

Análisis de Calidad de Vida de la Población Sanjuanina Aplicando Ciencia de Datos

Simón Pedro González¹, Myriam Herrera¹ y María Romagnano¹

Instituto de Informática, Facultad de Ciencias Exactas, Físicas y Naturales,
Universidad Nacional de San Juan, San Juan, Argentina

Resumen En el proceso de toma de decisiones resulta imprescindible la disponibilidad de información relevante, particularmente cuando las consecuencias de dichas decisiones tienen el potencial de impactar en la calidad de vida de un significativo número de personas. Para obtener dicha información, es importante contar con datos relevantes y un análisis adecuado. En este trabajo, se propone una metodología de análisis exploratorio, basada en análisis de correspondencia múltiple y agrupamiento jerárquico, para aplicar sobre datos de índole socio-económicos de las encuestas de individuo y hogar del INDEC. El objetivo del enfoque presentado es describir la población mediante un número reducido y manejable de factores y establecer perfiles, lo que, sumado a otros métodos, podría proveer de información concisa para la toma de decisiones, a nivel de administración pública regional y/o provincial.

Keywords: Ciencia de Datos · Población · Calidad de Vida

1. Introducción

En los tiempos que corren, los datos constituyen una nueva materia prima cuya explotación debe ser fuente de conocimiento para la comunidad y los organismos que la administran. Tradicionalmente, para realizar caracterizaciones socio-económicas de poblaciones se recurre al análisis de parámetros de tendencia, dispersión y distribuciones. Estas técnicas, a pesar de ser sin duda útiles, están propuestas desde una óptica mayormente determinista. Por tal razón se considera que pueden reducir la riqueza de la información en ciertos casos. Por otro lado, resultaría de gran utilidad una técnica que permita analizar los datos de forma global y dar de manera resumida y rápida una tipología de la población que indique cómo están situados los argentinos y en especial, los sanjuaninos. En el presente trabajo, el interés en el Gran San Juan responde, por un lado, al contexto geográfico del equipo de investigación y la institución académica que lo financia, y por otro, porque se considera que se requiere cubrir una ausencia en cuanto a análisis exploratorio y recuperación de perfiles poblacionales holísticos a partir de datos referentes a la población local.

En este trabajo se propone la utilización de herramientas y técnicas descriptivas propias de la Ciencia de Datos, campo interdisciplinario que agrupa teorías y métodos extraídos de la Matemática, la Estadística, la Inteligencia Artificial

(incluyendo el Aprendizaje de Máquina), la Minería de Datos, entre otros [1]. Además, involucra técnicas de procesamiento, análisis y visualización que permiten la extracción de información con un enfoque más holístico de los datos, para su posterior conversión a un mensaje manejable. Este mensaje sería entregado en forma de gráficas y tablas que resuman los datos, para luego extraer conclusiones generales. De esta manera, la información obtenida se pondría al servicio de analistas que soporten la toma de decisiones en políticas públicas [2]. Las técnicas planteadas en este trabajo se utilizaron como propuesta para el apoyo a la toma de decisiones en otros ámbitos, como instituciones financieras [3] o académicas [4] [5] [6], y se considera que su utilidad en estos casos podría extrapolarse al área de la administración pública local.

2. Metodología

2.1. Datos

Para este trabajo se utilizaron los datos publicados oficialmente por el Instituto Nacional de Estadística y Censos de la República Argentina (INDEC). Concretamente, se tomaron las bases de datos correspondientes a hogares e individuos, ambas resultantes de la Encuesta Permanente de Hogares del cuarto trimestre del año 2017 [7]. Se contó con 58182 respuestas de la encuesta de individuos de Argentina (de las cuales 1957 pertenecen al aglomerado¹ Gran San Juan) y 18793 respuestas de la encuesta de hogares de Argentina (de las cuales 566 pertenecen al aglomerado Gran San Juan).

