ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

FACULTAD DE ECONOMÍA Y NEGOCIOS

"CARACTERIZACIÓN ECONÓMICA Y EMPRESARIAL DE LAS PROVINCIAS DEL ECUADOR A TRAVÉS DE TÉCNICAS PARA EL ANÁLISIS DE DATOS MULTIVARIADOS, AL AÑO 2010"

TESIS DE GRADO:

Previo a la obtención del título de:

MASTER EN ECONOMÍA Y DIRECCIÓN DE EMPRESAS

Presentado por:

Holger Geovanny Cevallos Valdiviezo

DIRECTORA PROPUESTA:

Ing. Patricia Valdiviezo Valenzuela

Guayaquil –Ecuador

2013

DEDICATORIA

A Dios, a mi Salvador, quien me amó, y ha estado conmigo en todo momento dándome la sabiduría para poder culminar esta etapa tan importante en mi vida. De la misma manera a mis padres, hermanos, tíos, tías, y a mi abuelita; quienes con sus consejos, soporte, y compañía han sido una ayuda esencial para mí y una inspiración especial para proponerme grandes metas, ser siempre el mejor, y lograr durante mi vida todos los objetivos propuestos.

Holger Cevallos Valdiviezo

AGRADECIMIENTO

A Dios, a mis padres por su apoyo incondicional, a mis profesores por sus enseñanzas en el aula de clases, a Ariana R. por su cariño y gran apoyo desinteresado en la conformación de este documento, a mi directora de tesis, la Ing. Patricia Valdiviezo Valenzuela, por su ayuda y tiempo brindado.

TRIBUNAL DE GRADUACIÓN

Econ. Omar Maluk Uriguen
PRESIDENTE DEL TRIBUNAL

Ing. Patricia Valdiviezo Valenzuela
DIRECTORA DEL PROYECTO

Econ. Fabricio Zanzzi Díaz
VOCAL PRINCIPAL

DECLARACION EXPRESA

“La responsabilidad por los hechos, ideas y doctrinas expuestas en este proyecto me corresponden exclusivamente, y el patrimonio intelectual de la misma a la ESCUELA SUPERIOR POLITECNICA DEL LITORAL”

Holger Geovanny Cevallos Valdiviezo

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE GRÁFICOS

Caracterización económica y empresarial de las provincias del Ecuador a través de técnicas para el análisis de datos multivariados, al año 2010

1 Introducción

En este estudio, presentaremos un análisis de caracterización de las diferentes actividades económicas así como de las políticas y características empresariales en cada una de las provincias del Ecuador, usando procedimientos para el análisis estadístico de datos multivariados tales como: Análisis de Componentes Principales (PCA, por sus siglas en inglés) y Análisis de Clústers. Estas técnicas nos ayudan a ganar comprensión en la variabilidad de las variables de una base de datos multivariados y a agrupar observaciones o sujetos similares respectivamente, ambos empleando una reducción de dimensiones, del espacio original de p variables, a $q <<< p$ dimensiones.

El objetivo de usar estos procedimientos multivariados será entonces de resumir y contrastar de una manera efectiva y diferente los datos agregados para cada una de las provincias del Ecuador recolectados por el Instituto Nacional de Estadística y Censos (INEC) durante el Censo Económico del año 2010, en los rubros de caracterización de actividades económicas y de políticas/características empresariales, además de agrupar provincias con actividades económicas y políticas/características empresariales similares.

Las conclusiones de este proyecto podrán ser útil para los agentes encargados de tomar decisiones políticas y económicas a nivel macro y micro, así como para los empresarios, quienes desean tener este tipo de información para saber en dónde invertir, según su área de interés y/o especialización, y conocer sobre las políticas/características empresariales actuales en cada una de las provincias. Alternativamente, el presente análisis busca proveer un universo de investigación para el diseño de encuestas económicas y estructurales, luego de concluir con los sectores (provincias) en donde se realiza tal actividad o tema de interés. Este estudio resulta alternativo al clásico análisis estadístico descriptivo (el cual, ya ha sido brevemente presentado en los informes del INEC para el Censo Económico 2010) y que generalmente resulta difícil de entender e interpretar cuando se posee muchas variables e información.

2 Marco Teórico o Marco Referencial [1]

La economía de Ecuador depende de dos recursos básicos, la agricultura y el petróleo. Es un país pobre con una renta per cápita de unos 4.500 dólares estadounidenses. La agricultura contribuye con un 6% al PIB, y da trabajo al 38% de la población activa. No obstante, es esencial para capas mayores de la población. La industria aporta el 35% del PIB y acoge el 13% de los trabajadores, y los servicios suman el 59% del PIB y el 49% de la población activa.

La agricultura tiene un carácter dual, uno tradicional de subsistencia, para autoconsumo y para satisfacer las necesidades alimentarias del país, y otro de plantación para la exportación, en la que predominan las técnicas de la revolución verde. Los principales productos que se cultivan son arroz, trigo, cebada, maíz, arvejas (guisantes), frijoles (alubias), habas, lentejas, patatas, yuca, cebolla, col (repollo), tomate, aguacate, naranja, mandarina, naranjilla, piña, limón, higuerilla, maní, soja, palma africana (palmitos y aceite de palma), algodón, abacá, café, cacao, banano, plátano, caña de azúcar y tabaco. La región agrícola por excelencia es la Costa. La agricultura en la Sierra se distribuye según sus pisos climáticos. Es el ámbito de la agricultura tradicional.

Uno de los recursos más importantes de Ecuador es la ganadería, que se desarrolla, sobre todo, en los pisos medios de la Sierra. Se trata de una ganadería semi-extensiva que da servicio a las ciudades. Es uno de los sectores más dinámicos dentro de la producción agropecuaria. La mayor proporción corresponde a la ganadería bovina, tanto de carne como de leche. El ovino subsiste en situación muy precaria.

La silvicultura es un sector muy importante, gracias a los grandes bosques de maderas de gran calidad. Se localiza, principalmente, en el noroeste y en la región Oriental del país. Las principales especies disponibles son el canelo, chanul, mascarey, tangaré y fernansánchez, higuerón, árbol del algodón y balsa. Especies foráneas como el eucalipto y el pino se localizan en la región interandina. En la región de la Costa se producen por medio de siembra ochoma y caucho.

Los recursos pesqueros marítimos ecuatorianos son enormes, pero su flota pesquera es muy débil. Los principales productos son el atún, dorado, lenguado, corvina y pez espada. La acuicultura es un sector emergente.
Ecuador posee un indudable potencial minero. Cuenta con importantes recursos de oro, plata, cobre, antimonio, plomo, zinc, platino y otros elementos menores asociados. Se localizan, sobre todo, en la Sierra. El potencial aurífero se encuentra en las provincias de Cañar, Azuay y El Oro. El principal yacimiento es Portovelo. Pero el principal recursos es el petróleo, que se explota en el Oriente.

[1] Fuente: Website “La Guía” (http://geografia.laguia2000.com/economia/ecuador-economia)

La industria ecuatoriana es muy débil. Se desarrolló gracias a una política de substitución de las importaciones. Es un sector bastante protegido. Dos son los subsectores fundamentales: la industria agroalimentaria y la petroquímica. Los principales productos son: textiles, alimentos, bebidas, tabaco, fertilizantes, refinado de petróleo y producción de cemento. La actividad industrial se concentra en las dos provincias más pobladas: Guayas y Pichincha; y sobre todo Guayaquil y Quito.

El comercio es el sector servicios fundamental, tanto por lo que a comercio interno se refiere como al comercio exterior, ya que Ecuador es un país eminentemente exportador. Asociada a esta actividad están las remesas de divisas de los emigrantes y el turismo, que desarrollan el sector bancario.

La red viaria es muy débil, y no conecta adecuadamente todo el país. La más importante es la carretera Panamericana, que atraviesa de norte a sur desde Tulcán hasta Macará. Es la principal arteria de comunicación, tanto en el interior como con los países vecinos.

El turismo es un sector en auge, gracias a los innegables recursos naturales del país. Las islas Galápagos son el punto más demandado, pero por su alta protección no el más visitado. Se estima que el turismo es el cuarto sector económico de Ecuador.

A través de este estudio, basado en un análisis estadístico exploratorio, se desea conocer las actividades o sectores económicos que predominan en cada una de las provincias/regiones del Ecuador, así como la realidad empresarial en cada una de las provincias/regiones del país. Las políticas empresariales fueron medidas a través de ciertas variables que caracterizan el ámbito empresarial de las unidades económicas, disponibles en el Censo Económico efectuado por el INEC en el 2010 (ver definición de unidad económica según el Censo mencionado en la Sección 5.2). Asimismo, se desea agrupar provincias según actividades económicas y según políticas/características empresariales similares, de manera que los agentes de la política económica y empresarios puedan localizar sectores geográficos en el país en donde estos aspectos sean muy parecidos.

3 Justificación

Este proyecto se lo realiza ya que es necesario procesar el gran volumen de información (un gran número de variables) disponibles en la base de datos del INEC correspondiente al Censo Nacional Económico del año 2010, y entregar a los agentes de la política económica, empresarios y público en general, resultados resumidos y concretos acerca de una unidad de estudio en particular de esta base de datos. En nuestro caso, queremos investigar las actividades económicas y políticas/características empresariales de cada una de las provincias del Ecuador (unidad de estudio), como se diferencian las provincias (y/o regiones) las unas de las otras y como se relacionan y asemejan ciertas provincias (y/o regiones) en estos aspectos. Este proyecto realiza un análisis alternativo al descriptivo disponible a través del INEC en su sitio Web, usando técnicas multivariadas tales como el de Componentes Principales o el de Clúster, las cuales permiten analizar un gran número variables simultáneamente con el objetivo de resumir, presentar y agrupar los datos en q <<< p dimensiones (e.g. en el "biplot" q = 2) y así caracterizar a las provincias del Ecuador económica y empresarialmente.

Como fue mencionado anteriormente de manera breve, nuestras conclusiones podrán ser usadas por agentes encargados de tomar decisiones políticas y económicas a nivel macro y micro, quienes desean no solamente poseer información clara y oportuna, pero también obtener un gran volumen de información resumida, respaldada con gráficos fáciles de interpretar. En base a nuestras conclusiones, ellos podrán destinar recursos para la especialización de cada provincia en cada una de las actividades económicas identificadas, invertir en el desarrollo de industrias afines en cada grupo de provincias que desarrolle actividades y políticas empresariales similares, promover la formación y el desarrollo de empleados así como la investigación en provincias que necesiten prioritariamente de ello, conocer sobre cuáles son las provincias que aportan más con contribuciones (impuestos) en comparación a otras, conocer las provincias que usan más recursos energéticos en comparación a otras e implementar medidas para el uso eficiente de estos recursos, entre otras políticas. De la misma manera, los empresarios desearían poseer reportes con breves explicaciones acerca de gran cantidad de información y gráficos fáciles de interpretar, para decidir en qué provincias invertir según su área de interés y/o especialización, y conocer sobre las políticas empresariales actuales en cada una de las provincias del Ecuador a las que se enfrentarían si desearían invertir en tal o cual sector geográfico del Ecuador. Adicionalmente, nuestras conclusiones servirán como base para establecer universos de investigación para el diseño de posteriores encuestas económicas y estructurales.

4 Objetivos

4.1 Objetivo general:

Resumir y reportar las características económicas y empresariales de cada una de las provincias del país, así como identificar grupos de provincias con similitudes en cuanto a estos aspectos que ayuden a esclarecer y representar la distribución económica-geográfica del país, usando la base de datos del Censo Económico del INEC efectuado en el año 2010.

4.2 Objetivos específicos:

- Conocer las características económicas y empresariales de cada una de las provincias (y regiones) del Ecuador
- Identificar las semejanzas y diferencias entre las provincias del Ecuador, en el ámbito económico y empresarial
- Procesar y resumir una gran cantidad de información disponible en la base de datos del INEC correspondiente al Censo Económico 2010; a través del análisis multivariado, con el objetivo de proveer información interpretable y práctica a los agentes tomadores de decisión en la economía ecuatoriana, así como para los empresarios
- Esclarecer y representar la distribución económica-geográfica del país por provincias
- Proveer un universo de investigación para el diseño de encuestas económicas y estructurales, luego de concluir con los sectores (provincias/regiones) en donde se realiza tal actividad o tema de interés

5 Acerca del Censo Económico 2010 llevado a cabo por el INEC y su metodología

El INEC, en su sitio Web, puso a disposición del público en general las siguientes definiciones de términos censales así como información metodológica general acerca del Censo Económico llevado a cabo en el 2010:

5.1 ¿Cómo se definió al Censo Económico?

En el mencionado Censo, se definió a la población de la siguiente manera:

"Todas las unidades económicas que conforman el sector productivo, su ubicación, así como el registro de sus características principales. Se lleva a cabo mediante una serie de visitas a los establecimientos económicos".

5.2 ¿A quién se censó?

5.2.1 Locales auxiliares

Lugar determinado y VISIBLE que da soporte a un establecimiento pero no ejerce una actividad productiva. Ejemplo: bodega, parqueadero.

5.2.2 Establecimientos Visibles

Unidad económica que bajo una sola dirección o control combina actividades y recursos con la finalidad de producir bienes y servicios y está ubicada en lugar determinado. Ejemplo: sucursal de un supermercado, gasolinera, tienda de abarrotes.

5.2.3 Empresas

Persona natural o jurídica autónoma en sus decisiones financieras y de administración, propietaria o administradora de uno o más establecimientos. Ejemplo: cadena de supermercados, bancos, empresas públicas.

5.3 ¿Dónde se censó?

- Áreas amanzanadas (2.000 y más habitantes)
- Corredores viales principales
- Zonas de actividad económica especial

- Grandes empresas (eje transversal)

5.4 Algunos parámetros del proceso censal

- Sectores visitados: 22.684 sectores censados.
- Establecimientos registrados: 511.130 establecimientos analizados
- Durante 30 años los insumos para el cálculo de toda la estadística económica se basaron en la información del censo anterior (1980).
- Las fechas de empadronamiento fueron de Septiembre a Noviembre del 2010.
- El número de personal involucrado fue de 1.900 personas entre personal operativo y administrativo del proceso.
- El Universo Sectorial de investigación contiene los siguientes sectores: Manufactura, construcción, comercio, restaurantes y hoteles, transporte y comunicaciones, intermediación financiera, servicios inmobiliarios y a las empresas, administración pública, educación, salud, servicios sociales y personales.

5.5 Potencial información a partir del Censo

- El censo permite actualizar, después de 30 años, la información productiva y económica del Ecuador.
- El censo dará toda la información de base al Sistema de Cuentas Nacionales, el cual viene usando información proyectada a partir del censo de 1980.
- Esta información podrá unirse con los datos del censo de población y vivienda, disponibles desde agosto del 2011.
- El censo puede dar información a varios niveles de desagregación, tanto en el territorio como en ramas de actividad.

5.6 Etapas del Censo

Para esta fase existió una sinergia entre el Censo de Población y Vivienda y Censo Nacional Económico. La misma consistió en:

- Recabar información de la ubicación geográfica de los establecimientos económicos visibles (actividad que se realiza en un lugar fijo, separado del hogar) e invisibles (actividad que se realiza en un local que no es fijo o dentro del hogar) con referencia gráfica en mapas y planos.
- Identificar la actividad económica de los establecimientos.
- Realizar la digitalización de los establecimientos recogidos en la actualización cartográfica.
- Empadronamiento o Censo

5.7 Empadronamiento o Censo

Consiste en el barrido territorial y en los operativos especiales del universo y muestras a ser investigados con el fin de definir un directorio. Para el efecto se empleará un cuestionario ampliado.

5.8 Encuesta Exhaustiva

Para esta última fase se empleará una encuesta por muestra probabilística representativa del marco de empresas y establecimientos económicos empadronados, enfocada en la actividad económica sectorial, para luego elaborar una matriz de insumo producto y demás tipos de información.

5.9 Contenido de la base de datos correspondiente al Censo Económico 2010

A continuación, enlistaremos las variables que resumen la información recogida en el Censo Económico 2010 [2]:

#	Nombre de la variable	Rótulo	Tipo	Rango
1				
2		Nombre de provincia		
2.1	S1P2	Codigo de provincia	C	
2.2	PROVINCI	Nombre de provincia	C	
3		Nombre de canton		
3.1	S1P3	Codigo de canton	C	
3.2	CANTON	Nombre de canton	C	
4				
4.1	S2P41	Sexo del gerente o propietario	I	1-2
4.2	S2P6	Posee calificación artesanal el gerente o propietario del establecimiento	I	1-2
4.3	S2P7	Local propio o arrendado	I	1-2
4.4	S2P8	Tipo de establecimiento	I	1-4
4.5	S2P9	Tiene ruc el establecimiento	I	1-2
4.6	S4P6	Actividad de comercio al por mayor o menor	I	1-2
4.7	S4P7C1	Principal cliente a nivel local	I	1-3
4.8	S4P7C2	Principal cliente a nivel provincial	I	1-3
4.9	S4P7C3	Principal cliente a nivel nacional	I	1-3
4.10	S4P7C4	Principal cliente a nivel exterior	I	1-3
4.11	S5P1	Registros contables	I	1-2
4.12	S6P1	Establecimiento matriz sin fines de lucro	I	1-2
4.13	S6P2	Forma del establecimiento matriz	I	1-10
4.14	S6P3	Financiamiento para el establecimiento	I	1-2
4.15	S6P5	Establecimiento requiere financiamiento	I	1-2
4.16	S6P6	Establecimiento realizo investigaciones de mercado	I	1-2
4.17	S6P8	Gasto en manejo de desechos	I	1-2
4.18	S6P9	Gasto en investigación y desarrollo	I	1-2
4.19	S6P10	Gasto en capacitación y formación	I	1-2
4.20	S6P11	Uso de internet	I	1-2
4.21	S6P12	Afiliación a un gremio	I	1-2

[2] Esta información fue extraída del sitio web del INEC, específicamente de:
http://redatam.inec.gob.ec/cgibin/RpWebEngine.exe/PortalAction?&MODE=MAIN&BASE=CENEC&MAIN=WebServ
http://redatam.inec.gob.ec/cgibin/RpWebEngine.exe/PortalAction?&MODE=MAIN&BASE=CENEC&MAIN=WebServerMain.inl

4.22	S9P1	Información de todos los establecimientos de la empresa matriz	I	1-2
4.23	S2P5	Año de constitución del establecimiento	I	0-99999999
4.24	S3P11C2	Personal remunerado hombres	I	0-99999999
4.25	S3P11C3	Personal remunerado mujeres	I	0-99999999
4.26	S3P11C1	Total de personal remunerado	I	0-99999999
4.27	S3P12C2	Personal no remunerado hombres	I	0-99999999
4.28	S3P12C3	Personal no remunerado mujeres	I	0-99999999
4.29	S3P12C1	Total de personal no remunerado	I	0-99999999
4.30	S3P13C1	Total personal ocupado	I	0-99999999
4.31	S3P13C2	Total hombres ocupados	I	0-99999999
4.32	S3P13C3	Total mujeres ocupadas	I	0-99999999
4.33	S4P4C2	Año de inicio de la actividad principal	I	0-99999999
4.34	S5P61C1	Existencias al 01 de enero productos en proceso	R	
4.35	S5P61C2	Existencias al 31 de diciembre productos en proceso	R	
4.36	S5P62C1	Existencias al 01 de enero productos terminados	R	
4.37	S5P62C2	Existencias al 31 de diciembre productos terminados	R	
4.38	S5P63C1	Existencias al 01 de enero materias primas y auxiliares	R	
4.39	S5P63C2	Existencias al 31 de diciembre materias primas y auxiliares	R	
4.40	S5P64C1	Existencias al 01 de enero mercadería sin transformación	R	
4.41	S5P64C2	Existencias al 31 de diciembre mercadería sin transformación	R	
4.42	S5P71C1	Compras de activos fijos existencias al 31 de diciembre	R	
4.43	S5P72C1	Construcción de activos fijos existencias al 31 de diciembre	R	
4.44	S5P73C1	Ventas y/o bajas existencias al 31 de diciembre	R	
4.45	S5P74C1	Valor de activos fijos existencias al 01 de enero	R	
4.46	S5P74C2	Valor de activos fijos existencias al 31 de diciembre	R	
4.47	S6P31	Monto de financiamiento	R	
4.48	S6P51	Monto requerido de financiamiento	R	
4.49	S6P81	Monto de gasto en manejo de desechos	R	
4.50	S6P91	Monto de gasto en investigación y desarrollo	R	
4.51	S6P101	Monto de gasto en capacitación y formación	R	
4.52	CPCPE2D	Producto elaborado a 2 dígitos	C	
4.53	CPCPE3D	Producto elaborado a 3 dígitos	C	
4.54	CPCPE4D	Producto elaborado a 4 dígitos	C	
4.55	CPCPC2D	Producto comercializado a 2 dígitos	C	
4.56	CPCPC3D	Producto comercializado a 3 dígitos	C	
4.57	CPCPC4D	Producto comercializado a 4 dígitos	C	
4.58	CPCSO2D	Servicio ofrecido a 2 dígitos	C	
4.59	CPCSO3D	Servicio ofrecido a 3 dígitos	C	
4.60	CPCSO4D	Servicio ofrecido a 4 dígitos	C	
4.61	CPCMP2D	Materia prima a 2 dígitos	C	
4.62	CPCMP3D	Materia prima a 3 dígitos	C	
4.63	CPCMP4D	Materia prima a 4 dígitos.	C	
4.64	RGNATU	Regiones naturales	I	0-9
4.65	NCASE	Número de establecimientos	I	0-99999999
4.66	GASTREM	Gastos anuales en remuneraciones	R	
4.67	REMU	Gastos mensuales en remuneraciones	R	
4.68	GASTMAT	Gastos anuales en materia prima	R	
4.69	GASTRAC	Gastos anuales en repuestos y accesorios	R	
4.70	GASTEE	Gastos anuales en envases y embalajes	R	
4.71	GASTCOM	Gastos anuales en compras y mercadería	R	
4.72	GASTTER	Gastos anuales por servicios prestados por terceros y alquileres	R	
4.73	EGRESOS	Otros egresos anuales corrientes	R	
4.74	INTERES	Intereses anuales pagados	R	
4.75	TAX	Tasas, contribuciones y otros impuestos anuales (excluye iva, ice)	R	
4.76	INGRESOS	Total de ingresos anuales percibidos por ventas o prestación de servicios	R	
4.77	OTRING	Otros ingresos anuales	R	
4.78	INGEXT	Ingresos extraordinarios anuales	R	
4.79	FINANC	Fuentes de financiamiento	I	1-6
4.80	GENERG	Gastos anuales kilovatios/hora	R	
4.81	KWHA	Kilovatios/hora anual	R	
4.82	TRAPER	Estratos de personal ocupado	I	0-6
4.83	TRAING	Estratos de ingresos percibidos por ventas o prestación de servicios	I	0-8

4.84	GANOCON	Años de constitución del establecimiento	I	0-12
4.85	GANOINI	Años de inicio de la actividad principal	I	0-12
4.86	MCS	Sectores	I	1-4
4.87	RP1	Ciiu a un dígitos actividad principal	C	
4.88	RPD1	Clasificación ciiu 4.0 actividad principal	I	0-9999
4.89	RS1	Ciiu a un digito actividad secundaria	C	
4.90	RSD1	Clasificación ciiu 4.0 actividad secundaria	I	0-9999
4.91	RP2	Ciiu a dos dígitos actividad principal	C	
4.92	RP3	Ciiu a tres dígitos actividad principal	C	
4.93	RP4	Ciiu a cuatro dígitos actividad principal	C	
4.94	RS2	Ciiu a dos dígitos actividad secundaria	C	
4.95	RS3	Ciiu a tres dígitos actividad secundaria	C	
4.96	RS4	Ciiu a cuatro dígitos actividad secundaria	C	
4.97	RPD2	Descripción ciiu principal a dos dígitos	I	0-9999
4.98	RSD2	Descripción ciiu secundaria a dos digito	I	0-9999
4.99	RPD3	Descripción ciiu principal a tres dígitos	I	0-9999
4.100	RSD3	Descripción ciiu secundaria a tres dígitos	I	0-9999
4.101	RPD4	Descripción ciiu principal a cuatro dígitos	I	0-9999
4.102	RSD4	Descripción ciiu secundaria a cuatro digito	I	0-9999
4.103	NATJUR	Naturaleza jurídica	I	0-9
4.104	Z6P411	Monto de financiamiento con institución pública	R	
4.105	Z6P412	Monto de financiamiento con institución privada	R	
4.106	Z6P42	Monto de financiamiento con el gobierno	R	
4.107	Z6P43	Monto de financiamiento con institución no reguladas por el sbs	R	
4.108	Z6P441	Monto de financiamiento con otras fuentes con garantía	R	
4.109	Z6P442	Monto de financiamiento con otras fuentes sin garantía	R	
4.110	CANTON	Nombre de Cantón	C	
Procesado con Redatam+SP;CENEC 2010				

Tabla 1: Variables recogidas en el Censo Económico 2010.

La primera columna (“ # ”) de la Tabla corresponde al número de pregunta que dio origen a la variable correspondiente, la segunda columna corresponde al nombre con el que la variable fue registrada en el archivo de datos (.dat) y que aparecerá como encabezado cuando se proceda a leer el archivo de datos en el lenguaje estadístico R (software estadístico usado en este estudio), la tercera columna corresponde al nombre completo de la variable tratada en la pregunta correspondiente, la cuarta columna corresponde al tipo de variable asignado por el INEC: “C” (variable categórica), “R” (variable numérica), “I” (variable con entradas de números enteros), y la quinta y última columna representa los niveles posibles para las variables de tipo “I” (solamente).

5.10 Comentarios adicionales

El censo analizó 511.130 establecimientos en el país. Se podría argumentar que los datos de este censo no constituirían el total de la población descrita, ya que existe mucha informalidad en ciertas actividades económicas en el país y esto podría haber causado potenciales errores al momento del empadronamiento previo, realizado a las unidades económicas. Por tanto, esto nos haría alegar que los datos del presente censo no constituirían el universo de unidades económicas que conforman el sector productivo, y en consecuencia se debería considerar a esta como una muestra *“no*

representativa" de la población ya que la misma no sería el resultado de un proceso metodológico de un diseño experimental.

Adicionalmente, se conoce que 572.335 establecimientos fueron visitados, de los que sólo 511.130 fueron analizados. Las razones por las cuales no se analizó la diferencia de establecimientos (61,205) no fueron especificados por el INEC en sus reportes (e.g. del sitio web) e informes públicos.

Sin embargo, en este estudio, nos limitaremos al uso de estos datos recogidos por el INEC, por ser la única fuente en la que hemos encontrado esta información "censal", además por ser una fuente gratuita, de fácil y rápido acceso que representa datos oficiales al alcance de todos.

6 Metodología: Descripción breve de las técnicas estadísticas multivariadas a usar, Análisis de Componentes Principales (PCA), Análisis Clúster [3]

6.1 Análisis de Componentes Principales (PCA, por sus siglas en inglés)

El Análisis de Componentes Principales (PCA) es un método multivariado que es a menudo usado para reducción de dimensiones. Es un método que transforma las p variables originales $X_1, \dots, X_p$ a p *componentes principales* (PC) $Z_1, \dots, Z_p$ sin ninguna pérdida de información. Aunque la transformación desde p dimensiones a p dimensiones no parece una reducción de dimensiones, resulta que gracias a la construcción específica de los PC's, es a menudo suficiente considerar sólo $q < p$ PC's de tal manera que estos q PC's contengan mayor información que cualquier q de las variables originales. Ilustraremos que estos PC's no son construcciones arbitrarias, pero que a menudo pueden ser interpretados de tal manera que una buena comprensión de la variabilidad de la muestra (o de la población en si) pueda ser adquirida.

