

Final Análisis Inteligente de Datos

Diego Kozlowski

16 de Agosto 2017

Índice

| | |
|--|-----------|
| 1. Introducción | 3 |
| 2. La Encuesta Permanente de Hogares | 3 |
| 3. Técnicas implementadas | 4 |
| 3.1. Análisis de Correspondencia Multiple | 4 |
| 3.2. Análisis de Componentes Principales | 5 |
| 3.3. Clustering | 6 |
| 3.4. Logit | 6 |
| 3.5. El uso de ponderadores de replicación | 6 |
| 3.6. Software utilizado | 9 |
| 4. Principales resultados | 9 |
| 4.1. Base Hogar | 9 |
| 4.1.1. Estratégias de supervivencia de los hogares | 9 |
| 4.1.2. Características habitacionales de los hogares | 12 |
| 4.1.3. Características habitacionales de las viviendas | 13 |
| 4.2. Base individuos | 16 |
| 4.2.1. Características de los miembros del hogar | 16 |
| 4.2.2. Ingresos | 17 |
| 4.2.3. Tiempo de trabajo | 22 |
| 5. Conclusiones | 23 |
| Bibliografía | 25 |

1. Introducción

Es extendido el uso de las herramientas estadísticas y econométricas en el estudio de las condiciones socioeconómicas de los hogares, tanto a nivel mundial, como para el caso argentino. Por su parte la Encuesta Permanente de Hogares es la principal fuente de información para el estudio de las condiciones de vida de las personas en la Argentina. Sin embargo, la utilización de muchas de las técnicas estadísticas comprendidas en el presente curso tienen escasa repercusión para este tipo de trabajos, y en particular mediante la utilización de dicha encuesta. Por su parte, es amplio el uso de diferentes técnicas de *Data Mining* en otros aspectos de la vida económica de la sociedad, como es el caso del área de finanzas (Yang 2015, Loretan (1997)).

Se nos presenta de esta manera como un campo fértil para la investigación el uso de las técnicas aprendidas en un ámbito de la economía diferente al cual éstas suelen ser aplicadas, de forma tal que se logre potenciar el estudio de las sociedades mediante un nuevo herramiental, a la vez que se abre un nuevo espacio de aplicación, que conlleva nuevos desafíos para su implementación.

En este sentido, el presente trabajo propone simplemente ser un análisis exploratorio del potencial aporte analítico de las técnicas estudiadas en el campo de las ciencias sociales. De esta forma, la pregunta que guía el trabajo no es propia del objeto de estudio en sí, tal como sería preguntarse por las condiciones de reproducción de los hogares, la distribución del ingreso, o los movimientos de coyuntura en el mercado de trabajo, sino que tan sólo apunta a observar aquello que respecta al uso de la herramienta, y las potencias y dificultades que se presentan en una primera utilización de las mismas en este ámbito de estudio. El objetivo del presente trabajo es, por lo tanto, reconocer las potencias, los límites, y la superabilidad de los mismos, de las técnicas de Análisis de Correspondencia Múltiple, Análisis de Componentes Principales, Clustering y las regresiones Logit, en la respuesta de las preguntas propias del ámbito de las ciencias sociales.

El presente trabajo se estructura de la siguiente manera: luego de esta breve introducción, se realiza una descripción de la base de datos. Luego se comentan las técnicas utilizadas y las variables seleccionadas para cada técnica. A continuación se presentan los principales resultados obtenidos, las conclusiones y los posibles caminos a seguir.

2. La Encuesta Permanente de Hogares

Las bases utilizadas en el presente análisis provienen de los micro datos, para el segundo trimestre del 2016, provistos por la Encuesta Permanente de Hogares (EPH), elaborada por el INDEC. Según este organismo, la EPH

“es un programa nacional de producción sistemática y permanente de indicadores sociales. Releva las características sociodemográficas y socioeconómicas de la población, con énfasis en la medición de los niveles y características del empleo en la Argentina. Proporciona estimaciones válidas para los cuatro trimestres del año y cubre 31 aglomerados urbanos donde habita, aproximadamente, el 70 % de la población urbana del país. En su modalidad puntual, la EPH se realiza en Argentina desde el año 1973. A partir del 2003 y hasta la actualidad, la EPH aplica una modalidad continua.” (INDEC 2016)

Sobre la base de esta encuesta se elaboran las estadísticas oficiales respecto de la incidencia la pobreza y la indigencia, los principales indicadores de mercado de trabajo y distribución del ingreso (INDEC 2017). Esto la convierte en una de las principales fuentes de información para el análisis de las condiciones socioeconómicas de la Argentina.

Además de producir los indicadores mencionados, los micro datos de la encuesta son publicados de forma periódica en dos bases: La *base hogar* y la *base individual*. La primera contiene a nivel fila un registro por cada hogar, mientras que la segunda contiene una fila por cada componente de cada hogar. Ambas se pueden juntar por la conjunción del código único de vivienda, y el número de hogar, de forma tal de poder obtener las respuestas del hogar al que pertenece un individuo, o mediante algún tipo de agregación, las respuestas de los individuos que componen cada hogar en esta base. La base Hogar cuenta con 88 variables, mientras que la base individual cuenta con 177. La Base Hogar se encuentra delimitada en los siguientes módulos : *Identificación*, *Características de la vivienda*, *Características habitacionales del hogar*, *Estrategias de supervivencia del hogar*, *Ingresos* y *Organización del hogar*. Por su parte la Base individual tiene, además de las variables

identificatorias y de las características personales, una serie de preguntas concatenadas, de forma tal que en función de si la persona se encuentra ocupada, desocupada o inactiva, responderá un subconjunto distinto de preguntas, al igual que si el individuo ocupado trabaja en una relación de dependencia, de forma autónoma, o es patrón. Las preguntas subsecuentes refieren a sus ingresos, tiempos de trabajo, antigüedad, entre otros. Para el presente trabajo se decidió tomar sólo una pequeña porción de las variables existentes en ambas bases, además de crear la variable *nivel educativo* a partir de la conjunción de las siguientes cuatro preguntas:

- ¿Asiste o asistió a algún establecimiento educativo?
- ¿Cuál es el nivel más alto que cursa o cursó?
- ¿Finalizó ese nivel?
- ¿Cuál fue el último año que aprobó?

Cada variable por separado es en sí categórica, pero sabiendo la cantidad de años que implica cada nivel educativo, se puede sumar los años de educación de los niveles completos y la cantidad de años en el nivel incompleto, de forma tal de poder establecer la cantidad de años de educación formal del individuo, descontando los años que efectivamente le llevó (es decir, asumiendo que cumplió con los tiempos estipulados de cada nivel educativo).

3. Técnicas implementadas

En la presente sección se presenta la implementación de las técnicas utilizadas para realizar el análisis.

3.1. Análisis de Correspondencia Multiple

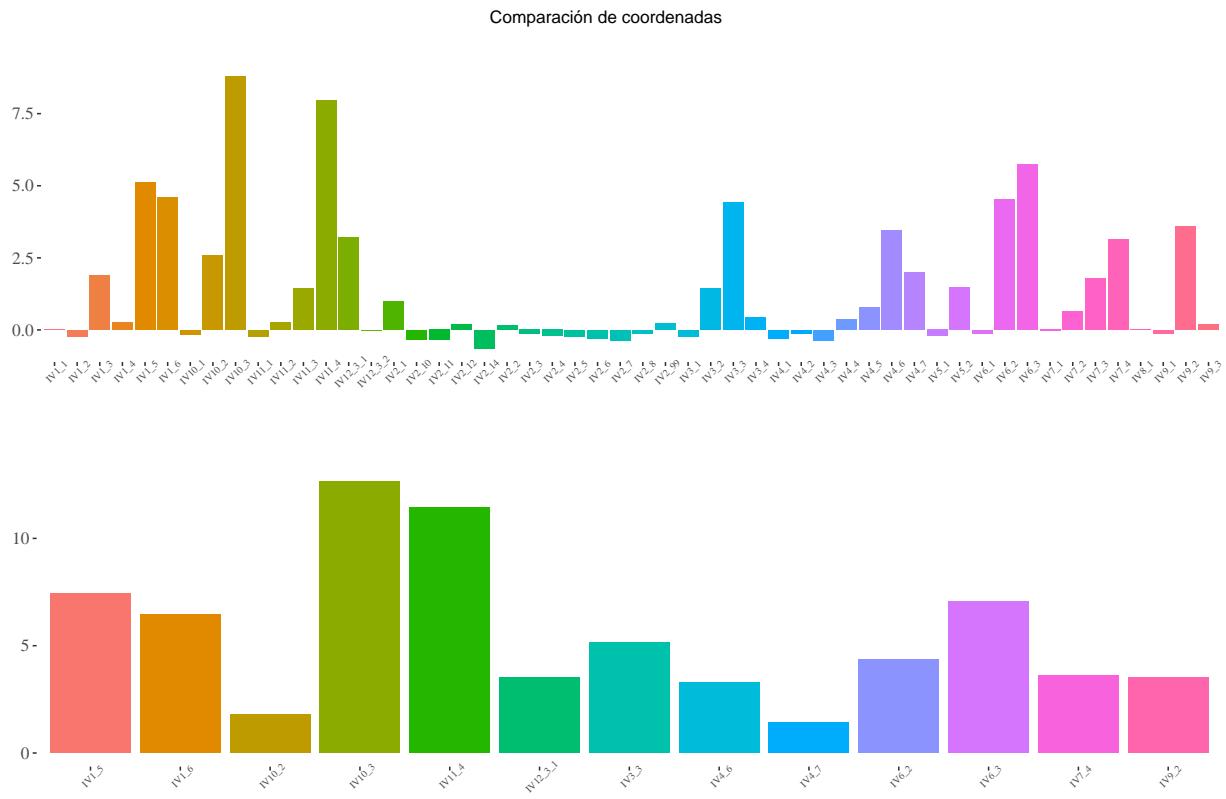
El análisis de Correspondencia Múltiple se presentó como una herramienta particularmente rica para un análisis exploratorio de una base como la descrita, dado que ésta presenta una cantidad considerable de variables discretas, pre agrupadas por tópicos. Es por esto que el análisis de Correspondencia Múltiple se segmentó en cuatro secciones en las cuales las variables apuntan a responder el mismo tipo de preguntas. Estas son:

