países no mediterráneos. También se percibe que los países LLDC del continente europeo tienen condiciones similares, en costos y tiempos de exportación, a los países no mediterráneos. En América, estas diferencias son porcentualmente más grandes que en Europa, pero los países africanos y asiáticos muestran diferencias significativas respecto a estas variables (Cuadro 5).

Un indicador importante para evaluar la homogeneidad y heterogeneidad de los datos es el coeficiente de variación (CV). Los cálculos para esta medida de tendencia central se entienden bajo los siguientes criterios:

- Si el CV es mayor a 25%, son datos heterogéneos o dispersos.
- Si el CV es menor a 25%, son datos homogéneos.

Cuadro 5
Volatilidad de los indicadores de comercio internacional
(en dólares y días)

Indicador	África		Asia		América		Europa		Oceanía		Total
	2005	2014	2005	2014	2005	2014	2005	2014	2005	2014	
Costo de exportación	916	1.384	903	1.704	400	556	271	388	355	297	1.034
Con litoral	487	700	552	577	405	563	290	403	355	297	535
Sin litoral	1.106	1.438	793	2.041	103	205	149	254			1.568
Tiempo para exportar	15	13	23	20	9	9	7	5	6	6	16
Con litoral	13	10	17	12	9	8	8	5	6	6	12
Sin litoral	13	14	18	22	6	4	5	4			21
Costo de importación	1.088	2.099	1.017	1.978	555	593	342	460	413	286	1.349
Con litoral	532	1.223	597	618	572	600	367	496	413	286	720
Sin litoral	1.084	2.123	924	2.516	26	265	159	243			2.062
Tiempo para importar	20	21	25	23	11	13	10	6	8	7	20
Con litoral	17	12	18	14	11	13	10	6	8	7	15
Sin litoral	18	27	25	27	2	1	5	4			28
Total	**1.115**	**1.737**	**916**	**1.583**	**686**	**845**	**562**	**683**	**587**	**487**	**1.140**

Fuente: Elaboración propia

En América y Europa los países mediterráneos presentan cifras mucho más homogéneas. Esto quiere decir que los datos están mucho más agrupados respecto a la media. En cambio, en el resto, se evidencia mayor dispersión debido al CV mayor a 25% (Cuadro 6).

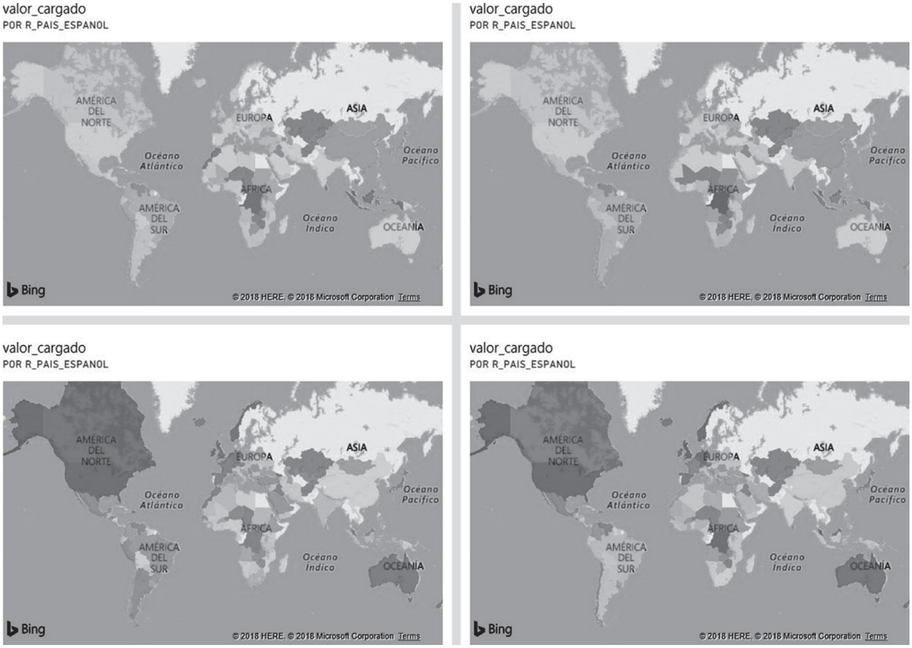
Cuadro 6
Coeficiente de variación de los indicadores de comercio internacional
(en porcentajes)

Variables	África		Asia		América		Europa	
	2005	2014	2005	2014	2005	2014	2005	2014
Costo de exportación								
Con litoral	40.7	49.6	69.3	58.3	40.1	44.1	28.7	34.7
Sin litoral	46.0	40.7	34.8	51.0	7.8	12.5	14.7	18.6
Tiempo para exportar								
Con litoral	40.6	40.2	73.7	68.0	44.7	54.3	47.8	40.5
Sin litoral	26.2	35.2	31.6	45.3	20.0	13.7	29.8	28.1
Costo de importación								
Con litoral	38.0	68.5	67.3	56.3	42.5	36.7	33.4	39.5
Sin litoral	34.6	43.6	36.0	55.9	1.8	13.2	15.3	18.1
Tiempo para importar								
Con litoral	43.5	43.1	68.6	68.4	47.7	73.5	62.0	51.1
Sin litoral	29.5	51.5	40.7	50.4	4.3	3.4	31.9	33.4

Fuente: Elaboración propia

Finalmente, para visualizar de manera integral los indicadores seleccionados, en el Gráfico 2 se presentan mapas de calor de las variables de comercio internacional:

Gráfico 2: Mapas de calor de los indicadores de comercio internacional



Fuente: Elaboración propia
Nota: De arriba hacia abajo: costos de exportación e importación; tiempo para exportar e importar.