2.2. Herramientas

Existen varios softwares de licencia paga con características útiles para el análisis de datos de tipo socio-demográficos, tal como Coheris Analytics Spad [8]. Sin embargo, respondiendo a un espíritu de accesibilidad y democratización de la información, se buscaron alternativas libres, optándose finalmente por R [9], un entorno y lenguaje de programación libre con enfoque estadístico, muy popular en las investigaciones biomédicas, bioinformáticas, financieras y demográficas. Las ventajas que proporciona, a cambio de una curva de aprendizaje más empinada, son la flexibilidad para tratar los datos de manera conveniente, una gran potencia y numerosas librerías útiles para el preprocesamiento, análisis y visualización. Entre las librerías utilizadas, cabe destacar la importancia de FactoMineR [10] para este trabajo, ya que proporciona una gran cantidad de técnicas de análisis multivariante, entre ellas, el agrupamiento jerárquico y el análisis de correspondencia múltiple.

¹ Aglomerado: región urbanizada que se extiende sin solución de continuidad a lo largo de varias circunscripciones administrativas.

2.3. Análisis

Siendo el objetivo principal encontrar estructuras internas en los datos, que permitan establecer perfiles o tipologías y teniendo en cuenta además, que las variables estudiadas son mayoritariamente de tipo nominal, se considera que el análisis de correspondencia múltiple [11] (ACM) es una buena opción para aplicar en este caso. El fundamento de esta técnica se encuentra en que las relaciones entre un elevado número de variables se deben a la presencia de factores subyacentes. Se busca identificar estos factores o variables comunes, de manera que sea más sencillo interpretar los datos y establecer perfiles. Los factores se definen como variables inobservadas o latentes, que están relacionadas con las observadas, y que, si bien se presentan en un menor número que estas últimas, pueden explicar la variabilidad de la totalidad de los datos. Esto les da la capacidad de resumir grandes tablas de datos.

En nuestro caso, cada variable representa una pregunta de la encuesta, y las posibles respuestas son las modalidades o categorías mutuamente excluyentes de cada variable. Al responder a la encuesta, un individuo asume una categoría para una variable determinada. En el Cuadro 1 y Cuadro 2 se describen los ámbitos, variables y categorías utilizadas en este trabajo. Cabe aclarar que se descartaron aquellas variables que se consideraron redundantes, con demasiada dependencia o bien, poco informativas (con pocas respuestas). Una vez seleccionadas las variables, se procede a elaborar la tabla de códigos condensados, una tabla lógica que representa los datos como una nube de puntos en un espacio de dimensión NP (N-individuos, P-modalidades), donde es posible identificar a los individuos (filas de la matriz) y las modalidades (columnas de la matriz) mediante coordenadas y calcular distancias con la métrica de χ^2 cuadrada. La distancia entre individuos refleja qué tan semejantes son, indicando si asumieron, en general, las mismas modalidades o no. Por otro lado, dos modalidades son próximas si fueron asumidas en gran medida por el mismo grupo de individuos.

La nube de puntos que describe los datos posee un centro de gravedad y una cierta dispersión con respecto a este, llamada inercia. El fundamento del ACM es la descomposición de la inercia en factores que explican, en mayor o menor grado, la dispersión de la nube de puntos. Esto resulta útil por dos razones, primero porque permite plantear visualizaciones entendibles, estableciendo un referencial de representación en dos dimensiones, cada una asociada a alguno de los factores obtenidos; y segundo, porque se pueden utilizar los factores para agrupar los datos en clases. Las clases son subconjuntos de individuos próximos en el espacio NP, es decir, que comparten ciertas características. Para agrupar los datos en clases no preestablecidas se utilizaron técnicas de aprendizaje no supervisado, agrupamiento o clustering. En este trabajo, se usaron los factores extraídos del ACM para agrupar los datos con clustering jerárquico aglomerativo [12]. Este último tiene una aproximación ascendente; cada observación empieza en un grupo propio, y los grupos más cercanos se fusionan a medida que se escala en la jerarquía, hasta formar un único grupo.