- **Base Hogar :**
 - Estrategias de supervivencia de los hogares:
 - 12 variables dicotómicas que responden respecto a las formas en que los hogares logran subsistir económicamente.
 - Características habitacionales de los hogares:
 - 5 variables respecto de la cantidad de ambientes, dormitorios, la condición de ocupación, el tipo de cocina y el tipo de baño.
 - Características habitacionales de las viviendas:
 - 9 variables respecto de los materiales utilizados en la vivienda, el acceso al agua, etc.
- **Base Individual:**
 - Características de los miembros del hogar:
 - 7 variables respecto de las características generales de los individuos: edad, categoría ocupacional, condición de actividad, relación de parentesco, etc.

Para llegar a dicha selección de variables, en el procesamiento de las mismas se comenzó realizando una primera aproximación con la totalidad de las variables de los módulos mencionado, para luego ver el aporte de cada uno de los pares atributo-valor en las cargas de la primera dimensión del Análisis de Correspondencia Múltiple. Se ordenó las cargas por su valor absoluto y se eliminó aquellos pares atributo-valor que tenían una carga menor a tres cuartos. Esto reduce a la vez la cantidad de variables y la cardinalidad de las mismas, dejando sólo aquellas que tienen mayor inercia, y llevando a un valor default a todas las demás. Este proceder resultó efectivo para aumentar la inercia de las primeras coordenadas. El Gráfico 1 muestra el cambio en las coordenadas antes y después de eliminar los valores de menor peso, para el caso del módulo de Vivienda. Tal como se observa allí, se reduce fuertemente la cardinalidad de las variables en estudio, lo que permite, en este

caso en particular, que la inercia de las primeras dos variables pase de 7,74 % y 4,67 % a 20,32 % y 11,56 %, respectivamente. Por lo que, a la hora de analizar las primeras dos coordenadas se explica más del doble de la variabilidad que en las dimensiones originales. Un procedimiento análogo se realizó para los otros tres ejercicios de Análisis de Correspondencia Múltiple.

Gráfico 1



3.2. Análisis de Componentes Principales

Una segunda línea de trabajo se realizó analizando mediante Componentes Principales una selección de variables continuas. En este caso, dado que las variables continuas pertenecen mayoritariamente a la base individuos, se decidió trabajar con ésta.

Se realizaron dos experimentos con ésta técnica: una primera aproximación se basó en tratar de analizar la relación entre los ingresos de los hogares y variables potencialmente correlacionadas, como las horas trabajadas, la antigüedad, la edad, el nivel educativo, entre otras. Por su parte, una segunda aproximación buscó ver la interacción entre los distintos tipos de ingresos de los hogares. La base registra los ingresos laborales provenientes de la ocupación principal y de ocupaciones secundarias, así como de diversos tipos de ingresos no laborales, tales como son las jubilaciones, indemnizaciones por despido, subsidios, becas, ganancias, prestamos, desahorro, etc. Dado que una alta correlación entre las variables es necesaria para que la reducción de la dimensionalidad mediante componentes principales sea efectiva, se decidió testear la significatividad de la correlación entre estas variables. La conclusión a la que se llegó es que no hay una correlación significativa entre las variables utilizadas en el segundo experimento, producto de que los hogares tienden a concentrar sus ingresos en unas pocas fuentes, y por lo tanto el individuo que, por caso, obtiene ingresos por subsidios o indemnizaciones, no lo hace por una ocupación principal o por jubilaciones, etc. De esta forma, la tendencia es a tener ingresos nulos de casi todas las fuentes excepto unas pocas, generando relaciones poco significativas entre las variables.

Un segundo enfoque para este ejercicio fue concentrarse en los ingresos laborales, tales como son el sueldo, los

tickets, comisiones, aguinaldos, propinas, etc. e incorporar dos variables ordinales, como son el tamaño de la empresa, y la antigüedad en el cargo. Nuevamente nos encontramos con que la relación entre las variables es no significativa en su gran mayoría, por motivos similares a los que se encontraban en los ingresos no laborales.

Estos análisis preliminares permitieron encontrar aquellas variables que sí mantenían una correlación significativa, y llevaron al tercer enfoque de este segundo ejercicio de Componentes Principales, que fue el efectivamente utilizado, donde concentraron los ingresos en los grupos de Ingreso Laboral de la ocupación principal, ingreso laboral de ocupaciones secundarias e Ingreso No Laboral, a los que se suman las variables antigüedad, horas trabajadas y tamaño del establecimiento.

De esta forma, si bien ambos ejercicios refieren a los ingresos y tienen variables en común, el primero apunta a analizar las características de los individuos respecto de sus ingresos, mientras que el segundo apunta a las diferentes formas de ingresos en relación a las características del puesto. La idea de realizar dichos ejercicios por separado también se basa en que al no tener una fuerte correlación entre las variables, trabajar con un número reducido de las mismas permite concentrar una buena cantidad de la variabilidad en los dos primeros componentes (49,3 % y 55,6 % respectivamente). Esto no quita que, dado que las diferencias en la variabilidad explicada entre el segundo y tercer componente del primer ejercicio eran bajas, se decidió seleccionar los primeros tres componentes para la presentación de resultados.

3.3. Clustering

Las técnicas de clustering se utilizaron como un segundo nivel de análisis sobre los biplots de Análisis de Correspondencia Múltiple, dividiendo entre dos o tres clusters, según el caso. El procedimiento consistió en utilizar el algoritmo k-means sobre las coordenadas de las observaciones obtenidas mediante el análisis de correspondencia múltiple, eligiendo diez puntos aleatorios de partida para los cluster, de tal manera de evitar los posibles máximos locales.

3.4. Logit

Se realizó también un pequeño ejercicio de modelado mediante una regresión logística, donde se buscó estudiar la relación entre la cantidad de horas trabajadas por las personas en la semana de referencia, respecto de si éstas tuvieron intenciones o no de trabajar más horas en dicha semana. Este ejercicio tiene una relación con el concepto de subocupación, que son aquellas personas que, trabajando menos de 35 horas semanales, desean trabajar más horas. El interés radicó sin embargo, no en analizar dicho concepto, sino en estudiar el *riesgo* de que una persona no quiera seguir trabajando más horas, a lo largo del rango de horas trabajadas, y ver como esta distribución se comporta de manera diferente según el género.

3.5. El uso de ponderadores de replicación

Dado que el objetivo central de la Encuesta Permanente de Hogares es realizar estimaciones periódicas sobre el mercado de trabajo y las condiciones de vida de la población Argentina, la base cuenta con un set de ponderadores que permiten proyectar los resultado muestrales al total de la población urbana. La ponderación de cada individuo de la base es la inversa de la probabilidad de ser seleccionado en la muestra. Es importante mencionar que no existe un diseño endógeno en la muestra que busque explícitamente sobrerepresentar a una porción de la población, como sí es el caso de otras encuestas donde se busca reducir la variabilidad de un subconjunto específico de la población¹. También se debe resaltar que los ponderadores se construyen con el objetivo de realizar estimaciones respecto de la población en indicadores del mercado de trabajo, y por lo tanto no fueron pensados desde su diseño para poder expresar la interacción entre las variables, que es lo que se realiza en el presente trabajo.

