Trabajo Final Análisis Multivariado
Análisis Casen

Diego Aravena
Alonso Campos

Profesor: Manuel Galea
Ayudante Emiliano Moreno

Resumen

En este estudio se propone segmentar la provincia de Santiago según características geográficas, el cual considerará los sectores poniente y oriente. Posterior a ello, mediante un análsis de componentes principales se determinará si las variables que fueron seleccionadas logran relacionar los aspectos socio-económicos y demográficos para poder clasificar correctamente los sectores propuestos, y así precisar el fenómeno junto con el respectivo respaldo estadístico. Para ello se disponen datos extraídos de la La Encuesta de Caracterización Socio-económica Nacional (CASEN) del año 2017.

El documento contiene 4 secciones. Primero una breve introducción sobre el tema y la hipótesis de interés. En la segunda, denominada Datos, se detallará la encuesta CASEN, principalmente los objetivos y, además se especificará la naturaleza de los datos utilizados y las variables seleccionadas. En la tercera sección, se explica la metodología e implementación de las componentes principales en el contexto del problema. Posteriormente, se presentarán los resultados y principales descubrimientos obtenidos a partir del análisis. Finalmente, se realizará una conclusión con una síntesis de lo estudiado en el informe y si se comprueba la hipótesis anteriormente planteada.

Introducción

La Región Metropolitana de Chile alberga a más de 7 millones de habitantes [1], siendo la región más poblada en todo el país. Además ésta se divide en 6 provincias, siendo la provincia de Santiago la que contiene el mayor número de comunas con un total de 32. Es bien sabido que las comunas de esta provincia, a pesar de estar cercanas unas con otras, pueden llegar a diferir fuertemente en cuanto a características socio-económicas, demográficas y geográficas.

Dada la relevancia de la región y, en particular la provincia de Santiago, para las autoridades del sector es importante poder caracterizar dichos comunas en función de pocos datos para así lograr distinguir ciertos elementos comunes y luego catalogarlas. Uno de las principales motivaciones suele ser, por ejemplo, para aplicar políticas públicas a determinados sectores.

Es por esto que, gracias a los datos recolectados por la encuesta Casen se propone la hipótesis de que es posible caracterizar a las comunas de la provincia de Santiago a través de la información del costo habitacional, los ingresos y longevidad de la población en cuestión.

Datos

Como se mencionó anteriormente los datos que se utilizaron pertenecen a la encuesta CASEN realizada el año 2017. A continuación se realizará un breve resumen de dicha encuesta.

Encuesta Casen

La Encuesta de Caracterización Socioeconómica Nacional es realizada por el Ministerio de Desarrollo Social y Familia) y tiene un carácter transversal y multipropósito. Además, se realiza de manera regular desde el año 1987.

La encuesta posee cinco objetivos, sin embargo, para los propósitos del informe, solo se centrará en el primero, es decir, "Conocer periódicamente la situación socio-económica de los hogares y de la población que reside en viviendas particulares, en aspectos como: composición de hogares y familias, educación, salud, vivienda, trabajo, e ingresos". Pues con este objetivo, permite obtener los datos de los habitantes de cada municipio y así obtener estadísticas de resumen por comuna, en categorías socio-económicas y demográficas.

El método de recolección de datos se realiza a partir de muestreo estratificado y multietápico donde las unidades de interés son los hogares a lo largo de todo el país, es decir, residentes de viviendas particulares, es importante mencionar que se excluyen los sectores más aislados. La entrevista se realiza a una persona, mayor de 18 años, dentro de hogar

[1] Información extraída de Censo 2017

y ésta debe responder por todos los miembros.

El cuestionario de la Encuesta CASEN para el año 2017 cuenta con 7 módulos temáticos, los que se considera: Registro de residentes, Educación, Trabajo, Ingresos, Salud, Identidades Redes y participación y, por último, Vivienda y entorno. Además, en dicha versión de la encuesta, se logró un tamaño de muestra de 70.948 hogares los que se implica 216.439 personas.