PCA es considerado como un método estadístico explorativo, i.e. la inferencia estadística (test de hipótesis e intervalos de confianza) no se encuentra involucrada. De esta manera, usaremos las verdaderas medias $\boldsymbol{\mu}$ y la verdadera matriz de varianza y covarianza $\boldsymbol{\Sigma}$, aunque en la práctica, cuando tenemos en nuestras manos una muestra, estas tienen que ser reemplazadas por sus estimados. Esta sustitución trae consecuencias a las propiedades del PCA, pero no iremos en detalle con respecto a esto en este proyecto ya que nuestros datos son censales.

6.1.1 Información versus Varianza

En un curso típico de Diseños de Experimento, se remarca que la varianza de un estimador es inversamente proporcional a la información del parámetro contenido en la muestra. Aquí, parecerá como si argumentáramos todo lo contrario, si no hay variabilidad en los datos (e.g. en la muestra), los datos no contienen información. Daremos un ejemplo simple: suponga que tenemos datos recolectados de árboles (e.g. la altura de los árboles) y datos del hábitat (e.g. nivel sobre el mar). Si no hubiera variabilidad en los datos (muestra), e.g. todos los árboles tienen la misma altura, y todos los hábitats se encuentran al mismo nivel sobre el mar, entonces no podríamos aprender nada de estos datos: los datos no contienen información.

[3] El texto presentado en la sección de Metodología fue tomado de [15], luego de traducirlo del inglés y de realizar ligeras modificaciones de acuerdo al contexto del presente estudio.

Consideremos el mismo ejemplo, pero ahora suponga que existe variabilidad. En particular, usted observa una correlación negativa entre las dos variables. En este caso, una de las dos variables puede ser predicha usando la información de la otra (por supuesto, habría varianza residual). Esto implica que una variable contiene información de la otra variable, y lo opuesto. Decimos entonces que existe cruce de información. Si, por otro lado, las dos variables fueran independientes, ellas tienen correlación igual a cero, y las variables no tienen poder de predicción la una de la otra.

6.1.2 Construcción de los Componentes Principales

Como fue mencionado anteriormente, PCA es básicamente una transformación de *p* variables originales a *p* nuevas variables que son conocidas como PC's. Empezaremos con algunos detalles en la construcción del primer PC. Luego, mostraremos cómo los otros son determinados.

6.1.2.1 El Primer Componente Principal (PC)

Un PC es una combinación lineal de *p* variables $\boldsymbol{X}^t = (X_1, \ldots, X_p)$. La variable multivariada $\boldsymbol{X}$ tiene la matriz de varianza y covarianza $\boldsymbol{\Sigma}$. El primer componente principal (PC) es representado de la siguiente manera:

$$Z = a_1X_1 + a_2X_2 + \ldots + a_pX_p = \boldsymbol{a}^t\boldsymbol{X}$$

Donde $\boldsymbol{a}^t = (a_1, \ldots, a_p)$ es el vector de coeficientes. Queremos que el primer PC contenga tanta información como sea posible, i.e. queremos que Z tenga la máxima varianza entre todas las combinaciones lineales de $\boldsymbol{X}$. Debido a que Z es una combinación lineal de los componentes de la variable multivariada, la varianza de Z puede ser calculada como:

$$\text{Var}\{Z\} = \text{Var}\{\boldsymbol{a}^t\boldsymbol{X}\} = \boldsymbol{a}^t\boldsymbol{\Sigma}\boldsymbol{a}$$

Si no imponemos ninguna restricción, entonces la solución es trivial: tomar todos los coeficientes igual a $+\infty$ y la varianza de Z es también $+\infty$. Por supuesto, esto no es una solución interesante. Por tanto, introducimos una restricción, la norma de $\boldsymbol{a}$ tiene que ser igual a uno, i.e.

$$\boldsymbol{a}^t\boldsymbol{a} = \sum_{i=1}^{p} \boldsymbol{a}_i^2 = 1$$

Para encontrar una solución, sustituimos $\boldsymbol{\Sigma}$ por su representación en descomposición espectral:

$$\text{Var}\{Z\} = \boldsymbol{a}^t\left(\sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i\, \boldsymbol{e}_i^t\right)\boldsymbol{a}$$

$$= \sum_{i=1}^{p} \lambda_i\, (\boldsymbol{a}^t\boldsymbol{e}_i)(\boldsymbol{e}_i^t\boldsymbol{a})$$

$$= \sum_{i=1}^{p} \lambda_i\, (\boldsymbol{a}^t\boldsymbol{e}_i)^2$$

Adoptamos la convención que los valores eigen se encuentran ordenados de la siguiente manera:

$$\lambda_1 > \lambda_2 > \ldots > \lambda_p > 0$$

Dado que $\boldsymbol{a}^t\boldsymbol{a} = 1$ y que $\boldsymbol{e}_i^t\,\boldsymbol{e}_i = 1$, se tiene que:

$$\sum_{i=1}^{p} (\boldsymbol{a}^t\boldsymbol{e}_i)^2 = 1$$

De esta manera, Var $\{Z\} \leq max_i\ \lambda_i = \lambda_1$, cumpliéndose la última igualdad si y sólo si $\boldsymbol{a} = \boldsymbol{e}_1$. Esto nos da inmediatamente la solución para el primer componente principal:

$$\boldsymbol{a} = \boldsymbol{e}_1,$$

y con esta selección, la varianza maximizada de la combinación lineal de los componentes de $\boldsymbol{X}$ es:

$$\text{Var}\{Z\} = \lambda_1$$

6.1.2.2 Interpretación geométrica

Consideramos a $\boldsymbol{X}$ y $\boldsymbol{a}$ como puntos o vectores en un espacio Euclidiano de *p* dimensiones. Entonces, $\boldsymbol{a}^t\,\boldsymbol{X}$ es una proyección ortogonal de $\boldsymbol{X}$ en $\boldsymbol{a}$.

Para obtener una comprensión geométrica más profunda sobre el primer PC *Z,* consideramos la forma cuadrática:

$$\boldsymbol{x}^t\,\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$$

Esto es exactamente la distancia Mahalanobis al cuadrado del punto $\boldsymbol{x}$ al origen $\boldsymbol{0}^t$ = (0,…,0) del espacio de *p* dimensiones. Podemos considerar el arreglo de todos los puntos $\boldsymbol{x}$ que satisfacen:

$$\boldsymbol{x}^t\,\boldsymbol{\Sigma}^{-1}\boldsymbol{x} = c,$$

con *c* como una constante. Sabemos que los puntos $\boldsymbol{x}$ forman una elipse con centro $\boldsymbol{0}$. A continuación, sustituimos $\boldsymbol{\Sigma}^{-1}$ por su representación de descomposición espectral:

$$c = \boldsymbol{x}^t \boldsymbol{\Sigma}^{-1} \boldsymbol{x} = \sum_{i=1}^{p} \frac{1}{\lambda_i} (\boldsymbol{x}^t \boldsymbol{e}_i)^2$$

$$= \sum_{i=1}^{p} \left(\frac{\boldsymbol{x}^t \boldsymbol{e}_i}{\sqrt{\lambda_i}} \right)^2$$

$$= \sum_{i=1}^{p} \left(\frac{y_i}{\sqrt{\lambda_i}} \right)^2 .$$

En donde $y_i = \boldsymbol{x}^t \boldsymbol{e}_i$. En la última ecuación reconocemos inmediatamente la ecuación de una elipse con ejes paralelos a los ejes de la base ($\boldsymbol{e}_1$, $\boldsymbol{e}_2$, ... , $\boldsymbol{e}_p$) en el cual, y_i representa las coordenadas. La elección de $\boldsymbol{x} = \boldsymbol{e}_1$ (primer PC) resulta en $y_1 = 1$, $y_2 = 0$, $y_3 = 0$, ... , $y_p = 0$. Esto es un punto en el eje mayor de la elipse. Este resultado también brinda una interpretación a los vectores eigen: son vectores que forman los ejes de la elipse de densidad constante. Por lo tanto, $\boldsymbol{e}_1$ señala en la dirección correspondiente a la varianza más grande en la muestra.

6.1.2.3 Más componentes principales

Una vez que el primer componente principal es encontrado (el cual por el momento es denotado como Z_1), podemos empezar a buscar al segundo PC. El objetivo final es de capturar tanta información como sea posible de los datos originales (e.g. muestra) en un número de componentes principales tan pequeño como sea posible. Si el segundo PC sería correlacionado con el primer PC, tendríamos un cruce de información, por lo que este no es el enfoque más eficiente. Por lo tanto, introducimos una restricción adicional en el segundo PC, requerimos que Cov $\{Z_2, Z_1\} = 0$. De hecho, lo que queremos es que los dos primeros componentes principales sean independientes, pero esta es una condición muy fuerte. De esta manera, demandaremos solamente que la covarianza o la correlación sea igual a 0 (las dos condiciones son equivalentes si los dos PC's son bivariados normalmente distribuidos).

El segundo componente principal es definido como:

$$Z_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p = \boldsymbol{a}_2^t \boldsymbol{X}$$

De tal manera que Z_2 tenga la máxima varianza entre todas las combinaciones lineales de los componentes de $\boldsymbol{X}$, con la restricción de que (1) $\boldsymbol{a}_2^t \boldsymbol{a}_2 = 1$, y (2) Cov $\{Z_2, Z_1\} = 0$.

La expresión para la varianza de Z_2 es obtenida nuevamente por sustitución de la representación de la descomposición espectral de $\boldsymbol{\Sigma}$,

$$\text{Var}\{Z_2\} = \sum_{i=1}^{p} \lambda_i (\boldsymbol{a}_2^t \boldsymbol{e}_i)^2$$

La covarianza es obtenida a través de operaciones similares:

$$\text{Cov}\{Z_2, Z_1\} = \text{Cov}\{\boldsymbol{a}_2^t \boldsymbol{X}, \boldsymbol{X}^t \boldsymbol{a}_1\}$$

$$= \boldsymbol{a}_2^t \boldsymbol{\Sigma} \boldsymbol{a}_1$$

$$= \sum_{i=1}^{p} \lambda_i \, (\boldsymbol{a}_2^t \boldsymbol{e}_i)(\boldsymbol{e}_i^t \boldsymbol{a}_1).$$

La última expresión inmediatamente sugiere la solución:

La covarianza puede ser solamente 0 si $\boldsymbol{a}_2$ es ortogonal a cada vector eigen $\boldsymbol{e}_i$ excepto para los vectores eigen perpendiculares a $\boldsymbol{a}_1 = \boldsymbol{e}_1$. Dado que los vectores eigen forman una base ortonormal, sabemos que $\boldsymbol{a}_1$ es ortogonal al vector eigen $\boldsymbol{e}_i$ $(i \neq 1)$. Por lo tanto, $\boldsymbol{a}_2$ tiene que ser igual a uno de los vectores eigen $\boldsymbol{e}_i$ $(i = 2, \dots, p)$. Por otro lado, la varianza de Z_2 debería ser máxima. De la ecuación de Var $\{Z_2\}$ arriba, deducimos que:

$$\boldsymbol{a}_2 = \boldsymbol{e}_2,$$

la cual, es la solución que estábamos buscando. La varianza del segundo componente principal es entonces:

$$\text{Var}\{Z_2\} = \lambda_2$$

Las soluciones para los PC's subsecuentes son análogas. En particular, para el *j-enésimo* componente principal (PC):

$$Z_j = a_{j1}X_1 + a_{j2}X_2 + \dots + a_{jp}X_p = \boldsymbol{a}_j^t \boldsymbol{X}$$

Buscamos los coeficientes $\boldsymbol{a}_j$ que maximicen la varianza de Z_j entre todas las combinaciones lineales para las cuales:

$$\text{Cov}\{Z_j, Z_1\} = 0 \text{ y Cov}\{Z_j, Z_2\} = 0 \dots \text{Cov}\{Z_j, Z_{j-1}\} = 0$$

$(j = 3, \dots, p)$. La solución es siempre $\boldsymbol{a}_j = \boldsymbol{e}_j$, y Var $\{Z_j\} = \lambda_j$. Note que hemos mostrado que los PC's son completamente determinados por los valores eigen y vectores eigen de $\boldsymbol{\Sigma}$.

La interpretación geométrica de los PC's se obtiene inmediatamente de los resultados de la sección: los vectores eigen son los ejes de la elipse de densidad constante.

6.1.2.4 ¿Covarianza o Correlación?

Para simplificar las cosas, podríamos decir que el primer PC señala en la dirección de la varianza más grande. Es fácil de ver que, si una de las *p* variables originales tiene una varianza grande relativa comparada con las otras, el componente Z_1 será dominado por esta variable. En algunos set de datos es efectivamente importante detectar esto, y allí no habrá entonces conflicto con el objetivo del análisis de componentes principales (e.g. esto ocurre cuando X_1 es el peso (en kg) en el día 1, y X_2 es el peso (también en kg) en el día 2, etc). Por otro lado, cuando las variables son medidas en diferentes unidades (e.g. X_1 es el peso (en kg), y X_2 es la longitud (en cm)) las proporciones mutuas de las varianzas dependen muy fuertemente en la escala en la cual las variables se encuentran medidas (e.g. el peso X_1 medido en g, y/o longitud X_2 en m). La solución del análisis de componentes principales es entonces sensible a la elección de la escala de las medidas, y esto puede oscurecer la interpretación de los PC's. En este caso, es mejor trabajar con variables estandarizadas:

$$X_{is} = \frac{X_i}{\sigma_i}$$

$(i = 1, \dots, p)$. Para estos encontramos que:

$$\text{Var}\{X_{is}\} = \frac{Var\{X_i\}}{\sigma_i^2} = 1$$

y que:

$$\text{Cov}\{X_{is}, X_{js}\} = \frac{Cov\{X_i, X_j\}}{\sigma_i \sigma_j} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} = \rho_{ij}.$$

Por tanto, la matriz de varianza y covarianza de las variables estandarizadas es igual a la matriz de correlación.

En conclusión, un análisis de componentes principales puede ser implementado usando la matriz de varianza y covarianza, así como usando la matriz de correlación.

Adicionalmente, en general, si queremos que nuestro análisis de PCA ponga un gran énfasis en la variables con las varianzas más grandes, el uso de la matriz de covarianza es la mejor opción. Por otro lado, si queremos tratar cada una de las variables de nuestra base como igualmente importantes, el uso de la matriz de correlación (variables estandarizadas) es la mejor opción.

6.1.3 Determinación del número de componentes principales (PC's)

En nuestra introducción de PCA, remarcamos que uno de los objetivos del análisis de Componentes Principales es cumplir con la reducción de dimensiones, pero hasta ahora, sólo se ha discutido sobre transformar el espacio de la variable multivariada $\boldsymbol{X}$ de p dimensiones a otro espacio de p dimensiones de la variable $\boldsymbol{Z}^t = (Z_1, \ldots, Z_p)$. Hasta ahora, sabemos que el primer componente principal tiene la varianza más alta (i.e. información), el segundo PC tiene la segunda varianza más alta, y así sucesivamente con los demás PC's. Por tanto, esperamos que sea suficiente mirar solamente a un número menor ($< p$) de componentes principales, mientras no se pierda mucha información.

En esta sección discutiremos algunas herramientas importantes que pueden ser usadas para decidir cuántos PC's deberían ser seleccionados, de tal manera que no se pierda mucha información.

6.1.3.1 Valores Eigen

Ya hemos argumentado que consideramos el contenido de información proporcional a la varianza. En consecuencia, parece razonable definir al *contenido total de información* como la varianza total. En particular, lo último viene dado por:

$$\sum_{i=1}^{p} \sigma_i^2 .$$

Por otro lado, podríamos usar una propiedad de los valores Eigen de $\boldsymbol{\Sigma}$:

$$\sum_{i=1}^{p} \lambda_i = \operatorname{tr}(\boldsymbol{\Sigma}) = \sum_{i=1}^{p} \sigma_i^2$$

Y debido a que $\lambda_i = \operatorname{Var}\{Z_i\}$, encontramos que:

$$\sum_{i=1}^{p} \sigma_i^2 = \sum_{i=1}^{p} \operatorname{Var}\{Z_i\}$$

Primeramente, esto muestra que la transformación realizada en el análisis de componentes principales no resulta en pérdida de información. Segundo, esto sugiere también que nosotros podríamos usar:

$$\frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$$

como una medida relativa de la importancia del i-enésimo PC. Esta fracción es la proporción de la información total contenida en el i-enésimo PC.

Basados en estos argumentos, podemos formular una regla práctica (cuando fuere posible): seleccionar PC's de tal manera que al menos el 80% de la varianza total esté contenida en estos PC's seleccionados.

6.1.3.2 Gráfico de Sedimentación (Scree Plot)

El gráfico de Sedimentación es una representación gráfica de los valores propios o valores Eigen. Su aplicación será mostrada a través del siguiente ejemplo:

Se ha aplicado el gráfico mencionado a una base de datos que contiene actividades económicas que caracterizan a un grupo de países europeos. No nos concentraremos en el contenido de esta base de datos, sólo mostraremos el Scree plot (Ilustración 1) para este ejemplo y los interpretaremos para describir su aplicación:

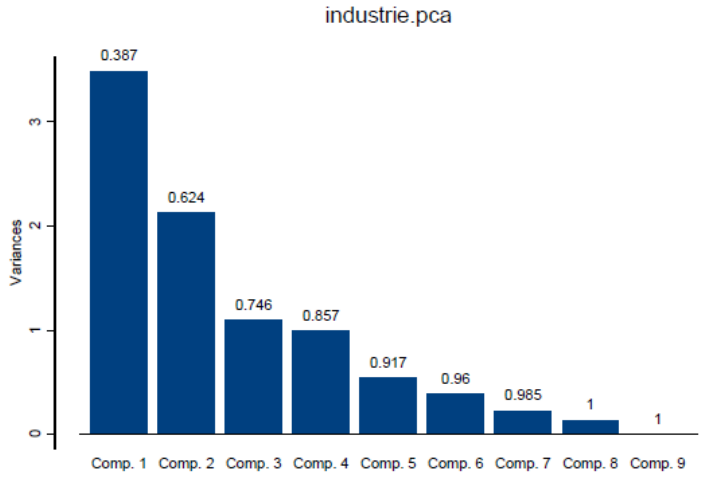


Ilustración 1: Ejemplo de Gráfico de Sedimentación (Scree Plot)

En un gráfico de Sedimentación, se busca esencialmente un codo (o una rodilla); este es el punto con la curvatura más grande. Puesto de otra manera, desde este punto de codo (o rodilla), la declinación de los valores propios cambia drásticamente. En nuestro ejemplo, este punto se encuentra entre PC2-PC3 o entre PC4-PC5. Este razonamiento se debe a lo siguiente: Si decidimos seleccionar PC3, entonces deberíamos también seleccionar PC4, ya que los valores Eigen (o valores propios) de estos componentes principales son casi iguales. Si, por otro lado, nos hubiéramos detenido después de seleccionar el PC2, entonces este problema no se hubiera suscitado, ya que el valor Eigen del PC3 es mucho más pequeño que el valor Eigen del PC2 (i.e. el codo o la

rodilla se encuentra entre PC2-PC3). El mismo razonamiento se puede aplicar para la rodilla formada en el PC4-PC5.

Finalmente, repetimos una vez más que el Gráfico de Sedimentación y la regla del 80% son sólo instrumentos a usar para detectar cuántos PC's deben seleccionarse.

6.1.4 Los "Scores"

Hemos considerado el análisis de componentes principales como una transformación. Esto quiere decir que todas las *n* observaciones son transformadas a *n* observaciones en los componentes principales (PC's). Las observaciones transformadas son referidas o llamadas 'scores'. Estas pueden ser tratadas luego como si fueran observaciones ordinarias, y su interpretación depende de la interpretación dada a los PC's.

A continuación, representaremos la transformación de las observaciones desde el espacio de las variables originales al espacio de los componentes principales, explícitamente como una transformación en notación matricial. Necesitaremos esta representación para comprender el 'biplot' en la próxima sección.

Definimos la matriz $\boldsymbol{M}$ como una matriz $p \times p$ con sus columnas dadas por los vectores eigen $\boldsymbol{e}_i$ de la matriz de varianza y covarianza $\widehat{\boldsymbol{\Sigma}}$ (note que aquí usamos explícitamente la matriz de covarianza estimada). Por tanto:

$$\boldsymbol{M} = (\boldsymbol{e}_1 \ldots \boldsymbol{e}_p).$$

Suponga que $\boldsymbol{X}$ es una matriz de datos con dimensiones $n \times p$. Con esta notación, la matriz de datos transformada con dimensiones $n \times p$ con respecto a los p PC's es dado por:

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{M}.$$

Esta es una notación muy compacta para la transformación. Primero mostraremos una importante propiedad de la matriz $\boldsymbol{M}$. Consideremos lo siguiente:

$$\boldsymbol{M}^{\boldsymbol{t}}\boldsymbol{M} = \begin{bmatrix} \boldsymbol{e}_1^t \\ \vdots \\ \boldsymbol{e}_p^t \end{bmatrix} \begin{pmatrix} \boldsymbol{e}_1 \ldots & \boldsymbol{e}_p \end{pmatrix} = \boldsymbol{I}.$$

El último paso se obtiene por la ortonormalidad de los vectores eigen. De la misma manera, se puede probar que también $\boldsymbol{M}\boldsymbol{M}^t = \boldsymbol{I}$. Decimos que $\boldsymbol{M}$ es una matriz idempotente. Esta propiedad nos da inmediatamente una expresión simple para la transformación inversa:

$$\boldsymbol{Z} = \boldsymbol{XM}$$

$$\boldsymbol{ZM^t = XMM^t}$$

$$\boldsymbol{ZM^t = X},$$

i.e. la matriz de datos original $\boldsymbol{X} = \boldsymbol{ZM^t}$. Previamente, denotamos a las columnas de $\boldsymbol{M}$ como los vectores eigen $\boldsymbol{e}_i$ $(i = 1, \dots, p)$. Ahora, introducimos también una notación para las filas de $\boldsymbol{M}$: que $\boldsymbol{m}_i^t$ denote la i-énesima fila de $\boldsymbol{M}$. Equivalentemente, $\boldsymbol{m}_i$ representa la i-enésima columna de $\boldsymbol{M^t}$. Por tanto:

$$\boldsymbol{X} = \boldsymbol{ZM^t} = \boldsymbol{Z}\,(\boldsymbol{m}_1 \dots \boldsymbol{m}_p)$$

Esta expresión muestra que los vectores $\boldsymbol{m}_i$ representan las coordenadas de las variables originales en el espacio de los componentes principales.

6.1.5 Biplot

6.1.5.1 Construcción

Todo lo que ha sido presentado en las secciones previas permanece válido si habríamos primero centrado las observaciones. En lo que presentamos a continuación, será más fácil trabajar con la matriz de datos centrada. Para reducir la sobrecarga de términos, tomaremos como convención el no cambiar la notación. Por tanto, de ahora en adelante la matriz $\boldsymbol{X}$ representa la matriz de datos centrada, i.e. de cada columna de la matriz de datos originales, la media (muestral) de la columna es substraída. En otras palabras, las medias (muestrales) de cada columna de la matriz de datos centrada es igual a cero. Para esta matriz de datos centrada, inmediatamente encontramos la relación:

$$\boldsymbol{X^t X} = \left[\sum_{i=1}^{n} \left(X_{ij} - \bar{X}_j\right)\left(X_{ik} - \bar{X}_k\right)\right]_{(jk)} = (n-1)\widehat{\boldsymbol{\Sigma}}$$

$(j, k = 1, \dots, p)$.

El “biplot” es un gráfico que resume el análisis completo de componentes principales. En particular, de este gráfico, usted puede leer la interpretación de los componentes principales (los ‘loadings’ y los ‘scores’ son graficados). Además, hasta cierto punto, este gráfico nos permite regenerar las observaciones en las variables originales. Del último párrafo de la sección previa, ya conocemos que la matriz de datos originales puede ser obtenida transformando ‘al reverso’ las observaciones en los *p* PC’s, i.e. $\boldsymbol{X} = \boldsymbol{ZM^t}$. Comenzando desde esta representación de la matriz de datos $\boldsymbol{X}$, primero

explicaremos el “biplot” de *p* dimensiones, el cual, es un gráfico en un espacio de *p* dimensiones. Por supuesto, si *p* > 2 o *p* > 3 no es posible hacer este gráfico. Por tanto, al final de esta sección, explicaremos como este “biplot” ‘virtual’ de *p* dimensiones puede ser aproximado a un “biplot” de 2 dimensiones.

Repetiremos la ecuación $\boldsymbol{X} = \boldsymbol{ZM}^t = \boldsymbol{Z}(\boldsymbol{m}_1 \ldots \boldsymbol{m}_p)$, pero ahora también escribimos la matriz $\boldsymbol{Z}$ con dimensiones $n \times p$, de tal manera que las *n* filas sean mostradas como vectores de *p* dimensiones. Denotemos como $\boldsymbol{z}_1^t \ldots \boldsymbol{z}_n^t$ a las *n* filas de $\boldsymbol{Z}$. El vector $\boldsymbol{z}_i$ por tanto denota la observación i-enésima en el espacio de las componentes principales. Podríamos entonces escribir $\boldsymbol{X} = \boldsymbol{ZM}^t$ de la siguiente manera:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{z}_1^t \\ \vdots \\ \boldsymbol{z}_n^t \end{bmatrix} \left(\boldsymbol{m}_1 \ldots \boldsymbol{m}_p\right) = \begin{bmatrix} \boldsymbol{z}_1^t\boldsymbol{m}_1 & \boldsymbol{z}_1^t\boldsymbol{m}_2 & \cdots & \boldsymbol{z}_1^t\boldsymbol{m}_p \\ \boldsymbol{z}_2^t\boldsymbol{m}_1 & \boldsymbol{z}_2^t\boldsymbol{m}_2 & \ldots & \boldsymbol{z}_2^t\boldsymbol{m}_p \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{z}_n^t\boldsymbol{m}_1 & \boldsymbol{z}_n^t\boldsymbol{m}_2 & \cdots & \boldsymbol{z}_n^t\boldsymbol{m}_p \end{bmatrix}$$

Esta representación claramente muestra que cada elemento X_{ij} de la matriz de datos $\boldsymbol{X}$ puede ser escrito como $\boldsymbol{z}_i^t\boldsymbol{m}_j$. Lo último, es exactamente la proyección ortogonal de los vectores de *p* dimensiones $\boldsymbol{z}_i$ y $\boldsymbol{m}_j$, que representa la i-enésima observación del espacio de componentes principales de *p* dimensiones, y las *p* coordenadas de la variable original *j*-enésima en el espacio de componentes principales, respectivamente.

Con los “scores”, es posible hacer un gráfico de “scores” que es nada más y nada menos que un gráfico "scatter" (o gráfico de dispersión) de los “scores” de los PC1 y PC2. Como fue mencionado anteriormente, dado que normalmente se espera que los dos primeros PC’s sean variables que contengan una gran cantidad de información (un gran porcentaje de la varianza total de las variables originales del set de datos), los dos primeros PC’s tienen ahora una interpretación clara, por lo que podríamos intentar aprender algo sobre los datos al chequear el diagrama de dispersión de los dos primeros PC’s.