Cuadro 1. Ámbitos, variables y categorías de la base de datos de individuos

Ámbito	Variabes	Categorías
Características generales	Relación de Parentesco	Jefe/a, Cónyuge/Pareja, Hijo/Hijastro/a, Yerno/Nuera, Nieto/a, Madre/Padre, Suegro/a, Hermano/a, Otros Familiares, No Familiar
	Sexo	Varón, Mujer
	Edad	Para análisis de todas las edades:[0,7), [7,13), [13,18), [18,26), [26,50), [50,70), [70,...] Para análisis de edades laborables:[16,26), [26,50), [50,75), [75, ...]
	Estado Civil	Unido/a, Casado/a, Separado/a o Divorciado/a, Soltero/a, Viudo/a
	Nivel Educativo	Sin instrucción, Primaria completa, Secundaria (Completa o Incompleta), Universitaria Incompleta, Universitaria Completa
Características de Ocupación	Estado Social	Ocupado, Desocupado, Inactivo, Menor de 10 años
	Actividad Laboral	Patrón, Cuenta propia, Obrero o empleado, Trabajador familiar sin remuneración
	Razón de Inactividad	Jubilado/Pensionado, Rentista, Estudiante, Ama de casa, Menor de 6 años, Discapacitado, Otros, No corresponde (no está inactivo)
	Busca Trabajo	Sí, No
	Tipo de Empresa (donde trabaja)	Privada, Estatal

3. Resultados

Existen dos aproximaciones posibles para realizar un ACM. Se puede tener en cuenta todos los datos, sin agrupamiento de variables, lo que lleva a una descripción general y proporciona una base para posteriores análisis. Por otro lado, se pueden analizar distintos ámbitos por separado, dejando activas las variables pertenecientes a un ámbito, e incorporando las restantes como suplementarios, lo que resulta útil para estudiar las relaciones entre dos ámbitos en particular. A continuación se presentarán algunos de los resultados obtenidos de ACM con ambos enfoques y clustering jerárquico, sobre los datos de Hogar e Individuos del aglomerado Gran San Juan, y el total de aglomerados de Argentina.

3.1. Sobre datos del aglomerado Gran San Juan

Primero se efectuó un análisis sobre individuos de todas las edades, el resultado se muestra en la Fig. 1. Para este gráfico, se tomó la variable “Estado

Cuadro 2. Ámbitos, variables y categorías de la base de datos de hogar

Ámbito	Variabes	Categorías
Características de Vivienda	Tipo de vivienda	Casa, Departamento, Pieza de inquilinato, Pieza en hotel/pensión
	Proveniencia del agua	Dentro de la vivienda, Fuera de la vivienda
	Baño/Letrina	Sí, No
	Tipo de baño	Inodoro con cadena, Inodoro a balde, Letrina, No tiene
	Desagüe	Cloaca, Pozo, No tiene
	Cercanía a basurales	Sí, No
	En zona inundable	Sí, No
	En villa de emergencia	Sí, No
	Cocina	Sí, No
	Lavadero	Sí, No
	Garaje	Sí, No
	Régimen de tenencia	Propietario, Inquilino/arrendatario de la vivienda, Ocupante gratuito, Otro
	Combustible usado para cocinar	Gas de red, Gas de tubo/garrafa, Otro
Ingresos	Ingreso per cápita familiar (IPCF)	[\$0,\$5.000),[\$5.000,\$10.000),[\$10.000,\$30.000), [\$30.000,\$60.000),[\$60.000,\$100.000), [\$100.000,\$200.000), [\$200.000, ...)
	Cantidad de miembros del hogar	1, 2, [3,5), [5,7), [7,10), [10, 15), [15, ...]