Procesamiento de los datos

La base de datos extraída desde el sitio web de la Encuesta CASEN contiene un total de 216.439 observaciones que corresponden al número de personas registradas en la encuesta pertenecientes a los 70.948 hogares, con un total de 804 variables que relacionadas con los 7 módulos temáticos, dichas variables corresponden a las repuestas registradas para cada individuo. En este informe se utilizaron los ejes de Ingresos, Registro de residentes y de Vienvienda y entorno.

La Tabla 1 presenta las variables que fueron seleccionadas desde la base de datos original junto con su respectiva descripción. Posteriormente se seleccionan todas aquellas viviendas que fueron entrevistadas en la provincia de Santiago.

Variable	Descripción	Unidad de medida
`v17`	Valor dividendo	Peso chileno
`v19`	Valor arriendo del sector	Peso chileno
`edad`	Edad del entrevistado	Años
`y1`	Salario líquido del trabajo principal el mes pasado	Peso chileno
`yaut`	Ingreso autónomo	Peso chileno
`ytot`	Ingreso total	Peso chileno
`ytoth`	Ingreso total del hogar	Peso chileno
`ypch`	Ingreso total per cápita del hogar	Peso chileno
`comuna`	Comuna de la vivienda	-
`provincia`	Provincia de la vivienda	-

Tabla 1: Variables seleccionadas de la base de datos

Variable	Min	Med	Max	NA's
`v17`	0	42000	5000000	22531
`v19`	0	280000	8000000	0
`y1`	0	400000	34000000	20189
`yaut`	83	401667	72691664	13241
`ytot`	83	370000	72691664	10689
`ytoth`	0	1113167	79193335	0
`ypch`	0	306957	30566664	0

Tabla 2: Breve resumen de las variables relacionadas a ingresos y gastos

La Tabla 2 presenta información sobre las variables que fueron medidas en pesos chilenos. Tal como se puede obsevar el rango en todas las variables es demasiado extenso, por lo tanto, se transformarán a escala logarítmica con el fin de reducir el impacto que puedan tener los datos extremos. Por otro lado, se debe recordar que del objetivo de este problema subyace considerar entrevistados que contribuyan con la información necesaria, es decir, se deben descartar los valores que no aporten información tales como: nulos, ceros y 99's (que es cuando el entrevistado no sabe).

De modo siguiente se agrupan todas las variables, incluyendo a `edad`, y luego se obtienen los promedios por comuna. De este modo la matriz de datos queda dimensión 32×8, donde cada fila corresponde a cada comuna de la provincia de Santiago y cada columna al promedio de cada variable en escala logarítmica. Una primera impresión de esta matriz está dada en la Tabla 3.

Comuna	y1	v17	v19	edad	yaut	ytot	ytoth	ypch
Santiago	13.17	12.55	12.56	35.23	13.14	13.03	13.88	12.87
Cerrillos	12.75	12.21	12.20	36.61	12.50	12.06	13.60	12.30
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

Tabla 3: Primeras dos filas de la matriz de datos

La separación de las comunas por sector está dada por la tabla 4 y para tener una mejor referencia se presenta la figura 1 que puede ayudar a entender mejor esta separación.

Sector	Comuna
Oriente	La Florida, La Reina, Las Condes, Lo Barnechea, Macul, Ñuñoa, Peñalolén, Providencia, Vitacura
Poniente	Santiago, Cerrillos, Cerro Navia, Conchalí, El Bosque, Estación Central, Huechuraba, Independencia, La Cisterna, La Granja, La Pintana, Lo Espejo, Lo Prado, Maipú, Pedro Aguirre Cerda, Pudahuel, Quilicura, Quinta Normal, Recoleta, Renca, San Joaquín, San Miguel, San Ramón