Si añadimos los *p* vectores $\boldsymbol{m}_j$ a este gráfico, podríamos obtener una especie de “biplot” que permite las proyecciones que acabamos de discutir. Estas proyecciones permiten una reconstrucción completa de la matriz de datos original $\boldsymbol{X}$ desde las *n* observaciones en el espacio de los componentes principales, y los *p* vectores $\boldsymbol{m}_j$ que contienen las coordenadas de la variable original *j*-enésima en el espacio de los componentes principales. Adicionalmente, de los vectores $\boldsymbol{m}_j$, podemos leer inmediatamente la interpretación de los componentes principales. Por ejemplo, si proyectamos ortogonalmente $\boldsymbol{m}_j$ en el primer vector base del espacio de componentes principales (e.g. el eje horizontal), obtenemos m_{j1} (i.e. el primer elemento del vector

$\boldsymbol{m}_j$). Recordemos que $\boldsymbol{m}_j$ es la *j*-enésima fila de $\boldsymbol{M}$, y que las columnas de $\boldsymbol{M}$ son los vectores eigen $\boldsymbol{e}$. Por tanto $m_{j1} = \boldsymbol{e}_{1j}$, el cual es el *j*-enésimo elemento del primer vector eigen. El valor $m_{j1} = \boldsymbol{e}_{1j}$ nos da entonces el peso o carga (loading) de la variable original *j*-enésima en el primer componente principal. De manera similar, si luego proyectamos otro $\boldsymbol{m}_k$ $(k \neq j)$ nuevamente en el primer vector base del espacio de componentes principales, encontramos que esto nos da $m_{k1} = \boldsymbol{e}_{1k}$, el cual representa al peso o carga ("loading") de la *k*-enésima variable original en el primer componente principal. Si continuamos haciendo esto para todas las *p* variables originales en el primer vector base, encontraremos los pesos o cargas (loadings) de todas las variables en el primer componente principal. Esto nos provee fácilmente la interpretación del primer componente principal. Razonamientos completamente análogos permanecen cuando se proyecta en el segundo vector base del espacio de componentes principales para encontrar las cargas o pesos (loadings) de las variables originales en el segundo componente principal.

Aunque la discusión del párrafo previo ya nos muestra que este tipo de gráficos nos entrega mucha información, la descripción dada, no es la descripción final del "biplot". En el siguiente párrafo, daremos la descripción correcta, la cual requiere solamente algún re-escalamiento de los vectores que son graficados. Este re-escalamiento incrementará incluso la interpretabilidad del "biplot". Antes de continuar, recomendamos leer la siguiente parte, que habla acerca de la Descomposición en Valores Singulares.

6.1.5.2 Descomposición en Valores Singulares

La Descomposición en Valores Singulares (SVD, por sus siglas en inglés) es un teorema que afirma que para cada matriz $\boldsymbol{X}$ de dimensiones $n \times p$, y rango *p*, existe:

- *p* vectores ortonormales $\boldsymbol{l}_i$ de longitud *n*
- *p* vectores ortonormales $\boldsymbol{e}_i$ de longitud *p*
- *p* números reales positivos $\delta_1 \geq \ldots \geq \delta_p$ (comúnmente llamados *valores singulares* de la matriz $\boldsymbol{X}$)

De tal manera que:

$$\boldsymbol{X} = \sum_{i=1}^{p} \delta_i \, \boldsymbol{l}_i \boldsymbol{e}_i^t.$$

Denotemos a $\boldsymbol{E}$ como la matriz $p \times p$ con la columna i-enésima $\boldsymbol{e}_i$, y $\boldsymbol{L}$ la matriz $n \times p$ con la columna i-enésima $\boldsymbol{l}_i$, y $\boldsymbol{\Delta}$ la matriz diagonal $p \times p$, con elemento i-enésimo en la diagonal igual a δ_i. Por tanto, la ecuación $\boldsymbol{X} = \sum_{i=1}^{p} \delta_i \boldsymbol{l}_i \boldsymbol{e}_i^t$ es equivalente a:

$$\boldsymbol{X} = \boldsymbol{L}\,\boldsymbol{\Delta}\,\boldsymbol{E}^t$$

Las columnas de $\boldsymbol{E}$ pueden ser obtenidas también de otra manera. Necesitamos calcular $\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{E}$ sustituyendo las dos matrices $\boldsymbol{X}$ por su Descomposición en Valores Singulares (SVD) mostrados en la ecuación $\boldsymbol{X} = \boldsymbol{L}\,\boldsymbol{\Delta}\,\boldsymbol{E}^t$:

$$\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{E} = \boldsymbol{E}\boldsymbol{\Delta}\boldsymbol{L}^t\boldsymbol{L}\boldsymbol{\Delta}\boldsymbol{E}^t\boldsymbol{E} = \boldsymbol{E}\boldsymbol{\Delta}^2$$

(en donde $\boldsymbol{L}^t\boldsymbol{L} = \boldsymbol{I}$ y $\boldsymbol{E}^t\boldsymbol{E} = \boldsymbol{I}$ debido a la ortonormalidad de las columnas, y $\boldsymbol{\Delta}\boldsymbol{\Delta} = \boldsymbol{\Delta}^2$, ya que ambas son matrices diagonales). Esto es lo miso que:

$$\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{e}_i = \delta_i^2\boldsymbol{e}_i$$

En consecuencia, δ_i^2 y $\boldsymbol{e}_i$ son los valores eigen y los vectores eigen de $\boldsymbol{X}^t\boldsymbol{X} = (n-1)\widehat{\boldsymbol{\Sigma}}$, hasta el factor constante $(n-1)$. Por la ortonormalidad de los dos, $\boldsymbol{e}_i$ y $\boldsymbol{e}_i$, encontramos que $\delta_i^2 = (n-1)\lambda_i$. De la misma manera encontramos que $\boldsymbol{l}_i$ $(i = 1, \dots, p)$ son los vectores eigen de la matriz $\boldsymbol{X}\boldsymbol{X}^t$ (de rango *p*). O, más simple, de la ecuación $\boldsymbol{X} = \boldsymbol{L}\,\boldsymbol{\Delta}\,\boldsymbol{E}^t$ encontramos inmediatamente, multiplicando ambos lados de la ecuación por $\boldsymbol{E}\boldsymbol{\Delta}^{-1}$:

$$\boldsymbol{L} = \boldsymbol{X}\boldsymbol{E}\boldsymbol{\Delta}^{-1}$$

Finalmente, notemos que la matriz $\boldsymbol{E}$ es exactamente la matriz $\boldsymbol{M}$ de las secciones anteriores.

6.1.5.3 Descripción correcta del "biplot"

Denotemos la matriz $\boldsymbol{\Delta}$, como una matriz diagonal que tiene los elementos $\sqrt{\lambda_1}, \dots, \sqrt{\lambda_p}$ en su diagonal. Escribamos:

$$\begin{aligned}
\boldsymbol{X} &= \boldsymbol{Z}\,\boldsymbol{M}^t \\
&= \boldsymbol{Z}\,\boldsymbol{\Delta}^{-1}\frac{\sqrt{n-1}}{\sqrt{n-1}}\,\boldsymbol{\Delta}\boldsymbol{M}^t \\
&= (\boldsymbol{Z}\,\boldsymbol{\Delta}^{-1}\sqrt{n-1})\,(\tfrac{\mathbf{1}}{\sqrt{n-1}}\boldsymbol{M}\boldsymbol{\Delta})^t
\end{aligned}$$

Si definimos:

$$\boldsymbol{G} = \boldsymbol{Z}\ \boldsymbol{\Delta}^{-1}\sqrt{n-1}\text{ y }\boldsymbol{H} = \frac{\mathbf{1}}{\sqrt{n-1}}\boldsymbol{M\Delta},$$

Obtendremos la siguiente representación:

$$\boldsymbol{X} = \ \boldsymbol{G}\,\boldsymbol{H}^t$$

la cual es similar a la igualdad original $\boldsymbol{X} = \boldsymbol{Z}\ \boldsymbol{M}^t$. Debido a que las matrices $\boldsymbol{G}$ y $\boldsymbol{H}$ tienen dimensiones $n \times p$ y $p \times p$, respectivamente, esta representación puede ser graficada exactamente como lo hicimos antes: *n* vectores con *p* dimensiones $\boldsymbol{g}_i$ ($\boldsymbol{g}_i^t$ sos las filas de $\boldsymbol{G}$), y *p* vectores con *p* dimensiones $\boldsymbol{h}_j$($\boldsymbol{h}_j$ son las filas de $\boldsymbol{H}$, o de manera equivalente, las columnas de $\boldsymbol{H}^t$). Debido a que $\boldsymbol{\Delta}$ es una matriz diagonal, y debido a que $\sqrt{n-1}$ es solamente un escalar, los vectores $\boldsymbol{g}_i$ y $\boldsymbol{h}_j$ son sólo versiones re-escaladas de los vectores $\boldsymbol{z}_i$ y $\boldsymbol{m}_j$. Sin embargo, la interpretación dada anteriormente acerca de las proyecciones todavía permanecen (ahora son proyecciones de vectores re-escalados).

Antes que demos una descripción detallada de las propiedades del "biplot" (i.e. lo que podemos concluir o aprender de un "biplot"), daremos dos expresiones más. Para comprender todos los pasos, es necesario comprender acerca de la Descomposición en Valores Singulares (SVD) descrita en la sección anterior:

$$\boldsymbol{HH}^t = \frac{1}{n-1}\boldsymbol{M\Delta}^2\boldsymbol{M}^t = \frac{1}{n-1}\boldsymbol{X}^t\boldsymbol{XMM}^t = \frac{1}{n-1}\boldsymbol{X}^t\boldsymbol{X} = \widehat{\boldsymbol{\Sigma}}$$

$$\boldsymbol{GG}^t = (n-1)\boldsymbol{LL}^t = (n-1)\boldsymbol{XM\Delta}^{-1}\boldsymbol{\Delta}^{-1}\boldsymbol{M}^t\boldsymbol{X}^t = (n-1)\boldsymbol{XM\Delta}^{-2}\boldsymbol{M}^t\boldsymbol{X}^t = \boldsymbol{X}\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}^t$$

La interpretación del "biplot" es basada en las siguientes propiedades:

1. Las proyecciones de los vectores $\boldsymbol{h}_i(i = 1, \ldots, p)$ en el *j*-enésimo eje son los pesos o cargas (loadings) del *j*-enésimo PC (hasta un factor $\sqrt{\lambda_j(n-1)}$)
   Por lo que, la interpretación de los PC's pueden ser leídas desde el "biplot".

2. La ecuación $\boldsymbol{HH}^t = \frac{1}{n-1}\boldsymbol{M\Delta}^2\boldsymbol{M}^t = \frac{1}{n-1}\boldsymbol{X}^t\boldsymbol{XMM}^t = \frac{1}{n-1}\boldsymbol{X}^t\boldsymbol{X} = \widehat{\boldsymbol{\Sigma}}$ implica inmediatamente que:

   $\|\boldsymbol{h}_i\|^2 = \hat{\sigma}_i^2$ i.e. el i-enésimo elemento de la diagonal de $\boldsymbol{HH}^t$

   y que:

   $\boldsymbol{h}_i^t\boldsymbol{h}_j = \hat{\sigma}_{ij}$ i.e. el elemento (i, j)-enésimo de $\boldsymbol{HH}^t$

   (lo último es la proyección ortogonal del vector $\boldsymbol{h}_i$ en el vector $\boldsymbol{h}_j$, o viceversa). Sin embargo, es más fácil mirar la correlación:

$$\hat{\rho}_{ij} = \frac{\hat{\sigma}_{ij}}{\hat{\sigma}_i \hat{\sigma}_j} = \frac{\boldsymbol{h}_i^t \boldsymbol{h}_j}{\|\boldsymbol{h}_i\| \|\boldsymbol{h}_j\|} = \cos(\boldsymbol{h}_i, \boldsymbol{h}_j)$$

Por tanto, el coseno del ángulo formado entre los dos vectores nos da la correlación entre las dos variables originales correspondientes.

3. Similarmente, de la ecuación $\boldsymbol{G}\boldsymbol{G}^t = \boldsymbol{X}\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}^t$ despejada anteriormente, encontramos que:

$$\boldsymbol{g}_i^t \boldsymbol{g}_j = \boldsymbol{X}_i^t \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{X}_j.$$

Por lo que,

$$\|\boldsymbol{g}_i - \boldsymbol{g}_j\|^2 = (\boldsymbol{g}_i^t \boldsymbol{g}_i + \boldsymbol{g}_j^t \boldsymbol{g}_j - 2\boldsymbol{g}_i^t \boldsymbol{g}_j)$$

$$= (\boldsymbol{X}_i - \boldsymbol{X}_j)^t \widehat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{X}_i - \boldsymbol{X}_j).$$

Esta última expresión permanece correcta cuando la observación $\boldsymbol{X}_i$ proviene de una matriz de datos no centrada. Las distancias euclidianas entre puntos en el "biplot" son por tanto iguales a las distancias Mahalanobis.

4. Finalmente, insistimos que cada observación original $\boldsymbol{X}_{ij}$ puede ser reconstruida proyectando $\boldsymbol{g}_i$ ortogonalmente en $\boldsymbol{h}_j$.

6.1.5.4 El "biplot" de dos dimensiones

En las secciones previas hemos argumentado poder hacer un gráfico en *p* dimensiones, como si fuera en realidad posible, pero en realidad, se grafica usualmente sólo en dos dimensiones. Quisiéramos encontrar las matrices $\boldsymbol{A}$ y $\boldsymbol{B}$ de dimensiones $n \times 2$ y $p \times 2$ respectivamente, de tal manera que $\boldsymbol{X} \approx \boldsymbol{A}\boldsymbol{B}^t$, y de tal manera que $\boldsymbol{A}$ y $\boldsymbol{B}$ sean cercanas a las dos primeras columnas de las matrices $\boldsymbol{G}$ y $\boldsymbol{H}$, para obtener la misma interpretación que en el "biplot" de *p* dimensiones. Por supuesto, si *p* > 2, alguna de la exactitud o precisión de la interpretación será perdida cuando analicemos la aproximación en 2 dimensiones.

Existe un teorema (factorización de Gabriel), que dice que si la matriz $n \times p$ $\boldsymbol{X}$ es de rango *r,* existen las matrices $\boldsymbol{A}$ $(n \times r)$ y $\boldsymbol{B}$ $(p \times r)$ de tal manera que $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{B}^t$. Sin embargo, para la mayoría de matrices de datos $\boldsymbol{X}$, el rango es igual a *p* < *n*, y por tanto regresamos a $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{M}^t$.

La solución existe aproximando la matriz $\boldsymbol{X}$ a través de la matriz $\boldsymbol{X}_2$ de rango 2. Esto puede ser logrado considerando la Descomposición en Valores Singulares (SVD) de la matriz $\boldsymbol{X}$ (ver en sección 6.1.5.2):

$$\boldsymbol{X} = \sum_{i=1}^{p} \delta_i \, \boldsymbol{l}_i \boldsymbol{m}_i^t$$

y truncando la descomposición luego del segundo término, resultando entonces en:

$$\boldsymbol{X}_2 = \sum_{i=1}^{2} \delta_i \, \boldsymbol{l}_i \boldsymbol{m}_i^t$$

Debido a la muy cercana relación entre la Descomposición en Valores Singulares (SVD) y el análisis de componentes principales (PCA), sabemos que esta es la aproximación de rango 2 de $\boldsymbol{X}$, que contiene la máxima información de $\boldsymbol{X}$. Por tanto, mientras más porcentaje (%) de la varianza total sea contenido por los dos primeros componentes principales, más podremos confiar en su interpretación.

En notación matricial, escribimos:

$$\boldsymbol{X}_2 = \boldsymbol{L}_2 \, \boldsymbol{\Delta}_2 \boldsymbol{M}_2^t.$$

En donde el subíndice "2" de las matrices $\boldsymbol{L}$ y $\boldsymbol{M}$ significa que solamente las dos primeras columnas de las matrices $\boldsymbol{L}$ y $\boldsymbol{M}$ son seleccionadas, y $\boldsymbol{\Delta}_2$ es la matriz diagonal 2×2 que contiene las raíces cuadradas de los dos primeros valores eigen.

En la Ilustración 2, mostramos un ejemplo gráfico de un "biplot" correspondiente a un set de datos demo llamado "industries".

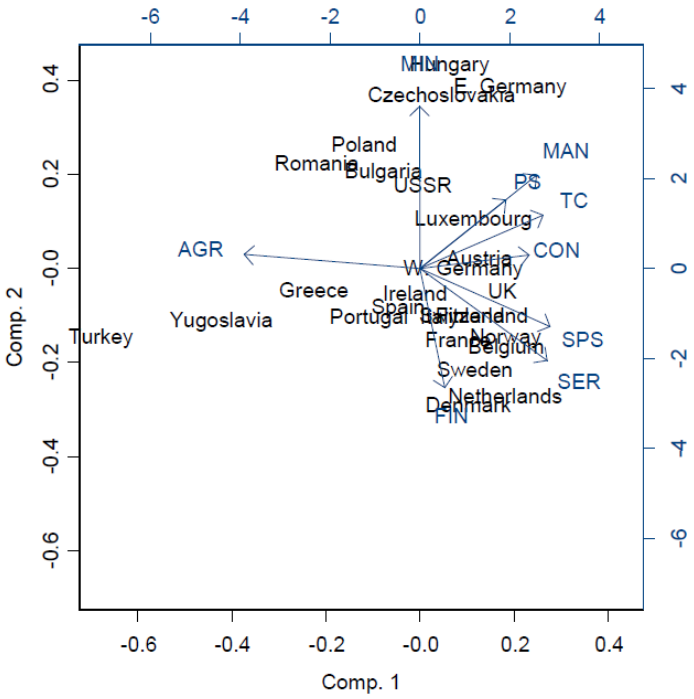


Ilustración 2: Ejemplo gráfico de un "biplot" correspondiente a un dataset demo llamado "industries"

6.2 Análisis Clúster (Clúster Analysis)

6.2.1 Introducción

El objetivo de un análisis clúster es el de agrupar las observaciones multivariadas en un limitado número de grupos (llamados clústers), de tal manera que las observaciones en un mismo clúster sean *similares*, y observaciones en diferentes clústers sean *distintas*. De esta descripción, es claro que las definiciones de *similar* y *distinto* son muy importantes.

Primero, introduciremos un par de ejemplos:

- Datos de microarreglos de DNA pueden ser representados como una matriz $n \times p$ de la cual, cada fila representa los niveles de expresión causados por *p* sondas de cADN. Cada fila corresponde a otro experimento (e.g. células de diferentes órganos, u originados de sujetos en diferentes grupos de tratamiento, …). Por tanto, las filas corresponden a las *muestras*, y las columnas a los *genes.*
  El número de variables (*genes*) puede ser bien extenso (típicamente miles). Una importante pregunta de investigación para la cual los microarreglos son usados con frecuencia es el de encontrar grupos en las muestras con patrones de expresión similares. Por otro lado, la pregunta de investigación puede también consistir en la búsqueda de los grupos de genes con niveles de expresión similares en todas las muestras. Esto podría incrementar la comprensión en la funcionalidad de los genes (e.g. coregulación o genes coexpresados).

  El Análisis Clúster es realizado con frecuencia en la fase exploratoria del análisis de datos, pero este también puede constituir un paso hacia encontrar genes de pronóstico con respecto al desarrollo de cáncer u otras enfermedades.

- En estudios de abundancia existen medidas del conteo de *n* especies de plantas en *p* diferentes campos o hábitats. La pregunta de investigación es: agrupar las especies en grupos que aparecen en los mismos hábitats, o, agrupar los hábitats de acuerdo a ocurrencias de especies de plantas similares. Esto resultaría en una descripción típica taxonómica de un hábitat.

Estos ejemplos ilustran que no solamente es importante que las *n* observaciones puedan ser agrupadas, pero que también es posible agrupar las *p* variables. Algunos métodos que serán discutidos en esta sección necesitan como entrada una matriz de disimilaridad o matriz de distancia. Estos métodos pueden ser usados para agrupar observaciones así como para agrupar variables, a condición que las matrices de disimilaridad estén a nuestra disponibilidad. A pesar de la dualidad de estos métodos

para efectuar el “clustering”, discutiremos en las próximas secciones los métodos como si nuestro objetivo principal es el de agrupar observaciones.

El análisis de Clúster es básicamente una familia de algoritmos que podrían ser clasificados en dos grandes grupos de métodos:

- Métodos Jerárquicos

Aquí, hacemos una distinción entre métodos aglomerativos y métodos divisivos. El primero empieza en la situación en el que cada observación individual forma su propio clúster (por lo que estos métodos empiezan con *n* clústers). En los próximos pasos, los clústers son fusionados secuencialmente, hasta que finalmente exista un sólo clúster con *n* observaciones. Los métodos divisivos funcionan justo al revés.

La solución de un “clustering” jerárquico es por tanto una secuencia de *n* soluciones de clústers anidados.

- Métodos basados en partición

Los métodos basados en partición empiezan típicamente con un número específico de clústers, digamos *k* clústers, y con una configuración de clúster inicial (i.e. una especificación de qué observaciones pertenecen a qué clúster). La técnica consiste entonces en la partición del espacio de observaciones de *p* dimensiones en *k* sub-espacios correspondientes a *k* clústers.

6.2.2 Herramientas gráficas de diagnóstico

Debido a que ningún método de agrupación o “clustering” garantiza que la solución sea una solución buena o correcta, existe la necesidad de tomar mano de herramientas de diagnóstico que puedan ser usadas para evaluar la calidad de la configuración del clúster. Describiremos entonces dos de estas herramientas: el “clusplot” y el gráfico “silhouette” o silueta.

Describiremos estos dos métodos gráficos basándonos en un ejemplo, usando el set de datos llamado “euro” en el cual 12 países europeos son agrupados en 2 grupos. Existen dos variables en el set de datos y son: el Producto Nacional Bruto (GNP por sus siglas en inglés) y el porcentaje del GNP que es atribuido a la agricultura. Los datos corresponden al año 1994.

6.2.2.1 El “Clusplot”

El “Clusplot” es básicamente un diagrama de dispersión de las observaciones en el plano de los dos primeros componentes principales (“scores” en los dos primeros PC’s). La lógica es que esta es la representación en dos dimensiones que retiene el máximo contenido de información. El “Clusplot” para el ejemplo del set de datos “euro” es ilustrado en la Ilustración 3. Basándonos en este gráfico de pocas dimensiones (2d) podemos efectuar una primera evaluación visual de la calidad del “clustering” (cuando $p > 3$ no es usualmente posible en términos de las variables originales).

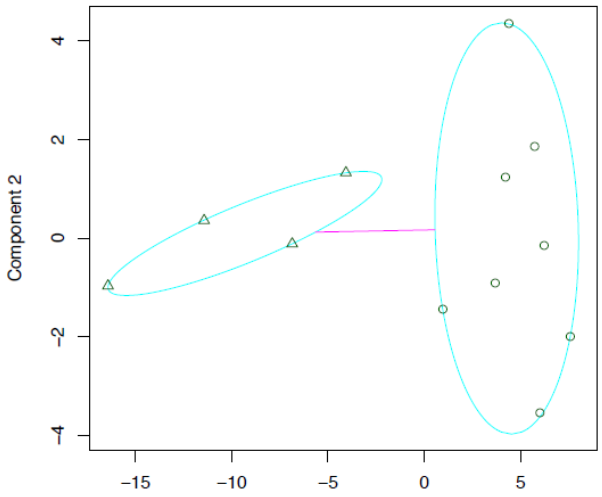


Ilustración 3: Ejemplo de Clusplot correspondiente al set de datos “euro”

6.2.2.2 Gráfico “silhouette” o silueta

Antes de ilustrarles sobre el gráfico “silhouette”, definiremos el valor “silhouette” (o valor silueta) para una observación i dentro del clúster *A*. Denotemos como $\bar{d}(i)$ a la media de disimilaridad de las disimilaridades entre la observación i y todas las observaciones en el clúster *A*, y denotemos como $\bar{d}(i, C)$ a la media de disimilaridad de las disimilaridades entre la observación i y todas las observaciones en el clúster $C \neq A$. Finalmente, denotemos:

$$\bar{d}_{\min}(i) = \min_{C \neq A} \bar{d}(i, C).$$

Luego, definimos el valor silueta (o valor “silhouette”) de la observación i como sigue:

$$s(i) = \frac{\bar{d}_{\min}(i) - \bar{d}(i)}{\max\{\bar{d}(i), \bar{d}_{\min}(i)\}}$$

Los extremos tienen interpretaciones simples:

$s(i) \approx +1 \rightarrow i$ está claramente en el clúster *A*

$s(i) \approx 0 \rightarrow i$ se encuentra entre los clústers *A* y *B*

$s(i) \approx -1 \rightarrow i$ se encuentra más cerca del clúster *B*

En el gráfico “silhouette” los valores “silhouette” son representados como barras para cada observación. Las observaciones son ranqueadas: las observaciones dentro del mismo clúster son graficadas juntas, y dentro de un clúster las observaciones son ranqueadas desde aquellas con valores “silhouette” más grandes (arriba) a aquellas con los valores “silhouette” más pequeños (abajo). La Ilustración 4 nos muestra un ejemplo de este gráfico implementado en nuestro set de datos demo “euro”.

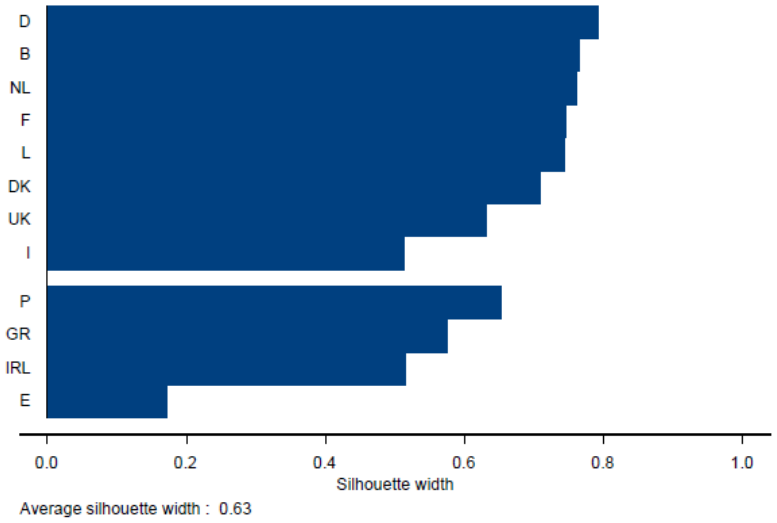


Ilustración 4: Ejemplo del Gráfico "silhouette" o silueta

Un buen “clustering” puede ser reconocido si todas las observaciones tienen altos valores “silhouette”. Si existe un clúster con una observación que posee un valor

"silhouette" negativo, entonces esto es un indicativo de un mal "clustering" o agrupación (o, al menos se puede decir que esta observación fue mal agrupada).

6.2.3 Análisis Clúster basados en partición

Los métodos clúster basados en partición requieren que el número de clústers (*k*) sea especificado previo al inicio del algoritmo. Luego discutiremos métodos para determinar un número apropiado de clústers.

6.2.3.1 Método de *K*-medias

6.2.3.1.1 Algoritmo

El algoritmo empieza con *k* centros de clúster iniciales $\boldsymbol{c}_1^{(0)}, \ldots, \boldsymbol{c}_k^{(0)}$, en el espacio de las variables de *p* dimensiones. Luego, un algoritmo iterativo comienza. El paso *s*-enésimo de este algoritmo consiste en:

1. asignar cada observación $\boldsymbol{x}_i$ al clúster *j* para el cual

$$d(\boldsymbol{x}_i, \boldsymbol{c}_j^{(s-1)}) \leq d(\boldsymbol{x}_i, \boldsymbol{c}_k^{(s-1)})$$

   $(k \neq j)$. Denotemos $m_i = j$ si la observación i es asignada al clúster j; llamamos m_i al clúster al cual la observación i pertenece.

2. Recalculamos los centros de los clústers $\boldsymbol{c}_j^{(s)}$. Esto es calculado como la media (muestral) multivariada de las observaciones asignadas al clúster.
3. Si los centros de los clústers son iguales a aquellos de la previa iteración, entonces el algoritmo ha convergido, y el mismo es detenido. Caso contrario, a s se le incrementa uno, y una nueva iteración empieza.

6.2.3.2 PAM

PAM son las siglas en inglés para el término *Particionamiento Alrededor de los Mediodes* (*Partitioning Around Mediods*). Es bastante similar al procedimiento previo. Las principales diferencias son las siguientes:

- En vez de centros de clúster (centroides), PAM usa Mediodes de clúster. El Mediode del clúster j es igual a una de las observaciones asignadas al clúster j (la observación más "representativa").
- Este método es menos sensible a la presencia de "outliers" o datos raros/atípicos que el método de *k*-medias. La sensibilidad del método de *k*-medias es una consecuencia del hecho que los centroides son calculados como medias (muestrales), medida que es generalmente conocida por ser sensible a datos raros/atípicos.