Social” como suplementaria. Esta elección tiene que ver con el hecho de que dicha variable puede dar un primer indicio de qué perfiles es posible encontrar en los datos (individuos ocupados, desocupados, inactivos y niños de corta edad). Resulta, por lo tanto, interesante determinar si existe un agrupamiento visible en esta variable, considerando únicamente las demás. Recordando que las categorías cercanas entre sí son aquellas que podrían presentar una relación más fuerte, se puede comenzar a realizar ciertas interpretaciones. Por ejemplo, las categorías “Sin instrucción”, “Menor de 10 años” y “Nieto/a” se encuentran relacionadas, y representan a los niños más pequeños. Por otro lado, nótese que los individuos con trabajo o desocupados se encuentran a la izquierda del eje vertical, mientras que los inactivos, a la derecha. Se puede apreciar como aquellos con trabajo tienen en general un nivel educativo más elevado. Finalmente, se efectuó un ACM con todas las variables activas sobre los datos de individuos, se tomaron los primeros 14 factores, que de forma acumulativa, explican más de un 60.00 % de la inercia global, lo que se considera suficiente para obtener resultados útiles. Con esos factores se efectuó el agrupamiento de los datos en 3 clusters. El número de clusters se decidió en función del análisis de dendogramas.

En el Cuadro 3, Cuadro 4 y Cuadro 5 se describieron los clusters en función de las modalidades más significativas para su definición. Estos cuadros pueden leerse de la siguiente forma: por ejemplo, en el Cuadro 3, la tercera fila indicaría

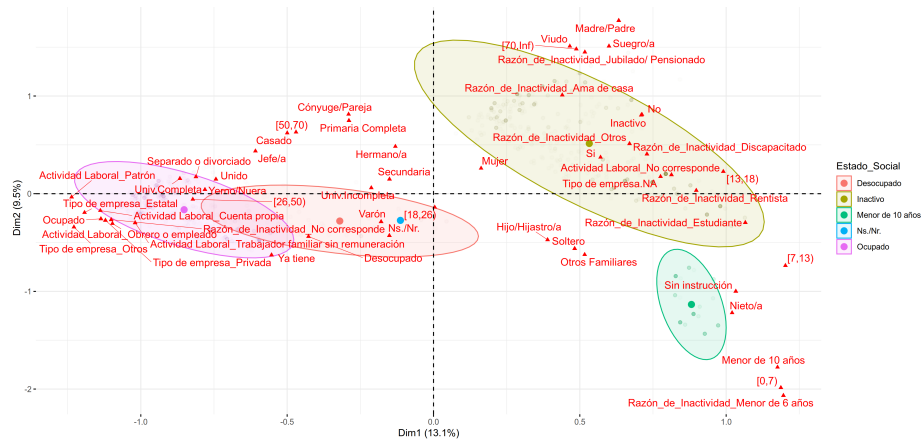


Figura 1. Individuos sobre las dos primeras dimensiones (San Juan).

que el 100.00 % de todos los individuos que están ocupados (trabajan) pertenecen al cluster, mientras que el 95.30 % del cluster está compuesto por individuos ocupados. Así, con cada cuadro se puede hacer una descripción general de la población incluida en el cluster. En el Cuadro 3 se puede observar que el cluster 1 agrupa individuos ocupados, mayoritariamente varones, entre 26 y 49 años, de nivel educativo secundario o universitario, jefes/as de hogar. Podemos considerar que esta sección de la población constituye la fuerza de trabajo remunerada. En el Cuadro 4 se puede observar que el cluster 2 incluye a individuos inactivos, mayoritariamente mujeres, de nivel educativo más modesto, jubilados, estudiantes o amos/as de casa. Podemos ver cómo el 37.90 % del cluster son estudiantes, el 38.95 % jubilados y el 17.00 % amos/os de casa. Esto causa que el rango etario varíe significativamente. Por último, en el Cuadro 5 se puede observar que los individuos incluidos en el cluster 3 son niños de entre 0 y 13 años, por lo general hijos o nietos del jefe del hogar. Cabe mencionar que esta presentación tabular tiene algunas ventajas con respecto a la representación de los clusters en una gráfica bidimensional. Al ser poca la variabilidad explicada por únicamente las dos dimensiones principales, una parte significativa de la información queda oculta al graficar los resultados. Debido a esto, se optó por utilizar las tablas de descripción para mostrar los perfiles obtenidos de un manera más completa.