Tabla 4: Distribución de comunas según sector

Figura 1: Mapa comunal de Santiago
Fuente: Wikipedia

Metodología

Antes de trabajar con los datos se revisa el supuesto de normalidad con el objetivo de obtener un respaldo de las conclusiones que se puedan obtener. Para esto se quiere construir un gráfico QQ-Plot obteniendo los cuantiles empíricos mediante la distancia de Mahalanobis

$$\delta_i = (\boldsymbol{X_i} - \bar{\boldsymbol{X}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{X_i} - \bar{\boldsymbol{X}_i}),\ i = 1, ..., n \tag{1}$$

se considerará a la matriz $\boldsymbol{X}$ como la matriz de datos, donde X_i corresponde a la i-ésima fila y $\bar{\boldsymbol{X}}$ corresponde al vector de medias de las columnas de la matriz $\boldsymbol{X}$. Por otro lado, Σ es la matriz de covarianzas de X. Luego se obtienen los cuantiles empíricos de los datos mediante la transformación

$$z_i = \frac{(\hat{\delta}_i/p)^{1/3} - (1 - 2/9p)}{\sqrt{2/9p}} \tag{2}$$

el cual se compara con los cuantiles teóricos de una distribución normal estándar de dimensión 8.

En la Figura 2 se presentan los QQ-Plots obtenidos para la distancia de Mahalanobis, que distribuye aproximadamente chi-cuadrado con 8 grados de libertad, y para la distancia transformada, que posee una distribución aproximada normal estándar. Se puede observar que la gran mayoría de los puntos están bien alieneados con respecto a la recta, sin embargo, en el extremo superior de ambos gráficos se puede notar un claro alejamiento, lo cual se puede estar debiendo a datos extremos que podrían estar influyendo en el supuesto. Para mayor tranquilidad se realizó un test de *Mardia*, cuya hipótesis plantea que los datos entregados distribuyen normal multivariado. La tabla 5 presenta los resultados entregados por el test, del cual se puede concluir que el supuesto de normalidad puede ser válido, es decir, no se rechaza la hipótesis para un nivel de significancia del 5 %.

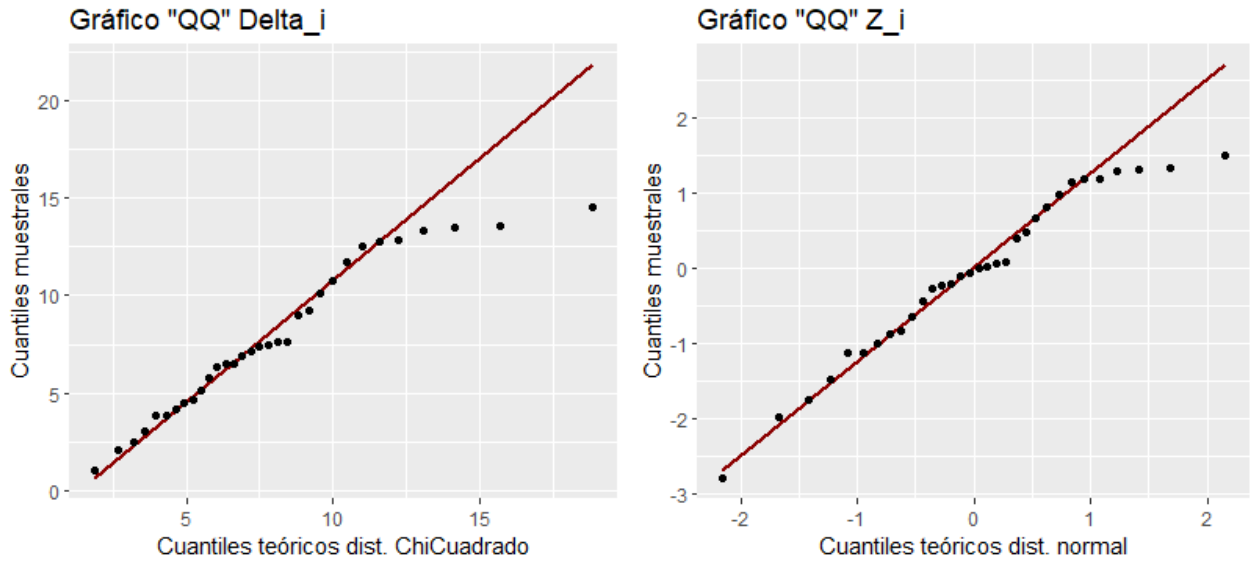

Figura 2: Gráficos QQ-Plot de las distancias con y sin transformar

Test	Estadístico	valor-p
Asímetria de Mardia	137.73	0.128
Kurtosis de Mardia	-0.172	0.863

Tabla 5: Resultados test de Mardia normal multivariada

Análisis de componentes principales

Una vez validado el supuesto de normalidad, se procede con el análisis de componentes principales.