Por su parte, no se encuentra en la bibliografía una respuesta única respecto al uso de ponderadores en

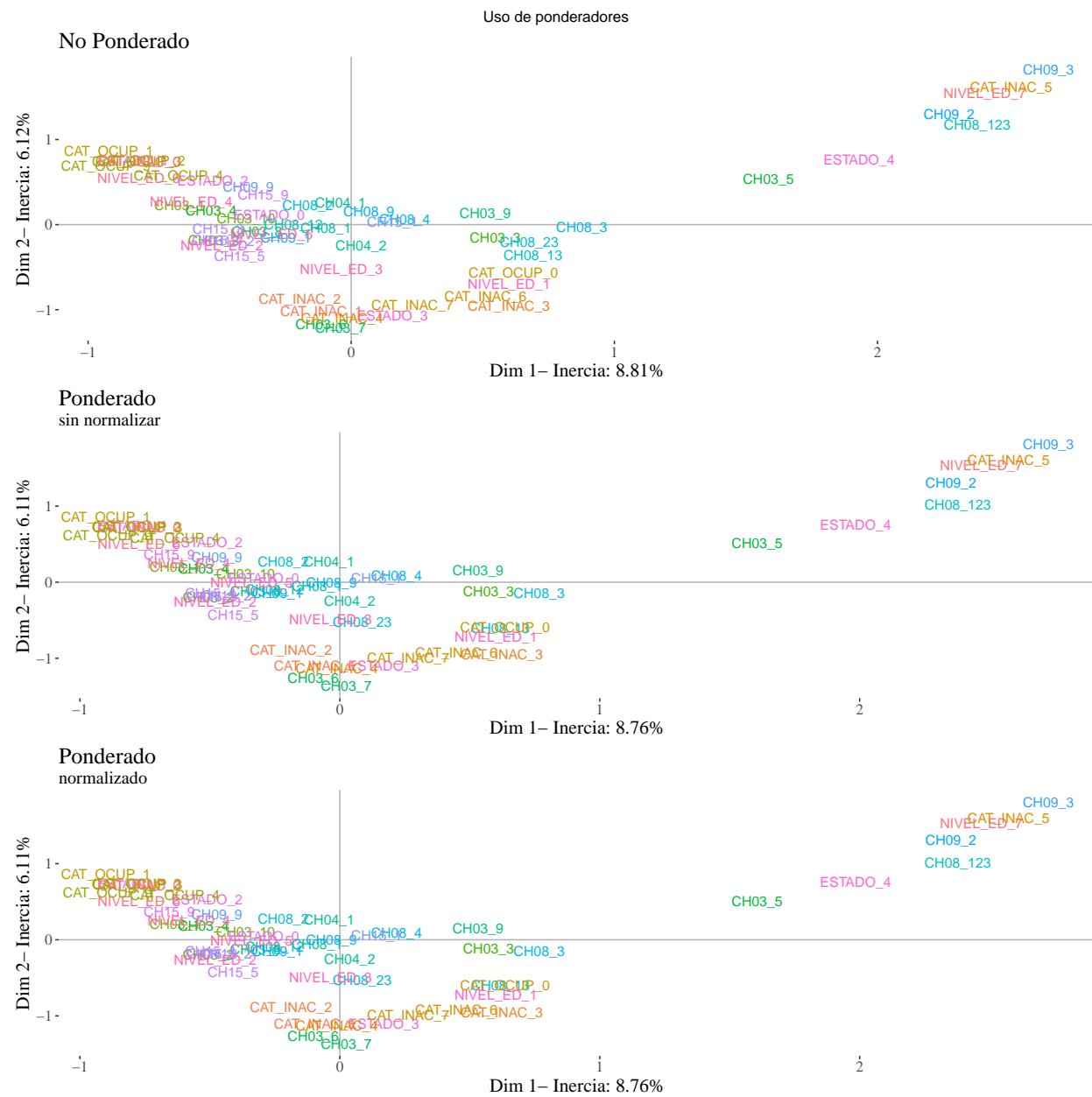
¹Tal es el caso de la *Current Population Survey* de Estados Unidos que sobrerepresenta los estados con menor población (Solon, Haider, y Wooldridge 2015)

diseños muestrales complejos(Young, s.f., Le Roux, Brigitte y Rouanet (2010), Cuadras (2014), Peña (2002) y Hastie, Tibshirani, y Friedman (2009)). Sin embargo, sí existe un extenso debate, aunque no consenso, respecto del uso de ponderadores en las regresiones lineales(Gelman 2007). En particular, Solon, Haider, y Wooldridge (2015) plantea que sólo en ciertos casos puntuales, como lo son la presencia de heterocedasticidad, o de un diseño endógeno, es correcto el uso de ponderadores, que por lo demás pueden llegar a generar una mayor variabilidad en los datos sin corregir sesgo alguno. Por su parte Pfeffermann (2011) plantea una serie de test para verificar las *condiciones de ignorabilidad*, donde se puede no hacer uso de los ponderadores. Sin embargo, estos test son propios de las regresiones lineales, y no son aplicables al Análisis de Correspondencia Múltiple o al Análisis de Componentes Principales.

Dada la falta de una respuesta clara respecto de la conveniencia del uso de dichos ponderadores, se realizó para algunos de los Análisis de Correspondencia Múltiple tanto una versión ponderada como una no ponderada, y también una ponderada y normalizada (es decir, no utilizando directamente los ponderadores de replicación, si no su versión normalizada), para observar las diferencias.

El Gráfico 2 presenta las tres versiones para el caso de las variables descriptivas de los miembros del hogar.

Gráfico 2



En este Gráfico se observa en primer lugar que es indistinto el uso de ponderadores de replicación y ponderadores normalizados. En segunda instancia, sí se aprecian diferencias entre el biplot que incorpora dichos ponderadores y aquel que no. Sin embargo, estas son, en principio, poco significativas. Esto es razonable dado el carácter aleatorio en la selección de la muestra.

Como conclusión, dado que los ponderadores no fueron diseñados para representar la interacción entre las variables, se continúa en el resto del presente trabajo sin hacer uso de los mismos. Sin embargo, se considera un punto pendiente para trabajos futuros comprender más profundamente la implicancia de su utilización.

3.6. Software utilizado

El presente trabajo fue realizado utilizando el software estadístico *R* (R Core Team 2017). Para la implementación del análisis de correspondencia múltiple se utilizó el paquete *FactoMineR* (Lê, Josse, y Husson 2008), mientras que para el análisis de componentes principales se utilizó el *R base* y el paquete *ggbiplot*(Vu 2015). Por su parte, para el modelo logit se utilizó la implementación del paquete *ggplot*(Wickham 2009). La decisión se basa en que se considera el proyecto *R* como una herramienta particularmente propicia para el análisis estadístico, que no requiere licencias, y que a su vez posee paquetes con las implementaciones necesarias para la aplicación de las técnicas seleccionadas.

4. Principales resultados

4.1. Base Hogar

4.1.1. Estratégias de supervivencia de los hogares

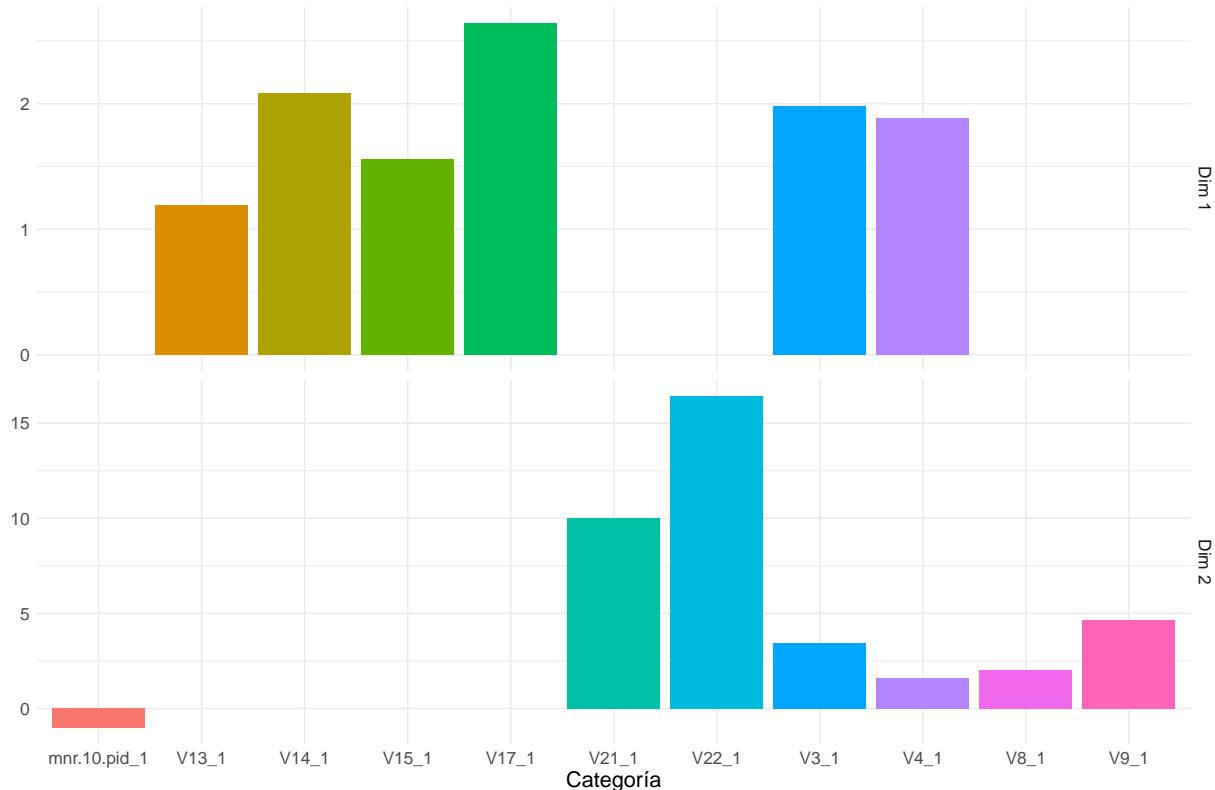
Tabla 1

| Variable | Descripción |
|------------|---------------------------|
| V1 | Trabajo |
| mnr.10.pid | Menor de 10 años pidiendo |
| V13 | Ahorros |
| V14 | Pedir a familiares |
| V15 | Pedir a bancos |
| V17 | vender pertenencias |
| V21 | Aguinaldo jubilación |
| V22 | Retroactivo jubilación |
| V3 | Indemnización despido |
| V4 | Seguro de desempleo |
| V8 | Alquiler |
| V9 | Ganancias |

El Gráfico 3 muestra las cargas de las variables en las primeras dos dimensiones, para aquellas variables que explicaban una mayor proporción de la inercia en el modulo de Estrategias de supervivencia de los hogares.