El algoritmo empieza con *k* observaciones "representativas" iniciales que son identificadas como la selección inicial para los mediodes: $\boldsymbol{c}_1^{(0)}, \dots, \boldsymbol{c}_k^{(0)}$. Luego, un algoritmo iterativo empieza. En el paso *s*-enésimo,

1. Cada observación $\boldsymbol{x}_i$ es asignada al clúster j para el cual

$$d(\boldsymbol{x}_i, \boldsymbol{c}_j^{(s-1)}) \leq d(\boldsymbol{x}_i, \boldsymbol{c}_k^{(s-1)})$$

   $(k \neq j)$. Con la notación m_i para la membresía de clúster. Lo anterior es equivalente a:

$$d(\boldsymbol{x}_i, \boldsymbol{c}_{m_i}^{(s-1)}) \leq d(\boldsymbol{x}_i, \boldsymbol{c}_j^{(s-1)})$$

   Para todo $j = 1, \dots, k$.

2. Nuevos mediodes $\boldsymbol{c}_j^{(s)}$ $(j = 1, \dots, k)$ son seleccionados de las *n* observaciones de tal manera que:

$$Q^{(s)} = \sum_{i=1}^{n} d(\boldsymbol{x}_i, \boldsymbol{c}_{m_i}^{(s)})$$

   sea mínimo. Esta función es la función objetivo que tiene que ser minimizada.

3. El algoritmo se detiene si todos los mediodes han convergido.

Notemos que PAM es efectivamente muy similar al método de *k*-medias. La única diferencia conceptual es la función objetivo en el paso 2 de las iteraciones.

6.2.3.3 Clara

Clara son las siglas para el término en idioma inglés "Clustering large applications", i.e. Clara es un método que puede ser usado para set de datos muy grandes. Un problema importante con los set de datos que son muy extensos, es que muchos de

los métodos de “clustering” requieren la matriz de disimilaridad completa, que contiene $n(n-1)/2$ valores diferentes. Por ejemplo, si $n = 1000$ (y eso que este número de observaciones no es extremadamente grande), el programa del computador (que se esté usando para efectuar el análisis) necesitará suficiente memoria para 499500 disimilaridades. Además que el tiempo de cálculo puede convertirse en un problema con grandes bases de datos.

No daremos más detalles sobre este método, ya que no será de empleo en el análisis que se presentará en las siguientes secciones.

6.2.4 Análisis Clúster Jerárquico

6.2.4.1 Algoritmo General

En esta pequeña sección teórica, nos enfocaremos principalmente en los *métodos clúster jerárquicos aglomerativos*. En general, los algoritmos funcionan de la siguiente manera:

En el paso 0 cada observación es un clúster. Existen por tanto *n* clústers. Los siguientes pasos consisten en:

1. Fusionar aquellos dos clústers que tengan la más pequeña disimilaridad inter-clúster.
2. Recalcule las disimilaridades inter-clúster

En el paso 0 las disimilaridades inter-clúster son por supuesto igual a las disimilaridades entre las observaciones correspondientes, pero tan pronto como un clúster de más de una observación exista, existe la necesidad de una definición más general de disimilaridad. En particular, debería ser extendido al concepto de *disimilaridades inter-clúster*. A continuación discutiremos algunas definiciones. Cada definición determina otro tipo de Análisis Clúster.

6.2.4.2 Disimilaridades Inter-Clúster

Representamos clústers (e.g. C_1 y C_2) como sets de observaciones pertenecientes a estos clústers. La disimilaridad inter-clúster es denotada por $d(C_1, C_2)$. Consideramos las siguientes definiciones:

- **Vinculación simple, enlace simple, o por el vecino más próximo:**

$$d(C_1, C_2) = \min_{\boldsymbol{x}_1 \in C_1;\, \boldsymbol{x}_2 \in C_2} d(\boldsymbol{x}_1, \boldsymbol{x}_2),$$

i.e. la disimilaridad entre C_1 y C_2 es definida como la disimilaridad más pequeña entre una observación de C_1 y una observación de C_2. Esto es ilustrado en la Ilustración 5.

El agrupamiento usando Vinculación simple es sensible a *encadenamientos*.

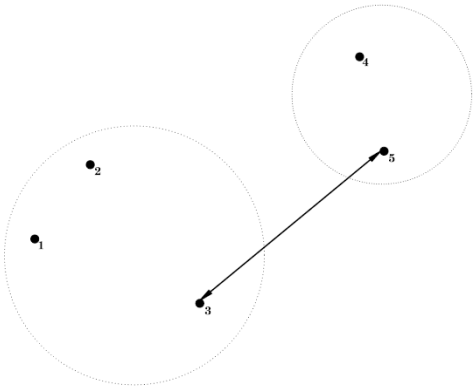


Ilustración 5: Ejemplo gráfico de la definición de disimilaridad de Vinculación simple (disimilaridad = distancia Euclidiana)

- **Vinculación completa, enlace completo o por el vecino más lejano:**

$$d(C_1, C_2) = \max_{\boldsymbol{x}_1 \in C_1;\, \boldsymbol{x}_2 \in C_2} d(\boldsymbol{x}_1, \boldsymbol{x}_2),$$

i.e. la disimilaridad entre C_1 y C_2 es definida como como la disimilaridad más grande entre una observación de C_1 y una observación de C_2. Esto es ilustrado en la Ilustración 6.

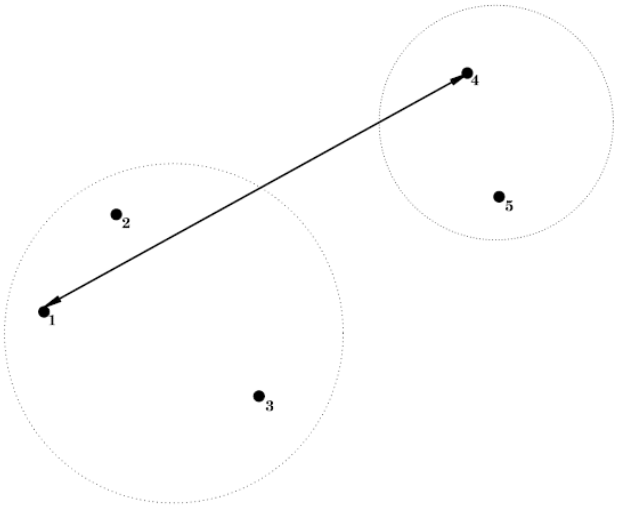


Ilustración 6: Ejemplo gráfico de la definición de disimilaridad de Vinculación completa (disimilaridad = distancia Euclidiana)

- **Vinculación Promedio: Promedio de grupo:**

$$d(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{\boldsymbol{x}_1 \in C_1;\, \boldsymbol{x}_2 \in C_2} d(\boldsymbol{x}_1, \boldsymbol{x}_2),$$

i.e. la disimilaridad entre C_1 y C_2 es definida como la disimilaridad promedio entre todas las observaciones de C_1 y todas las observaciones de C_2. Esto es ilustrado en la Ilustración 7.

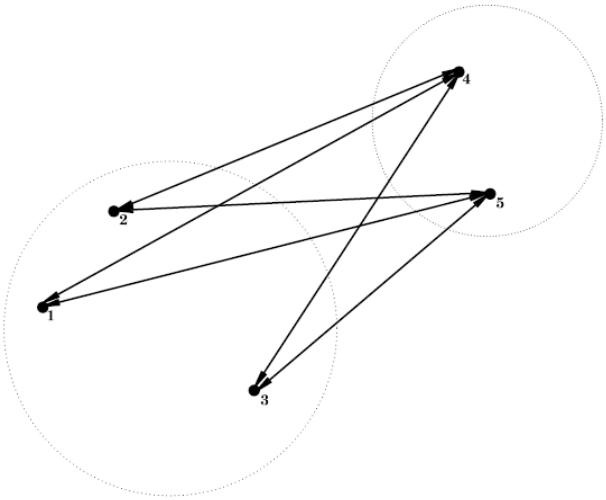


Ilustración 7: Ejemplo gráfico de la definición de disimilaridad de Vinculación promedio (disimilaridad = distancia Euclidiana)

6.2.4.3 Árbol de Clústers

Debido a la naturaleza jerárquica del algoritmo, este tipo de método de agrupación (clúster) nos da una secuencia de clústers anidados, i.e. la configuración de los clústers con c clústers se produce de la configuración con $c + 1$ clústers por la fusión de dos de estos $c + 1$ clústers.

Gracias a esta estructura restrictiva, las soluciones de un análisis de clúster jerárquico pueden ser fácilmente presentadas en una estructura parecida a un árbol. La altura de las ramas corresponde a la disimilaridad inter-clúster a la cual los dos clústers correspondientes fueron fusionados. Esta representación de árbol, que muestra todas las soluciones (de 1 hasta *n* clústers) simultáneamente, puede ser útil y de ayuda en decidir cuántos clústers existen realmente en los datos.

En la sección de análisis e implementación (específicamente en las Secciones 8.2.1 y 8.4.1) se mostrará el uso de este procedimiento con nuestros datos.

6.2.4.4 Método de Ward

El método de agrupamiento de Ward es un método de clúster jerárquico aglomerativo que no es basado en disimilaridades inter-clúster. La fusión de dos clústers en este método es basado en la suma de los cuadrados dentro de los clústers. Denotemos a $\boldsymbol{x} = (x_1 \ldots x_p) \in C$ como una observación de p dimensiones en el clúster C. Para el clúster C definimos la suma de cuadrados:

$$S(C) = \sum_{\boldsymbol{x} \in C} \sum_{i=1}^{p} (x_i - \bar{x}_{Ci})^2$$

en donde $\bar{x}_{Ci}$ es la media (muestral) de la variable i dentro del clúster C. En particular, en cada paso del algoritmo, dos clústers C_i y C_j son fusionados si esta fusión implica que el incremento de la suma de cuadrados

$$S(C_i \cup C_j) - \big(S(C_i) + S(C_j)\big),$$

es mínimo entre todas las posibles fusiones en este paso.

6.2.4.5 Mona

Mona es el único método de agrupamiento jerárquico divisivo que mencionaremos en este trabajo. Este es un método que es aplicable a datos con p variables binarias (por lo que no es aplicable a nuestro trabajo). Es además un método *monotético*, lo que quiere decir que en cada paso sólo una variable es usada para dividir un clúster en dos nuevos clústers (los otros métodos discutidos anteriormente son todos *politéticos*).

Suponga que en el paso *s* la variable X_i es usada para dividir un clúster. Debido a que cada variable es binaria, esta división implica que uno de los dos nuevos clústers contendrá ahora solamente observaciones para las cuales $X_i = 1$, y el otro contendrá sólo observaciones con $X_i = 0$. Esto implica además que en los próximos pasos la variable X_i no puede ser usada nuevamente para definir una división de clúster.

El algoritmo funciona de la siguiente manera. Suponga que en el paso *s* existen aún c clústers: $C_1, \dots, C_c$. Cada uno de estos c clústers es recursivamente partido hasta que solamente clústers con exactamente una observación queden disponibles.

Para cada clúster C_j $(j = 1, \dots, c)$ las observaciones en las variables X_u y X_v $(u \neq v = 1, \dots, p)$ son resumidas en una tabla de contingencia 2×2:

X_u / X_v	1	0
1	a_{uv}	b_{uv}
0	c_{uv}	d_{uv}

(Por tanto $a_{uv} + b_{uv} + c_{uv} + d_{uv} = |C_j|$). Una medida de la fortaleza de la relación entre las dos variables X_u y X_v dentro del clúster j viene dado por:

$$A_{uv} = |a_{uv}d_{uv} - b_{uv}c_{uv}|.$$

La variable que es seleccionada para partir el clúster C_j en el paso *s*, es aquella variable que muestra la asociación más grande con todas las demás variables. La fuerza total de asociación entre la variable X_u y las demás es calculada de la siguiente manera:

$$A_u = \sum_{v \neq u} A_{uv}$$

La variable seleccionada para partir el clúster C_j es por tanto la variable X_i para la cual $A_i = \max_u A_u$.

6.2.5 Determinación del Número de Clústers

6.2.5.1 Introducción

El objetivo de un análisis clúster es encontrar grupos en los datos. Esto implica que el número de clústers es también una parte de la solución. En el análisis de clúster jerárquico encontramos en la solución una secuencia entera de configuraciones de clúster, empezando desde un clúster, hasta *n* clústers. Dependiendo de las estructura del árbol, podemos escoger una o más soluciones para un análisis más profundo. En

esta sección discutiremos algunas otras herramientas para evaluar la calidad de una configuración de clúster.

En el análisis clúster basado en partición, el número de clústers debe ser especificado antes de empezar el algoritmo. Por supuesto, podemos correr varios análisis de clúster basados en partición empezando con algunas opciones para *k* (el número de clústers). Después, las soluciones pueden ser comparadas y evaluadas a través de e.g. "clusplots" y gráficos "silhouette" o silueta. En esta sección, introduciremos todavía otra herramienta para comparar configuraciones de clúster con respecto al número de clústers (o grupos). El método es basado en el Lambda de Wilk (Wilk's Lambda).

Una metodología general consiste en correr primero el método de clúster jerárquico ya que este método inmediatamente nos da soluciones para muchas *k* (de hecho nos brinda soluciones para $k = 1, \dots, n$). En particular, el árbol de clúster nos da un resumen muy bueno de las soluciones para muchos valores de *k*. Basados en el árbol de clúster, un rango de posibles valores para *k* es seleccionado. Para cada uno de estos posibles valores para *k*, se implementa el método basado en partición para obtener (mejores) soluciones para cada *k* seleccionada.

6.2.5.2 Gráfico de Sedimentación ("Scree Plot") basado en el Lambda de Wilk

En la literatura de estadística se discute sobre el Lambda de Wilk como un test estadístico para detectar diferencias entre medias multivariadas. Un razonamiento lógico a ser usado en la evaluación de una configuración de clúster es que se espera que los centros de los clústers (medias de clústers) estén bien separados. O, de manera equivalente, que la varianza de las observaciones dentro de un clúster sea pequeña en comparación a la varianza de los centros de los clústers (i.e. la varianza entre los grupos). Esto es exactamente lo que el Lambda de Wilk mide:

$$\Lambda = \left(\frac{|\boldsymbol{E}|}{|\boldsymbol{B} + \boldsymbol{E}|} \right)^{2/n},$$

siendo $\boldsymbol{E}$ y $\boldsymbol{B}$ las matrices de los cuadrados y productos cruzados dentro y entre los clústers, respectivamente (note que el exponente $2/n$ no es de hecho necesario). En donde:

$$\boldsymbol{E} = (n - K)\widehat{\boldsymbol{\Sigma}} = \sum_{k=1}^{K} \sum_{l=1}^{n_k} (\boldsymbol{X}_{kl} - \overline{\boldsymbol{X}}_k)^t (\boldsymbol{X}_{kl} - \overline{\boldsymbol{X}}_k)$$

$$\boldsymbol{B} = \sum_{k=1}^{K}\sum_{l=1}^{n_k}(\bar{\boldsymbol{X}}_k - \bar{\boldsymbol{X}})^t(\bar{\boldsymbol{X}}_k - \bar{\boldsymbol{X}}) = \sum_{k=1}^{K} n_k\,(\bar{\boldsymbol{X}}_k - \bar{\boldsymbol{X}})^t(\bar{\boldsymbol{X}}_k - \bar{\boldsymbol{X}})$$

$$\boldsymbol{B} + \boldsymbol{E} = \sum_{k=1}^{K}\sum_{l=1}^{n_k}(\boldsymbol{X}_{kl} - \bar{\boldsymbol{X}})^t(\boldsymbol{X}_{kl} - \bar{\boldsymbol{X}})$$

($k = 1, \dots K$; n_k = número de observaciones en clúster k; $\bar{\boldsymbol{X}}_k$ es el vector multivariado que contiene a las medias de cada variable en el clúster k; $\bar{\boldsymbol{X}}$ es el vector multivariado que contiene las medias de cada variable tomando todas las observaciones del set de datos en consideración; y $\boldsymbol{X}_{kl}$ es el vector multivariado correspondiente a la observación l del clúster k).

Si tuviéramos algunas soluciones de clúster (por decir $k = 1, \dots K$) podríamos calcular el Lambda de Wilk para cada k, en notación Λ_k. Con estos estadísticos, podemos obtener un gráfico de sedimentación ("scree plot") en el cual Λ_k es graficado contra k. La Ilustración 8 nos muestra un ejemplo de este gráfico. El gráfico proviene de un análisis de clúster realizado a un set de datos llamado "abundance" sobre especies de plantas y hábitats.

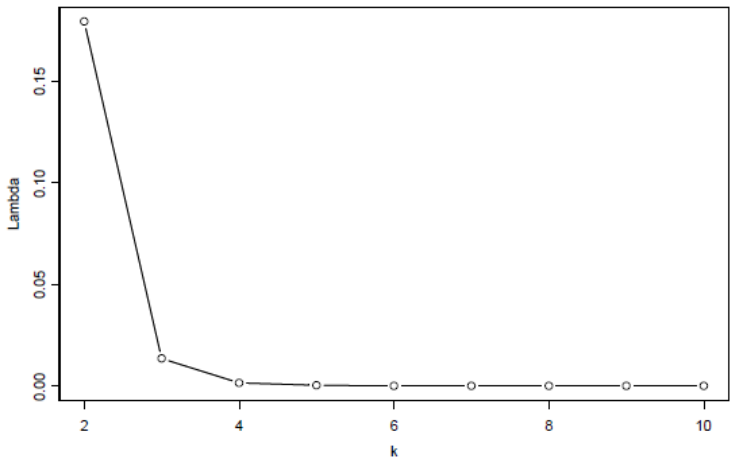


Ilustración 8: Gráfico de Sedimentación (Scree Plot) basado en el Lambda de Wilk realizado a un set de datos llamado "abundance" sobre especies de plantas y hábitats

En este caso particular, el gráfico de sedimentación nos muestra un decrecimiento muy pronunciado en el valor del Lambda de Wilk Λ_k cuando vamos de $k = 2$ a $k = 3$ y un decremento menos pronunciado cuando vamos de $k = 3$ a $k = 4$. Sin embargo, luego de $k = 4$, el valor del Lambda de Wilk se estabiliza. Por tanto, para nuestro pequeño ejemplo, concluimos que $k = 3$ o $k = 4$ clústers podrían ser una buena solución.

7 Aplicación de las técnicas de análisis en la base de datos

7.1 Manipulación de los datos

Como se ha dicho anteriormente, para efecto de nuestro análisis estadístico en el presente estudio, se usó el lenguaje estadístico R y su IDE (por sus siglas en inglés - Entorno de desarrollo integrado-) RStudio. La base de datos descargada de [10], "CENEC_2010.dat", posee 511,130 observaciones y 114 variables. Una vez importada la base de datos a R, podemos darnos cuenta que existen muchos "NAN" (caracter usado por R para representar valores/datos perdidos). Esto pudo haberse producido por algunas razones, entre las cuales mencionamos como ejemplo: la no respuesta a la pregunta por parte del encuestado, dicha pregunta no aplicaba por no haber pasado una pregunta "filtro" previa, errores por parte de los encuestadores al momento de registrar los datos, entre otros relacionados. Sin embargo, esta información (sobre el origen de estos "NAN") no es provista por el INEC.

Para realizar nuestro *primer* análisis acerca de las actividades económicas predominantes en cada provincia (o región), se tomó los datos de la variable categórica "Provincias" (S1P2) que indica la provincia en donde cada unidad económica encuestada opera y los datos de la variable "Actividad Principal a dos Dígitos CIIU" (CIIU2.P) que fue una de las clasificaciones para actividades económicas usadas por el INEC para especificar a qué actividad económica cada unidad se dedicaba. Se escogió esta variable "a dos dígitos" ya que queríamos obtener información más específica acerca de la actividad de cada unidad económica (también existían clasificaciones "a tres dígitos" y "a cuatro dígitos", sin embargo si se escogía una de ellas habríamos tenido una gran lista de actividades y habríamos perdido generalidad en ellas). Con el objetivo de hacer aún más general el análisis y el de evitar tener muchos grupos de actividades económicas (que todavía se tenía con la clasificación "a dos dígitos" escogida), se fusionaron ciertas "Actividades Principales a dos Dígitos CIIU" formando así un menor número de grupos de actividades económicas, siguiendo el criterio racional de fusionar actividades similares. De esa manera no se redundaba en ellas y se hacía más consistente el análisis. Además, ciertas actividades económicas de la clasificación original a las cuales muy pocas unidades económicas se dedicaban *relativamente* fueron desechadas del análisis, e.g. Fabricación de papel y de productos de papel, Impresión y reproducción de grabaciones, Fabricación de coque y de productos de la refinación del petróleo, Otras industrias manufactureras, Reparación e instalación de maquinaria y equipo, entre otras. En consecuencia, se formaron los siguientes nuevos grupos de actividades económicas resumidos en la Tabla 2:

Nombre Actividad Económica para Análisis	Actividades a ser fusionadas
Agricultura, ganadería	• Agricultura, ganadería, caza y actividades de servicios cone…
Silvicultura y Actividades con la madera	• Silvicultura y extracción de madera. • Producción de madera y fabricación de productos de madera y…..
Pesca y acuicultura	• Pesca y acuicultura
Actividades relacionadas con la minería y petróleo	• Extracción de carbón de piedra y lignito • Extracción de petróleo crudo y gas natural • Extracción de minerales metalíferos • Explotación de otras minas y canteras • Actividades de servicios de apoyo para la explotación de minas….
Elaboración de alimentos/bebidas	• Elaboración de productos alimenticios • Elaboración de bebidas
Fabricación de productos textiles, de vestir, cueros	• Fabricación de productos textiles • Fabricación de prendas de vestir • Fabricación de cueros y productos conexos
Actividades relacionadas a producir productos químicos, farmacéuticos	• Fabricación de substancias y productos químicos • Fabricación de productos farmacéuticos, sustancias químicas
Fabricación de productos de caucho, plástico, minerales, metales	• Fabricación de productos de caucho y plástico • Fabricación de otros productos minerales no metálicos • Fabricación de metales comunes • Fabricación de productos elaborados de metal, excepto maquinas… • Fabricación de muebles
Fabricación de productos de informática, electrónicos, eléctricos, maquinarias y equipos	• Fabricación de productos de informática, electrónica y óptico • Fabricación de equipo eléctrico • Fabricación de maquinaria y equipo n.c.p.
Fabricación de vehículos, equipo de transporte	• Fabricación de vehículos automotores, remolques y semirremolque • Fabricación de otros tipos de equipos de transporte
Actividades de suministro de energía	• Suministro de electricidad, gas, vapor y aire acondicionado
Actividades relacionadas a la Construcción	• Construcción de edificios • Obras de ingeniería civil • Actividades especializadas de la construcción
Actividades relacionadas al Comercio	• Comercio al por mayor y al por menor; reparación de vehículo • Comercio al por mayor, excepto el de vehículos automotores … • Comercio al por menor, excepto el de vehículos automotores

Nombre Actividad Económica para Análisis	Actividades a ser fusionadas
Actividades de transporte	• Transporte por vía terrestre y por tuberías • Transporte por vía acuática • Transporte por vía aérea • Almacenamiento y actividades de apoyo al transporte • Actividades postales y de mensajería
Actividades de Hostelería	• Actividades de alojamiento • Servicio de alimento y bebida
Actividades de Comunicación, informática e información	• Actividades de programación y transmisión • Telecomunicaciones • Programación informática, consultoría de informática… • Actividades de servicios de información
Finanzas y Seguros	• Actividades de servicios financieros, excepto las de seguros • Seguros, reaseguros y fondos de pensiones… • Actividades auxiliares de las actividades de servicios financieros
Actividades de investigación, científicas y otras profesionales	• Investigación científica y desarrollo • Publicidad y estudios de mercado • Otras actividades profesionales, científicas y técnicas • Actividades veterinarias
Actividades de la Salud y Asistencia Social	• Actividades de atención de la salud humana • Actividades de atención en instituciones • Actividades de asistencia social sin alojamiento
Actividades de Arte, entretenimiento y recreación	• Actividades creativas, artísticas y de entretenimiento • Actividades de bibliotecas, archivos, museos y otras actividades • Actividades de juegos de azar y apuestas • Actividades deportivas, de esparcimiento y recreativas

Tabla 2: Grupos de actividades económicas definitivos usados para nuestro análisis de Caracterización económica de las provincias del Ecuador, luego de fusionar ciertas actividades de la variable "Actividades Principales a dos Dígitos CIIU".

Posteriormente, se formó una tabla de contingencia, recolectando las frecuencias de las celdas que eran producto de la clasificación cruzada entre las variables "Provincias" y "prov_activ", la última siendo la variable que se formó luego de fusionar algunas actividades económicas de la variable "Actividades Principales a dos Dígitos CIIU" según lo descrito en el párrafo anterior y en la Tabla 2. En otras palabras, el valor de cada celda en la nueva tabla de contingencia describía el número de unidades económicas que operaban en la provincia i y que se dedicaban a la actividad económica j , en donde $i = 1, \dots, 24$ y $j = 1, \dots, 20$. Para obtener la tabla final se obtuvo las frecuencias relativas, es decir, el porcentaje de unidades económicas en la provincia i que se dedicaban a la actividad j con relación al total de unidades económicas encuestadas en la provincia i. En consecuencia, se obtuvo una tabla de contingencia de dimensión 24×20, la cual fue usada para implementar las técnicas

de Análisis de Componentes Principales y Análisis de Clúster. La Tabla 3 siguiente nos muestra el "output" de R correspondiente a la tabla final usada en el *primer* análisis de caracterización económica de las provincias (o regiones) del Ecuador.