El objetivo del informe es poder clasificar las comunas según sector de la provincia, en base a un menor número de variables de las que fueron presentadas. Considerando lo anterior y dado que los datos disponibles presentan alta correlación (ver figura 3) es que se propone un análisis de componentes principales.

Las ventajas de utilizar este método son: en primer lugar es posible reducir las cantidad de variables, pues uno de los objetivos de este método es la reducción de dimensionalidad. Además, como los datos presentan alta correlación entre ellos, aplicar esta metodología se obtendrán nuevas variables ortogonales entre sí, por lo que se resuelve el problema de datos altamente correlacionados. También, al agrupar las variables mediante combinaciones lineales, en ocasiones, es posible clasificar observaciones, pues dichas combinaciones funcionan como medida de resumen de las variables, así en el caso bivariado, al graficar dos componentes principales se pueden determinar si el método logra discriminar correctamente, en este caso, cada sector.

Un punto a considerar es que no todas las variables están en la misma unidad de medida, por lo que se descarta la utilización de la matriz de covarianzas de $\boldsymbol{X}$, lo que deja a disposición la matriz de correlación R. La figura 3 presenta esta última matriz, donde se puede observar rápidamente una alta tasa correlación entre todas las variables a excepción de `edad`.

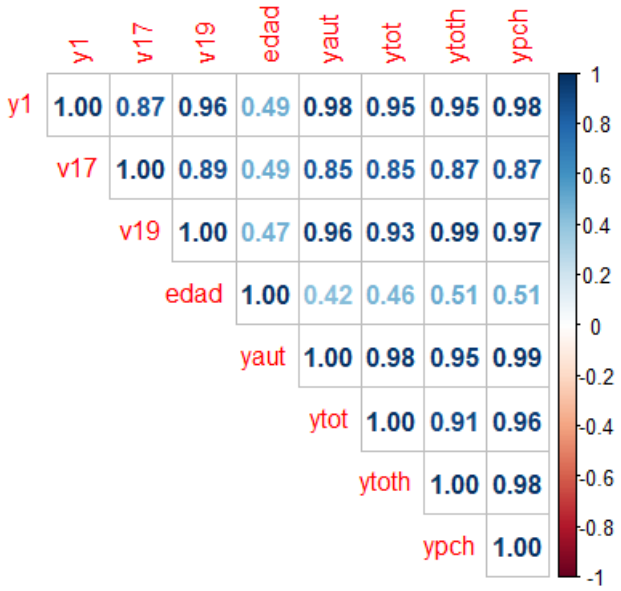

Figura 3: Correlación entre las variables de la matriz X

En primer lugar se obtienen los valores y vectores propios de la matriz de correlación, los cuales corresponden a los entregados en la ecuación (3) y (4) respectivamente.

$$\boldsymbol{l} = \begin{pmatrix} 6{,}88 & 0{,}74 & 0{,}20 & 0{,}11 & 0{,}04 & 0{,}02 & 0{,}005 & 0{,}003 \end{pmatrix}^T \tag{3}$$