4. Estrategia empírica

4.1. Machine Learning

El aprendizaje automático, o *Machine Learning*, es un área de investigación en constante expansión. Derivada de la inteligencia artificial, puede ser aplicada en diversos campos, como son las ciencias computacionales, estadística y en el caso de este trabajo de investigación, con la economía a través del comercio exterior. Como definición, el aprendizaje automático, mediante un proceso de inducción del conocimiento, busca generalizar comportamientos y reconocer patrones a partir de los datos. Los diferentes algoritmos de aprendizaje automático, de acuerdo a la salida o resultado al que llegan los mismos, son agrupados en:

- Aprendizaje supervisado. Pretende determinar una función que puede mapear una entrada de datos a una salida basada en ejemplos anteriores.
- Aprendizaje no supervisado. Se tienen conjuntos de datos de entrada y se busca establecer patrones para realizar el etiquetado de los nuevos datos. Uno de los métodos más comunes es el análisis de conglomerados (*clustering*).
- Aprendizaje semi-supervisado. Es una combinación de los dos algoritmos anteriores, teniendo en cuenta ejemplos clasificados y no clasificados.
- Aprendizaje por refuerzo. Los algoritmos aprenden observando el mundo que les rodea y con un continuo flujo de información en las dos direcciones (del mundo a la máquina, y de la máquina al mundo), realizando un proceso de ensayo-error y reforzando aquellas acciones que reciben una respuesta positiva en el mundo.
- Transducción. Similar al aprendizaje supervisado, pero su objetivo no es construir de forma explícita una función, sino únicamente tratar de predecir las categorías en las que caen los posteriores ejemplos, basándose en los ejemplos de entrada, sus respectivas categorías y los ejemplos nuevos al sistema. Es decir, estaría más cerca del concepto de aprendizaje supervisado dinámico.
- Aprendizaje multitarea. Engloba todos aquellos métodos de aprendizaje que usan conocimiento previamente aprendido por el sistema, de cara a enfrentarse a problemas parecidos a los ya vistos.

Tomando en cuenta el marco de la presente investigación nos enfocaremos en los algoritmos de clasificación no supervisados, ya que, a partir de datos e indicadores de un set de datos generado, se pretende lograr una agrupación (*clustering*).

4.2. Agrupamiento "*Clustering*"

El análisis por agrupamiento, a través de los algoritmos, pretende realizar la clasificación de observaciones en subgrupos -*clusters*- para que las observaciones en cada grupo se asemejen entre sí según ciertos criterios. Teóricamente, los puntos de datos que están en el mismo grupo deben tener propiedades y/o características similares, mientras que los puntos de datos en diferentes grupos deben tener propiedades y/o características muy diferentes. La agrupación

es un método de aprendizaje no supervisado y es una técnica común para el análisis estadístico de datos que se utiliza en muchos campos.

Dentro de los algoritmos de clasificación no supervisados tenemos al *K-means* y al *K-mediods*, ambos algoritmos de partición cuyos datos de ingreso (observaciones) pueden ser vectores reales d-dimensionales.

4.2.1. *K-means*

Uno de los algoritmos más utilizados para realizar agrupamientos es el *K-means,* o, en su traducción al español, K-medias. MacQueen (1967) indica que este algoritmo tiene el propósito central de particionar un conjunto de observaciones (n) en k agrupaciones, donde cada observación es asignada a un grupo cuyo valor medio es más cercano a un centroide.

Dado un conjunto de observaciones:

$$\left(x_1, x_2, \ldots, x_n\right) \tag{1}$$

donde cada observación es un vector real de d dimensiones, k-medias construye una partición de las observaciones en k conjuntos donde ($k \leq n$), a fin de minimizar la suma de los cuadrados dentro de cada grupo (WCSS):

$$\frac{\arg\min}{s} \sum_{i=1}^{k} \sum_{x \in Si} \| x - \mu_{\text{i}} \|^2 = \frac{\arg\min}{s} \sum_{\text{i}=1}^{k} \left|s_i\right| Vars_{\text{i}} \tag{2}$$

Donde μ_i es la media de los puntos en S_i. Esto es equivalente a minimizar las desviaciones cuadradas por pares de puntos en el mismo *cluster*:

$$\frac{\arg\min}{s} \sum_{i=1}^{k} \frac{1}{2\left|s_i\right|} \sum_{x,y \in S_i} \| x - y \|^2 \tag{3}$$

La equivalencia se puede deducir de la identidad:

$$\sum_{x \in S_i} \| x - \mu_i \|^2 = \sum_{x \neq y \in S_i} (x - \mu_i)\left(\mu_{i-y}\right) \tag{4}$$

Debido a que la varianza total es constante, esto también es equivalente a maximizar el BCSS, que se desprende fácilmente de la ley de la varianza total.

a) Interacción del algoritmo

El algoritmo se divide en cuatro pasos al momento de su ejecución; el primer paso se enfoca en realizar las asignaciones de las observaciones a los grupos y el segundo al cálculo de los nuevos centroides.

Paso 1

Selección arbitraria de los k objetos que serán los centroides iniciales.

Paso 2

Asignaciones. Se asigna a cada observación con el centroide más cercano, con base en el valor medio de las observaciones.

$$S_i^{(t)} = \left\{ x_p : \| x_p - m_i^{(t)} \| \leq \| x_p - m_j^{(t)} \| \; \forall 1 \leq j \leq k \right\} \tag{5}$$

Paso 3

Actualización. Se recalculan los centros de los subgrupos y se actualiza la media.

$$m_i^{(t+1)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_j \in S_i^{(t)}} x_j \tag{6}$$

Paso 4

Se iteran los pasos 1 y 2, hasta que el algoritmo llegue a una convergencia, que se da cuando las asignaciones generadas ya no cambian.

b) K-medoids

Los algoritmos basados en el método de *K-medoids* tienen el propósito de dividir un conjunto de observaciones en grupos, teniendo como principal diferencia con el algoritmo *K-means* la utilización de los datos que forman parte del conjunto de datos a ser analizados como representantes de las agrupaciones, que son denominados *"medoids"*. Cada observación restante es agrupada con el *medoid* más cercano. Estos algoritmos tienen características más robustas ante el ruido que se puede dar en las observaciones, siendo uno de los más representativos el algoritmo Partición Alrededor de *Medoids* (PAM).

4.2.2. Algoritmo PAM (*Partitioning Around Medoids*)

Este algoritmo pretende determinar las k agrupaciones de las n observaciones, identificando los objetos representativos de cada agrupación. La identificación de los k *medoids* inicia con la selección arbitraria de k objetos representativos. Cada interacción del algoritmo busca mejorar la calidad del agrupamiento. El algoritmo cuenta con cinco pasos en su ejecución:

Paso 1

Selección arbitraria de los *k-medoids* iniciales.

Paso 2

Cálculo de TCij para todos los pares de objetos Oi, Oh donde Oi es el *medoid* actual y Oh no lo es.

Paso 3

Seleccionar el par Oi, Oh que corresponda al mínimo Oi, Oh (TCih). Si el mínimo TCih es negativo, se intercambia Oi con Oh; y **se** vuelve al paso 2.

Paso 4

Repetir los pasos 2 y 3 hasta que no se presenten cambios.

Paso 5

Asignar cada objeto a su *medoid* más cercano. La convergencia del algoritmo se da cuando se obtiene el menor costo y las asignaciones a las agrupaciones no se modifican.