Gráfico 3

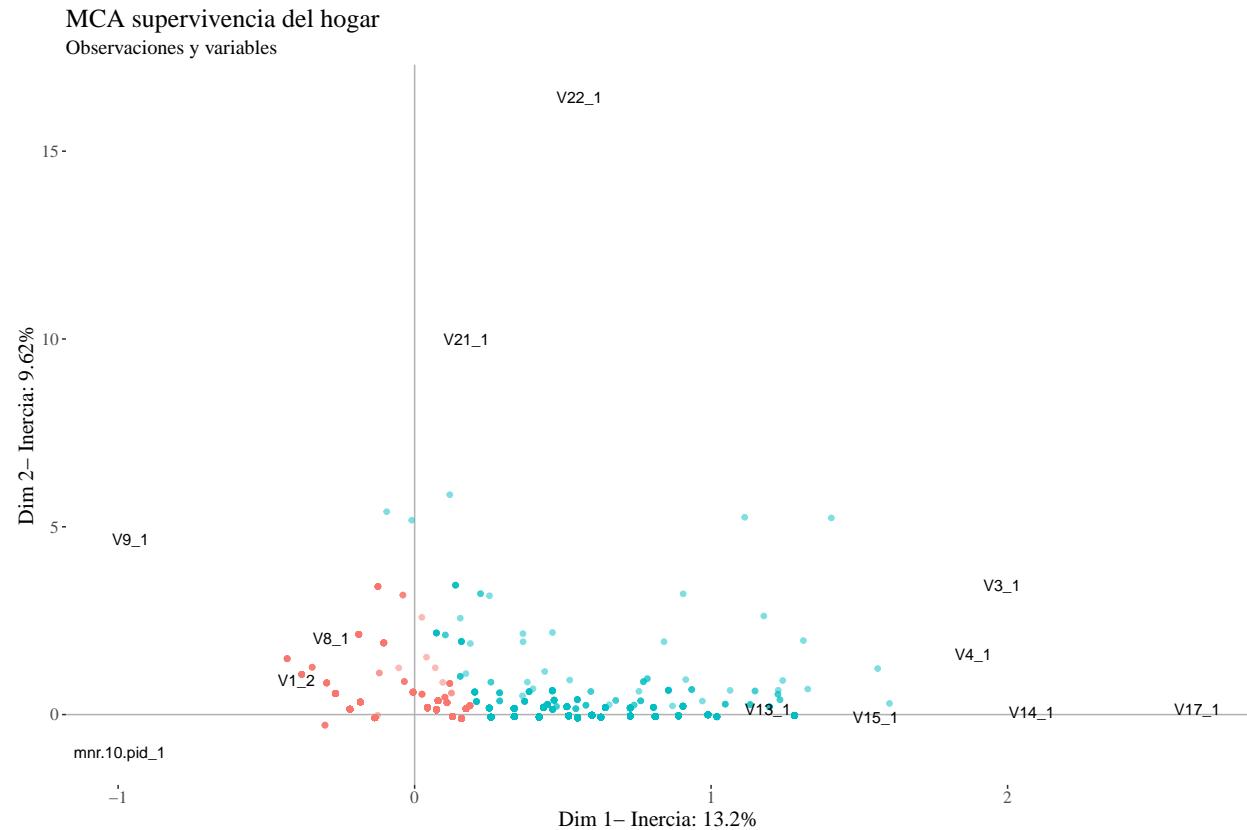
Cargas



Ayudados de las aclaraciones de la Tabla 1, podemos observar en el Gráfico 3 que la primera dimensión es explicada por las estrategias de supervivencia relacionadas con hogares sin una fuente estable de ingresos, como son los ahorros, pedir prestamos, vender pertenencias, indemnizaciones y seguros de desempleo. Por su parte, la segunda dimensión contrasta a los hogares que tienen niños menores de diez años pidiendo, respecto de los ingresos extraordinarios relacionados con las jubilaciones, y los alquileres y ganancias. Es claro el contraste en la segunda dimensión entre los hogares en una peor situación económica y aquellos en una situación económica más holgada. Es importante recordar que las cargas que se observan en el Gráfico 3 son producto de eliminar aquellos pares atributo-valor con menor inercia, y por lo tanto más cercanos al perfil medio. Es decir, son dimensiones que hablan de una porción de la muestra que se aleja del perfil medio.

El gráfico 4 muestra el biplot de variables y observaciones para las coordenadas vistas en el gráfico anterior.

Gráfico 4



En el Gráfico 4 se puede ver una clara separación entre las variables representadas por cada una de las coordenadas. Se observa también como en el segundo cuadrante se ubican quienes no viven del trabajo, y lo hacen de alquileres (más cercano al perfil medio) y Ganancias por empresas (más alejado del perfil medio), mientras que sobre el primer cuadrante se encuentran ordenadas a lo largo del eje horizontal aquellos hogares que sobreviven de sus ahorros, aquellos que en una peor situación económica deben acudir no sólo a desahorrar sino a pedir prestamos a entidades bancarias, quienes no tienen acceso al crédito y deben acceder a pedir a familiares, llegando hasta aquellos que deben vender sus pertenencias, como una situación de carencia extrema. La primera dimensión se puede interpretar como distinguiendo el acceso al mercado de trabajo. A la izquierda se ubican los hogares que no necesitan vivir el ingreso laboral porque tienen otros ingresos provenientes de ganancias y rentas. A la derecha, se ubican los hogares que necesitan ingresos laborales para subsistir, pero no logran ingresar al mercado de trabajo, y deben recurrir a otros mecanismos de subsistencia, como el desahorro, venta de pertenencias, u otros.

Sobre el tercer cuadrante se ubican aquellos hogares con niños que deben salir a pedir. En el otro extremo del eje, están los hogares que tienen retroactivos en la jubilación, es decir, la segunda dimensión discrimina entre hogares con adultos mayores respecto de aquellos con niños pequeños. El cluster realizado divide a la población siguiendo la primera dimensión, separando entre quienes tienen una condición de mayor precariedad en sus ingresos respecto del perfil medio y quienes tienen una mejor situación económica.

Es importante marcar que, si bien parecieran ser pocas observaciones, esto no es en realidad así, dado que los hogares tienden a ubicarse en los mismos lugares, por tratarse de variables categóricas. Aquellos puntos con una mayor transparencia indican una menor cantidad de hogares ubicados en esa tupla específica de pares atributo-valor. Este comentario aplica, naturalmente, para el resto del trabajo.

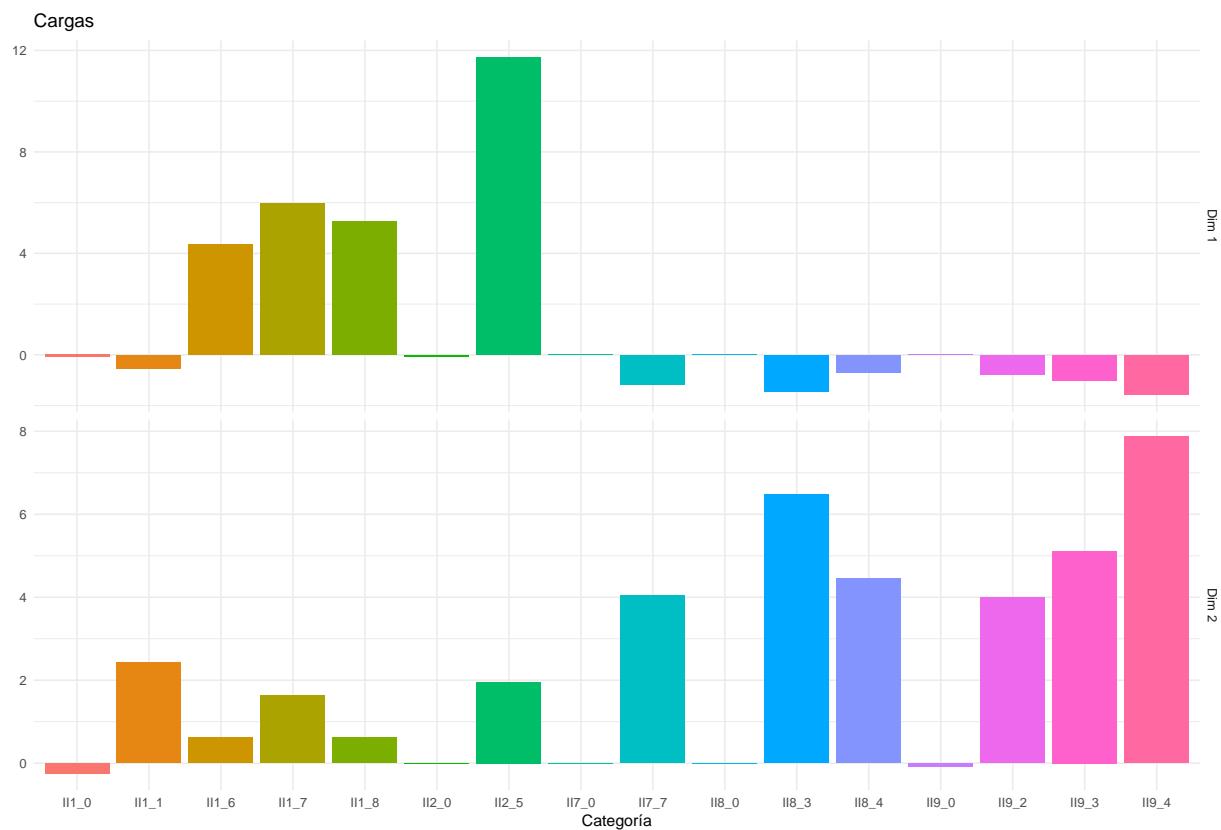
4.1.2. Características habitacionales de los hogares

Tabla 2

| Variable | Descripción |
|----------|--|
| II1 | Ambientes de uso exclusivo |
| II2 | Dormitorios |
| II7_7 | Ocupante, sin permiso |
| II8_3 | cocina a leña/carbón |
| II8_4 | cocina no convencional |
| II9_2 | Baño de uso compartido con otro hogar |
| II9_3 | Baño de uso compartido con otra vivienda |
| II9_4 | No tiene baño |

El Gráfico 5 muestra las cargas de las primeras dos dimensiones de las variables colapsadas para el módulo de características habitacionales de los hogares. Aquellos pares atributo valor que están denominados en 0 corresponden a las variables colapsadas, que se les asignó dicho valor default.