$$\boldsymbol{G} = \begin{pmatrix} -0{,}374 & -0{,}075 & -0{,}139 & -0{,}060 & 0{,}821 & 0{,}320 & 0{,}230 & 0{,}050 \\ -0{,}347 & 0{,}009 & 0{,}914 & -0{,}172 & 0{,}036 & -0{,}107 & -0{,}023 & -0{,}041 \\ -0{,}374 & -0{,}095 & 0{,}021 & 0{,}402 & -0{,}298 & 0{,}616 & -0{,}377 & 0{,}281 \\ -0{,}210 & 0{,}969 & -0{,}097 & -0{,}036 & -0{,}014 & 0{,}020 & -0{,}064 & -0{,}040 \\ -0{,}373 & -0{,}171 & -0{,}243 & -0{,}180 & 0{,}055 & -0{,}227 & -0{,}618 & -0{,}551 \\ -0{,}367 & -0{,}109 & -0{,}208 & -0{,}646 & -0{,}448 & 0{,}201 & 0{,}389 & 0{,}020 \\ -0{,}373 & -0{,}045 & -0{,}052 & 0{,}593 & -0{,}178 & -0{,}202 & 0{,}512 & -0{,}413 \\ -0{,}377 & -0{,}061 & -0{,}176 & 0{,}061 & 0{,}014 & -0{,}611 & -0{,}069 & 0{,}664 \end{pmatrix} \tag{4}$$

Los vectores propios de la ecuación (4) cumplen las restricciones de ortogonalidad y cuya norma euclidena sea igual a 1. Además, como se está trabajando con la matriz de correlación se cumple lo siguiente

$$tr(R) = \sum_{i=2}^{8} l_i = 8$$

Para transformar la matriz X usando los componentes principales primero se debe estandarizar la matriz. Esta matriz se denota como $\widetilde{X}$ y la tabla 6 presenta las primeras 2 filas que resultaron luego de estandarizar las variables.

Comuna	y1	v17	v19	edad	yaut	ytot	ytoth	ypch
Santiago	0.087	0.109	0.020	-0.173	0.161	0.195	-0.008	0.102
Cerrillos	-0.117	0.001	-0.149	-0.052	-0.140	-0.156	-0.132	-0.121
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

Tabla 6: Primeras dos filas de la matriz de datos estandarizada

Luego la matriz de datos de los componentes principales se obtienen de la ecuación (5)

$$Z = \widetilde{X}G \tag{5}$$

La matriz Z cumple con la condición de que sus columnas sean ortogonales y, por lo tanto, se cumple que $cor(Z) = I_8$, donde I_8 corresponde a la matriz identidad de dimensión 8 (veáse figura 4).

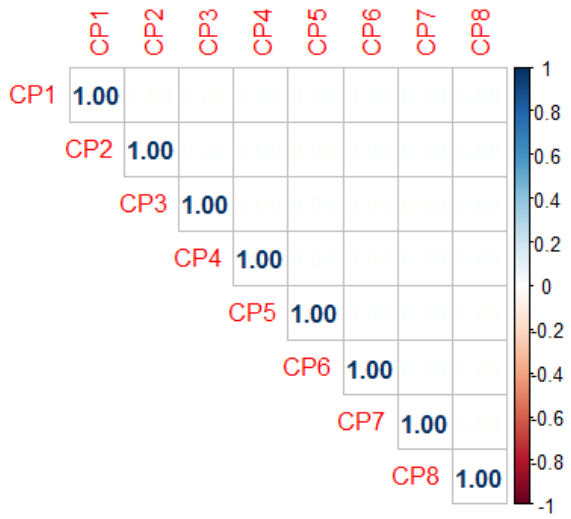

Figura 4: Matriz de correlación de componentes principales

Resultados

La tabla 7 presenta un resumen de la varianza explicada por cada componente principal correspondiente a los valores propios de la matriz R. Se puede ver que el primer componente representa un 86 % de la variabilidad de los datos, lo cual es bastante alto y representativo, y junto con el segundo componente principal ya logra captar un 95 % de la varianza.