5. Aplicación y resultados

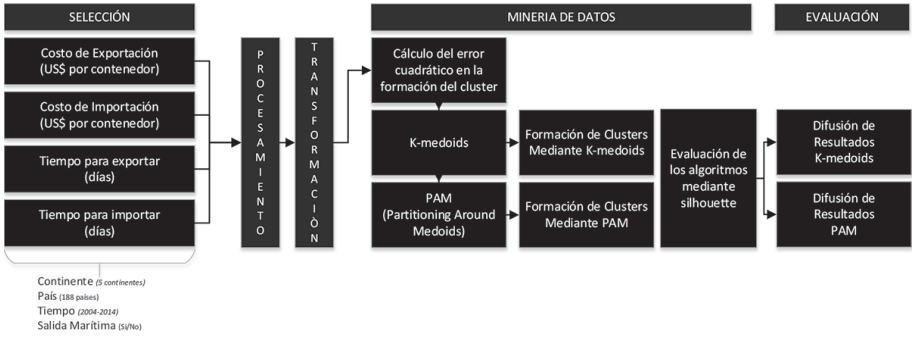
5.1. Datos y fuentes

Una fuente de datos importante y pública, utilizada por los investigadores, es la del Banco Mundial, entidad que a través de la aplicación *DataBank*[4] permite el acceso a 70 bases de datos. La base de datos seleccionada en este trabajo es la *World Development Indicators,* que comprende información de 264 países y con series temporales mayores a los 20 años.

5.2. Metodología empleada

La metodología empleada comprende la utilización del proceso de extracción de conocimiento, KDD[5], que para esta investigación tiene como núcleo a la minería de datos mediante la metodología de agrupamiento o *Clustering* (Gráfico 3).

Gráfico 3: Proceso de extracción de conocimiento utilizado



Fuente: Elaboración propia

4 http://databank.worldbank.org/data/reports.aspx?source=world-development-indicators

5 *Knowledge Discovery in Databases* (KDD), que se refiere al proceso no trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información.

5.2.1. Selección

La serie temporal de los datos utilizados en este trabajo comprende un periodo de diez años (2004 a 2014) de 217 países, con indicadores relacionados al comercio y otros indicadores macroeconómicos. Sin embargo, ha sido necesaria la depuración, en el set de datos, de países que no tienen información o presentan muchos datos nulos, quedándonos con 188 países de cinco continentes.

5.2.2. Procesamiento

El cargado y procesamiento de los datos ha sido realizado con la herramienta *data integration*[6] de *Pentaho*, para lo cual se crearon varios procesos ETL para integrar los datos descargados de *DataBank* y almacenarlos en un solo repositorio de datos. Asimismo, se diseñó y generó una base de datos de países y continentes con dimensiones adicionales con las que no contaba el set de datos del Banco Mundial; una de las dimensiones adicionales e importantes es la relacionada con la mediterraneidad de los países.

5.2.3. Transformación

La transformación de los datos también fue realizada mediante procesos ETL, debido a que las bases de datos del Banco Mundial identifican los nombres de los países en inglés, y para hacer la integración con otras bases de datos se tuvo que generar enlaces a partir del código de tres caracteres que tienen los países. Asimismo, para facilitar la georreferenciación, la visualización de los países y las agrupaciones, se ha integrado información georreferenciada de los países que son objeto de estudio.

5.2.4. Minería de datos

Es importante, al momento de realizar un proceso de análisis y generación de conocimiento a través de la exploración de volúmenes grandes de información, contar con una herramienta que facilite el proceso de generación de conocimientos. En este sentido, un lenguaje de programación que tiene un entorno de fácil acceso y a su vez un enfoque de análisis estadístico, es el lenguaje R, que fue desarrollado inicialmente por Robert Gentleman y Ross Ihaka, del Departamento de Estadística de la Universidad de Auckland, en 1993. R brinda un entorno

6 https://www.hitachivantara.com/en-us/products/big-data-integration-analytics/pentaho-data-integration.html

colaborativo y abierto, por lo cual se cuenta con una amplia gama de herramientas estadísticas a través de librerías de acceso gratuito.

a) Cálculo del error cuadrático en la formación del *cluster*

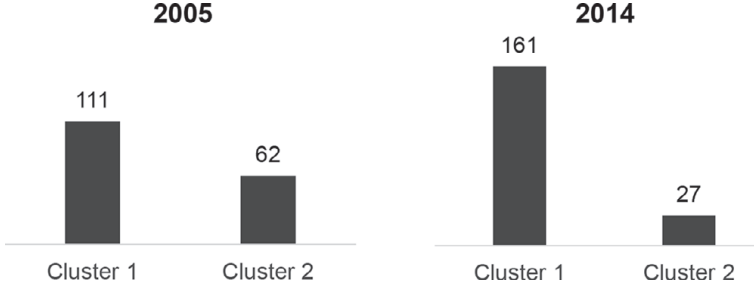
Es necesario identificar el número de *clusters* a ser formados antes de ejecutar los algoritmos *K-means* y PAM, por lo que se apeló al cálculo del error cuadrático SSE de cada agrupación y su visualización, para determinar la cantidad de agrupaciones a ser generadas (Anexo 1). Producto del cálculo realizado, se observa que el incremento menos significativo para determinar el número de agrupaciones se da entre 2 a 3 *clusters*. Las disminuciones siguientes son cada vez menores, por lo que se identifica que la formación óptima se da para la generación de dos agrupaciones.

b) Aprendizaje no supervisado – *Clustering*

i. Algoritmo K-medoids, PAM (*Partitioning Around Medoids*)

Los *clusters* generados con el algoritmo PAM para la gestión 2005 agrupan a 111 países dentro del *cluster* 1 y los restantes 62 países en el *cluster* 2. Para la gestión 2014 se incrementa el número de países que integran el *cluster* 1, llegando a 161, afectando de esta manera a la cantidad de países que integran el *cluster* 2, que se reduce a 27 (Gráfico 4).