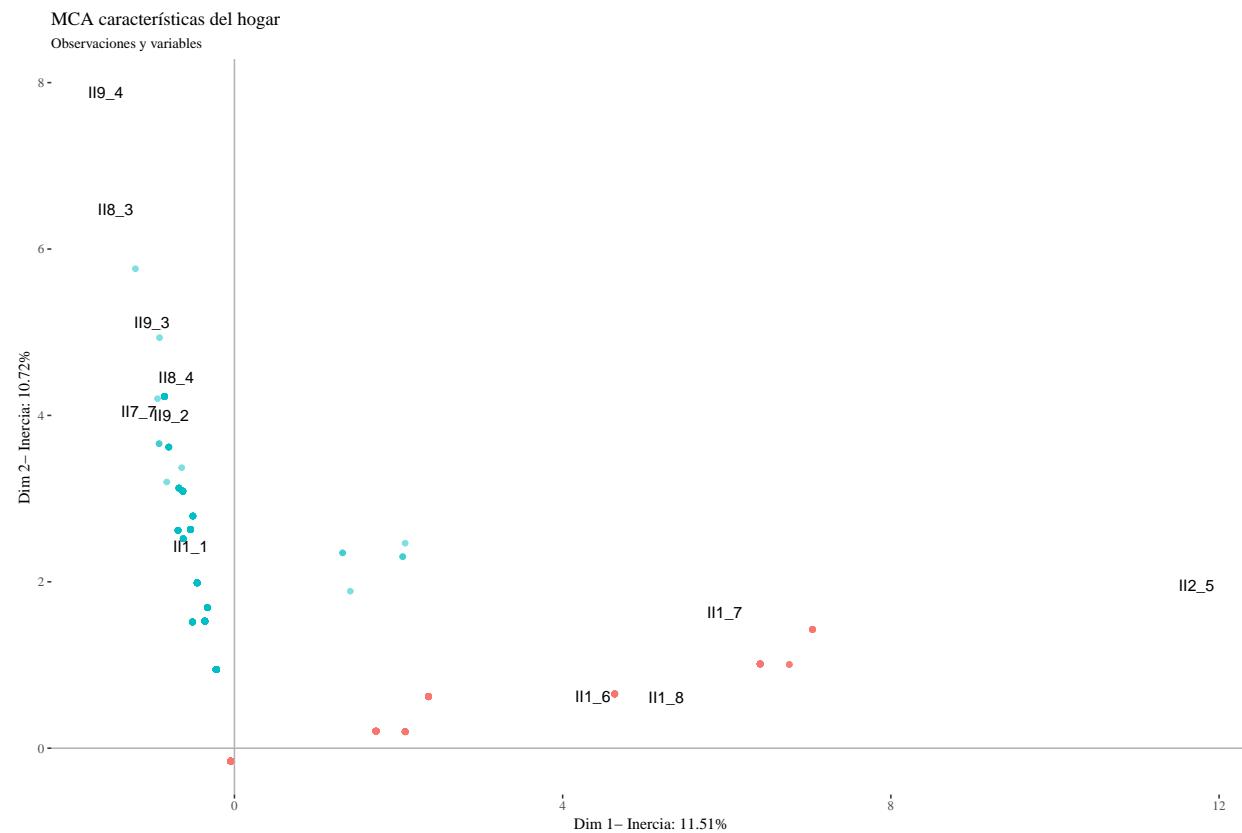
Gráfico 5



En este gráfico se ve como la primera dimensión contrasta entre aquellos hogares con 6,7 y 8 ambientes y 5 dormitorios, respecto de aquellos hogares con uno o ningún ambiente, y con problemas estructurales. Es decir, es un contraste entre hogares con viviendas grandes respecto de aquellos con viviendas más humildes. Por su parte, la segunda dimensión puede ser considerada de fuerza, y da mucha relevancia a los problemas habitacionales en lo que respecta al tipo de cocina y baño utilizado, así como la condición de ocupación de la vivienda.

El Gráfico 6 muestra el biplot asociado a las coordenadas del gráfico anterior.

Gráfico 6



Aquí se observa sobre el segundo cuadrante como aquellos hogares con un dormitorio se encuentran cerca del perfil medio, y a medida que presentan condiciones de mayor carencia respecto a al tipo de cocina y baño utilizado, se alejan de éste, siguiendo el eje vertical. Por su parte, en el segundo cuadrante se ubican los hogares con viviendas de mayor tamaño, donde destacan aquellos que además de tener muchos ambientes, tiene muchas habitaciones. Nótese que el Cluster divide a los hogares entre aquellos con una mejor condición habitacional y aquellos con mayores carencias, y que el perfil medio queda ubicado en el primer grupo. Se puede considerar a la primera dimensión como tamaño de los hogares y a la segunda como precariedad de la vivienda. Se observa la relación entre ambas dimensiones en que los hogares de gran tamaño no tienen condiciones de precariedad, mientras que los de menor tamaño se asocian más a este tipo de problemáticas.

4.1.3. Características habitacionales de las viviendas

Tabla 3

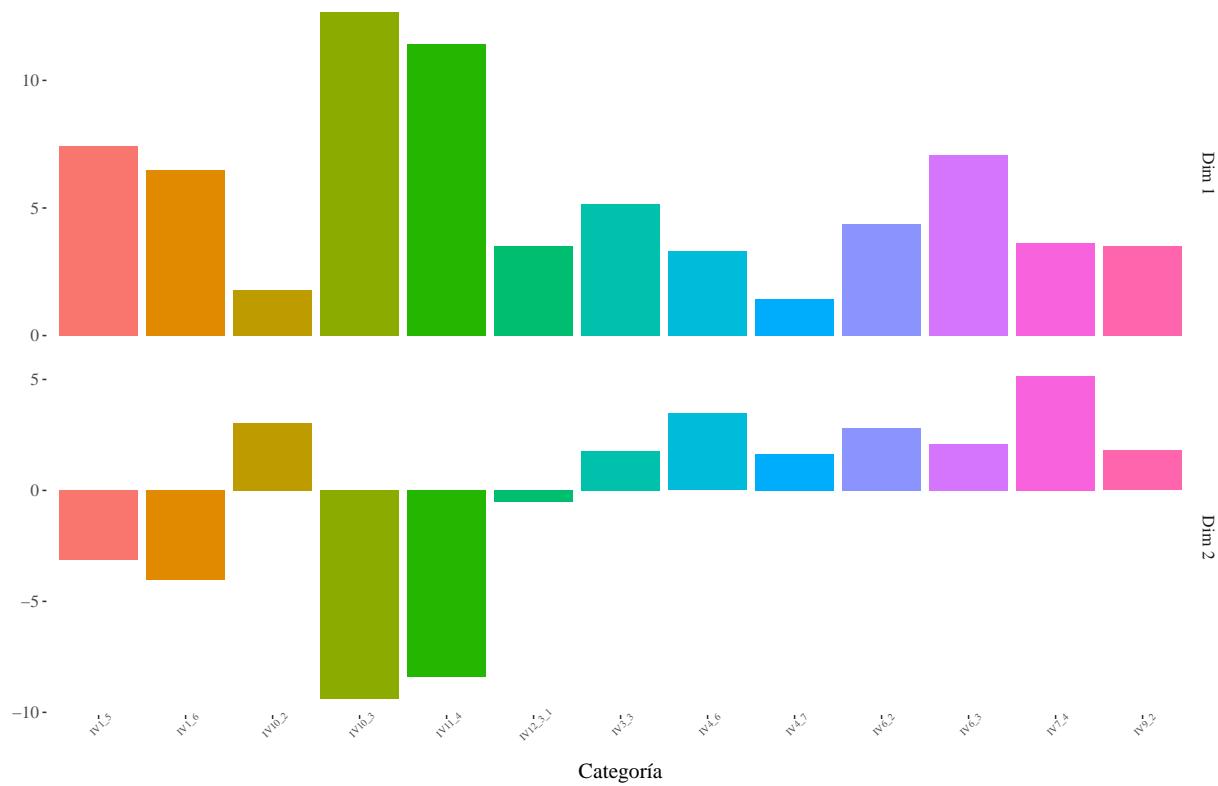
| Variable | Descripción |
|----------|-------------------------------------|
| IV1_5 | Local no construido para habitación |
| IV1_6 | Vivienda no convencional |
| IV10_2 | Inodoro a balde |
| IV10_3 | Letrina sin arrastre de agua |
| IV11_4 | Desague a excavación a tierra |
| IV12_3 | Villa de emergencia |
| IV3_3 | Piso de tierra |
| IV4_6 | Techo de chapa metal |

| Variable | Descripción |
|----------|--------------------------------|
| IV4_7 | Techo de paja |
| IV6_2 | Agua fuera de la vivienda |
| IV6_3 | Agua fuera del terreno |
| IV7_4 | Agua de fuente no convencional |
| IV9_2 | Baño fuera de la vivienda |

El Gráfico 7 presenta las cargas de las primeras dos dimensiones de las variables colapsadas para el módulo de Características habitacionales de las viviendas.