Valor propio	1	2	3	4	5	6	7	8
Varianza	6.8842	0.7386	0.1992	0.1071	0.0384	0.0231	0.0059	0.0034
Prop. varianza	0.8605	0.0923	0.0249	0.0134	0.0048	0.0029	0.0007	0.0004
Var. acumulada	0.8605	0.9528	0.9777	0.9911	0.9959	0.9988	0.9996	1.0000

Tabla 7: Tabla resumen componentes principales

Dada la información entregada por la tabla 7, se vuelve atractivo querer considerar las primeras 2 CP. Para esto primero se debe probar si es posible descartar las últimas 6 CP, por lo que se busca rechazar la siguiente hipótesis

$$H_0 : f(l) = \frac{l_3 + l_4 + l_5 + l_6 + l_7 + l_8}{\sum_{i=1}^{8} l_i} > c_0$$

Considerando $c_0 = 0{,}05$, H_0 indica el supuesto de que las últimas 6 CP logran explicar más del 5 % de la variabilidad total. Para un nivel de significancia α del 5 %, H_0 se rechaza si

$$\sqrt{n-1}(f(l) - c_0) < z_\alpha \sqrt{Var(f(l))}$$

donde se obtiene que

- $\sqrt{n-1}(f(l) - 0{,}05) = -0{,}159$

- $z_\alpha \sqrt{Var(f(l))} = -0{,}1195$

Como no se cumple la condición no es posible rechazar H_0, por lo tanto, se decide considerar un intervalo de confianza para c_0 y así estudiar mejor su comportamiento aproximado para lograr tomar una decisión.

El IC para c_0 está dado por

$$\left[f(l) - z_\alpha \sqrt{\frac{Var(f(l))}{n-1}}; f(l) + z_\alpha \sqrt{\frac{Var(f(l))}{n-1}} \right] = [0,0256; 0,0686]$$

como se puede ver, el 0.05 pertenece al intervalo, por lo que tiene sentido que H_0 no se rechace. Sin embargo, el rango del intervalo no excede más allá de 2 unidades de c_0, lo cual significa que las primeras 2 CP estarán explicando entre un 93 % y 97 % aproximadamente. Por lo tanto, no se considerarán las últimas 6 CP debido a su poca relevancia y para obtener una mejor interpretabilidad.

La figura 5 presenta los gráficos de correlación entre las primeras 2 componentes principales, donde es posible observar claramente una diferenciación entre las comunas del sector oriente y poniente, no obstante, aún así existen municipios como Macul, La Florida, Peñalolén y Lo Barnechea que se clasifican de manera similar que las comunas del sector poniente. Además es importante notar que las comunas de este último sector quedan bien agrupadas, a excepción de San Miguel, San Joaquín, San Ramón y Qulicura, aunque para las primeras 3 tiene sentido, ya que, son comunas vecinas y se estarían comportando de manera similar cuando la segunda CP es positiva y cuando es negativa entra en juego la comuna de Qulicura, la cual se ve más afectada que el resto por esta componente. Por otra parte, es posible notar que la influencia de la primera CP afecta en mayor medida a las comunas del sector oriente, específicamente Vitacura, Las Condes, Providencia, Ñuñoa y La Reina, pues estas se alejan más del cero en comparación con la segunda CP.