Se realizó un proceso similar sobre los datos de hogares, y se muestran los resultados los cuadros siguientes. En el Cuadro 6 se puede observar que el cluster 1 agrupa los hogares de buenas condiciones, generalmente con gas de red, cloaca, inodoro con cadena y lavadero. La cantidad de miembros por hogar es relativamente reducida, generalmente entre 1 y 3. En el Cuadro 7 se puede observar que el cluster 2 incluye los hogares de condición modesta, generalmente con pozo como desagüe, gas de garrafa y sin lavadero. Aumenta la cantidad de miembros por hogar, viéndose un 31.50 % de hogares en este cluster con una cantidad entre 5 y 7, y un 12.16 % entre 7 y 10. Finalmente, en el Cuadro 8 puede observarse que el cluster 3 agrupa los hogares con condiciones de vida precarias y un alto

Cuadro 3. Cluster 1: 42.57% de individuos de San Juan

Categoría	% de categoría en cluster	% de cluster asume categoría
Tipo de empresa=Privada	100.00	75.03
Estado Social=Ocupado	100.00	95.32
Actividad Laboral=Obrero o empleado	100.00	71.43
Edad=[26,50)	78.25	53.54
Actividad Laboral=Cuenta propia	100.00	24.73
Tipo de empresa=Estatal	100.00	19.57
Relación de Parentesco=Jefe/a	63.60	43.22
Nivel Educativo=Univ.Completa	77.03	19.33
Sexo=Varón	53.11	59.42
Estado civil=Unido	68.20	19.57
Estado civil=Casado	54.89	35.05
Edad=[50,70)	54.61	27.73
Nivel Educativo=Secundaria	50.13	46.46
Edad=[18,26)	49.28	16.33
Nivel Educativo=Univ.Incompleta	50.52	11.65

Cuadro 4. Cluster 2: 43.94% de individuos de San Juan

Categoría	% de categoría en cluster	% de cluster asume categoría
Busca trabajo=No	100.00	98.26
Estado Social=Inactivo	100.00	99.77
Razón de Inactividad=Jubilado	100.00	38.95
Razón de Inactividad=Ama de casa	100.00	17.09
Edad=[13,18)	96.88	18.02
Razón de Inactividad=Estudiante	74.26	37.91
Edad=[70,Inf)	91.38	18.49
Sexo=Mujer	55.80	66.51
Estado civil=Viudo	86.36	11.05
Relación de Parentesco=Cónyuge	57.36	22.21
Nivel Educativo=Secundaria	49.87	44.77
Edad=[18,26)	50.72	16.28
Estado civil=Unido	31.80	8.84
Relación de Parentesco=Jefe/a	36.40	23.95
Nivel Educativo=Sin instrucción	33.49	16.63

riesgo ambiental sanitario. En este cluster se presentan todos los hogares sin baño y los que están en zonas inundables. Además, un 75.00% de estos hogares están cerca de basurales, y un 76.60%, cerca de villas de emergencia.