	Agr	Silv.madera	Pesca.ac	Mineria	alim.beb	text.vest	quim	plast.cauch.met	elct.eq.maq	veh	sumEner	construc	comercio	transp	serv.aloj.com	inf.comunic	finanz	act.prof.cient	asist.social	art.entr.recr
1	0.43	1.17	0.02	0.06	2.47	4.31	0.11	6.30	0.30	0.12	0.04	0.52	58.84	1.83	11.21	3.27	1.08	1.08	5.52	1.33
2	0.00	0.97	0.00	0.00	2.54	2.54	0.06	3.45	0.03	0.03	0.09	0.06	66.57	0.82	11.03	5.53	1.17	0.70	3.16	1.26
3	0.60	0.93	0.01	0.01	1.86	3.44	0.04	4.96	0.01	0.21	0.10	0.29	65.98	1.76	9.37	3.63	0.93	0.80	3.39	1.68
4	0.18	0.62	0.00	0.02	2.07	1.75	0.02	3.09	0.00	0.07	0.07	0.05	68.31	2.24	11.13	5.88	0.71	0.94	2.14	0.69
5	0.14	0.87	0.00	0.01	2.84	3.62	0.08	6.42	0.12	0.13	0.09	0.26	60.44	1.14	11.13	5.91	1.44	1.02	3.09	1.23
6	0.33	1.13	0.00	0.04	2.51	4.04	0.03	4.69	0.19	0.15	0.07	0.30	59.83	0.70	12.56	6.40	1.01	0.79	3.64	1.60
7	0.37	0.42	0.21	0.15	1.58	2.20	0.02	3.79	0.12	0.09	0.12	0.22	66.97	1.39	11.55	4.56	0.60	0.67	3.74	1.24
8	0.05	0.69	0.02	0.01	1.72	1.62	0.02	3.07	0.01	0.04	0.11	0.17	67.20	1.21	14.62	3.83	0.52	0.57	3.23	1.28
9	0.13	0.27	0.05	0.02	2.69	1.59	0.16	3.21	0.16	0.07	0.04	0.31	67.97	1.18	11.62	4.20	0.49	0.83	3.69	1.33
10	0.10	1.61	0.00	0.00	2.04	4.28	0.04	3.83	0.09	0.10	0.04	0.14	63.05	1.04	13.08	5.41	0.77	0.78	2.55	1.05
11	0.50	0.79	0.01	0.01	1.94	2.19	0.05	4.83	0.14	0.09	0.11	0.35	65.65	1.22	11.64	4.09	1.02	0.85	3.50	1.03
12	0.11	0.44	0.01	0.00	2.62	1.93	0.03	3.44	0.11	0.09	0.05	0.10	65.62	0.98	12.83	4.79	0.57	0.61	3.93	1.71
13	0.14	0.73	0.07	0.00	2.92	1.63	0.05	3.31	0.04	0.07	0.12	0.29	66.52	1.13	12.17	3.27	0.62	0.68	4.21	2.02
14	0.55	2.24	0.11	0.13	2.11	2.58	0.00	4.00	0.00	0.08	0.21	0.29	60.45	2.42	14.29	3.37	1.05	1.00	2.89	2.24
15	0.13	1.09	0.04	0.00	2.14	2.49	0.00	2.92	0.00	0.04	0.09	0.17	60.02	0.83	18.25	5.46	0.83	0.79	2.84	1.88
16	0.10	2.01	0.00	0.00	1.81	2.32	0.00	3.34	0.10	0.14	0.07	0.34	58.71	1.74	17.14	5.66	0.85	0.95	3.24	1.50
17	0.14	0.81	0.00	0.07	2.49	2.93	0.18	4.37	0.20	0.11	0.04	0.67	60.67	1.20	13.30	5.32	1.00	1.19	4.22	1.10
18	0.12	1.16	0.00	0.00	1.84	4.86	0.06	4.08	0.18	0.29	0.06	0.34	62.37	0.82	12.59	4.16	1.24	0.99	3.67	1.18
19	0.19	1.59	0.11	0.04	1.78	2.67	0.04	4.70	0.15	0.04	0.22	0.19	61.42	2.30	13.99	3.85	1.26	0.30	2.89	2.30
20	0.00	0.09	0.09	0.00	2.37	1.52	0.00	3.50	0.00	0.09	0.38	0.00	51.61	4.83	23.39	3.79	1.89	0.57	3.12	2.75
21	0.08	0.70	0.00	0.10	1.30	2.12	0.05	4.17	0.13	0.10	0.05	0.31	64.70	1.27	14.06	4.90	0.67	0.75	2.59	1.94
22	0.08	1.37	0.00	0.11	1.41	2.10	0.08	3.59	0.04	0.11	0.08	0.38	64.82	2.25	14.42	4.08	0.84	0.72	2.21	1.30
23	0.23	0.88	0.00	0.00	2.03	2.82	0.06	3.98	0.12	0.34	0.02	0.24	67.66	1.37	9.93	4.88	0.66	0.72	2.93	1.12
24	0.23	0.25	0.05	0.03	2.62	1.21	0.01	3.90	0.04	0.06	0.15	0.12	68.41	0.70	13.11	4.78	0.28	0.36	2.08	1.60

Tabla 3: Tabla usada en el primer análisis de caracterización económica de las provincias (o regiones) del Ecuador.

Para realizar nuestro *segundo* análisis de caracterización empresarial de las provincias del país, se formó asimismo una tabla de contingencia que cruzaba las 24 provincias del Ecuador con las siguientes variables económicas: Monto de Financiamiento con Institución Privada, Monto de Financiamiento con el Gobierno, Monto requerido de financiamiento, Monto de gasto en manejo de desechos, Monto de gasto en investigación y desarrollo, Monto de gasto en capacitación y formación, Gastos anuales en remuneraciones, Gastos anuales en materia prima, Gastos anuales en envases y embalajes, Gastos anuales en compras y mercadería, Tasas-contribuciones y otros impuestos anuales (excluye IVA…), Total de ingresos anuales percibidos por ventas o prestación, Gastos anuales kilovatios/hora; las cuales se escogió de la base de datos del Censo Nacional Económico 2010 por describir de alguna manera las políticas empresariales de los establecimientos económicos encuestados. Por supuesto, los montos de estas 13 variables se encontraban disponibles a nivel de cada unidad económica en la base original, por lo que fue necesario manipular la tabla inicial de tal manera que se obtenga la tabla final cruzada a nivel agregado, es decir, que los montos correspondan ahora para cada provincia. Finalmente, para obtener valores más uniformes, se dividió cada celda para el número total de encuestados en cada provincia, obteniendo así valores para cada celda que representaban ahora el monto promedio correspondiente a la variable económica-empresarial X_j en la provincia i. La Tabla 4 siguiente nos muestra el "output" de R correspondiente a la tabla final usada en el *segundo* análisis de caracterización de las políticas empresariales en las provincias (o regiones) del Ecuador.

	FinanPriv	FinanGob	Finanreq	Manedesech	GasInvDesar	GasCapacForm	GAST_REM	GAST_MAT	GAST_EE	GAST_COM	TAX	INGRESOS	G_ENERG
1	7077.1617	670.396759	17064.982	96.321865	276.277947	131.42378	18834.707	27488.778	1280.5492	58818.947	2693.2385	286968.70	863.2811
2	48983.6826	190.839464	55643.779	62.136663	50.247016	72.75177	13829.545	3588.724	366.7456	7179.635	132.5527	37108.31	267.0614
3	1121.4487	38.627747	11078.506	152.850332	121.331004	49.99019	11561.951	8966.929	518.7650	17771.412	2112.3849	71665.15	710.9585
4	1957.3532	23.683143	10332.189	57.407801	9.090944	153.23170	11670.284	10594.277	218.6027	31115.965	1256.4741	75152.29	351.7725
5	1861.1551	3.936741	8402.995	29.513555	81.341305	270.72643	13996.898	8249.620	310.7883	18579.494	4695.9326	86731.77	521.7810
6	1603.2488	1489.152062	10455.801	60.081079	83.729776	79.30437	13818.209	12025.152	365.2488	16411.095	509.2246	82201.87	470.8110
7	1449.9077	80.491417	14903.234	107.845293	9.085746	87.53228	13955.096	15298.887	10244.9611	43609.674	1174.8831	207084.77	502.4146
8	1099.7123	95.882671	11091.516	22.239328	15.608193	84.78435	13646.478	9125.977	468.3997	16428.371	558.5160	109540.70	523.0550
9	6839.1433	486.915373	17324.015	307.346542	314.049169	228.43125	25554.859	36689.967	5310.4368	72547.550	2635.9013	330948.26	1631.7495
10	2424.1565	30.868463	6877.136	81.969382	41.421984	73.44605	12317.409	12013.521	548.0404	57358.212	1181.9112	151556.85	330.4339
11	1744.3178	94.806035	11661.316	90.489766	32.675484	170.91898	14005.703	5247.355	479.9612	23211.407	979.1287	93205.27	338.5497
12	1874.5876	47.706278	17206.138	88.938370	118.906158	229.72872	14562.867	7220.825	987.9536	26624.059	1086.9385	103978.52	660.8310
13	3327.0477	25.734075	31370.548	68.326702	55.429982	285.93054	15996.279	19553.048	1947.7917	41158.392	1541.5837	141883.51	680.2689
14	1763.4958	17.625053	18765.795	23.324884	46.378749	54.62379	9691.067	9354.408	177.2779	15514.293	221.9171	49148.95	384.3526
15	1424.6995	26.784801	12365.440	31.311655	15.976617	50.67373	24338.856	5716.612	778.2320	11546.142	190.0177	61660.17	481.1144
16	2043.5920	54.129353	12787.200	37.407684	3.522664	84.21365	9593.291	11792.636	226.3311	18089.799	1151.8496	82236.77	405.1396
17	16314.3114	781.033803	36416.412	229.214892	1461.516784	611.71534	58090.119	43535.643	5980.4790	92566.029	5526.9858	590009.49	2414.1368
18	3537.6147	74.687912	11749.288	97.375825	112.763802	96.69584	15703.951	11276.035	500.5355	43958.341	1880.0808	130966.91	544.6954
19	1122.6889	378.223041	28066.733	79.056433	629.464912	58.79064	13439.586	7220.609	288.2746	11739.777	174.5747	114950.26	442.1959
20	988.5899	775.642588	111792.132	637.154251	146.338600	177.92476	48096.990	11272.179	480.5010	16411.417	3265.8227	146198.12	1373.7442
21	2363.1670	485.337162	10917.746	27.648798	18.704744	71.97575	15748.243	10414.956	277.3416	19670.647	152.3780	82215.94	557.7339
22	9199.0165	496.316853	8367.966	8.911692	84.790734	109.80224	11835.833	7692.148	359.7711	19750.811	2068.9076	85903.90	495.2369
23	2153.9039	14.447754	8974.056	98.784299	175.905415	55.80465	10600.528	8649.568	553.2472	33253.388	1662.5607	140299.77	560.5491
24	689.1898	1391.698344	27670.423	27.952206	52.098922	34.44348	11246.966	4921.005	557.5099	45457.990	1777.9120	107146.80	482.6326

Tabla 4: Tabla final usada en el segundo análisis de caracterización de las políticas empresariales en las provincias (o regiones) del Ecuador.

8 Análisis y Resultados

8.1 Caracterización Económica de las provincias del Ecuador a través del Análisis de Componentes Principales (PCA)

En primer lugar, procederemos a realizar el análisis de caracterización económica de las provincias del Ecuador con el fin de investigar las actividades económicas predominantes en las provincias (o regiones) del Ecuador. Como fue mencionado anteriormente, para este fin usaremos el paquete estadístico R.

El análisis de esta sección será un análisis alternativo al descriptivo tradicional. Debido a que hemos creado una tabla de frecuencias relativas para las actividades económicas en cada provincia (ver Tabla 3), las variables correspondientes a dichas actividades se tratarían como diferentes "niveles de categorías" en la composición total para cada provincia, por lo que un correcto análisis descriptivo consistiría en e.g. un gráfico de barras (o de pastel) para cada provincia mostrando la composición porcentual de cada actividad económica, un gráfico de barras confrontando dos provincias (o más) en el cual se compare las frecuencias relativas correspondientes a las actividades económicas de ambas. Aunque algo ilustrativo, ya el último gráfico mencionado sería más difícil de interpretar al incluir más de una barra en el gráfico (cada barra representa a una provincia). Sería complicado interpretar las 20 variables (actividades) económicas de todas las observaciones (provincias) *simultáneamente*, y obtener una conclusión general acerca de cómo las provincias difieren con respecto a las actividades realizadas en cada una de ellas. Por ello, usaremos el Análisis de Componentes Principales (PCA) para este fin, el cual nos permite analizar a las 24 provincias de manera simultánea e investigar contrastes entre ellas con respecto a las 20 variables económicas a analizar. Esperamos entonces poder reducir el número de dimensiones (de $p = 20$ a $q <<< 20$) de tal manera que podamos formular nuestras conclusiones empleando algunos gráficos en el espacio de componentes principales reducido.

Como mencionamos anteriormente, nuestro set de datos contiene datos para las 24 provincias del Ecuador. En particular, hemos arreglado nuestros set de datos de tal manera que obtengamos frecuencias relativas de diferentes actividades económicas en cada provincia (datos expresados como porcentaje del total de personas "censadas" en cada provincia).

Las variables definidas en la sección previa son: Agricultura, ganadería; Silvicultura y Actividades con la madera; Pesca y acuicultura; Actividades relacionadas con la minería y petróleo; Elaboración de alimentos/bebidas; Fabricación de productos textiles, de vestir, cueros; Actividades relacionadas a producir productos químicos, farmacéuticos; Fabricación de productos de caucho, plástico, minerales, metales;

Fabricación de productos de informática, electrónicos, eléctricos, maquinarias y equipos; Fabricación de vehículos, equipo de transporte; Actividades de suministro de energía; Actividades relacionadas a la Construcción; Actividades relacionadas al Comercio; Actividades de transporte; Actividades de Hostelería; Actividades de Comunicación, informática e información; Finanzas y Seguros; Actividades de investigación, científicas y otras profesionales; Actividades de la Salud y Asistencia Social; Actividades de Arte, entretenimiento y recreación.

Los *p* = 20 PC's son entonces calculados, basados en las variables no estandarizadas:

```
Call:
princomp(x = prov_activ_rel, cor = FALSE)

Standard deviations:
    Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8
4.68180083 2.19441335 1.05138604 0.73333887 0.68140838 0.49578587 0.42863992 0.36986151
    Comp.9    Comp.10    Comp.11    Comp.12    Comp.13    Comp.14    Comp.15    Comp.16
0.26448809 0.17074864 0.15337357 0.11662185 0.07582276 0.05949429 0.04151814 0.03176691
   Comp.17    Comp.18    Comp.19    Comp.20
0.01879206 0.01502224 0.00858742 0.00000000

 20  variables and  24 observations.
```

Ilustración 9: "Output" de R correspondiente a las desviaciones estándar de los 20 PC's de nuestro set de datos para el análisis de caracterización económica de las provincias del Ecuador.

La Ilustración 9 muestra el output de R correspondiente a las desviaciones estándar de los 20 PC's. Si eleváramos cada una de estos valores al cuadrado obtendríamos los valores Eigen.

Luego, los "loadings" pueden ser llamados por R:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11	Comp.12	Comp.13
Agr										0.339	-0.200	0.791	0.189
Silv.madera				-0.280	-0.426	0.115	0.635	-0.314	-0.181	-0.163	0.297		
Pesca.ac												0.133	0.156
Mineria													
alim.beb				0.339	0.303		0.307	0.534	-0.549	0.119	0.170		
text.vest		-0.373	0.211	-0.117	-0.569		-0.432	0.418		0.129	0.141		
quim												-0.209	0.149
plast.cauch.met		-0.348	-0.155	-0.161	0.289	-0.717	-0.144	-0.240	-0.129		0.211		-0.184
elct.eq.maq												-0.177	0.278
veh												-0.101	0.140
sumEner												0.170	0.126
construc								-0.206		0.130	-0.307	-0.394	0.534
comercio	0.839	0.427					-0.122				0.100		-0.100
transp	-0.113		-0.391	-0.494	0.330	0.506	-0.210			0.140	0.259		
serv.aloj.com	-0.525	0.703			-0.174	-0.230	-0.187		-0.113		0.117		
inf.comunic			0.805		0.394	0.146		-0.138	0.256		0.139		
finanz				-0.109				0.175		-0.848	-0.303	0.171	
act.prof.cient						0.146		-0.192	-0.254	0.183	-0.616	-0.101	-0.597
asist.social		-0.184	-0.285	0.703		0.243	-0.129	-0.300	0.219	-0.132	0.253		-0.150
art.entr.recr			-0.172			-0.211	0.404	0.386	0.658		-0.137	-0.191	-0.217

Ilustración 10: "Output" de R correspondiente a los "loadings" de los primeros 13 PC's de nuestro set de datos para el análisis de caracterización económica de las provincias del Ecuador.

La Ilustración 10 es el output de R correspondiente a los coeficientes de los componentes principales (en este caso, presentamos solamente los primeros 13). En otras palabras, estos coeficientes representan a los vectores Eigen. Los coeficientes menores al valor 0.1 no son desplegados por R ya que estos no contribuyen mucho a los PC's, y por esta razón podrían ser omitidos para su interpretación. El output nos muestra para el primer componente principal:

$$Z_1 = 0.839\text{comercio} - 0.113\text{transp} - 0.525\text{serv. aloj. com},$$

que Z_1 contrasta el porcentaje de personas que se dedican a las "Actividades relacionadas al Comercio" con el porcentaje de personas que se dedican a las "Actividades de transporte" y a las "Actividades de Hostelería". Es particularmente con respecto a este contraste que las 24 provincias son diferentes las unas de las otras. El segundo PC viene dado por:

$$Z_2 = -0.373\text{text. vest} - 0.348\text{plast. cauch. met} + 0.427\text{comercio} + 0.703\text{serv. aloj. com},$$

que es también un contraste. Ahora, el segundo PC contrasta las actividades de "Fabricación de productos textiles, de vestir, cueros" y de "Fabricación de productos de caucho, plástico, minerales, metales" *contra* las "Actividades relacionadas al Comercio" y las "Actividades de Hostelería". Podríamos entonces interpretar esto como el contraste entre los porcentajes de personas que se dedican a actividades industriales/manufactureras *contra* los porcentajes de personas dedicadas al comercio o actividades de intercambio. Note que el coeficiente de la variable "asist.social" ha sido removido para facilitar la interpretación, ya que su valor es relativamente pequeño.

Como conocemos, el primer PC es la dirección que posee la mayor cantidad de información (i.e. varianza) de los datos luego de proyectar ortogonalmente cada observación a este vector, que constituye el primer vector eigen $\boldsymbol{e}_1$; el segundo PC corresponde al segundo vector eigen $\boldsymbol{e}_2$ ortogonal al primer vector eigen $\boldsymbol{e}_1$ y que posee la segunda mayor cantidad de información, sin que haya cruce de información con la del primer PC, i.e. $\text{Cov}\{Z_1, Z_2\} = 0$. Queremos lograr una reducción de dimensiones, de tal manera que sea suficiente mirar solamente a un número pequeño (< *p*) de componentes principales, mientras no se pierda mucha información. El siguiente paso por tanto, es el de seleccionar un número apropiado de componentes principales para lograr esto último, sea a través de la regla del 80% o a través del gráfico de sedimentación.

```
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4     Comp.5      Comp.6      Comp.7      Comp.8      Comp.9
Standard deviation     4.6818008 2.1944133 1.05138604 0.73333887 0.68140838 0.495785872 0.428639919 0.369861514 0.26448809
Proportion of Variance 0.7415811 0.1629182 0.03739876 0.01819459 0.01570897 0.008316127 0.006216101 0.004628189 0.00236671
Cumulative Proportion  0.7415811 0.9044994 0.94189813 0.96009272 0.97580169 0.984117820 0.990333921 0.994962110 0.99732882
                          Comp.10      Comp.11      Comp.12      Comp.13      Comp.14      Comp.15      Comp.16      Comp.17
Standard deviation     0.170748641 0.1533735651 0.1166218489 0.0758227559 0.0594942903 0.0415181409 0.0317669056 1.879206e-02
Proportion of Variance 0.000986387 0.0007958548 0.0004601428 0.0001945055 0.0001197522 0.0000583188 0.0000341415 1.194763e-05
Cumulative Proportion  0.998315207 0.9991110617 0.9995712046 0.9997657101 0.9998854623 0.9999437811 0.9999779226 9.999899e-01
                          Comp.18      Comp.19 Comp.20
Standard deviation     1.502224e-02 8.587420e-03       0
Proportion of Variance 7.634877e-06 2.494929e-06       0
Cumulative Proportion  9.999975e-01 1.000000e+00       1
```

Ilustración 11: "Output" de R correspondiente al análisis de importancia de los 20 PC's de nuestro set de datos para el análisis de caracterización económica de las provincias del Ecuador.

La ilustración 11 nos muestra el output de R correspondiente al análisis de importancia de los componentes. R nos da para cada PC Z_i la fracción $\frac{\lambda_i}{\sum_{j=1}^{p}\lambda_j}$, así como $\frac{\sum_{j=1}^{i}\lambda_j}{\sum_{j=1}^{p}\lambda_j}$ (cumulativo). Basándonos en la regla del 80%, podríamos concluir que seleccionar los dos primeros PC's es una muy buena apuesta, ya que seleccionando ambos retenemos el 90.44% de la información total contenida en nuestros datos "censales". La Ilustración 12 nos muestra la otra herramienta usada para la selección de componentes, el gráfico de sedimentación:

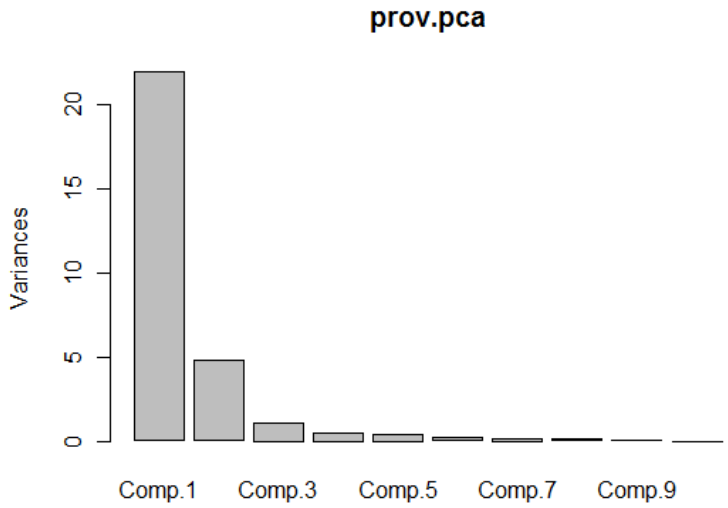


Ilustración 12: Gráfico de sedimentación de nuestro set de datos para el análisis de caracterización económica de las provincias del Ecuador.

El gráfico de sedimentación es una representación gráfica de los valores Eigen, que representan a la varianza de los datos en cada uno de los PC's (i.e. información contenida en cada uno de los PC's). Buscamos "un codo" en el gráfico. El mismo se observa entre los PC's 2 y 3. Por tanto, basándonos en el gráfico, se seleccionaría los dos primeros PC's, ya que si se selecciona el 3er PC se debería seleccionar también el

4to ya que los valores Eigen de estos dos PC's son casi iguales. Los dos criterios en este caso coinciden en la selección de 2PC's con el objetivo de lograr una reducción de dimensiones de nuestro espacio.

Precisamente, el "biplot" es el gráfico que nos muestra un gráfico de dispersión de los "scores" correspondientes al PC1 y al PC2. Recordemos que los "scores" eran nada más y nada menos que las *n* observaciones transformadas desde el espacio original de *p* dimensiones al espacio de *p* dimensiones de los componentes principales. Para la construcción del "biplot" se toman los scores de las variables PC1 y PC2 solamente. La ilustración 13 nos muestra el "biplot" correspondiente a nuestro análisis:

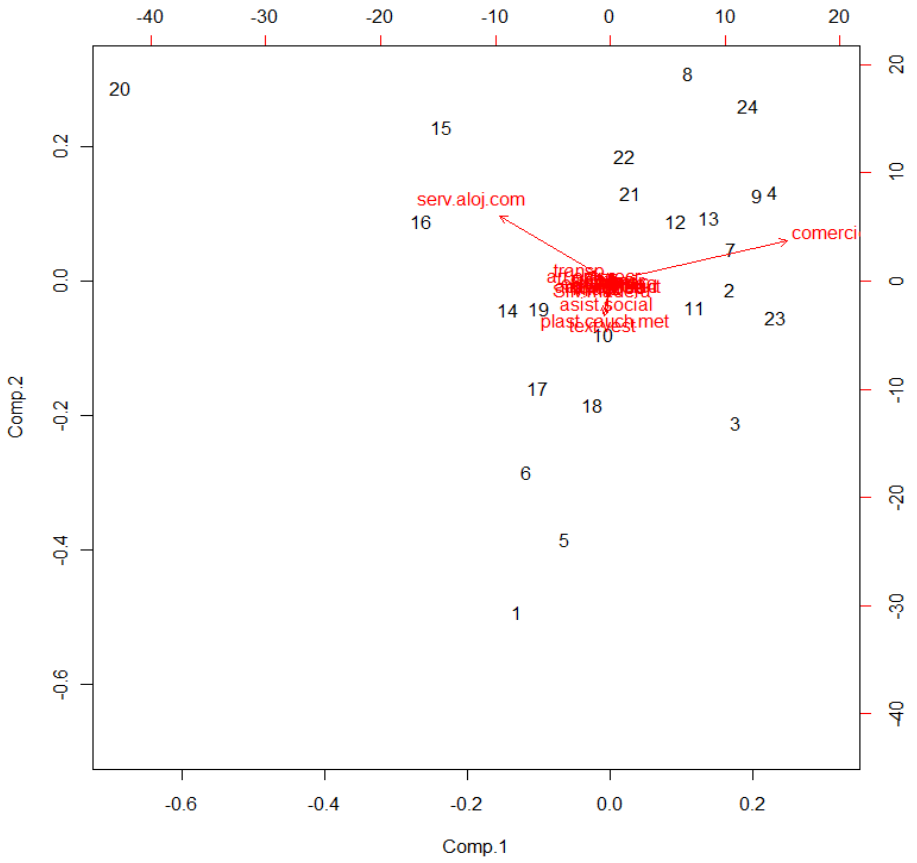


Ilustración 13: "Biplot" generado por R de nuestro set de datos para el análisis de caracterización económica de las provincias del Ecuador.

¿Qué podemos entonces aprender de nuestro "biplot"?

Primero, a cada uno de los vectores en el "biplot" se le ha asignado el nombre de la variable correspondiente. Estos vectores representados en la matriz $\boldsymbol{B}$ de dimensión $p \times 2$ representan una aproximación a las 2 primeras columnas de la matriz $\boldsymbol{H}$ (la versión re-escalada de la matriz $\boldsymbol{M}$). Recordemos e.g. que el vector $\boldsymbol{m}_i$ (la i-enésima columna de $\boldsymbol{M}^t$) representan las coordenadas de la variable i-enésima original en el subespacio de los dos primeros componentes principales. Por lo que por ejemplo, si proyectamos el vector correspondiente a la primera variable, "Agricultura, ganadería",

ortogonalmente en el PC1, obtenemos el coeficiente asignado a esta variable en la combinación lineal del PC1. De la misma manera, si proyectamos este vector en el segundo PC, se obtiene el coeficiente asignado a la variable en cuestión ("Agricultura, ganadería") en la combinación lineal del segundo PC. Por tanto, si pudiéramos graficar las *p* dimensiones del espacio de componentes principales, y proyectar este vector ortogonalmente en todos los PC's, obtendríamos todos los "loadings" o coeficientes de esta variable en los PC's. Este razonamiento aplica también para los demás vectores mostrados en el gráfico. *Segundo*, si proyectamos los *p* vectores en el primer componente principal, se obtendría los "loadings" de cada una de las variables en el primer PC, y por tanto obtendríamos el primer vector Eigen nuevamente ($\boldsymbol{e}_1$). El mismo razonamiento aplicaría si proyectáramos los *p* vectores en el segundo componente principal, se reconstruiría el segundo vector Eigen ($\boldsymbol{e}_2$). *Tercero*, si proyectamos las observaciones del "biplot", que se encuentran en el subespacio de los dos primeros PC's, ortogonalmente en los vectores ($\boldsymbol{b}_1$, $\boldsymbol{b}_2$, $\boldsymbol{b}_3, \ldots, \boldsymbol{b}_p$), las columna de $\boldsymbol{B}^t$, encontraremos a las observaciones originales. Por tanto, si proyectamos e.g. la observación "8" (que corresponde a la provincia de Esmeraldas) en la flecha de la variable "comercio", la longitud de esta proyección nos da el porcentaje original de actividades económicas de Esmeraldas que trabaja en el sector del comercio. Según se observa, el porcentaje es alto y es mayor que el promedio. Lo mismo podríamos decir de "3" (Cañar) y de "24" (Santa Elena). Pero si miramos a la observación e.g. "15" (Napo) y la proyectamos ortogonalmente en esa dirección, daremos en el otro lado, lo que significa que un porcentaje menor al promedio se encuentra trabajando en el comercio en esta provincia. *Cuarto*, el ángulo entre las flechas determina la correlación entre las variables correspondientes a dichas flechas. Suponga que θ_{ij} sea el ángulo entre dos flechas correspondientes a las variables i y j, entonces el $\cos\theta_{ij}$ representa la correlación entre estas dos variables. Por ejemplo, en nuestro "biplot" el ángulo entre las flechas correspondientes a "comercio" y "serv.aloj.com" es cercano a 135°, y por tanto el coseno de tal ángulo será un número negativo algo cercano a -1, lo que quiere decir que estas dos variables poseen una alta correlación negativa. *Quinto*, la longitud de las flechas es proporcional a las varianzas de las variables originales. Por tanto, podríamos concluir que "comercio" y "serv.aloj.com" poseen una varianza más alta en comparación con las demás variable. Y *sexto*, podemos dar una interpretación lógica a nuestros datos en base a este "biplot" y los coeficientes de Z_1 y Z_2 presentados anteriormente. Como se dijo anteriormente, existe un contraste en la dirección del segundo componente principal entre las actividades de "Fabricación de productos textiles, de vestir, cueros" y de "Fabricación de productos de caucho, plástico, minerales, metales" *contra* las "Actividades relacionadas al Comercio" y las "Actividades de Hostelería", lo que se podría ver como un contraste entre el porcentaje de personas que se dedican a actividades industriales/manufactureras contra las personas dedicadas al comercio o actividades de intercambio. Sólo el PC2 nos brindó coeficientes con interpretación:

$$Z_2 = -0.373\text{text.vest} - 0.348\text{plast.cauch.met} + 0.427\text{comercio} + 0.703\text{serv.aloj.com.}$$

Este contraste se lo puede observar también en nuestro “biplot”, en la dirección del PC2. No se encuentra una interpretación por provincia usando la técnica del PCA. Sin embargo, en general, podríamos decir lo siguiente:

- ✓ Las provincias situadas en la región Costa, Oriente e Insular poseen un mayor porcentaje de unidades económicas dedicadas a la actividad del Comercio o de Intercambio en comparación con las provincias situadas en la región Sierra (en la dirección del PC2).
- ✓ Las provincias situadas en la región Sierra poseen un mayor porcentaje de unidades económicas dedicadas a actividades industriales/manufactureras (“Fabricación de productos textiles, de vestir, cueros”, “Fabricación de productos de caucho, plástico, minerales, metales”) en comparación con las provincias situadas en las demás regiones (en la dirección del PC2).
- ✓ Las provincias del Napo, Pastaza y Galápagos poseen un mayor porcentaje de unidades económicas dedicadas a “Actividades de Hostelería” en comparación con las demás provincias del país.