Gráfico 7

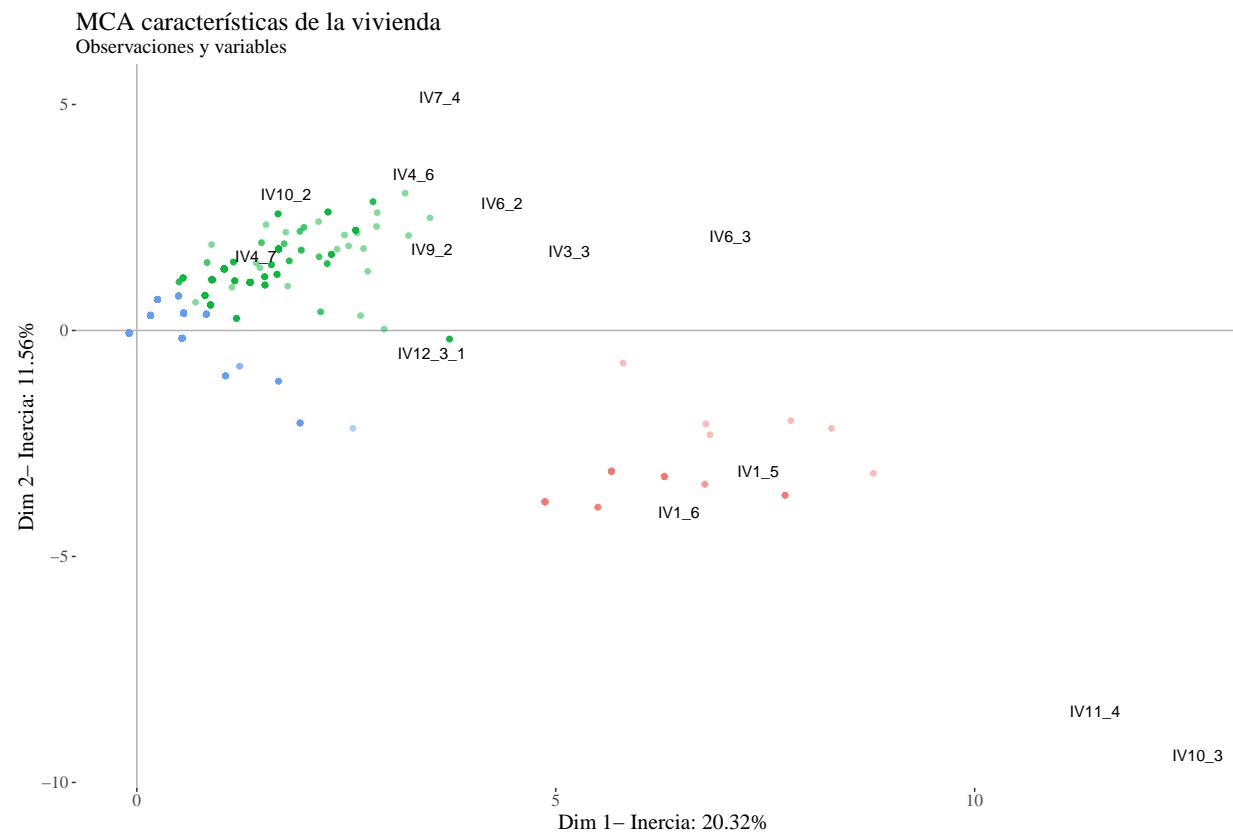
Cargas



En este gráfico se destaca la primera dimensión como de fuerza y la segunda como de forma. Las variables más importantes en ambas dimensiones son respecto del acceso al agua de los hogares. En particular, tienen más inercia las respuestas de aquellos hogares cuyas letrinas no tienen arrastre de agua, y aquellos cuyo desagüe del baño es a una excavación a tierra. La segunda dimensión contrasta a las viviendas no convencionales y sin acceso al agua respecto del resto. En términos generales, las variables que quedaron luego de colapsar aquellas con menor inercia, son de los hogares con distintos tipos de déficit habitacionales.

El Gráfico 8 muestra el biplot asociado a las coordenadas del gráfico anterior.

Gráfico 8



- **Primera dimensión:** Problemas de acceso al agua
 - Es una dimensión de fuerza, a mayor valor, menor acceso al agua
- **Segunda dimensión:** Tipo de Vivienda
 - Valores positivos: casas y departamentos precarios
 - Valores negativos: casillas, construcciones no hechas para habitar, mayormente ubicados en villas de emergencia.
- Se crean tres clusters
 - Azul: Entorno del **perfil medio**: Sin problemas particulares para acceder al agua, casas, departamentos, etc.
 - Verde: **Construcciones precarias**: Pisos de tierra, techos de chapa, pero con algún acceso al agua corriente, no necesariamente ubicados en villas de emergencia.
 - Rojo: **casillas**: construcciones de cilo bolsa, casas precarias, ubicadas en villas de emergencia

4.2. Base individuos

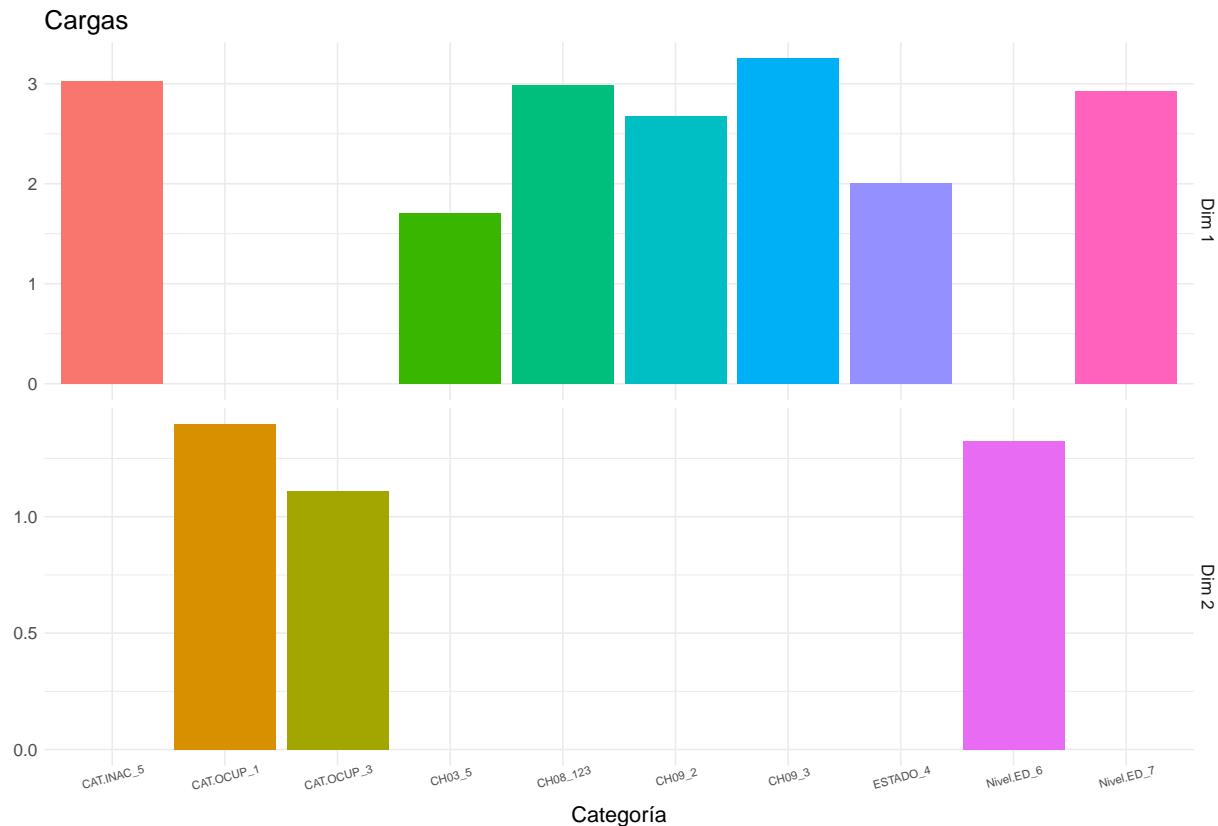
4.2.1. Características de los miembros del hogar

Tabla 4

| Variable | Descripción |
|------------|---|
| CAT.INAC_5 | Menor de 6 años |
| CAT.OCUP_1 | Partón |
| CAT.OCUP_3 | Asalariado |
| CAT.OCUP_9 | Ns.Nr. |
| CH03_5 | Nieto del jefe de hogar |
| CH08_3 | Seguro Público |
| CH08_123 | Tiene obra social, prepaga y Seguro público |
| CH09_2 | No sabe leer y escribir |
| CH09_3 | Menor de 2 años |
| ESTADO_1 | Ocupado |
| ESTADO_4 | Menor de 10 años |
| Nivel.ED_6 | Universitario completo |
| Nivel.ED_7 | Sin instrucción |

El último ejercicio de Análisis de Correspondencia Múltiple se realizó sobre la base de individuos. El Gráfico 9 presenta las cargas de las primeras dos dimensiones de las variables colapsadas para las Características de los miembros de los hogares.