Gráfico 4: Formación de clusters mediante PAM



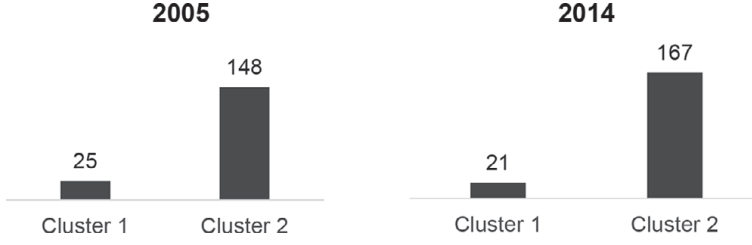
Fuente: Elaboración propia

En una década, muchos países que fueron parte del *cluster* 2, pasaron a integrar el *cluster* 1. Este hecho será parte de un análisis posterior, debido a la importancia que suscita para el estudio.

ii. Algoritmo *K-means*

Los resultados de la ejecución del algoritmo *K-means* para el periodo 2005-2014 generan una distribución de las observaciones representada en dos grupos de datos o *clusters*. Para el periodo 2005, el *cluster* 2 se forma con 148 países, cantidad que se incrementa en la gestión 2014, llegando a 167 países. El *cluster* 1 reduce su composición de 25 a 21 países (Gráfico **5**).

Gráfico 5: Formación de *clusters* mediante PAM



Fuente: Elaboración propia

En esta etapa, la composición de los *clusters* ya muestra una segmentación de países que debe ser evaluada para nutrir el análisis de este estudio.

5.2.5. Evaluación de los algoritmos mediante *silhouette*

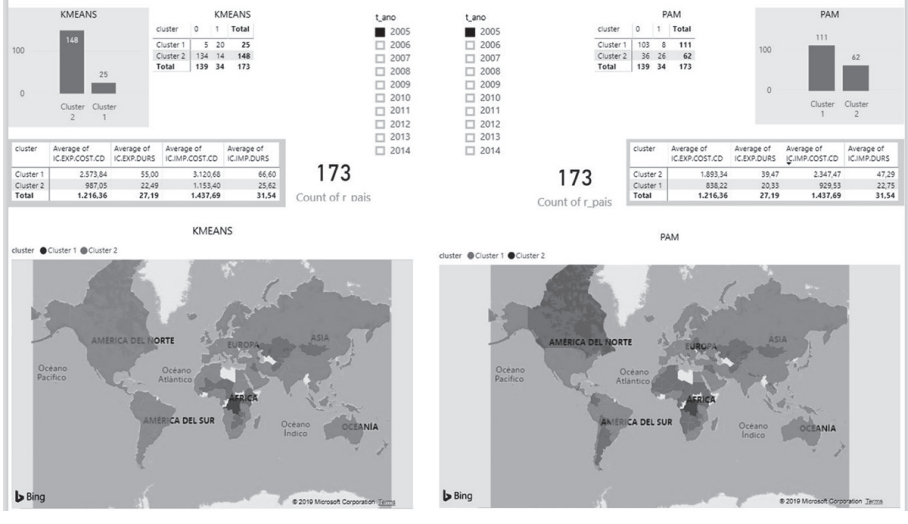
Los resultados del algoritmo PAM son evaluados en dos periodos, dando como producto un promedio de *silhouette* de 0.54 para la gestión 2005, siendo éste aceptable para el emparejamiento de las observaciones. Sin embargo, para el periodo 2014 el emparejamiento muestra un promedio que llega a 0.74, que denota un escenario mucho más favorable en la asignación de las observaciones a los *clusters* (Anexo 2).

En la evaluación del algoritmo *K-means*, el promedio *silhouette* para la gestión 2005 es de 0.60, por lo que se puede decir que la solución de los *clusters* de emparejamiento de cada observación es adecuada. Los *clusters* formados para la gestión 2014 tienen un promedio más alto, llegando a 0.77, por lo que se puede concluir que el emparejamiento es mejor que el del primer periodo (Anexo 2).

5.2.6. Resultados

Se han seleccionado cuatro variables de 188 países por un periodo de diez años; mediante el cálculo del error cuadrático en la formación de *clusters*, se ha identificado que el óptimo se da para la generación de dos agrupaciones (Anexo 1). Los *clusters* formados mediante el algoritmo *K-means* separan en el año 2014 a 167 países en el *cluster* 1 como los países con costos más bajos y tiempos más cortos de exportación e importación, y agrupa en el *cluster* 2 a 21 países cuyos tiempos y costos de exportación e importación son más altos.

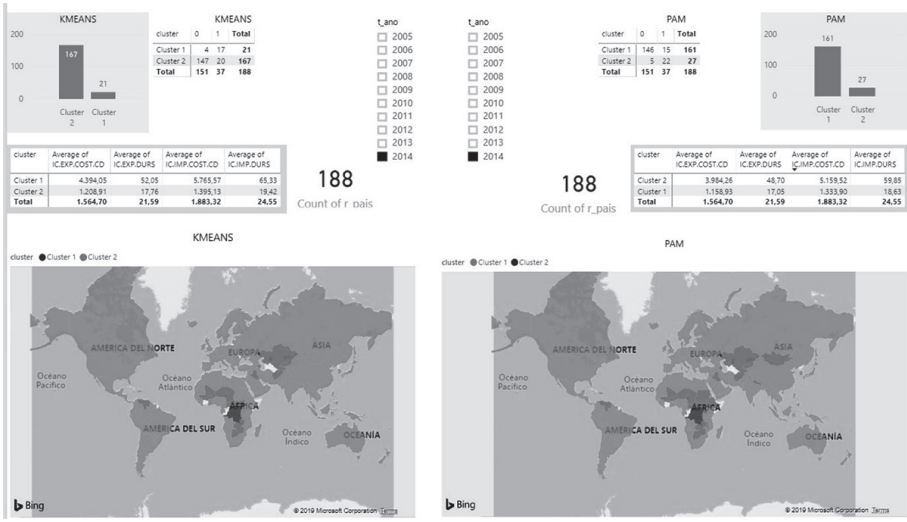
Gráfico 6: *Clusters* del algoritmo *K-means* Y PAM (2005)



Fuente: Elaboración propia

Los *clusters* formados mediante el algoritmo PAM separan en 2014 a 161 países en el *cluster* 1, como los países con costos más bajos y tiempos más cortos de exportación e importación, y agrupan en el *cluster* 2 a 27 países cuyos tiempos y costos de exportación e importación son más altos.

Gráfico 7: *Clusters* del algoritmo *K-means* y PAM (2014)



Fuente: Elaboración propia

En un periodo de diez años, según los resultados del algoritmo PAM, algunos países que formaron parte del *cluster* 2 (países con tiempos y costos más altos de exportación e importación) en el año 2005, como Colombia, Bolivia, Argentina y Paraguay, dejaron de ser asignados a esta agrupación y se trasladaron al *cluster* 1. En el caso de Bolivia, durante el periodo 2005-2014 mejoraron sus indicadores en tiempos de exportación, pasando de 24 a 19 días. Un avance aún más importante se manifiesta en los tiempos de importación, pasando de 36 a 23 días, mostrando una reducción de 13 días en un periodo de 10 años.