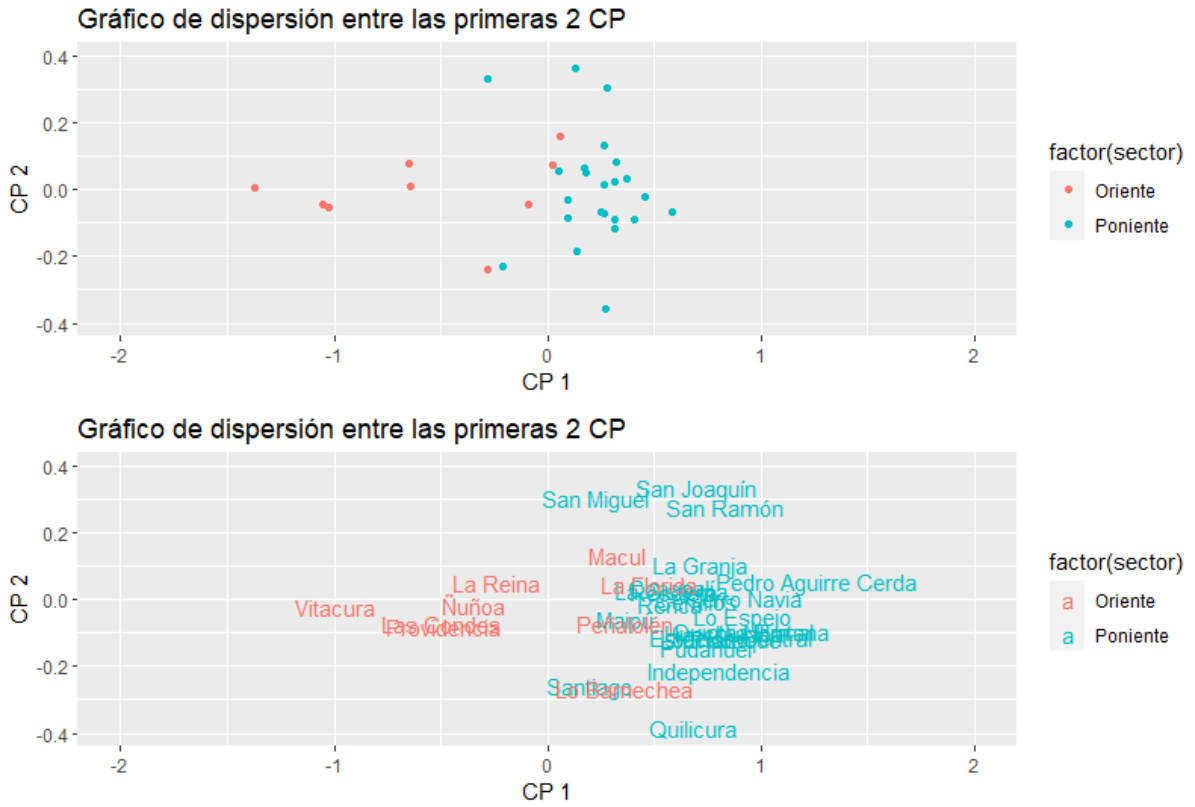

Figura 5: Gráficos de dispersión entre las dos primeras componentes principales filtrando por sector

De la figura 6 se puede concluir que todas las variables quedan bien explicadas por las primeras dos CP debido a la cercanía que tienen con la periferia. La primera CP se puede interpretar, escencialmente, como la diferencia entre todas las variables relacionadas con gastos e ingresos, mientras que la segunda CP queda expresada, escencialmente, en términos de la variable `edad`. Además, con la información de la tabla 8 es posible confirmar estos resultados, ya que se puede observar que la primera componente está más fuertemente correlacionada con las variables que tienen que ver con gastos e ingresos, mientras que la segunda componente tiene una correlación prácticamente de cero con estas variables y una muy alta con la variable `edad`.

Además, el primer vector propio de (4) indica básicamente que se está considerando un promedio de todas las variables menos `edad`, puesto que tienen pesos muy similares, sin embargo, esta última variable tiene un mayor peso en la segunda CP, donde el resto de variables posee un peso cercano a cero. Esto último se puede confirmar viendo el segundo vector propio.

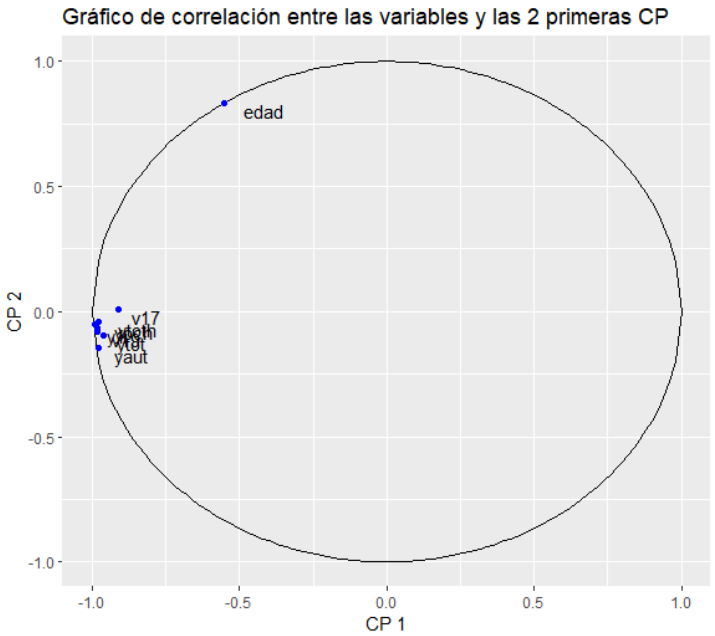

Figura 6: Gráficos de correlación entre las dos primeras componentes principales y las variables de X