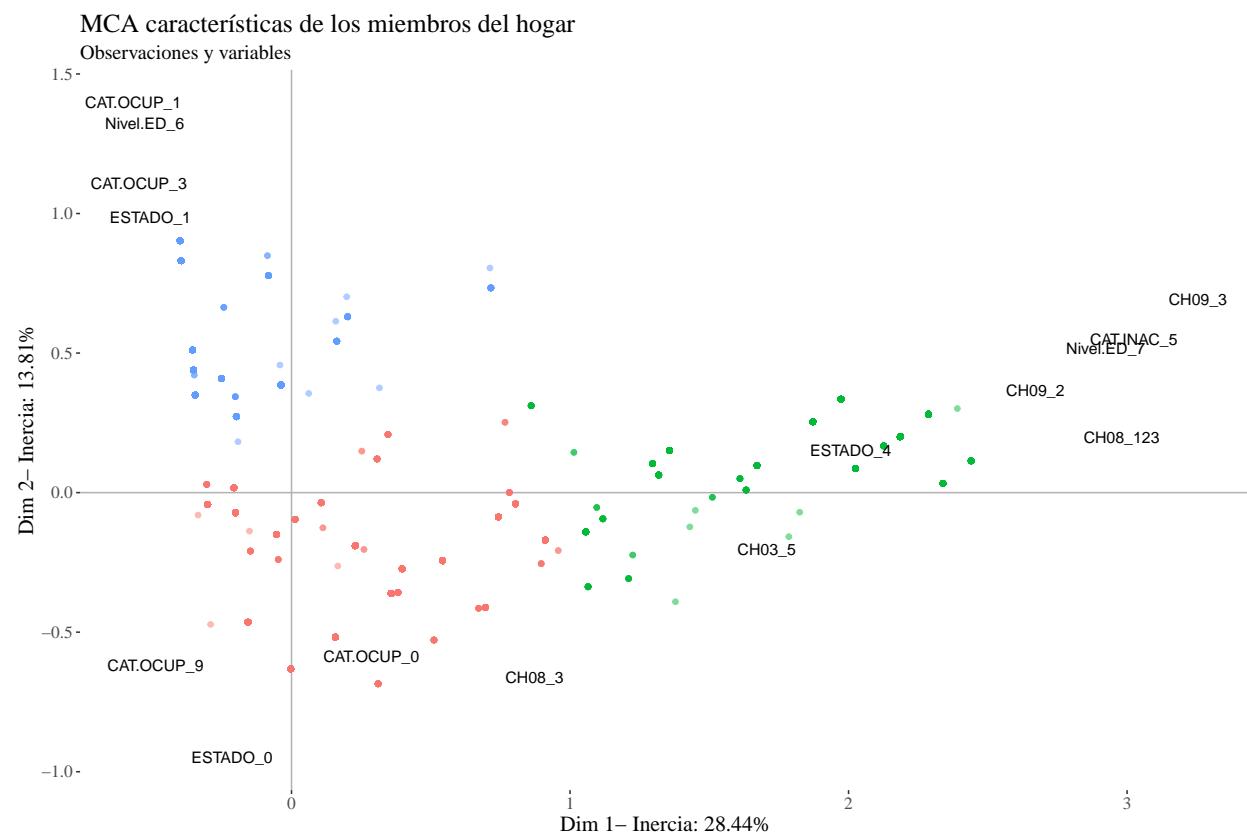
Gráfico 9



En este gráfico se observa como las dimensiones no comparten categorías. En la primera dimensión tienen mayor inercia los menores de edad, nietos del jefe de hogar, y las variables asociadas a estas características, como ser que tengan una cobertura social de varios tipos, dada la mayor protección social sobre los menores, que no tienen instrucción ni saben leer y escribir, etc. La segunda dimensión destaca a los asalariados y patrones, y a aquellos con una formación universitaria completa o superior. Es decir, distingue a la porción de la población con una inserción en el mercado de trabajo.

El Gráfico 10 muestra el biplot asociado a las coordenadas del gráfico anterior.

Gráfico 10



Este biplot divide a la población en tres clusters. El primero (verde) reúne a los menores de edad con las características mencionadas, el segundo (rojo) reúne a otros miembros del hogar, cuyas variables aparecen en 0 porque fueron colapsadas. En dicho grupo se encuentran los demás miembros del hogar que no se encuentran ocupados. En el tercer grupo (azul) se encuentran los asalariados, patrones u ocupados de otro tipo, que son quienes tienen mayor nivel educativo formal. Se puede percibir en este gráfico la estructura familiar de los hogares, desde aquellos que los sustentan económicamente, hasta quienes son más dependientes del mismo, como son los niños pequeños. En el medio, se encuentra el cluster de quienes a la vez no están insertos en el mercado de trabajo, pero se acercan más en su perfil demográfico a quienes sí lo están.

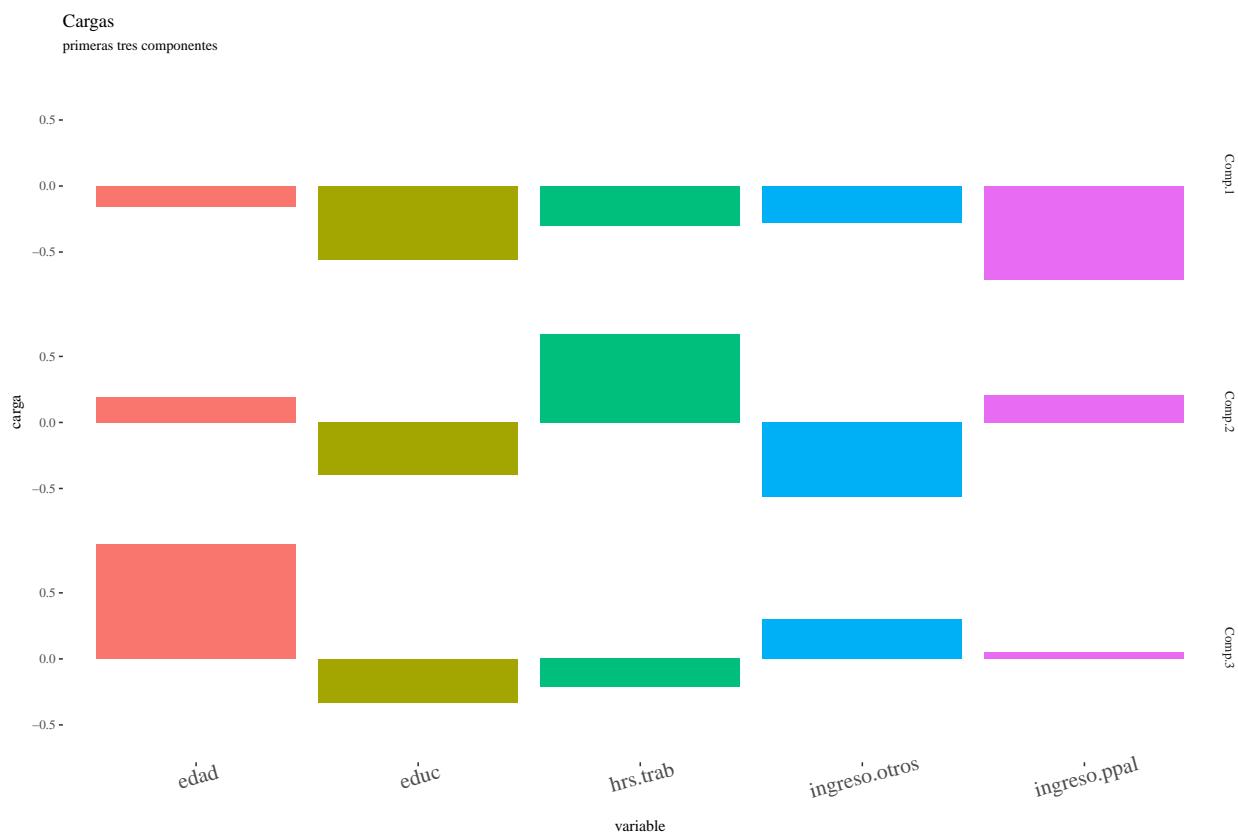
4.2.2. Ingresos

4.2.2.1. Ingresos y nivel educativo.

El primer ejercicio realizado de Análisis de Componentes Principales reúne a las variables de ingresos laborales por ocupación principal, otros tipos de ingresos no laborales, junto con el nivel educativo alcanzado y la edad.

El Gráfico 11 muestra las cargas de las primeras tres componentes principales.