El estudio de las características y la frecuencia con la que se presenten los hogares de uno u otro perfil puede ser muy útil a la hora de determinar las necesidades más urgentes en cuanto a calidad de vida de la población. Por otro lado, si se contara con el área administrativa de la que proviene cada una de las entradas

Cuadro 5. Cluster 3: 13.49% de individuos de San Juan

Categoría	% de categoría en cluster	% de cluster asume categoría
Estado Social=Menor de 10 años	100.00	100.00
Nivel Educativo=Sin instrucción	61.83	100.00
Edad=[0,7)	100.00	69.32
Relación de Parentesco=Nieto/a	55.32	29.55
Edad=[7,13)	47.37	30.68
Relación de Parentesco=Hijo/astro/a	21.79	63.64

Cuadro 6. Cluster 1: 58.30% de hogares de San Juan

Categoría	% de categoría en cluster	% de cluster asume categoría
Combustible para cocinar=Gas de red	77.25	93.64
Tipo de desagüe del baño=Cloaca	83.86	56.67
Garage=Tiene garaje	86.34	42.12
IPCF=[\$10.000,\$30.000)	88.57	37.58
Lavadero=Tiene lavadero	72.94	75.15
Régimen de tenencia=Propietario	68.97	78.79
Cantidad de miembros=[2,3)	85.59	30.61
Cocina=Tiene cocina	62.75	96.97
Tipo de baño=Inodoro con cadena	60.48	99.70
Tipo de vivienda=Departamento	88.00	13.33
IPCF=[\$5.000,\$10.000)	68.40	47.88
Cantidad de miembros=[1,2)	70.24	17.88
Cercanía a basurales=Si	16.00	1.21
Tipo de vivienda=Casa	55.43	86.67

de datos, se podría, inclusive, evaluar en qué zonas de la provincia se encuentran los hogares pertenecientes a perfiles prioritarios, que requieran de mayores esfuerzos para el incremento de las condiciones de vida.

3.2. Sobre datos del total de aglomerados de Argentina

Se realizó el mismo procedimiento sobre los datos de todos los aglomerados de Argentina, con el fin de poder realizar comparaciones. En la Fig. 2 se observan los resultados de individuos. Se tomó la variable Estado Social distinguir a los individuos en el gráfico. El clustering se realizó teniendo en cuenta las mismas consideraciones que en el caso de San Juan, y pueden verse los resultados en el Anexo, en el Cuadro A1, Cuadro A2 y Cuadro A3. En el Cuadro A1 se puede observar que el cluster 1 agrupa a individuos ocupados, mayoritariamente varones, entre 26 y 49 años, de nivel educativo secundario o universitario, jefes/as de hogar. En el Cuadro A2 se puede observar que el cluster 2 incluye individuos inactivos, mayoritariamente mujeres, de nivel educativo más modesto, jubilados,

Cuadro 7. Cluster 2: 39,22% de hogares de San Juan

Categoría	% de categoría en cluster	% de cluster asume categoría
Combustible para cocinar=Garrafa	82.32	60.81
IPCF=[\$0,\$5.000)	73.91	61.26
Tipo de desagüe del baño=Pozo	54.55	83.78
Garage=No tiene garaje	50.12	91.44
Lavadero=No tiene lavadero	60.62	61.71
Cocina=No tiene cocina	80.36	20.27
Régimen de tenencia=Ocupante gratuito	83.33	15.77
Cantidad de miembros=[7,10)	90.00	12.16
Cantidad de miembros=[5,7)	63.06	31.53
Tipo de vivienda=Casa	42.05	97.75
Tipo de baño=Inodoro a balde	85.00	7.66
Cantidad de miembros=[10,15)	100.00	2.25
Tiene agua=Fuera de la vivienda	85.71	2.70
Tiene agua=Dentro de la vivienda	38.64	97.30
Tipo de baño=Inodoro con cadena	37.68	92.34
IPCF=[\$5.000,\$10.000)	29.44	30.63

Cuadro 8. Cluster 3: 2,48% de hogares de San Juan

Categoría	% de categoría en cluster	% de cluster asume categoría
Zona inundable=Si	100.00	100.00
Cercanía a basurales=Si	36.00	75.00
Villa de emergencia=Si	100.00	76.67
Baño/letrina=No	100.00	100.00

estudiantes o amos/as de casa. En el Cuadro A3 se puede ver que el cluster 3 agrupa a los niños de entre 0 y 13 años. Como puede observarse, las estructuras que se presentan en los datos no difieren significativamente de las halladas en el análisis del Gran San Juan.