Durante el periodo de estudio, también se han encontrado disminuciones en los costos de exportación e importación, pasando de US$ 1440 a 1425 en la exportación de un contenedor, y para la importación de US$ 1747 a 1452, durante el periodo 2005-2014. Las mejoras en estos indicadores han provocado que Bolivia pase del *cluster* 2 durante las gestiones 2005 y 2006, al *cluster* 1 de la gestión 2007 en adelante.

Los mapas de *clusters* (Gráficos 6 y 7) sugieren que existirían barreras más fuertes que el enclaustramiento marítimo, dado que, en algunos casos, países que sí tienen salida al mar presentan características en comercio internacional similares a los países mediterráneos,

o incluso menos favorables que este último grupo. Así lo evidencia la transición de *clusters*, comparando los años 2005 y 2014. Tal es el caso de Canadá, Colombia, Argentina y algunos países de Europa, los cuales en 2005 fueron clasificados por el algoritmo PAM en el *cluster* donde radicaban la mayoría de los países mediterráneos.

También se encuentra que existen países sin litoral que han sido capaces de aliviar los efectos negativos de la condición de mediterraneidad; como son los casos de Bolivia y Paraguay en Latinoamérica. Estos países, hasta el año 2006, pertenecían al *cluster* de países clasificados como mediterráneos. A partir de la gestión 2007, los algoritmos sugieren que Bolivia, particularmente, habría acortado las brechas en las variables de comercio internacional de mejor manera que la mayoría de los países mediterráneos. También resulta interesante la incorporación de Venezuela, mediante el algoritmo *K-means* (Gráfico 7), al *cluster* de países mediterráneos en 2014; sugiriendo que la situación económica actual que se vive en ese país habría reducido las ventajas con las que contaba en materia de costos y tiempos en el comercio exterior.

Estos resultados son coherentes con las relaciones de variables de comercio propuestas por el modelo de gravedad de comercio internacional. Encuadrado dentro de la economía internacional, el modelo establece que el comercio entre dos países (o conjuntos de países, para efectos del documento) es proporcional al tamaño económico de ambos, medido por el PIB e inversamente proporcional a la distancia que existe entre ambos. Matemáticamente, se expresa de la siguiente forma:

$$F_{ij} = G * \frac{M_i * M_j}{D_{ij}} \tag{7}$$

Donde F_{ij} representa el flujo comercial entre el país "i" y el país "j"; G es una constante; D es la distancia entre los dos países; y M representa el tamaño de la economía de los países. En esta línea, siguiendo a Raballand (2003), el modelo de gravedad puede extenderse de la siguiente forma:

$$F_{ij} = f\left(PIB_i,\ D_{ij}, Instit_i, Acceso_i, Medit_i, Infra_i, Tarifa_i\right) \tag{8}$$

Donde PIB_i es el Producto Interno Bruto del país i; $Instit_i$ es el desarrollo institucional; $Acceso_i$ representa la menor distancia entre el país i y un mercado mundial mayor; $Medit_i$ identifica la condición de mediterraneidad; $Infra_i$ es la calidad de infraestructura vial; y $Tarifa_i$ es una medida del costo de exportaciones e importaciones.

En este sentido, la clasificación de Bolivia y Paraguay como países no mediterráneos, además de ser explicada por los costos del comercio internacional, podría haberse suscitado por la mejora en términos económicos de estos países. Bolivia ha tenido tasas de crecimiento del PIB por encima del promedio regional, principalmente por su fuerte componente de demanda interna e inversión en infraestructura; Paraguay, por su parte, es una de las economías emergentes que también ha mostrado un notable crecimiento, principalmente por la inversión extranjera que atrae, lo cual es resultado de la mejora de su institucionalidad. El caso de Venezuela es particular y atípico para la muestra, por lo que la recesión que enfrenta en este periodo ha influido notablemente en su desempeño en comercio internacional, a tal punto que ha mostrado características de enclaustrado.

5.2.7. Análisis discriminante

El análisis discriminante es una técnica estadística multivariante cuya finalidad es describir las diferencias significativas (si existen) entre g grupos de objetos ($g > 1$) sobre los que se observan p variables (variables discriminantes). Más concretamente, se comparan y describen las medias de las p variables clasificadoras a través de los g grupos.

En caso de que estas diferencias existan, se intentará explicar en qué sentido se dan, y proporcionar procedimientos de asignación sistemática de nuevas observaciones con grupo desconocido a uno de los grupos analizados, utilizando para ello sus valores en las p variables clasificadoras (éstos sí son conocidos).

Cuadro 7
Análisis discriminante PAM

Indicador	Resultado	Indicador	Resultado
Accuracy (exactitud)	74.57%	*Accuracy* (exactitud)	89.36%
Sensibilidad o precisión	41.94%	Sensibilidad o precisión	81.48%
Recall	76.47%	*Recall*	59.46%
Especificidad	74.10%	Especificidad	96.69%
Prevalencia o tasa de incidencia	19.65%	Prevalencia o tasa de incidencia	19.68%

Fuente: Elaboración propia

La exactitud, o "*accuracy*" en inglés, permite calcular la relación de aciertos respecto al total de observaciones realizadas, siendo para el caso más óptimo el valor 1 (100%) o en el otro extremo 0. Para el algoritmo PAM, los niveles de exactitud al momento de comparar los resultados de los *clusters* generados con la realidad de los países mediterráneos, muestran para los años 2005 y 2014 un porcentaje de 74.57% y 89.36%, respectivamente (Cuadro 7).

Para el caso del algoritmo *K-means* (Cuadro 8), los porcentajes de exactitud en la formación de *clusters* relacionados a los países mediterráneos y no mediterráneos, llegan a un 89.02% en 2005 y 87.23% en 2014. Los resultados de la evaluación discriminante de ambos algoritmos permiten apreciar que la formación de los *clusters* a partir de los datos de costos y tiempos de exportación, tiene un alto porcentaje de exactitud, llegando a aproximadamente a un 89% de efectividad en el caso de PAM para la gestión 2014 y *K-means* para la gestión 2005.