Variable	$CP1$	$CP2$	$CP1^2 + CP2^2$
y1	-0.98	-0.06	0.97
v17	-0.91	0.01	0.83
v19	-0.98	-0.08	0.97
edad	-0.55	0.83	1.00
yaut	-0.98	-0.15	0.98
ytot	-0.96	-0.09	0.94
ytoth	-0.98	-0.04	0.96
ypch	-0.99	-0.05	0.98

Tabla 8: Tabla de correlación entre las variables y las 2 primeras CP

Conclusión

Teniendo en consideración todo lo comentado anteriormente, se puede concluir que en el año 2017 las comunas del sector oriente, en específico Vitacura, Las Condes, Providencia, Ñuñoa y La Reina, se caracterizan escencialmente por tener un promedio alto de ingresos y un promedio alto del costo de arriendo y/o dividendo de sus viviendas. No obstante, las comunas de San Miguel, Santiago, Lo Barnechea y Peñalolén también se caracterizan por tener un promedio de ingresos y gastos mayor pero en menor medida que las comunas anteriores, con la diferencia de que la edad promedio de San Miguel será mayor que la de Santiago y Lo Barnechea. Todas las comunas en la figura 6 que están cercanas al cero en el eje de la segunda CP no logran ser caracterizadas por el promedio de edad. Las

comunas del sector poniente, por su parte, se despliegan a lo largo de la segunda CP lo que da indicio de su caracterización, que sería el promedio de edad. Además, se puede observar que la gran mayoría se encuentra en el lado positivo del eje de la primera CP, lo cual significa que estas poseen un promedio de gastos e ingresos menor que los del sector oriente.

A modo de resumen, se obtuvo que las comunas del sector oriente se desglozan a lo largo del eje de la primera CP, mientras que, al ir acercándose al cero estas tienden a desglozarse a través del eje de la segunda CP. Caso contrario a lo que ocurre con las comunas del sector poniente, las cuales se distribuyen a lo largo del eje de la segunda CP pero sin desglozarse demasiado en el eje de la primera CP. Por otro lado, una de las sorpresas fue que la comuna de Lo Barnechea tuviese una caracterización similar a la comuna de Santiago, lo cual rompe el esquema de lo esperado, y también la poca diferencia que hay entre las comunas del sector poniente considerando además las comunas de Macul, La Florida y Peñalolén. Mientras que lo que estuvo dentro de lo esperado fue el haber captado la diferencia entre el resto de comunas del sector nor-oriente con las del sector poniente.

Referencias

- 1. Wikipedia contributors. Archivo:Comunas de Santiago (nombres).svg [Internet]. Wikipedia, The Free Encyclopedia. Disponible en: https://es.m.wikipedia.org/wiki/Archivo:Comunas_de_Santiago_%28nombres%29.svg
- Observatorio Social. Libro de Códigos Base de Datos [Internet]. 2018 sep. Disponible en: http://observatorio.ministeriodesarrollosocial.gob.cl/storage/docs/casen/2017/Libro_de_Codigos_Casen_2017.pdf
- Observatorio Social. Manual del investigador Guía práctica para el uso y análisis de información [Internet]. 2017. Disponible en: http://observatorio.ministeriodesarrollosocial.gob.cl/storage/docs/casen/2017/Manual_del_Investigador_Casen_2017.pdf
- Censo [Internet]. Default. [citado el 7 de diciembre de 2022]. Disponible en: https://www.ine.gob.cl/ine-ciudadano/definiciones-estadisticas/censo
- Korkmaz S, Goksuluk D, Zararsiz G. MVN: An R Package for Assessing Multivariate Normality. The R Journal. 2014 6(2):151-162.