En el Anexo, en el Cuadro A4, Cuadro A5 y Cuadro A6 se presentan los resultados del análisis sobre los datos de hogares de Argentina. En el Cuadro A4 puede observarse que el cluster 1 agrupa los hogares de condiciones precarias. Este cluster incluye a mayor parte de aquellos que viven en piezas de inquilinato. Se destacan las categorías de inodoro a balde, letrina, agua fuera de la vivienda y la falta de lavadero, garaje e incluso cocina. En el Cuadro A5 puede observarse que el cluster 2 agrupa los hogares de condiciones bajas/modestas. Ciertas características como el uso de pozos para desagüe y gas de garrafa se presentan como predominantes, al igual que en cluster 1, pero hay menor proporción de hogares sin garaje, lavadero o cocina. Los ingresos per cápita son un poco más altos, y los hogares en su mayor parte tienen inodoro con cadena y cañerías con

10 S. P. González et al.

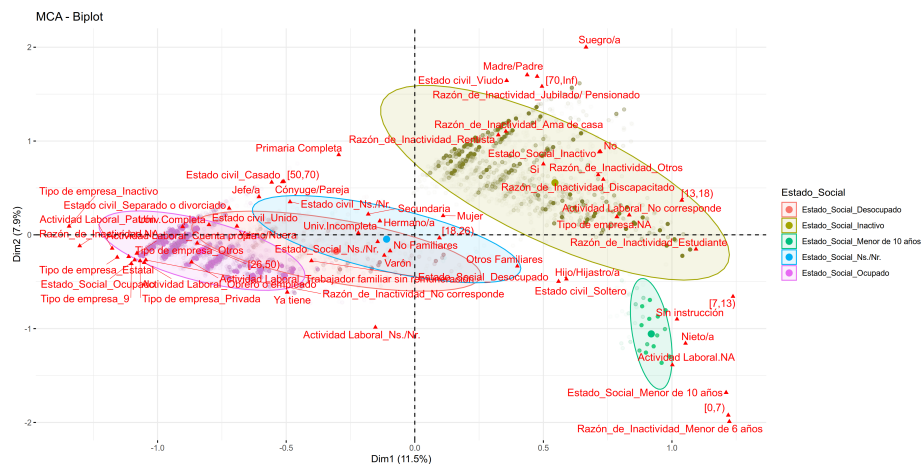


Figura 2. Individuos sobre las dos primeras dimensiones (Argentina).

agua corriente dentro de la vivienda. En el Cuadro A6 puede observarse que el cluster 3 agrupa los hogares de condiciones medias y altas. Tienen generalmente agua corriente, gas de red, cloaca, inodoro con cadena, cocina, lavadero y garaje. Además, vemos los ingresos más altos y la menor proporción de miembros por hogar.

Considerando todos los aglomerados de Argentina, podemos ver que el grupo de condición más baja es más numeroso que si se considera únicamente el Gran San Juan, pero reúne un rango más amplio de condiciones, y no se ve tan fuertemente caracterizado por factores de riesgo ambiental sanitario como cercanía a basurales o zonas inundables. En cuanto al resto de los hogares, podemos indicar que el agrupamiento sobre los datos de Argentina revela una estructura con una mayoría importante de hogares de condiciones medias y altas (81.86%), pero con un rango de condiciones más variable dentro del grupo de lo que se observa en el Gran San Juan.