Cuadro 8
Análisis discriminante *K-means*

Indicador	Resultado	Indicador	Resultado (%)
Accuracy (exactitud)	89.02%	*Accuracy* (exactitud)	87.23
Sensibilidad o precisión	80.00%	Sensibilidad o precisión	80.95
Recall	58.82%	*Recall*	45.95
Especificidad	96.40%	Especificidad	97.35
Prevalencia o tasa de incidencia	19.7%	Prevalencia o tasa de incidencia	19.68

Fuente: Elaboración propia

6. Conclusiones

Los países mediterráneos enfrentan diferentes restricciones económicas identificadas en la literatura, debido principalmente a las grandes distancias que tienen para transar en los grandes mercados, la dependencia en la política exterior de los países vecinos con salida al mar y los altos costos del comercio internacional. En el modelo propuesto de *clustering* de las variables costo y tiempo para la exportación e importación mediante los algoritmos *K-means* y PAM, se explora la dinámica de los países sin litoral para evidenciar si efectivamente la posición mediterránea condiciona a estos países a enfrentar permanentemente brechas significativas en los costos y tiempos para el comercio internacional.

Con ambos algoritmos se ha llegado a determinar la formación óptima de dos *clusters,* con información de 188 países; posteriormente, mediante pruebas de silueta, se ha comprobado esta optimización. El primer *cluster* agrupa a una gran cantidad de países no mediterráneos, entre desarrollados y en desarrollo; mientras que el segundo *cluster* agrupa a la mayoría de economías mediterráneas en desarrollo (LLDC).

Las soluciones de los algoritmos han sido probadas y evaluadas en su consistencia. Para ello, se recurrió al análisis discriminante, el cual indica, mediante la tasa de exactitud, precisión y especificidad, que es significativo no rechazar la hipótesis de que los países identificados por los *clusters* 1 y 2 pertenecen en gran medida a tales grupos; en otras palabras, la "clusterización" fue eficiente. Es preciso enfatizar que, para este proceso, no se introdujo al modelo información *a priori* sobre si cada país es mediterráneo o no.

De esta forma, los resultados sugieren que existirían barreras más fuertes que el enclaustramiento marítimo, dado que, en algunos casos, países que sí tienen salida al mar presentan características similares a los países mediterráneos, o incluso menos favorables que este último grupo. Así lo evidencia la transición de *clusters* comparando los años 2005 y 2014. Tal es el caso de Canadá, Colombia, Argentina y algunos países de Europa, los cuales en 2005 fueron clasificados por el algoritmo PAM en el *cluster* donde radicaban la mayoría de los países mediterráneos.

Por otra parte, se evidencia que existen países sin litoral que han sido capaces de aliviar los efectos negativos de la condición de mediterraneidad; como son los casos de Bolivia y Paraguay

en Latinoamérica. Estos países pertenecían hasta el año 2006 al *cluster* de países clasificados como mediterráneos. A partir de 2007, los algoritmos sugieren que Bolivia, particularmente, habría acortado las brechas en las variables de comercio internacional de mejor manera que la mayoría de los países mediterráneos. También es notable la incorporación de Venezuela, mediante el algoritmo *K-means*, al *cluster* de países mediterráneos en 2014; sugiriendo que la situación económica actual que se vive en ese país, habría reducido las ventajas con las que contaba en materia de costos y tiempos en comercio exterior.

Para Bolivia, esta reducción de brechas se explicaría por razones cuyo estudio puede dar continuidad a esta investigación. La estabilidad del tipo de cambio, el subsidio a los hidrocarburos, la mayor apertura comercial por los acuerdos de integración comercial y el impulso a la demanda interna, incluida la inversión en infraestructura que se dio a partir de 2007, habrían repercutido en mejoras en costos y tiempos para la exportación e importación de bienes y servicios. De hecho, en el ámbito mundial, en 2004, de 188 países, Bolivia se situaba en el puesto 130 respecto al indicador de costos de exportación (un mayor ranking equivale a menores costos), mientras que en 2014 ascendió al puesto 124. En el ámbito regional, respecto al mismo indicador, de 34 países de América, Bolivia habría escalado del puesto 29 al 26 en 2014.

Fecha de recepción: 3 de abril de 2019
Fecha de aceptación: 16 de septiembre de 2019
Manejado por ABCE/SEBOL/IISEC

Referencias

1. Arvis, J. F., Marteau, J. F. y Raballand, G. (2010). "The cost of being landlocked: logistics costs and supply chain reliability". *The World Bank*.
2. De, Prabir. (2006). "Trade, infrastructure and transaction costs: the imperatives for Asian economic cooperation". *Journal of Economic Integration*, 21(4), 708-735.
3. Driffield, N. y Jones, C. (2013). "Impact of FDI, ODA and migrant remittances on economic growth in developing countries: A systems approach". *The European Journal of Development Research*, 25(2), 173-196.
4. Faster, D. S. (2014). *Pentaho Data Integration*.
5. Faye, M. L., McArthur, J. W., Sachs, J. D. y Snow, T. (2004). "The challenges facing landlocked developing countries". *Journal of Human Development*, 5(1), 31-68.
6. Fayyad, U., Piatetsky-Shapiro, G. y Smyth, P. (1996). "The KDD process for extracting useful knowledge from volumes of data". *Communications of the ACM*, 39(11), 27-34.
7. Grigoriou, C. (2007). "Landlockedness, infrastructure and trade: new estimates for central Asian countries". *The World Bank, Development, Research Group.*
8. Kaufman, L. y Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis* (Vol. 344). Nueva York: John Wiley & Sons.
9. Lahiri, B. y Masjidi, F. K. (2012). "Landlocked countries: A way to integrate with coastal economies". *Journal of Economic Integration*, 27(4), 505-519.
10. Leiva-Valdebenito, S. A. y Torres-Avilés, F. J. (2010). "Una revisión de los algoritmos de partición más comunes en el análisis de conglomerados: un estudio comparativo". *Revista Colombiana de Estadística*, 33(2), 321-339.
11. MacKellar, L., Wörgötter, A. y Wörz, J. (2000). "Economic development problems of landlocked countries". *Transition Economic Series*, Nº 14.
12. MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations". *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* 1(14), 281-297.
13. Mendoza, R., Céspedes, A., Ticona, U. *et al.* (2018). "Restricciones al comercio y al desarrollo económico en países mediterráneos: Impacto en el crecimiento, la pobreza y el comercio, el caso de Bolivia". *Revista de Análisis del BCB*, 28(1), 231-301.

14. Paudel, R. C. (2014). "Economic Growth in Developing Countries: Is Landlockedness Destiny?" *Economic Papers: A journal of applied economics and policy*, 33(4), 339-361.

15. Pérez-López, C. (2008). *Minería de datos: técnicas y herramientas*. Madrid, España: Thomson Ediciones.

16. Radelet, S. y Sachs, J. D. (1998). "Shipping costs, manufactured exports, and economic growth". *Annual Meeting of the American Economics Association,* Chicago.