Recordemos que todas estas interpretaciones y propiedades no son exactamente verdaderas, lo serían si y sólo si hubiéramos hecho la transformación desde el espacio original de *p* dimensiones al espacio nuevo de los PC’s de *p* dimensiones. Sin embargo, se espera que nuestro “biplot” sea una buena representación de los datos en un número menor de dimensiones (que sea una muy buena aproximación del espacio original de *p* dimensiones), ya que se escogió los dos primeros PC’s de tal manera que retengamos el 90.44% de la información total contenida en nuestros datos originales. Adicionalmente, debido a que se quería tomar en cuenta las varianzas reales de las variables, se usó la matriz de covarianza para nuestro análisis.

8.2 Caracterización Económica de las provincias del Ecuador a través del Análisis Clúster

En esta sección, tendremos como objetivo caracterizar económicamente a las provincias del Ecuador a través del agrupamiento de provincias que tengan patrones de porcentajes similares en las actividades económicas, o en otras palabras, agrupar provincias en grupos que trabajan en sectores o actividades económicas similares.

Para aquello, usando la metodología detallada en la Sección 6.2, se escogió la siguiente estrategia:

1. Decidiremos primero el rango de posible número de clústers con un Análisis de Clúster Jerárquico aglomerativo, usando como disimilaridad inter-clúster sea la

vinculación simple, la vinculación completa, la vinculación promedio o la vinculación de Ward; usando como criterio de elección del tipo de disimilaridad inter-clúster al “coeficiente aglomerativo” que provee R al desplegar el árbol de clúster.

2. Luego, implementaremos PAM (Particionamiento Alrededor de los Mediodes) para cada número de clústers que se encuentra dentro del rango de posible número de clústers escogido en el paso previo. Para cada número de clústers obtendremos una solución que será evaluada a través de las herramientas gráficas de diagnóstico ya descritas: el gráfico “silhouette” y el “clusplot”. La configuración de clúster escogida, será aquella que presente valores “silhouette” positivos para las observaciones en los clústers (en caso de presencia de “outliers” o datos atípicos, los mismos necesitarán más investigación) y aquella que presente una configuración de clúster definida (grupos separados, si es posible) al visualizar el “clusplot”.

8.2.1 Elección del número de clústers

Como mencionamos en nuestra estrategia, usaremos el método de Clúster Jerárquico aglomerativo para decidir el rango de posible número de clústers. Usar como disimilaridad inter-clúster a la vinculación simple no siempre es una buena estrategia, debido al problema del “chaining” (encadenamiento). Por ello, nos movemos a decidir entre los demás métodos de disimilaridad. Usando al “coeficiente aglomerativo” como criterio de elección del tipo de disimilaridad inter-clúster, optamos por usar a la vinculación de Ward por tener un Coeficiente aglomerativo = 0.88, el más alto de entre todas las definiciones de disimilaridad consideradas.

Debido a la naturaleza del algoritmo de Clúster Jerárquico aglomerativo, este método nos da una secuencia de clústers anidados. Como fue mencionado anteriormente, gracias a la estructura restrictiva del algoritmo, este nos permite presentar las soluciones en una estructura que se asemeja a un árbol, llamada árbol de clúster. El árbol de clúster para el análisis de Clúster Jerárquico aglomerativo usando como definición de disimilaridad inter-clúster al método de Ward se muestra en la Ilustración 14:

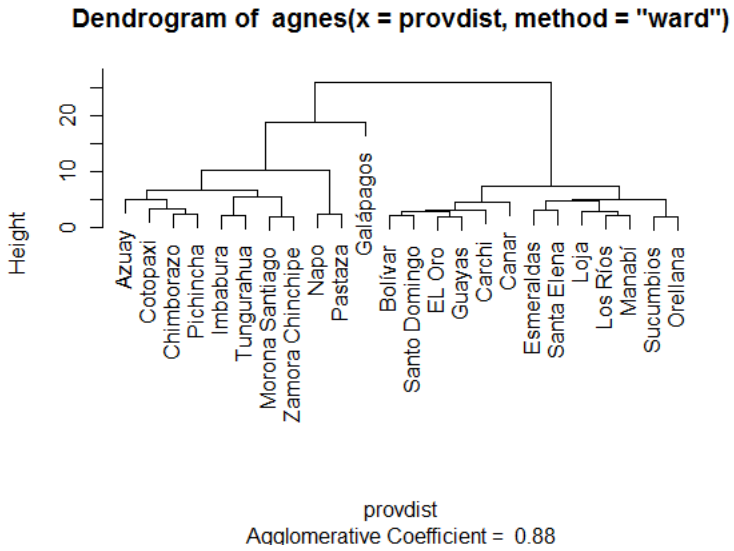


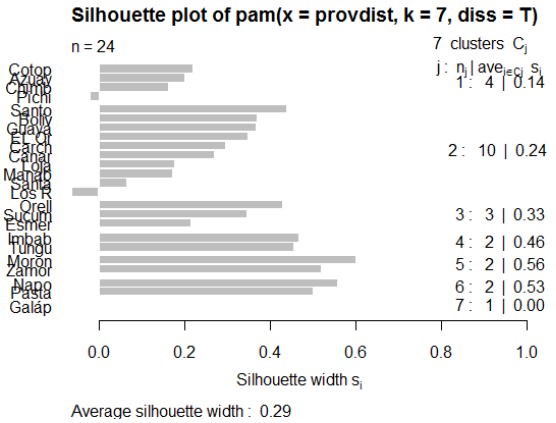
Ilustración 14: Árbol de clúster correspondiente al método Ward de nuestro set de datos para el análisis de caracterización económica.

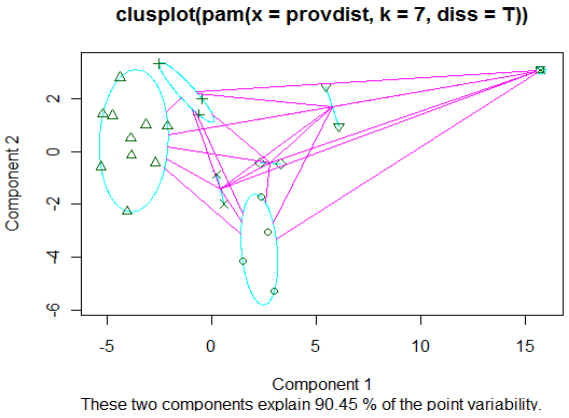
Es claro que se prefiere tener un rango con posibles números de clústers que simplemente decir existen entre 1 y 24 clústers (el número de observaciones de nuestro set de datos). Del árbol de clúster presentado en la Ilustración 14 de arriba, podríamos concluir que en nuestros datos existen $3 \leq k \leq 7$ clústers, cuyas configuraciones podrían ser analizadas luego de realizar un análisis PAM y evaluadas con herramientas gráficas de diagnóstico.

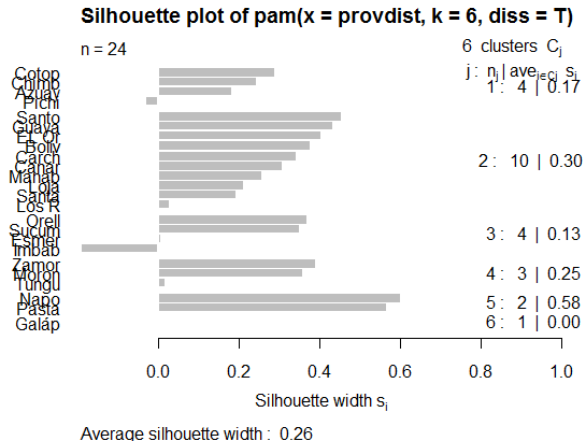
8.2.2 Aplicación de PAM y de las herramientas gráficas de diagnóstico

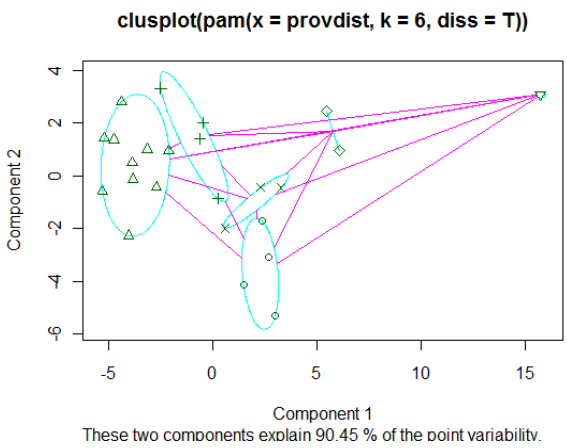
Se decidió usar el algoritmo de Particionamiento Alrededor de los Mediodes y no e.g. el de *K*-medias ya que este último es sensible a la presencia de datos atípicos. En caso de existir datos atípicos que no hayan sido identificados todavía, el uso del método de *K*-medias podría causar que los centroides (o centros), calculados como medias multivariadas, se encuentren en zonas del espacio en donde las observaciones no son físicamente posibles, ya que los datos atípicos atraen a las medias hacia si, y por ende la configuración de clúster solución no sería confiable ya que en la siguiente iteración estos "nuevos" centroides son nuevamente utilizados para reasignar cada observación a uno de los clústers "nuevos" definidos por los "nuevos" centroides.

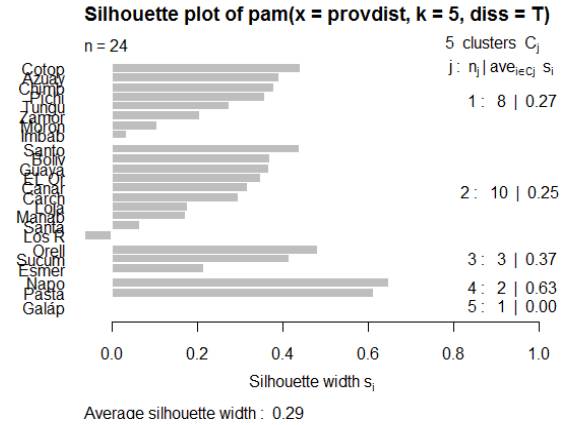
Luego de implementar PAM en R para $k = 3, 4, 5, 6, 7$; se evaluó cada solución usando el gráfico silueta y el "clusplot". Estos gráficos, para cada solución de k son mostrados en la Ilustración 15.

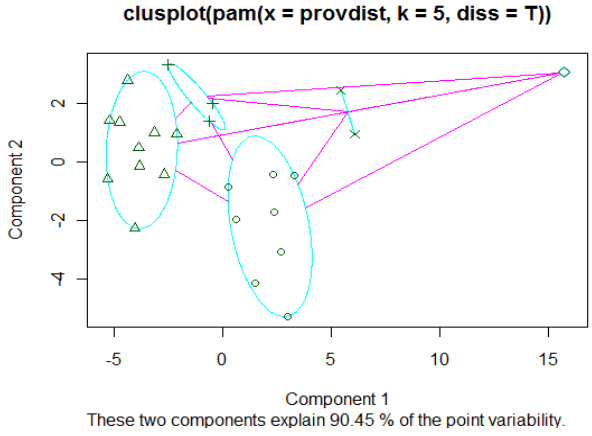

Silhouette plot of pam(x = provdist, k = 7, diss = T)
n = 24
7 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 4 | 0.14
2 : 10 | 0.24
3 : 3 | 0.33
4 : 2 | 0.46
5 : 2 | 0.56
6 : 2 | 0.53
7 : 1 | 0.00
Cotop
Azuay
Chimb
Pichi
Santo
Bolív
Guaya
El Or
Carch
Cañar
Loja
Manab
Santa
Los R
Orell
Sucum
Esmer
Imbab
Tungu
Moron
Zamor
Napo
Pasta
Galáp
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.29


clusplot(pam(x = provdist, k = 7, diss = T))
Component 2
Component 1
These two components explain 90.45 % of the point variability.


Silhouette plot of pam(x = provdist, k = 6, diss = T)
n = 24
6 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 4 | 0.17
2 : 10 | 0.30
3 : 4 | 0.13
4 : 3 | 0.25
5 : 2 | 0.58
6 : 1 | 0.00
Cotop
Chimb
Azuay
Pichi
Santo
Guaya
El Or
Bolív
Carch
Cañar
Manab
Loja
Santa
Los R
Orell
Sucum
Esmer
Imbab
Zamor
Moron
Tungu
Napo
Pasta
Galáp
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.26


clusplot(pam(x = provdist, k = 6, diss = T))
Component 2
Component 1
These two components explain 90.45 % of the point variability.


Silhouette plot of pam(x = provdist, k = 5, diss = T)
n = 24
5 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 8 | 0.27
2 : 10 | 0.25
3 : 3 | 0.37
4 : 2 | 0.63
5 : 1 | 0.00
Cotop
Azuay
Chimb
Pichi
Tungu
Zamor
Moron
Imbab
Santo
Bolív
Guaya
El Or
Cañar
Carch
Loja
Manab
Santa
Los R
Orell
Sucum
Esmer
Napo
Pasta
Galáp
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.29


clusplot(pam(x = provdist, k = 5, diss = T))
Component 2
Component 1
These two components explain 90.45 % of the point variability.

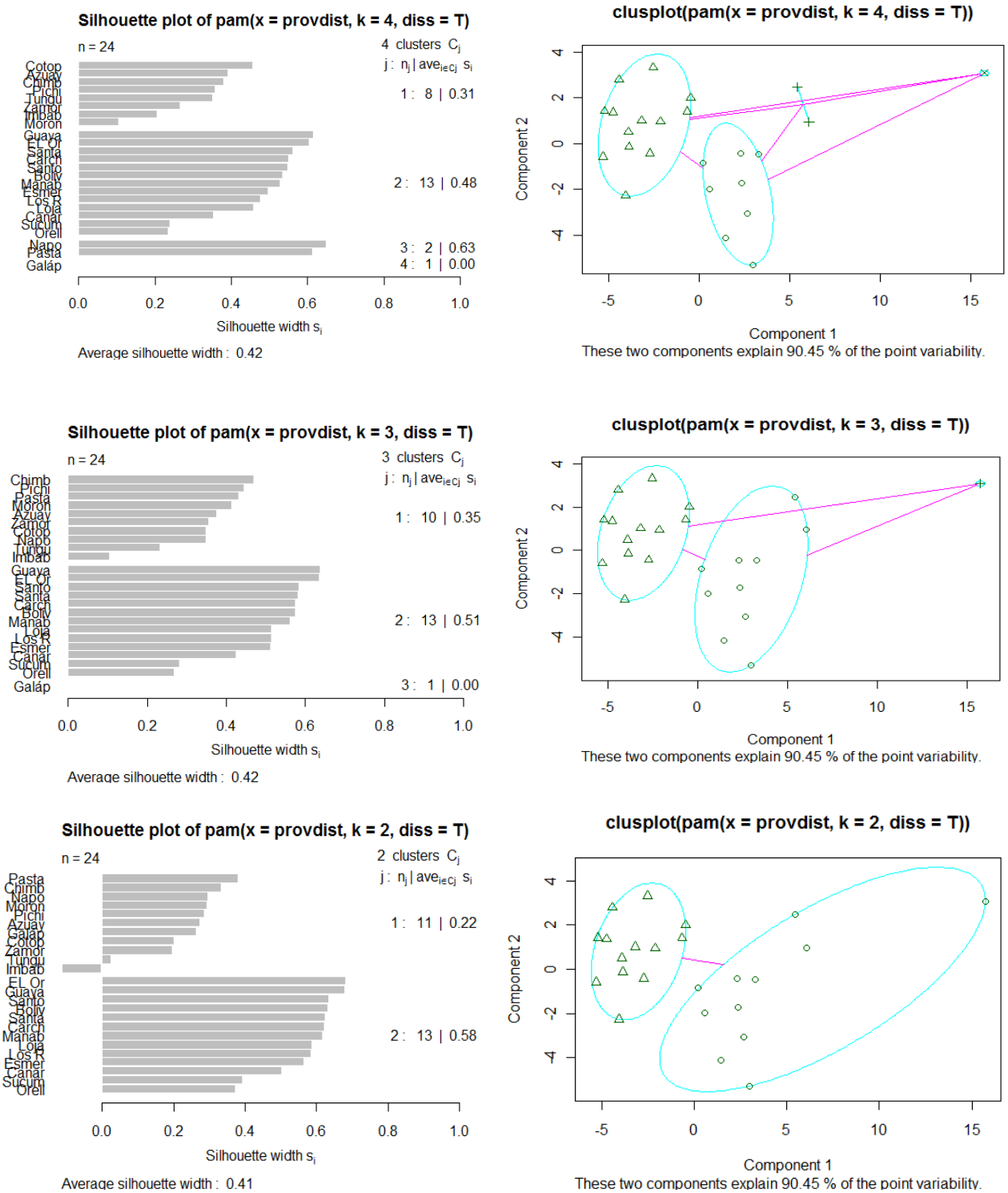


Ilustración 15: Gráficos siluetas y gráficos "clusplot" para $k = 3, 4, 5, 6, 7$; de nuestro set de datos para el análisis de caracterización económica.

De acuerdo con los gráficos de la Ilustración 15, las configuraciones de clústers con $k = 3$ y $k = 4$ poseen valores "silhouette" positivos en cada grupo formado y ambas poseen una clara separación entre los grupos al graficar (en forma de diagrama de dispersión) los "scores" de las $n = 24$ observaciones pertenecientes al PC1 y al PC2. Recordemos que estos representan las observaciones proyectadas en el subespacio de los componentes principales compuesto por los dos primeros PC's. Por ello, las configuraciones de clúster "óptimas" en las cuales cada observación de cada clúster se encuentra en promedio más cercana a las observaciones compañeras de su clúster que a las observaciones de otros clústers (valores "silhouette" positivos) y cuyos grupos se encuentran completamente separados los unos de los otros en 2-d son las configuraciones de $k = 3$ y $k = 4$ grupos (si en 2-d los grupos se encuentran separados, en *p*-d los grupos estarán igualmente separados). Entre las dos opciones finales, *optaremos por elegir la configuración de clúster con* $k = 3$*, ya que posee valores "silhouette" más altos que cuando* $k = 4$. Adicionalmente, cuando $k = 4$, nos encontramos con un clúster de dos observaciones (provincias Napo y Pastaza), que no es generalmente común tener (y a veces puede ser un síntoma de la presencia de datos atípicos). A pesar de esto, al adoptar como solución $k = 3$ grupos, consideramos posible el hecho de que Galápagos haya resultado como única observación en el clúster_3, ya que esta provincia por si sóla constituye una región, la región Insular y provincia de Galápagos. Por tanto, pasaremos por alto esto ya que esta solución nos brinda una interpretación interesante. El hecho que el 90.45% de la información total de los datos es retenida en los dos primeros PC's hace que nuestra elección sea la más confiable.

En consecuencia, el análisis clúster presente, que agrupa provincias cuyas unidades económicas se dedican a más o menos las mismas actividades económicas, resulta en 3 grupos conformados de la siguiente manera (ver Tabla 5 e Ilustración 16):

Clúster 1	Clúster 2	Clúster 3
•Chimborazo •Pichincha •Pastaza •Morona Santiago •Azuay •Zamora Chinchipe •Cotopaxi •Napo •Tungurahua •Imbabura	•Guayas •El Oro •Santo Domingo •Santa Elena •Carchi •Bolívar •Manabí •Loja •Los Ríos •Esmeraldas •Cañar •Sucumbíos •Orellana	•Galápagos

Tabla 5: Solución final ($k = 3$ grupos) del Análisis Clúster para el estudio de caracterización económica de las provincias del Ecuador.

En la tabla 5 se detallan los grupos y las provincias que contienen cada uno de ellos. Las provincias que se encuentran en un mismo grupo poseen unidades económicas que se dedican a más o menos las mismas actividades económicas.

A partir de los clústers formados según las actividades económicas (Tabla 5), podemos también representar la distribución económica-geográfica del país (Ilustración 16):

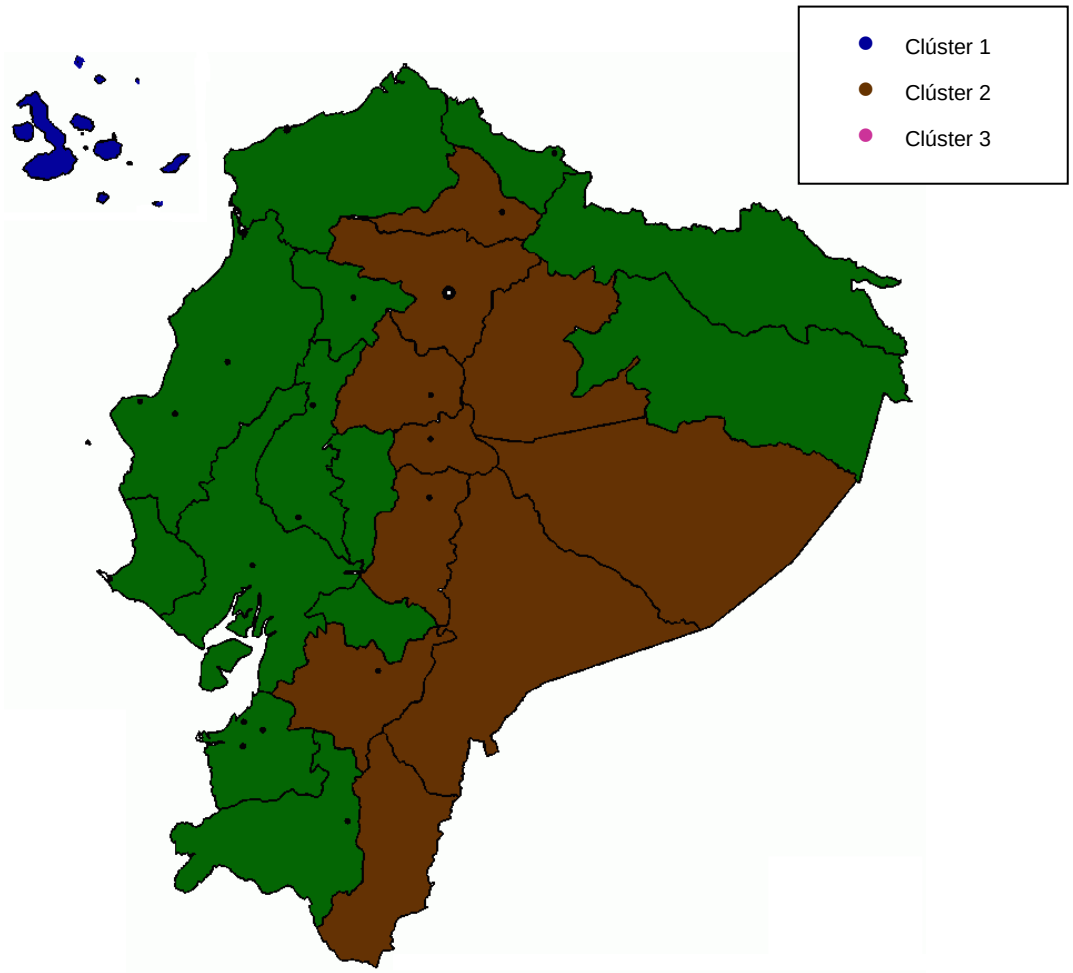


Ilustración 16: Distribución económica-geográfica del país a partir de los clústers formados según las actividades económicas

Estos agrupamientos como fue mencionado inicialmente, pueden ser utilizados por agentes responsables de la toma de decisiones a nivel macroeconómico en planeaciones de ordenamiento territorial, la aplicación de políticas comunes en grupos de provincias dedicados a las mismas (similares) actividades económicas; así como por empresarios, quienes desearían saber en qué provincias invertir en base a esta información.

8.3 Caracterización empresarial de las provincias del Ecuador a través del Análisis de Componentes Principales (PCA)

Como mencionamos anteriormente, las (13) variables económicas recogidas del Censo Económico 2010 del INEC relacionadas con las políticas empresariales de las unidades económicas encuestadas fueron: Monto de Financiamiento con Institución Privada; Monto de Financiamiento con el Gobierno; Monto requerido de financiamiento; Monto de gasto en manejo de desechos; Monto de gasto en investigación y desarrollo; Monto de gasto en capacitación y formación; Gastos anuales en remuneraciones; Gastos anuales en materia prima; Gastos anuales en envases y embalajes; Gastos anuales en compras y mercadería; Tasas, contribuciones y otros impuestos anuales (excluye IVA…); Total de ingresos anuales percibidos por ventas o prestación; Gastos anuales kilovatios/hora.

Finalmente, se obtuvo los datos como se muestra en la Tabla 4, con las provincias del Ecuador como observaciones, como lo fue en el análisis previo, y como variables las 13 que fueron mencionadas en el párrafo anterior. Los valores para cada celda representan el monto promedio correspondiente a la variable económica-empresarial X_j en la provincia i. Como mencionamos en la Sección 8.1, es difícil interpretar las 13 variables de manera simultánea usando el simple análisis descriptivo, por tanto, para obtener una conclusión general acerca de cómo las provincias se diferencian las unas de las otras con respecto a los montos promedio de las variables de política empresarial aplicaremos nuevamente el análisis de componentes principales. Se espera entonces poder reducir las dimensiones de tal manera que se puedan formular conclusiones al mirar algunos gráficos en el espacio reducido de los PC's.