3.3. Conclusiones

En el presente trabajo se aplicó ACM y clustering jerárquico como técnica de análisis exploratorio para datos socio-económicos de las encuestas de individuo y hogar del INDEC. Se obtuvieron agrupamientos y se establecieron perfiles que describen los datos de una manera más completa e integradora. Tales perfiles pueden servir para llevar a cabo análisis posteriores más específicos, e inclusive, pueden traducirse fácilmente en mensajes manejables para ya comenzar a identificar los grupos poblacionales donde se deben enfocar esfuerzos para el mejoramiento de la calidad de vida. A futuro, se planea realizar un estudio comparativo entre el índice de calidad de vida calculado de la manera tradicional y los perfiles poblacionales obtenidos, para determinar relaciones. Además, se pretende analizar la evolución de dichos perfiles en el tiempo, utilizando los datos de las encuestas de individuos y hogares de semestres y años siguientes.

Referencias

1. V. Dhar, “Data science and prediction”, 2013. Communications of the ACM. vol. 56, no. 12, pp. 64-73. Recuperado de: <https://dl.acm.org/doi/abs/10.1145/2500499>
2. Z. Engin y P. Treleaven, “Algorithmic Government: Automating Public Services and Supporting Civil Servants in using Data Science Technologies”, 2019. The Computer Journal, vol. 62, no. 3, pp. 448-460, Recuperado de: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8852885>
3. L. Ganga, M. Herrera y M. Romagnano, “Clasificación de clientes de tarjeta de crédito de una institución financiera a partir de Weka”, 2018. Congreso Internacional de Ciencias de la Computación y Sistemas de Información (CICCSI 2018) – Universidad de Campagnat, Mendoza y Universidad Nacional de San Juan, San Juan, Argentina.
4. M. Herrera, M. Romagnano, L. Ganga, “Aporte de la Ciencia de Datos a la toma de decisiones institucionales”, 2019. II Congreso de Matemática Aplicada, Computacional e Industrial, (MACI 2019), Río Cuarto, Córdoba. Recuperado de: <https://amcaonline.org.ar/maci/index.php/maci2019/maci/paper/viewFile/5112/344>
5. M. Herrera, A. Mallea, E. Torres, F. Millán, M. I. Lund, “Features of the recent college graduate of the Faculty of Exact, Physical and Natural Sciences of the National University of San Juan”, 2017. XLIII CLEI (Conferencia Latinoamericana de Informática), Simposio Latinoamericano de Manejo de Datos en Información (SLMDI). Universidad Tecnológica Nacional - Facultad Regional de Córdoba, Ciudad de Córdoba, Argentina. ISBN 978-1-5386- 3057-0. Recuperado de: <http://www.clei2017-46jaiio.sadio.org.ar/sites/default/files/Mem/SIESC/SIESC-20.pdf>
6. S. Ruiz, M. Herrera, M. Romagnano, L. Mallea y M. I. Lund, “Tipología Influyente en el Rendimiento Académico de Alumnos Universitarios”, 2018. Revista Entre Ciencia e Ingeniería, vol. 12, no. 23, pp. 109 – 116.
7. Instituto Nacional de Estadística y Censos de la República Argentina, “Encuesta Permanente de hogares”. Recuperado de: <https://www.indec.gob.ar/indec/web/Institucional-Indec-BasesDeDatos>
8. Equipo de desarrollo de Coheris Analytics Spad, “Coheris Analytics Spad: Le Logiciel Data-Mining dédié aux Data Scientists”, 2021. Recuperado de: <https://ia-data-analytics.fr/logiciel-data-mining/>
9. Equipo de desarrollo de R, “R: A language and environment for statistical computing”, 2018. R Foundation for Statistical Computing, Vienna, Austria. Recuperado de <https://www.R-project.org/>
10. S. Lê, J. Josse y F. Husson, “FactoMineR: An R Package for Multivariate Analysis”, 2008. Journal of Statistical Software, vol. 25. no. 1, pp. 1-18. Recuperado de: <http://dx.doi.org/10.18637/jss.v025.i01>
11. P. Daniel, “Análisis de Datos Multivariantes”, 2002. Madrid, España, Editorial de la Universidad Carlos III.
12. S. Theodoridis, K. Koutroumbas. “Pattern Recognition”, 2003. San Diego, Estados Unidos, Editorial Elsevier.