17. Raballand, G. (2003). "Determinants of the negative impact of being landlocked on trade: an empirical investigation through the Central Asian case". *Comparative Economic Studies*, 45(4), 520-536.

18. Shrestha, H. y Heffley, D. (2003). "Regional Integration and Industrial Location in a Landlocked Spatial Economy". Economics Working Papers, University of Connecticut.

19. Smith, A. (1796) *An Inquiry into the Nature and Causes of the Wealth of Nations*, 2 vols. Editado por Edwin Caanan. University of Chicago, Chicago, IL.

20. UN-OHRLLS. (2013). *The development economics of landlockedness: understanding the development costs of being landlocked*. Nueva York: United Nations.

21. Wagstaff, K., Cardie, C., Rogers, S. y Schrödl, S. (2001, June). "Constrained k-means clustering with background knowledge". *ICML*, (1), 577-584.

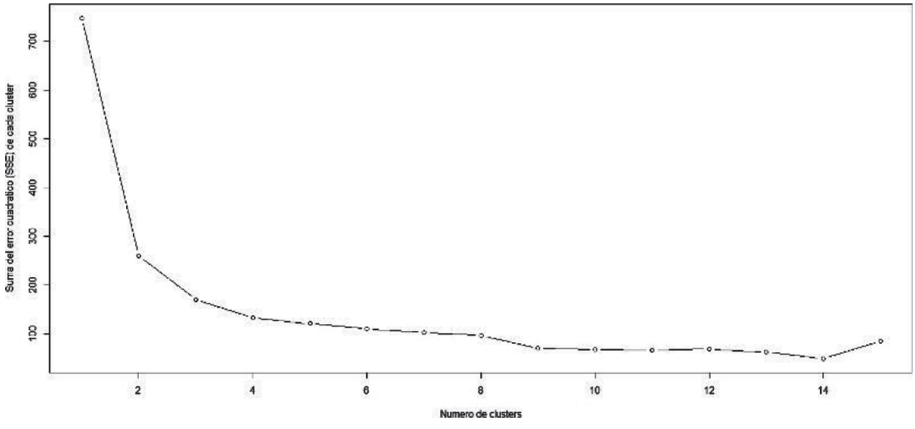
22. Wamboye, E. (2012). *External debt, trade and FDI on economic growth of least developed countries*. Pennsylvania State University.

Anexos

Anexo 1

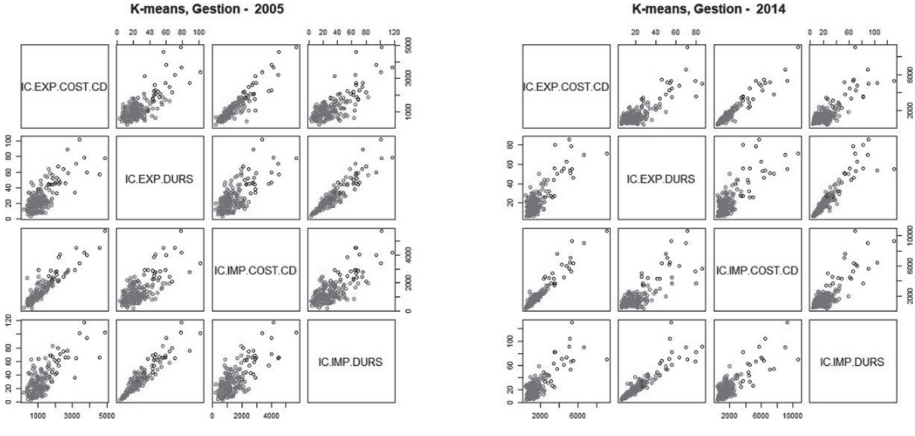
Optimización de *clusters*

Gráfico A1: Error cuadrático de los clusters



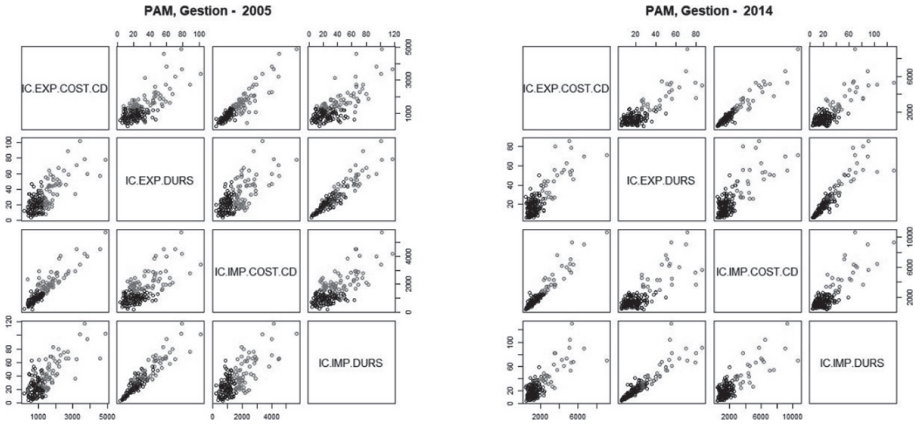
Nota: Este esquema permite observar que el incremento menos significativo para determinar el número de agrupaciones se da entre 2 a 3 *clusters*.
Fuente: Elaboración propia

Gráfico A2: Optimización PAM del número de *clusters*



Fuente: Elaboración propia

Gráfico A3: Optimización *K-means* del número de *clusters*



Fuente: Elaboración propia

Anexo 2

Evaluación de los algoritmos mediante *silhouette*

Para la evaluación se recurrió a un método de interpretación y validación de consistencia dentro de grupos de datos, denominado "silueta". La técnica proporciona una representación gráfica sucinta de la exactitud en la que cada objeto se encuentra dentro de su grupo.

El valor de silueta es una medida de cuán similar es un objeto a su propio *cluster* (cohesión) en comparación con otros *clusters* (separación). La silueta varía de -1 a +1, donde un valor alto indica que el objeto está bien adaptado a su propio *cluster* y está poco relacionado con los *clusters* vecinos. Si la mayoría de los objetos tienen un valor alto, entonces la configuración de agrupamiento es apropiada. Si muchos puntos tienen un valor bajo o negativo, entonces la configuración del *cluster* puede tener demasiados o muy pocos *clusters*[7].

Supongamos que los datos se han agrupado a través de cualquier técnica, como *K-means*, en k *clusters*. Para cada dato i, permita que a(i) sea la distancia promedio entre i y todos los demás datos dentro del mismo *cluster*. Se puede interpretar a(i) como una medida de qué tan bien i está asignado a su *cluster* (cuanto menor es el valor, mejor es la asignación). Luego definimos la

7 La silueta se puede calcular con cualquier medida de distancia, como la distancia euclidiana o la distancia de Manhattan.

disimilitud promedio (diferencia / varianza) del punto i a un *cluster* c como el promedio de la distancia desde i a todos los puntos en c.