Los gráficos de sedimentación luego de implementar el análisis de componentes principales usando la matriz de covarianza y correlación respectivamente (para cuando cor = TRUE y para cuando cor = FALSE en R) se muestran en la Ilustración 17:

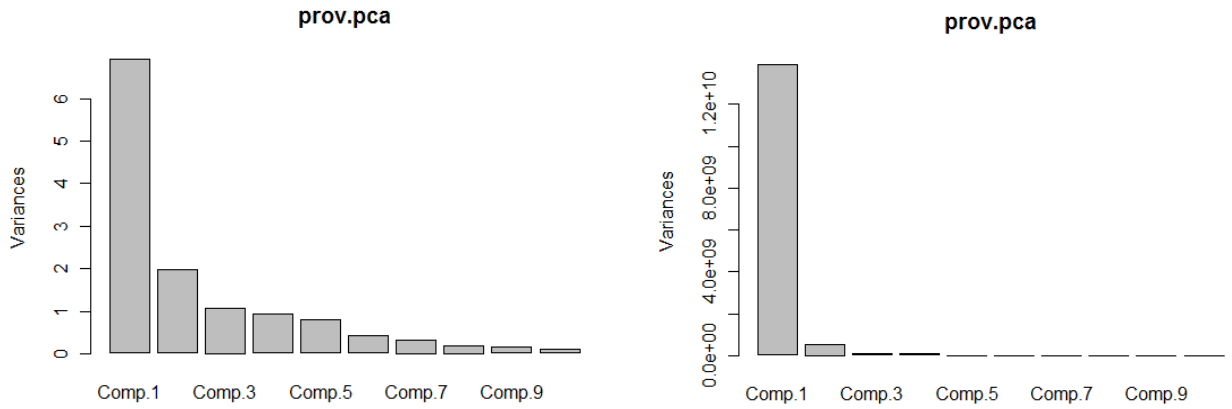


Ilustración 17: Gráficos de Sedimentación para el análisis de componentes principales usando la matriz de covarianza (izquierda) y correlación (derecha), aplicado a nuestro set de datos (Tabla 4) para el análisis de caracterización empresarial.

Seleccionar los dos primeros componentes principales en ambos casos sería razonable según los gráficos de sedimentación correspondientes. Al demandar el "biplot" obtendremos los siguientes gráficos (para cuando cor = TRUE y para cuando cor = FALSE respectivamente en R), mostrados en la Ilustración 18:

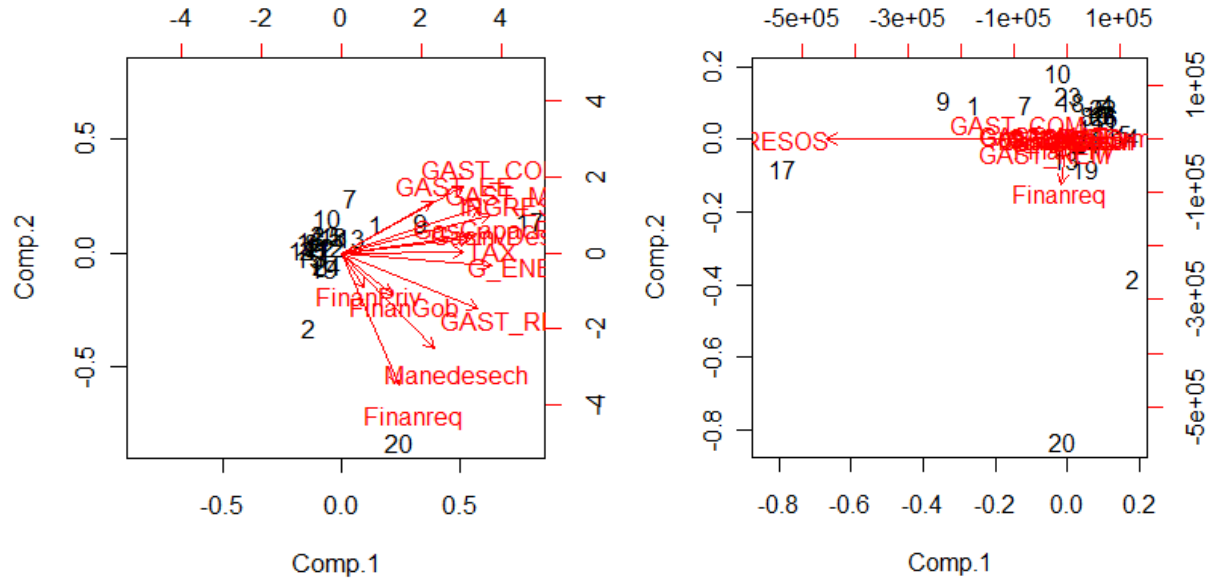


Ilustración 18: "Biplots" correspondientes al PCA implementado con la matriz de correlación (izquierda) y con la matriz de covarianza (derecha), aplicado a nuestro set de datos para el análisis de caracterización empresarial.

Recordemos que cuando usamos la matriz de covarianza para realizar el PCA permitimos que las variables con mayor variabilidad sean las dominantes en las direcciones de los componentes principales, por otro lado cuando usamos la matriz de correlación se da igual importancia a las variables en las direcciones de los PC's.

De la ilustración 18 muy pocas conclusiones interesantes podemos extraer. No se pueden encontrar claros contrastes. Para realizar el PCA creemos más adecuado el uso de la matriz de correlación debido a la gran varianza que poseen las variables de estudio (Ilustración 18, izquierda). De la Ilustración 18, las conclusiones más claras que podríamos extraer son las siguientes:

- ✓ La provincia de Pichincha y la provincia del Guayas poseen un mayor ingreso promedio por local, un mayor gasto anual en compras y mercadería promedio por local, un mayor gasto anual promedio en materia prima por local en comparación con las demás provincias
- ✓ Las provincias de Galápagos y Bolívar son aquellas que requieren un mayor financiamiento en promedio por local en comparación con las demás provincias

No ahondaremos en más detalles en cuanto a encontrar más contrastes, ya que los mismos no son claros en ambos gráficos. Cuando esto sucede, lo mejor es no emplear esta técnica excesivamente para sacar conclusiones.

8.4 Caracterización empresarial de las provincias del Ecuador a través del Análisis Clúster

En esta sección, tendremos como objetivo caracterizar empresarialmente a las provincias del Ecuador a través del agrupamiento de provincias que tengan patrones de políticas empresariales similares, o en otras palabras, agrupamiento de provincias que tengan montos promedios similares entre las variables que caracterizan a las políticas empresariales. Para aquello, la estrategia que usaremos será esencialmente la misma que se usó en la Sección 8.2.:

1. Decidiremos primero el rango de posible número de clústers con un Análisis de Clúster Jerárquico aglomerativo, usando como disimilaridad inter-clúster sea la vinculación simple, la vinculación completa, la vinculación promedio o la vinculación de Ward; usando como criterio de elección del tipo de disimilaridad inter-clúster al "coeficiente aglomerativo" que provee R al desplegar el árbol de clúster.
2. Luego, implementaremos PAM (Particionamiento Alrededor de los Mediodes) para cada número de clústers que se encuentra dentro del rango de posible número de clústers escogido en el paso previo. Para cada número de clústers

obtendremos una solución que será evaluada a través de las herramientas gráficas de diagnóstico: el gráfico “silhouette” y el “clusplot”. La configuración de clúster escogida será aquella que presente valores “silhouette” positivos para las observaciones en los clústers (en caso de presencia de “outliers” o datos atípicos, los mismos necesitarán más investigación) y aquella que presente una configuración de clústers separados (si es posible) al visualizar el “clusplot”.

8.4.1 Elección del número de clústers

Como mencionamos en nuestra estrategia, usaremos el método de Clúster Jerárquico aglomerativo para decidir el rango de posible número de clústers. Usando como criterio de elección del tipo de disimilaridad inter-clúster al “coeficiente aglomerativo” desplegado por R, optamos por usar a la vinculación de Ward por tener un Coeficiente aglomerativo = 0.94, el más alto de entre todas las definiciones de disimilaridad consideradas.

Debido a la naturaleza del algoritmo de Clúster Jerárquico aglomerativo, este método nos da una secuencia de clústers anidados. Gracias a la estructura restrictiva del algoritmo, es posible presentar las soluciones en un árbol de clúster. El árbol de clúster para el análisis de Clúster Jerárquico aglomerativo de nuestros datos (Tabla 4) usando como definición de disimilaridad inter-clúster al método de Ward se muestra en la ilustración 19:

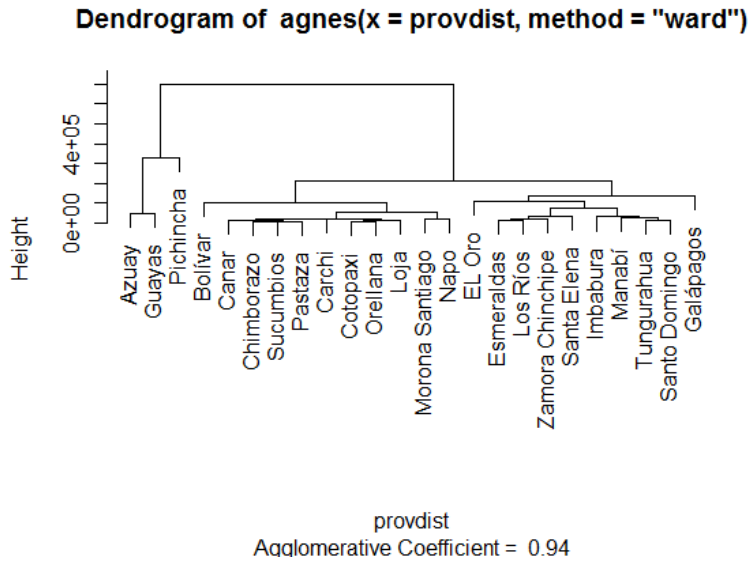
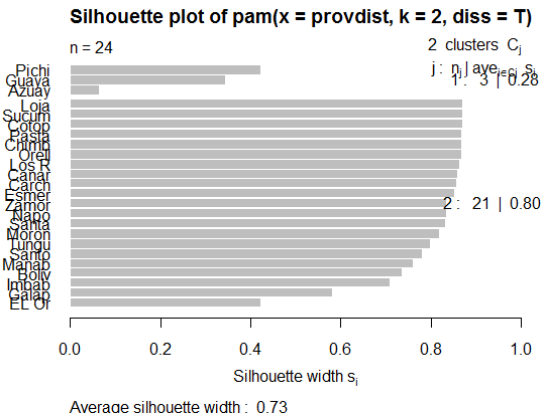
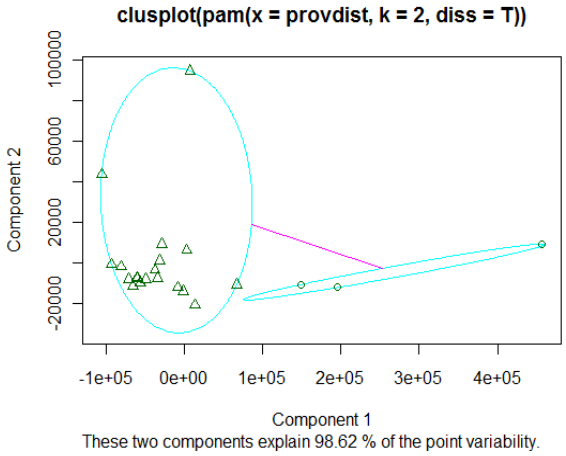


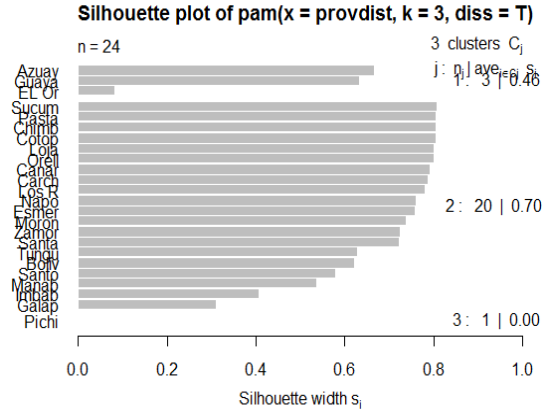
Ilustración 19: Árbol de clúster empleando el método de Ward como definición de disimilaridad, aplicado a nuestro set de datos para el análisis de caracterización empresarial.

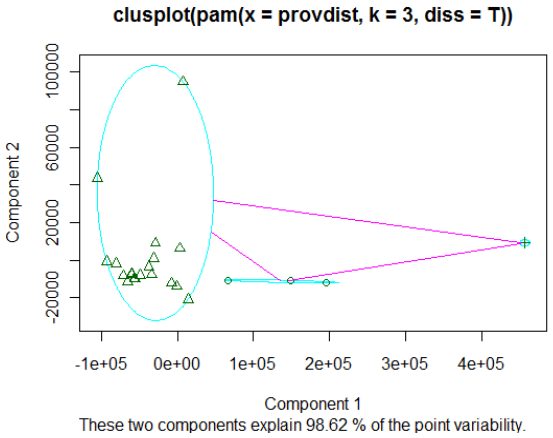
De este gráfico y en base a la definición de disimilaridad, que es representada en el gráfico de árbol de clúster a través de la altura (Height), podríamos concluir que

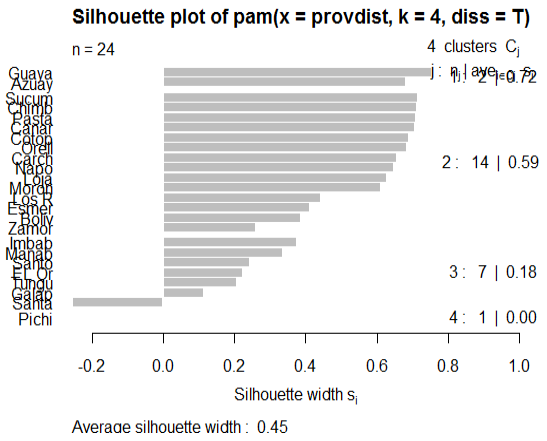
existirían entre $2 \leq k \leq 6$ grupos. Una vez que tenemos este rango de posible número de clústers, usamos esto como entrada para el algoritmo PAM y evaluaremos las diferentes configuraciones de clúster usando el gráfico “silhouette” y el gráfico “clusplot”. A continuación mostramos estos gráficos de diagnóstico para $k = 2$, $k = 3$, $k = 4$, $k = 5$ y $k = 6$.

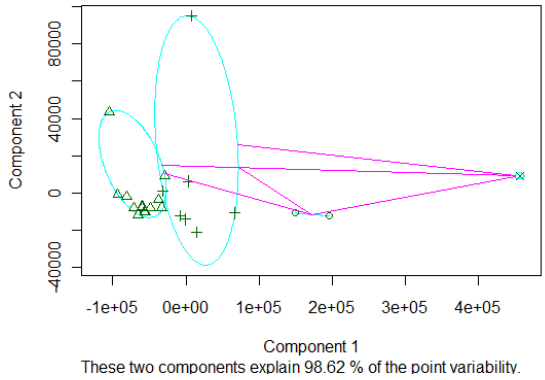












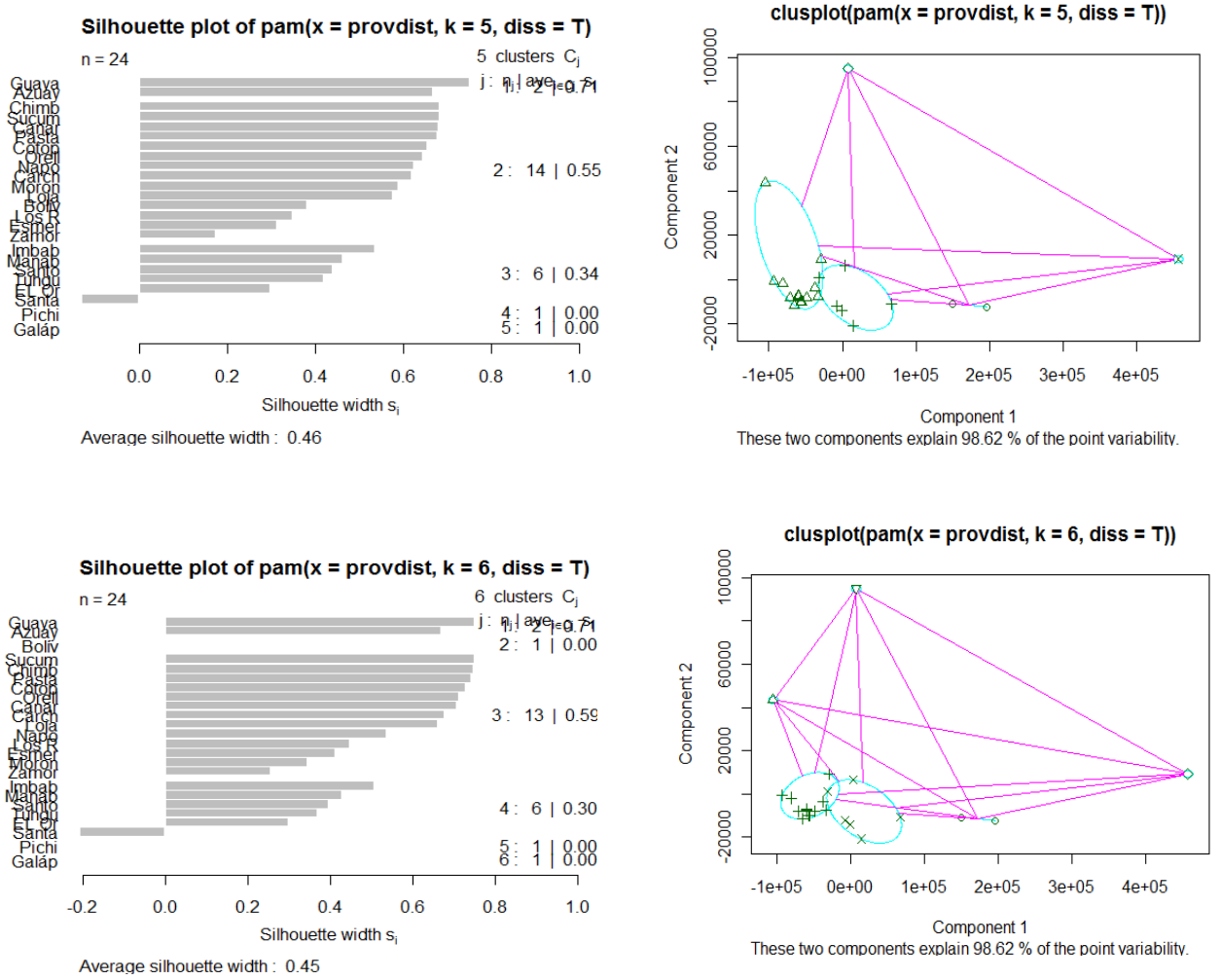


Ilustración 20: Gráficos siluetas y gráficos "clusplot" para $k = 2, 3, 4, 5, 6$ en nuestro set de datos para el análisis de caracterización empresarial.

Según los gráficos de diagnóstico mostrados concluiríamos que las configuraciones de clúster "óptimas", en las que cada observación de cada clúster se encuentra en promedio más cercana a las observaciones compañeras de su clúster que a las observaciones de otros clústers (valores "silhouette" positivos) y cuyos grupos se encuentran completamente separados los unos de los otros en el subespacio conformado por los dos primeros PC's, suceden cuando $k = 2$ y $k = 3$ (si en 2-*d* los grupos se encuentran separados, en *p*-d los grupos estarán igualmente separados). Optaremos entonces como solución la configuración clúster con $k = 3$ por ser una solución interesante.

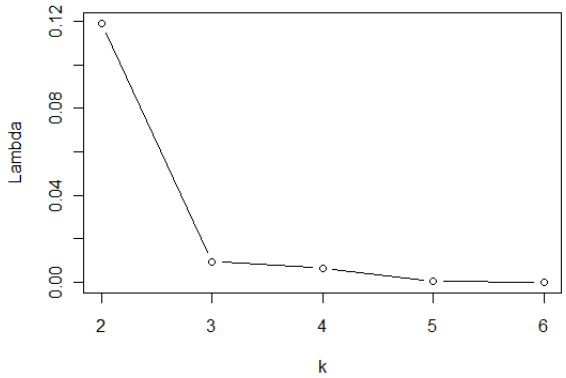


Ilustración 21: Gráfico de sedimentación del Lambda de Wilk para $k = 2, 3, 4, 5, 6$; de nuestro set de datos para el análisis de caracterización empresarial.

Además, el gráfico del Lambda de Wilk (que mide la varianza intra-grupos contra la varianza inter-grupos) también soporta nuestra decisión de tomar $k = 3$.

El hecho que el 98.62% de la información total de los datos es retenida en los dos primeros PC's (considerando las varianzas reales de las variables empresariales), hace que nuestra elección sea más confiable. A pesar de que tener clústers con una sóla observación no es muy común, ya que muchas veces puede ser un síntoma de presencia de datos atípicos, hemos retenido al clúster_3 compuesto por la observación correspondiente a la provincia de Pichincha únicamente. Observando los datos en nuestro set este hecho puede ser una elección razonable, ya que Pichincha posee montos promedios muy altos en todas las variables empresariales consideradas (de hecho en casi todas es la que posee el monto promedio más alto). Esto reflejaría el gran desarrollo empresarial y económico que posee la provincia de Pichincha, siendo la provincia en la que e.g. el Ingreso promedio por unidad económica (u.e.) es el más alto, el gasto promedio en investigación y desarrollo por u.e. es el más alto, el gasto promedio en capacitación y formación por u.e. es el más alto, el monto promedio de Tasas, contribuciones y otros impuestos anuales por u.e. es el más alto, y así con casi todas las variables consideradas (13).

Luego, aparecen las provincias de Azuay, Guayas y El Oro; quienes se ubican en el clúster_1. Por tanto, estas provincias poseen montos promedios similares entre las variables económicas consideradas. Estos clústers y las provincias que conforman el clúster_2 son resumidos a continuación en la siguiente tabla:

Clúster 1	Clúster 2	Clúster 3
• Azuay • Guayas • El Oro	• Sucumbios • Pastaza • Chimborazo • Cotopaxi • Loja • Orellana • Cañar • Carchi • Los Ríos • Napo • Esmeraldas • Morona Santiago • Zamora Chinchipe • Santa Elena • Tungurahua • Bolívar • Santo Domingo de los Tsáchilas • Manabí • Imbabura • Galápagos	• Pichincha

Tabla 6: Solución ($k = 3$ grupos) del Análisis Clúster para el estudio de caracterización empresarial de las provincias del Ecuador.

En la tabla 6 se detallan los grupos y las provincias que contienen cada uno de ellos. Las provincias que se encuentran en un mismo grupo poseen montos promedios cercanos por local entre las variables de caracterización empresarial consideradas.

Como se puede apreciar en la Tabla 5, bien se podría decir que el presente Análisis Clúster agrupa en el $clúster_3$ a la provincia más importante económicamente hablando del Ecuador por albergar a su capital, Quito, y por ser el mayor centro administrativo, financiero y comercial del país[4]. En el $clúster_1$ se agrupan a las provincias de tal vez mayor importancia económica del país luego de Pichincha: Azuay, Guayas y El Oro; y en el $clúster_2$ las restantes. Nuestra configuración clúster solución agrupa provincias con similar fuerza económica en el país.

Dado que el $clúster_2$ está compuesto por muchas provincias, decidimos implementar las estrategia inicial de nuestro Análisis Clúster pero ahora sólo en este grupo, con el objetivo de formar “sub-clústers”. En otras palabras, excluyendo las provincias: Pichincha, Azuay, Guayas y El Oro. Por tanto, primero se escoge un rango

[4] Fuente: Wikipedia (http://es.wikipedia.org/wiki/Provincia_de_Pichincha)

convincente para el posible número de “sub-clústers” existentes en los datos de estas provincias (restantes) a través del método de clúster jerárquico aglomerativo, seleccionando al método de Ward como la definición de disimilaridad. La solución de lo último se lo puede visualizar en el árbol de clúster presentado en la Ilustración 22 siguiente:

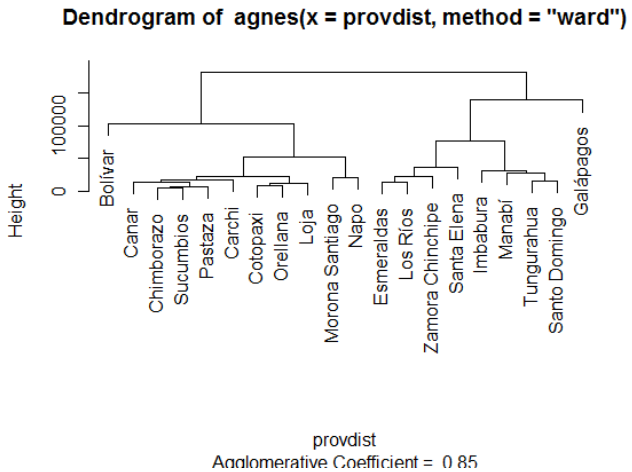


Ilustración 22: Árbol de clúster empleando el método de Ward como definición de disimilaridad, aplicado a nuestro set de datos para el análisis de caracterización empresarial excluyendo las observaciones (provincias) Pichincha, Azuay, Guayas y El Oro.

De la ilustración 22 anterior podríamos decir que existen $2 \leq sub.k \leq 5$ sub-grupos, que vendrían a ser “sub-clústers” en el análisis original de esta sección. Por tanto, aplicamos PAM y usamos el gráfico “silhouette” y el gráfico “clusplot” para evaluar las diferentes configuraciones de clúster. La configuración de clúster “óptima”, luego de implementar nuestra estrategia, fue la de $sub.k = 3$. A continuación, mostraremos el gráfico “silhouette” y el gráfico “clusplot” para $sub.k = 3$ (Ilustración 23). Esta vez, no mostraremos estos gráficos para todos los valores de $sub.k$ en el rango seleccionado.

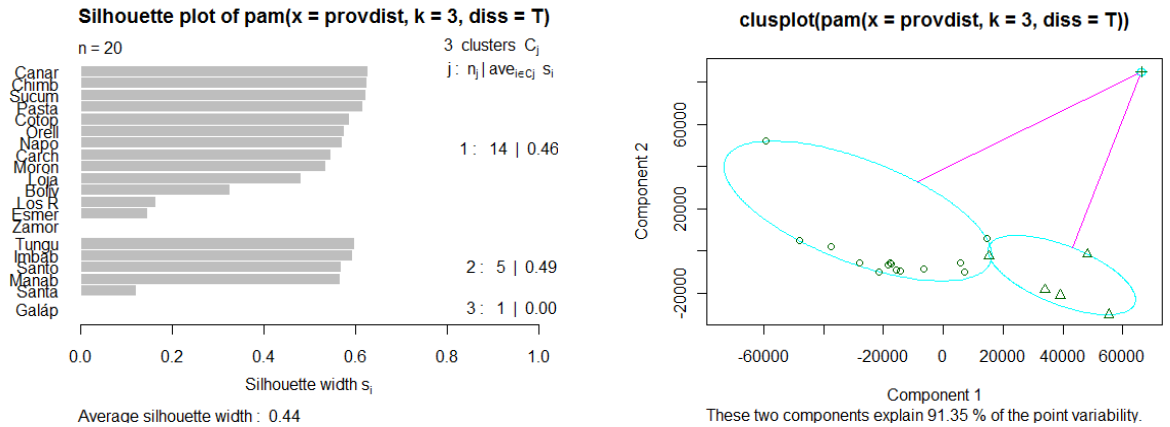


Ilustración 23: Gráfico silueta y gráfico “clusplot” para $sub.k = 3$; aplicado a nuestro set de datos para el análisis de caracterización empresarial excluyendo las observaciones (provincias) Pichincha, Azuay, Guayas y El Oro.

La configuración clúster de $sub.k = 4$ también presentó valores “silhouettes” positivos y completa separación de los sub-grupos en el sub-espacio de los dos primeros PC’s, sin embargo, mostraba a la provincia de Bolívar sóla en un clúster, algo que queremos evitar por razones ya explicadas anteriormente.

En consecuencia, el análisis clúster presente, que agrupa provincias en grupos con montos promedios similares entre las variables empresariales, nos dio como resultado final lo siguiente (ahora incluyendo a los “sub-clústers obtenidos):

Clúster 1	Clúster 2	Clúster 3

• Azuay
• Guayas
• El Oro
	• **Sub-clúster 2.1**
	  • Cañar
	  • Chimborazo
	  • Sucumbios
	  • Pastaza
	  • Cotopaxi
	  • Orellana
	  • Napo
	  • Carchi
	  • Morona Santiago
	  • Loja
	  • Bolívar
	  • Los Ríos
	  • Esmeraldas
	  • Zamora Chinchipe
	• **Sub-clúster 2.2**
	  • Tungurahua
	  • Imbabura
	  • Santo Domingo de los Tsáchilas
	  • Manabí
	  • Santa Elena
	• **Sub-clúster 2.3**
	  • Galápagos
		• Pichincha

Tabla 7: Solución final del Análisis Clúster para el estudio de caracterización empresarial de las provincias del Ecuador.

En la tabla 7 se detallan los grupos y las provincias que contienen cada uno de ellos. Las provincias que se encuentran en un mismo grupo poseen montos promedios cercanos por local entre las variables de caracterización empresarial consideradas.

A partir de los clústers formados según las variables de caracterización empresarial (Tabla 7), podemos también representar la distribución empresarial-geográfica del país según lo muestra la Ilustración 24:

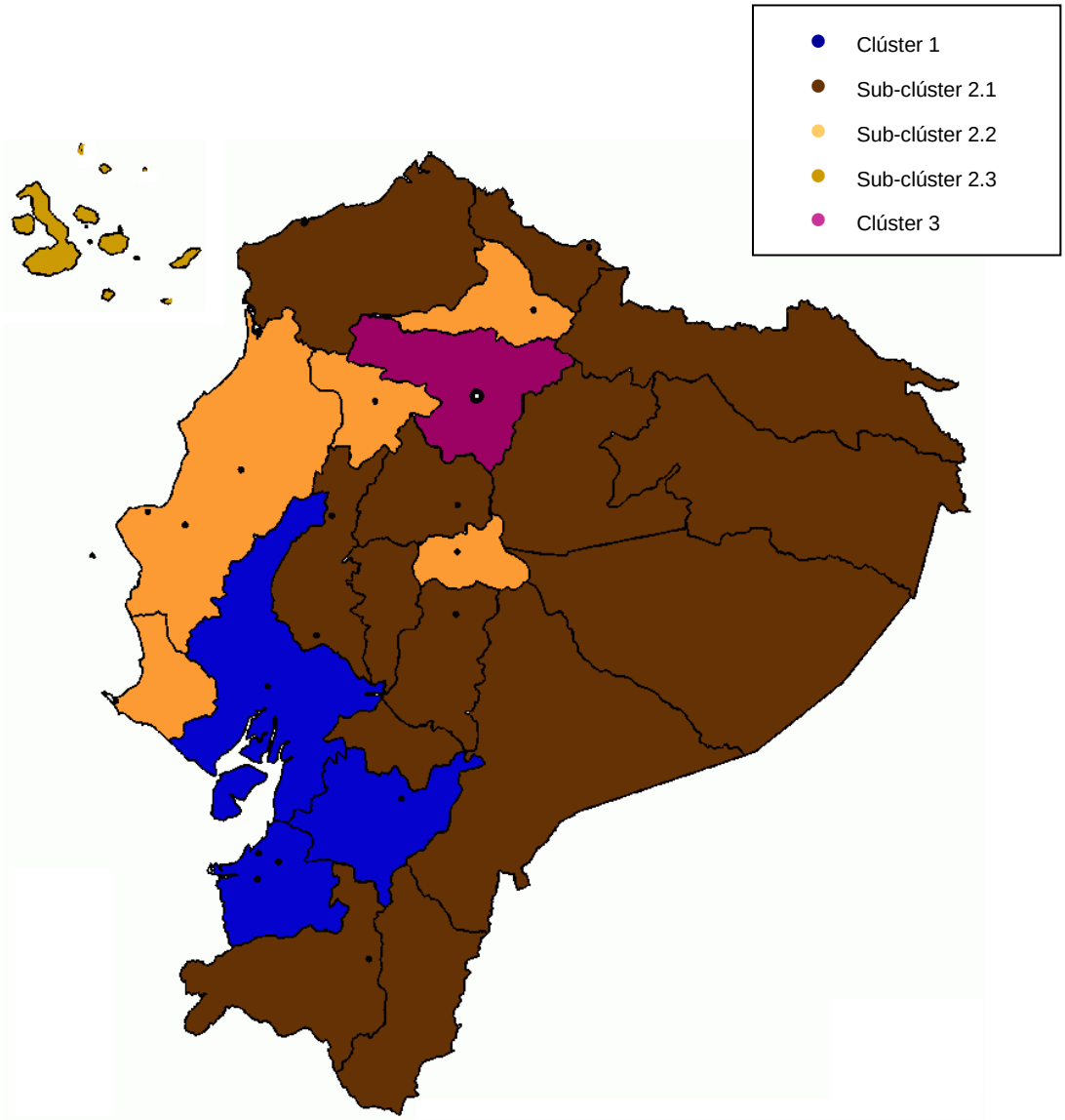


Ilustración 24: Distribución empresarial-geográfica del país a partir de los clústers formados según las variables de caracterización empresarial consideradas.

Por tanto, nuestra solución muestra una conclusión interesante: las provincias con similar importancia económica en el país poseen también políticas empresariales similares (no interpretamos una relación causal ni relación alguna entre estos dos aspectos, esclarecemos que manejamos estos argumentos con cuidado, ya que estas técnicas son de índole exploratorio).

9 Conclusiones y Recomendaciones

Nuestro análisis sobre la caracterización económica de las provincias del Ecuador tenía por objetivo brindar un espectro general sobre cómo las 24 provincias del Ecuador difieren con respecto a las actividades económicas desarrolladas en ellas. De este análisis, concluimos que: *primero*, en general las provincias situadas en la región Costa, Oriente e Insular poseen un mayor porcentaje de unidades económicas dedicadas a la actividad del Comercio o de Intercambio en comparación con las provincias situadas en la región Sierra; *segundo*, que las provincias situadas en la región Sierra poseen un mayor porcentaje de unidades económicas dedicadas a actividades industriales/manufactureras ("Fabricación de productos textiles, de vestir, cueros", "Fabricación de productos de caucho, plástico, minerales, metales") en comparación con las provincias situadas en las demás regiones; y *tercero*, que las provincias del Napo, Pastaza y Galápagos poseen un mayor porcentaje de unidades económicas dedicadas a "Actividades de Hostelería" en comparación con las demás provincias del país.

La siguiente técnica que fue usada para el análisis de caracterización económica fue la del Análisis clúster. En este análisis, se concluyó que existen tres grupos de provincias en nuestro set de datos censales, cada uno con patrones de porcentajes similares en las actividades económicas definidas. Es decir, en nuestros datos existen 3 grupos de provincias, y en cada uno se agrupan provincias cuyas unidades económicas se dedican a similares actividades económicas. Los grupos resultantes fueron:

Clúster 1	Clúster 2	Clúster 3
• Chimborazo • Pichincha • Pastaza • Morona Santiago • Azuay • Zamora Chinchipe • Cotopaxi • Napo • Tungurahua • Imbabura	• Guayas • El Oro • Santo Domingo • Santa Elena • Carchi • Bolívar • Manabí • Loja • Los Ríos • Esmeraldas • Cañar • Sucumbíos • Orellana	• Galápagos

Tabla 8: Réplica de la Tabla 5 que muestra la Solución final ($k = 3$ grupos) del Análisis Clúster para el estudio de caracterización económica de las provincias del Ecuador

En el clúster $_1$, las provincias conforman la mayoría la región Sierra, en el clúster 2 las provincias conforman la mayoría la región Costa y el clúster 3 lo conforma la única provincia de la región Insular. Las provincias del la región Oriente se encuentran distribuidas entre el clúster 1 y 2. Esta solución es consistente con la del análisis PCA. Esto podría indicar que las actividades económicas desarrolladas en el Ecuador difieren según las regiones naturales del país. Esta información podría ser útil para los inversionistas quienes desean saber en qué provincias podrían invertir según el área de especialización de las mismas, así como para los agentes tomadores de decisión quienes podrán destinar recursos para la especialización de cada provincia en cada una de las actividades económicas primarias identificadas, invertir en el desarrollo de industrias afines en cada grupo de provincias, entre otras políticas relacionadas.

En nuestro análisis de caracterización empresarial de las provincias del Ecuador, usamos PCA por un lado para conocer de manera general cómo las provincias se diferencian con respecto a los montos promedio de las variables de política empresarial consideradas por local. La interpretación del "biplot" generado no fue muy clara. Sin embargo, podríamos nombrar un par de hallazgos: *primero*, la provincia de Pichincha y la provincia del Guayas poseen un mayor ingreso promedio por local, un mayor gasto anual en compras y mercadería promedio por local, un mayor gasto anual en materia prima promedio por local, en comparación con las demás provincias; y *segundo*, las provincias de Galápagos y Bolívar son aquellas que requieren un mayor financiamiento en promedio por local en comparación con las demás provincias.

Clúster 1	Clúster 2	Clúster 3

• Azuay
• Guayas
• El Oro
	• **Sub-clúster 2.1**
	  • Cañar
	  • Chimborazo
	  • Sucumbios
	  • Pastaza
	  • Cotopaxi
	  • Orellana
	  • Napo
	  • Carchi
	  • Morona Santiago
	  • Loja
	  • Bolívar
	  • Los Ríos
	  • Esmeraldas
	  • Zamora Chinchipe
	• **Sub-clúster 2.2**
	  • Tungurahua
	  • Imbabura
	  • Santo Domingo de los Tsáchilas
	  • Manabí
	  • Santa Elena
	• **Sub-clúster 2.3**
	  • Galápagos
		• Pichincha

Tabla 9: Réplica de la Tabla 7 que muestra la Solución final del Análisis Clúster para el estudio de caracterización empresarial de las provincias del Ecuador.

Por otro lado, usamos también el Análisis Clúster para agrupar provincias que tengan patrones de políticas empresariales similares, o en otras palabras, para agrupar provincias en grupos que tengan montos promedios similares entre las variables escogidas que caracterizan a las políticas empresariales. Obtuvimos entonces los siguientes “clústers” y “sub-clústers”:

Este Análisis Clúster agrupa en el $clúster_3$ a la provincia más importante económicamente hablando del Ecuador, Pichincha. En el $clúster_1$ se agrupan a las provincias de mayor importancia económica del país luego de Pichincha: Azuay, Guayas y El Oro; y en el $clúster_2$ las restantes. Nuestra solución muestra una conclusión interesante: las provincias con similar importancia económica en el país poseen también políticas empresariales similares (no interpretamos una relación causal ni relación alguna entre estos dos aspectos, esclarecemos que manejamos estos argumentos con cuidado, ya que estas técnicas son de índole exploratorio)

Podemos también concluir que Guayas y Pichincha poseen políticas empresariales de mucha inversión y gasto, aunque también de muchos ingresos. Además, que las políticas empresariales de Pichincha difieren a aquellas de otras provincias desarrolladas del país como Guayas, Azuay y El Oro, y que estas tres últimas poseen políticas empresariales similares. El resumen arriba nos muestra también “sub-clústers” conteniendo provincias que poseen políticas empresariales similares.

Estas conclusiones podrían ser útiles para los agentes tomadores de decisión macroeconómicas quienes desearían conocer esta información, así como para inversionistas quienes desearían saber sobre provincias con políticas empresariales comunes para así hacer su decisión de inversión o saber en qué grupo de provincias con políticas empresariales afines se podría operar.

10 Referencias

(1) Abramowitz, M., and I. A. Stegun; eds. *Handbook of Mathematical Functions*. U.S. Department of Commerce, National Bureau of Standards Applied Mathematical Series

(2) Adriaans, P., and D. Zantinge. *Data Mining*. Harlow, England: Addison-Wesley, 1996

(3) Anderberg, M. R. *Cluster Analysis for Applications*. New York: Academic Press, 1973

(4) Berry, M. J. A., and G. Linoff. *Data Mining Techniques: For marketing, Sales and Customer Relationship Management* (2nd ed.) (paperback). New York: John Wiley, 2004.

(5) Everitt, B. S., S. Landau and M. Leese. *Cluster Analysis* (4th ed.). London: Hodder Arnold, 2001

(6) Fraley, C., and A. E. Raftery. "Model-Based Clustering, Discriminant Analysis and Density Estimation". *Journal of the American Statistical Association* (2002)

(7) Gower, J. C., and D. J. Hand. *Biplots*. London: Chapman and Hall, 1996

(8) Hartigan, J. A. *Clustering Algorithms*. New York: John Wiley, 1975

(9) Hastie T., Tibshirani R. and Friedman J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Berlin, Germany: Springer – Verlag, 2001.

10) INEC, Base de datos del Censo Nacional Económico 2010: http://www.inec.gob.ec/cenec/index.php?option=com_content&view=article&id=231&Itemid=125&lang=es

11) Johnson R. and Wichern D. (1998). *Applied Multivariate Statistical Analysis*. Prentice Hall, 816pp.

12) Kaufman L. and Rousseeuw P. (1990). *Finding groups in data. An introduction to cluster analysis*. Wiley.

13) Mardia, K. V., J. T. Kent, and J. M. Bibby. *Multivariate Analysis* (Paperback). Londonde: Academic Press, 2003

14) Ramsay J. and Silverman B. (2002). *Applied Functional Data-analysis*. Springer-Verlag.

15) Thas O. (2011) *Course Notes Multivariate Data Analysis* - Department of Applied Mathematics, Biometrics and Process Control, Ghent University

11 Anexos

Códigos del software estadístico R

```
################################
setwd("C:/Users/holger/Documents/Maestría Economía ESPOL Tesis")
censo=read.table("CENEC_2010.dat",sep="\t",header=TRUE,dec=',')
censo[1:10,]
names(censo)
dim(censo)
censo[50000,1]


#############################################################
# Análisis según las actividades económicas
#############################################################

# Análisis de Componentes Principales
# Ahora, usando la clasificación: Actividad Principal a dos Dígitos CIIU (68)
# Poniendo ciertas categorías en grupos

# Agricultura, ganadería, caza y actividades de servicios cone
# Silvicultura y extracción de madera/Producción de madera y fabricación de productos de madera
# Pesca y acuicultura
# Minería
# Elaboración de productos alimnticios y bebidas
# Fabricación de prod textiles, de vestir, cueros
# Fabricación de prod, sustancias químicas, farmacéuticos
# Fabricacion de prod plastico, caucho, metales, no metales, elaborados de metal, muebles
# Fabricación de productos electronicos, maqinarias y equipos
# Fabricacion vehiculos, otros tipos de transporte
# Suministro de energía
# Construcción
# Comercio
# Transporte y almacenamiento
# Actividades de alojamiento y Serv de alimento y bebida.
# Información y comunicación
# Actividades financieras y de seguros
# Actividades profesionales, científicas y técnicas
```

```
# Actividades de atención de la salud humana y de asistencia social
# Artes, entretenimiento y recreación


prov_act=censo[,c(1,68)]
prov_act[1:20,]
is.data.frame(prov_act)
dim(prov_act)
prov_act<-prov_act[prov_act$CIIU2.P!=" ",]

dim(prov_act)

names(prov_act)[1]<-""
names(prov_act)[2]<-""
prov_act=table(prov_act)
prov_act=as.data.frame.matrix(prov_act)
prov_act=prov_act[,c(-1)]



#Dividir las variables en grupos de categorías

Agr=prov_act[,1]
Silv.madera=rowSums(prov_act[,c(2,15)])
Pesca.ac=prov_act[,3]
Mineria=rowSums(prov_act[,4:8])
alim.beb=rowSums(prov_act[,9:10])
text.vest=rowSums(prov_act[,12:14])
quim=rowSums(prov_act[,19:20])
plast.cauch.met=rowSums(prov_act[,c(21,22,23,24,30)])
elct.eq.maq=rowSums(prov_act[,c(25,26,27)])
veh=rowSums(prov_act[,c(28,29)])
sumEner=prov_act[,33]
construc=rowSums(prov_act[,38:40])
comercio=rowSums(prov_act[,41:43])
transp=rowSums(prov_act[,44:48])
serv.aloj.com=rowSums(prov_act[,49:50])
inf.comunic=rowSums(prov_act[,53:56])
finanz=rowSums(prov_act[,57:59])
act.prof.cient=rowSums(prov_act[,64:67])
asist.social=rowSums(prov_act[,76:78])
art.entr.recr=rowSums(prov_act[,79:82])
```

```
prov_activ=cbind(Agr,Silv.madera,Pesca.ac,Mineria,alim.beb,text.vest,quim,plast.cauch.met,

elct.eq.maq,veh,sumEner,construc,comercio,transp,serv.aloj.com,inf.comunic,
            finanz,act.prof.cient,asist.social,art.entr.recr)
prov_activ=as.data.frame(prov_activ)
dim(prov_activ)
prov_activ_rel=(prov_activ/rowSums(prov_activ))*100
prov_activ_rel=prov_activ_rel[-25,]
prov_activ=prov_activ[-25,] # Sin 'zonas delimitadas'


provincia=c("Azuay","Bolívar","Canar", "Carchi","Cotopaxi","Chimborazo","EL Oro","Esmeraldas","Guayas","Imbabura","Loja",
        "Los Ríos","Manabí","Morona Santiago","Napo","Pastaza","Pichincha","Tungurahua","Zamora Chinchipe","Galápagos",
        "Sucumbios","Orellana","Santo Domingo", "Santa Elena")

cor(prov_activ_rel)

# Análisis Componentes principales, usando Matriz covarianza
#Usando %
prov.pca=princomp(prov_activ_rel,cor=FALSE)
prov.pca
prov.pca$loadings
print(prov.pca$loadings,cutoff=0)
summary(prov.pca)

screeplot(prov.pca)
prov.pca$scores

biplot(prov.pca)

# Análisis Componentes principales, usando Matriz correlación (variables estandarizadas)
prov.pca=princomp(prov_activ_rel,cor=TRUE)
prov.pca
prov.pca$loadings
print(prov.pca$loadings,cutoff=0)
summary(prov.pca)

screeplot(prov.pca)
```

```
prov.pca$scores

par(mar=c(5, 4, 4, 2) + 0.1)
biplot(prov.pca)

#Análisis Cluster
dim(prov_activ_rel)
rownames(prov_activ_rel) <- provincia

provdist <- dist(prov_activ_rel,"euclidean")
library(cluster )

par(mfrow=c(1,1))
prov.agnes.sin <- agnes(provdist, method="single")
plot(prov.agnes.sin,which.plots=2)
prov.agnes.ave <- agnes(provdist, method="average")
plot(prov.agnes.ave,which.plots=2)
prov.agnes.com <- agnes(provdist, method="complete")
plot(prov.agnes.com,which.plots=2)
prov.agnes.war <- agnes(provdist, method="ward")
plot(prov.agnes.war,which.plots=2)

prov.pam<-pam(provdist,7,diss=T)
plot(prov.pam)
clusplot(prov.pam)

prov.pam<-pam(provdist,6,diss=T)
plot(prov.pam)
clusplot(prov.pam)

prov.pam<-pam(provdist,5,diss=T)
plot(prov.pam)
clusplot(prov.pam)

prov.pam<-pam(provdist,4,diss=T)
plot(prov.pam)
clusplot(prov.pam)


prov.pam<-pam(provdist,3,diss=T)
plot(prov.pam)
clusplot(prov.pam)
```

```
prov.pam<-pam(provdist,2,diss=T)
plot(prov.pam)
clusplot(prov.pam)

#Wilki Lambda
prov.k.scree <- as.data.frame(matrix(ncol=2,nrow=7))
names(prov.k.scree) <- c("k","Lambda")
prov.k.scree$k <- 2:8



formulahelp<-names(prov_activ_rel)[1]
for (i in 1:19) {
formulahelp<-paste(formulahelp,names(prov_activ_rel)[i+1],sep=",")
}

formulahelp<-paste("cbind(",formulahelp,sep="")
formulahelp<-paste(formulahelp,")~clust",sep="")
formulahelp<-as.formula(formulahelp)

for(k in 2:8) {
prov.pam <- pam(provdist,k,diss=T)
prov_activ_rel$clust <- as.factor(prov.pam$clust)
tmp <- manova(formulahelp,data=prov_activ_rel)
prov.k.scree$Lambda[k-1] <- summary(tmp,test="Wilks")$stats[3]
}

plot(prov.k.scree$k,prov.k.scree$Lambda,type="b",xlab="k",ylab="Lambda")

############################################################

############################################################
# Análisis según aspectos económicos en las empresas
############################################################

prov_act2=censo[,c(1,49,50,55,58,60,62,95,96,98,99,103,104,108)]
prov_act2[1:10,]
prov_act2[40000:40008,]
dim(prov_act2)
table(prov_act2[,1])
is.data.frame(prov_act2)

prov_act3=rowsum(data.matrix(prov_act2[,-1]),data.matrix(prov_act2[,1]),na.rm=TRUE)
```

```
prov_act3=as.data.frame(prov_act3)
prov_act3=prov_act3[-25,]

#provincia=c("Azuay","Bolívar","Canar", "Carchi","Cotopaxi","Chimborazo","EL Oro","Esmeraldas","Guayas","Imbabura","Loja",
 # "Los Ríos","Manabí","Morona Santiago","Napo","Pastaza","Pichincha","Tungurahua","Zamora Chinchipe","Galápagos",
  # "Sucumbios","Orellana","Santo Domingo", "Santa Elena", "Zonas No delimitadas")

#prov_act3=cbind(prov_act3,provincia)
#rownames(prov_act3)<-prov_act3[,10]
#prov_act3=prov_act3[,-10]


names(prov_act3)[1]<-"FinanPriv"
names(prov_act3)[2]<-"FinanGob"
names(prov_act3)[3]<-"Finanreq"
names(prov_act3)[4]<-"Manedesech"
names(prov_act3)[5]<-"GasInvDesar"
names(prov_act3)[6]<-"GasCapacForm"


prov_act3
dim(prov_act3)

# prov_act3_rel=(prov_act3/rowSums(prov_act3))*100


dim(prov_act3)

# Usando Matriz de correlación
prov.pca=princomp(prov_act3,cor=T)
prov.pca
prov.pca$loadings
print(prov.pca$loadings,cutoff=0)
summary(prov.pca)

screeplot(prov.pca)
prov.pca$scores

biplot(prov.pca)
```

#Dividido número encuestados en cada provincia: Monto promedio por local en la respectiva provincia

```
prov_act4=prov_act3/table(prov_act2[,1])[-25]

dim(prov_act4)

prov.pca=princomp(prov_act4,cor=FALSE)
prov.pca
prov.pca$loadings
print(prov.pca$loadings,cutoff=0)
summary(prov.pca)

screeplot(prov.pca)
prov.pca$scores

biplot(prov.pca)
```

##

#Sin Guayas, Pichincha, Azuay, El Oro

```
prov_act5=prov_act3[c(-1,-7,-9,-17), ]/table(prov_act2[,1])[c(-1,-7,-9,-17,-25)]

dim(prov_act5)

prov.pca=princomp(prov_act5,cor=T)
prov.pca
prov.pca$loadings
print(prov.pca$loadings,cutoff=0)
summary(prov.pca)

screeplot(prov.pca)
prov.pca$scores

biplot(prov.pca)
```

```
#Análisis Cluster


provincia=c("Azuay","Bolívar","Canar",  "Carchi","Cotopaxi","Chimborazo","EL
Oro","Esmeraldas","Guayas","Imbabura","Loja",
        "Los  Ríos","Manabí","Morona
Santiago","Napo","Pastaza","Pichincha","Tungurahua","Zamora
Chinchipe","Galápagos",
       "Sucumbios","Orellana","Santo Domingo", "Santa Elena")

prov_act4=cbind(prov_act4,provincia)
rownames(prov_act4)<-prov_act4[,14]
prov_act4=prov_act4[,-14]

provdist <- dist(prov_act4,"euclidean")
library(cluster )

par(mfrow=c(1,1))
prov.agnes.sin <- agnes(provdist, method="single")
plot(prov.agnes.sin,which.plots=2)
prov.agnes.ave <- agnes(provdist, method="average")
plot(prov.agnes.ave,which.plots=2)
prov.agnes.com <- agnes(provdist, method="complete")
plot(prov.agnes.com,which.plots=2)
prov.agnes.war <- agnes(provdist, method="ward")
plot(prov.agnes.war,which.plots=2)


prov.pam<-pam(provdist,2,diss=T)
plot(prov.pam)
clusplot(prov.pam)

prov.pam<-pam(provdist,3,diss=T)
plot(prov.pam)
clusplot(prov.pam)


prov.pam<-pam(provdist,4,diss=T)
plot(prov.pam)
clusplot(prov.pam)

prov.pam<-pam(provdist,5,diss=T)
plot(prov.pam)
```

```
clusplot(prov.pam)

prov.pam<-pam(provdist,6,diss=T)
plot(prov.pam)
clusplot(prov.pam)

#Wilki Lambda
prov.k.scree <- as.data.frame(matrix(ncol=2,nrow=5))
names(prov.k.scree) <- c("k","Lambda")
prov.k.scree$k <- 2:6



formulahelp<-names(prov_act4)[1]
for (i in 1:8) {
formulahelp<-paste(formulahelp,names(prov_act4)[i+1],sep=",")
}

formulahelp<-paste("cbind(",formulahelp,sep="")
formulahelp<-paste(formulahelp,")~clust",sep="")
formulahelp<-as.formula(formulahelp)

for(k in 2:6) {
prov.pam <- pam(provdist,k,diss=T)
prov_act4$clust <- as.factor(prov.pam$clust)
tmp <- manova(formulahelp,data=prov_act4)
prov.k.scree$Lambda[k-1] <- summary(tmp,test="Wilks")$stats[3]
}

plot(prov.k.scree$k,prov.k.scree$Lambda,type="b",xlab="k",ylab="Lambda")



#Análisis Cluster : SIN GUAYAS, PICHINCHA, AZUAY, EL ORO


provincia=c("Bolívar","Canar",
"Carchi","Cotopaxi","Chimborazo","Esmeraldas","Imbabura","Loja",
        "Los                                                    Ríos","Manabí","Morona
Santiago","Napo","Pastaza","Tungurahua","Zamora Chinchipe","Galápagos",
       "Sucumbios","Orellana","Santo Domingo", "Santa Elena")

prov_act5=cbind(prov_act5,provincia)
```

```
rownames(prov_act5)<-prov_act5[,14]
prov_act5=prov_act5[,-14]

provdist <- dist(prov_act5,"euclidean")
library(cluster )

par(mfrow=c(1,1))
prov.agnes.sin <- agnes(provdist, method="single")
plot(prov.agnes.sin,which.plots=2)
prov.agnes.ave <- agnes(provdist, method="average")
plot(prov.agnes.ave,which.plots=2)
prov.agnes.com <- agnes(provdist, method="complete")
plot(prov.agnes.com,which.plots=2)
prov.agnes.war <- agnes(provdist, method="ward")
plot(prov.agnes.war,which.plots=2)


prov.pam<-pam(provdist,2,diss=T)
plot(prov.pam)
clusplot(prov.pam)

prov.pam<-pam(provdist,3,diss=T)
plot(prov.pam)
clusplot(prov.pam)


prov.pam<-pam(provdist,4,diss=T)
plot(prov.pam)
clusplot(prov.pam)

prov.pam<-pam(provdist,5,diss=T)
plot(prov.pam)
clusplot(prov.pam)


#Wilki Lambda
prov.k.scree <- as.data.frame(matrix(ncol=2,nrow=5))
names(prov.k.scree) <- c("k","Lambda")
prov.k.scree$k <- 2:6



formulahelp<-names(prov_act5)[1]
```

```
for (i in 1:8) {
formulahelp<-paste(formulahelp,names(prov_act5)[i+1],sep="+")
}

formulahelp<-paste("cbind(",formulahelp,sep="")
formulahelp<-paste(formulahelp,")~clust",sep="")
formulahelp<-as.formula(formulahelp)

for(k in 2:6) {
prov.pam <- pam(provdist,k,diss=T)
prov_act5$clust <- as.factor(prov.pam$clust)
tmp <- manova(formulahelp,data=prov_act5)
prov.k.scree$Lambda[k-1] <- summary(tmp,test="Wilks")$stats[3]
}

plot(prov.k.scree$k,prov.k.scree$Lambda,type="b",xlab="k",ylab="Lambda")
```