Sea b(i) la distancia promedio más pequeña de i a todos los puntos en cualquier otro *cluster*, de los cuales i no es miembro. Se dice que el *cluster* con la menor diferencia promedio (diferencia/varianza) es el "*cluster* vecino" de i porque es el siguiente *cluster* que mejor se ajusta para el punto i. Ahora se define una silueta:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{B1}$$

La ecuación (7) puede también escribirse como:

$$\begin{gathered} s(i) = 1 - \frac{a(i)}{b(i)} \;\; ; \\ si\, a(i) < b(i); \\ si\, a(i) = b(i)\frac{b(i)}{a(i)} - 1; \\ si\, a(i) > b(i) \end{gathered} \tag{B2}$$

De la definición anterior en (8), es evidente que:

$$-1 \le s(i) \le 1 \tag{B3}$$

Además, se debe tener en cuenta que la puntuación es 0 para *clusters* con tamaño = 1. Esta restricción se agrega para evitar que la cantidad de *clusters* aumente significativamente.

Para s(i) estar cerca de 1, se requiere a(i)<b(i). Como a(i) es una medida de cuán diferente es i para su propio *cluster*, un valor pequeño significa que está bien emparejado. Además, un b(i) grande implica que i está mal adaptado a su *cluster* vecino. Por lo tanto, un s(i) cercano

a uno significa que los datos están agrupados apropiadamente. Si s(i) está cerca del negativo, entonces, con la misma lógica, se ve que i sería más apropiado si estuviera agrupado en su *cluster* vecino. Un s(i) cerca de cero significa que el dato está en el borde de dos *clusters* naturales.

El s(i) promedio de todos los puntos de un *cluster* es una medida de cuán estrechamente agrupados están todos los puntos del *cluster*. Por lo tanto, el s(i) promedio sobre todos los datos es una medida de qué tan apropiadamente se han agrupado los datos. Si hay demasiados o muy pocos conglomerados, como puede ocurrir cuando se utiliza una mala elección de k en el algoritmo de agrupación, algunos de los *clusters* típicamente mostrarán siluetas mucho más estrechas que el descanso. Por lo tanto, los gráficos de silueta y los promedios se pueden usar para determinar el número natural de conglomerados dentro de un conjunto de datos. También se puede aumentar la probabilidad de que la silueta se maximice en el número correcto de *clusters* al volver a escalar los datos utilizando pesos de entidades que son específicos del *cluster*.

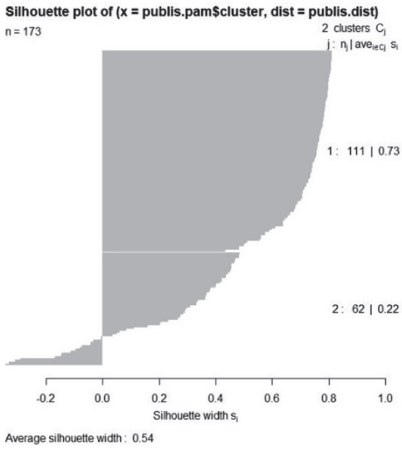
Una forma de evaluar los datos de salida del algoritmo es generando un gráfico de silueta de *K-means* y de PAM. Este cálculo es realizado para cada dato, de manera de visualizar qué tan bien se adapta al *cluster* al que fue asignado. Esto se hace comparando la cercanía de la observación con las otras observaciones dentro del mismo *cluster*.

Los valores cerca de 1 indican que el dato está bien ubicado en su *cluster*, mientras que los valores cercanos a 0 indican que es probable que el dato realmente deba pertenecer a otro *cluster*. Dentro de cada *cluster*, el valor se muestra de menor a mayor. En caso de que la mayoría de los valores estén cerca de 1, se concluye que el ajuste es bueno, pero si hay muchas observaciones cercanas a 0, es una indicación de la deficiencia del ajuste.

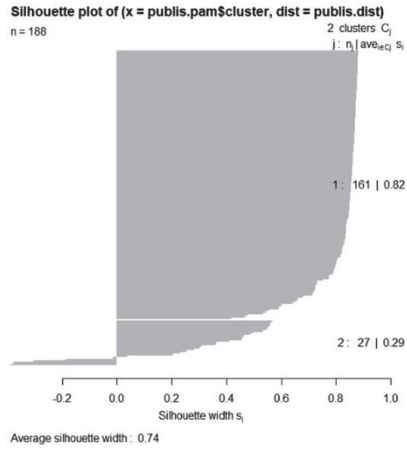
Las soluciones encontradas por los algoritmos *K-means* y PAM son evaluadas en dos momentos: la primera evaluación para el periodo 2005 y la otra para el periodo 2014, cuyas representaciones gráficas permiten observar lo siguiente (Gráficos B1 y B2):

Gráfico B1: Solución de PAM mediante *Silhouette*

2005



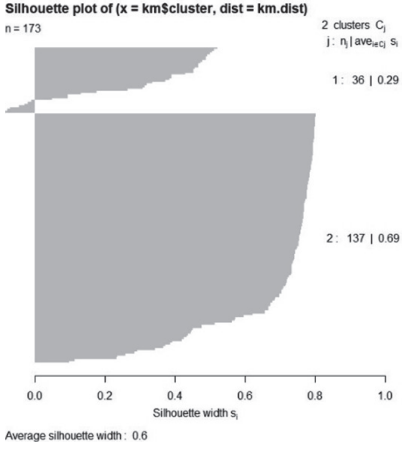
2014



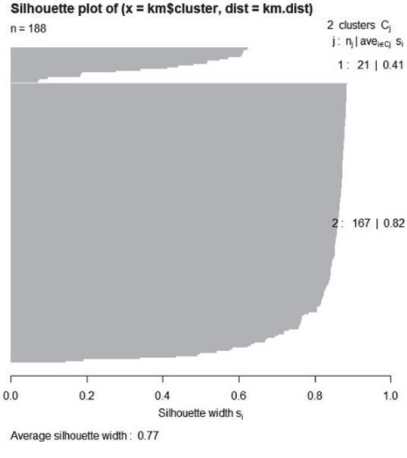
Fuente: Elaboración propia

Gráfico B2: Solución de *K-means* mediante *Silhouette*

2005



2014



Fuente: Elaboración propia