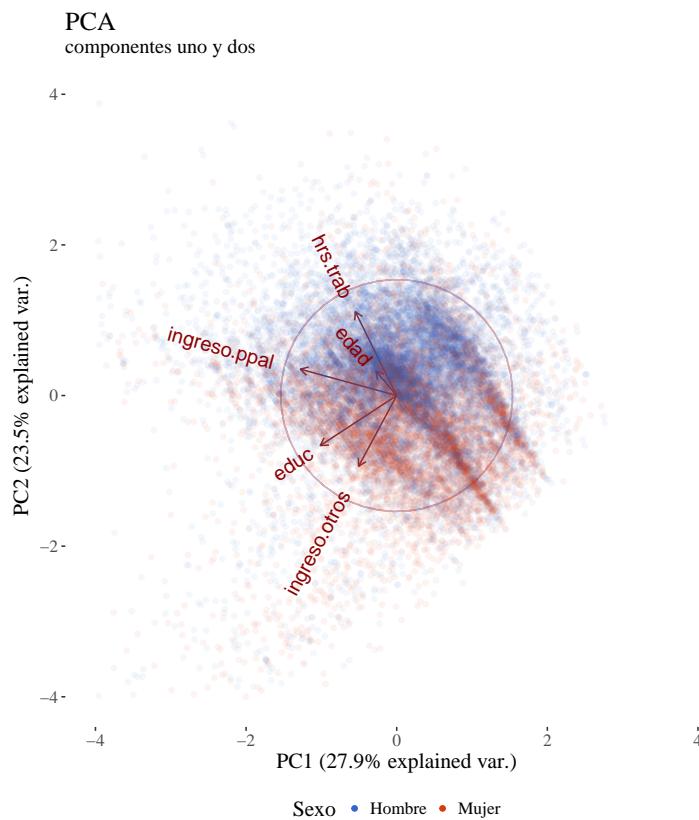
Gráfico 11



Aquí se observa que la primera componente es de fuerza, y lo más importante es el nivel educativo y el ingreso por ocupación principal, mientras que la segunda componente contrasta las horas trabajadas en la ocupación principal respecto de la educación y los ingresos no laborales. Por su parte la tercera componente contrasta la edad con la educación y las horas trabajadas.

El Gráfico 12 muestra el biplot asociado a las primeras dos componentes, y en color el género.

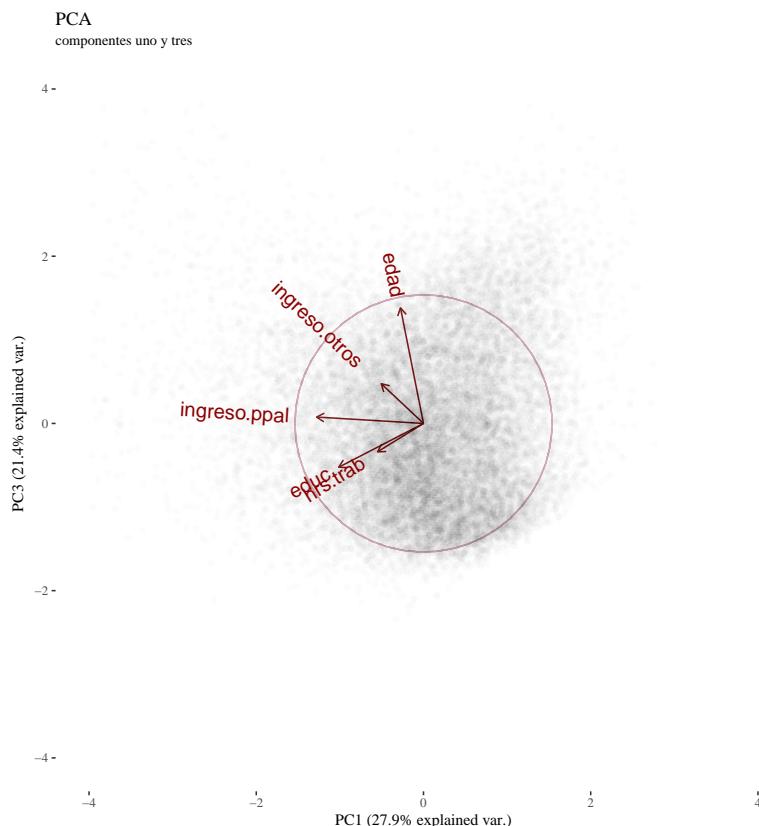
Gráfico 12



En el Gráfico 12 se observa que las variables edad y horas trabajadas en la ocupación principal están muy asociadas entre sí, a la vez que se asocian negativamente con ingresos no laborales. Sin embargo, es necesario remarcar que la variable edad no explica una porción importante de la variabilidad de estas componentes. Por su parte, resulta interesante marcar como la segunda componente principal pareciera dividir mejor el género que la primera, por lo que se puede asociar parcialmente al género masculino con las horas trabajadas en la ocupación principal, y a las mujeres con los ingresos derivados de otro tipo de ocupaciones no laborales. En suma, se puede intuir de este gráfico la estructura patriarcal, con un jefe de familia varón, que trabaja más horas, y la mujer obteniendo ingresos de tipo no laboral. Por su parte, la educación, en estas dos primeras componentes, se asocia negativamente a la cantidad de horas trabajadas, y se asocia más fuertemente al género femenino.

El Gráfico 13 muestra el biplot asociado a la primera y la tercera componente principal.

Gráfico 13



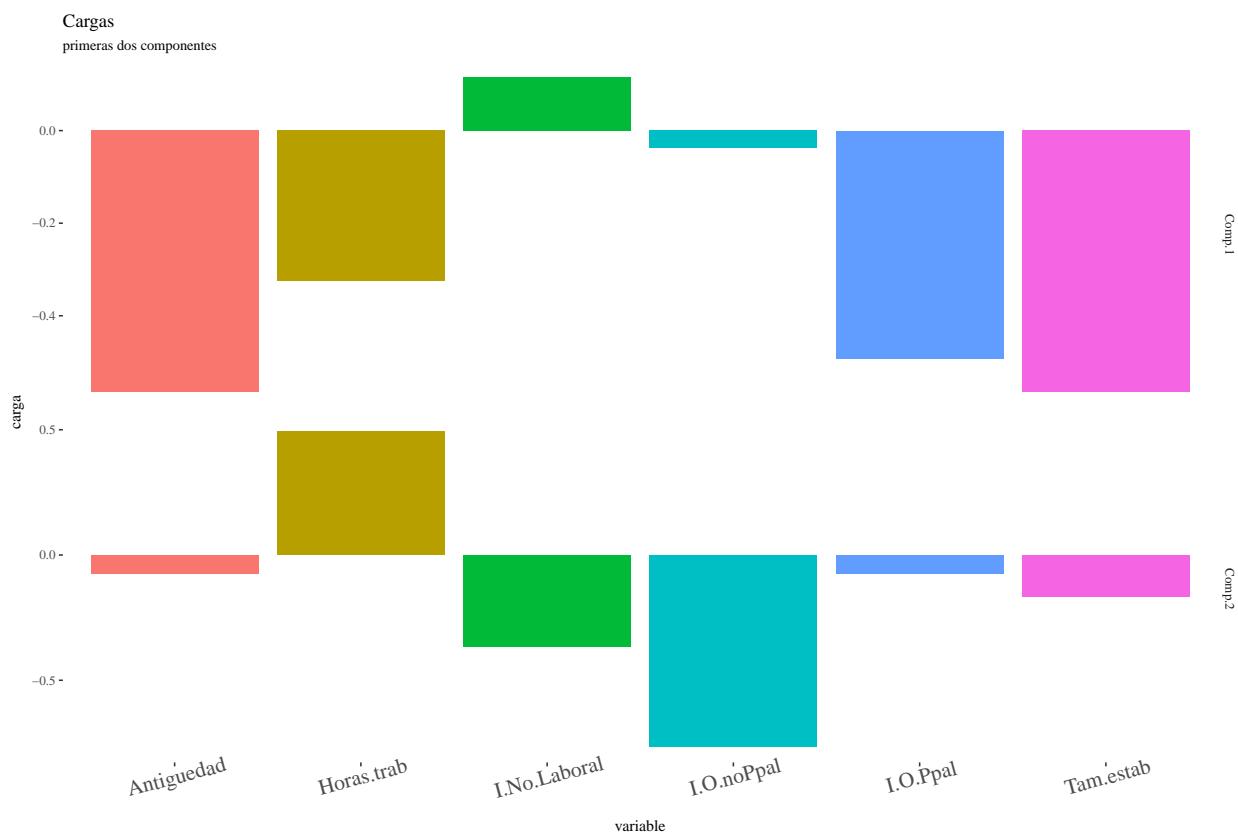
En este gráfico se destaca la asociación entre la educación y las horas trabajadas, y éstas asociadas a los ingresos por la ocupación principal. La relación observada entre edad y horas trabajadas en el Gráfico 12 se pierde en éste biplot, donde la relación es negativa. Así también, la relación negativa entre horas trabajadas y educación que se observaba en el gráfico anterior, ahora es positiva. Las tres componentes con mayor variabilidad explicada son edad, ingreso de la ocupación principal y educación. La edad es independiente de ingreso por ocupación principal.

4.2.2.2. Tipos de ingreso

El segundo ejercicio realizado de Análisis de Componentes Principales procura explorar la relación entre el lugar de trabajo, el rol del individuo y sus ingresos. Para ello, toma variables propias del puesto de trabajo, como el tamaño del establecimiento, la antigüedad en el puesto de trabajo y las horas trabajadas; y variables respecto de los distintos tipos de ingreso: provenientes de la ocupación principal, de una ocupación secundaria y los ingresos no laborales.

El Gráfico 14 muestra los loadings de los dos primeros componentes principales para las variables mencionadas.

Gráfico 14

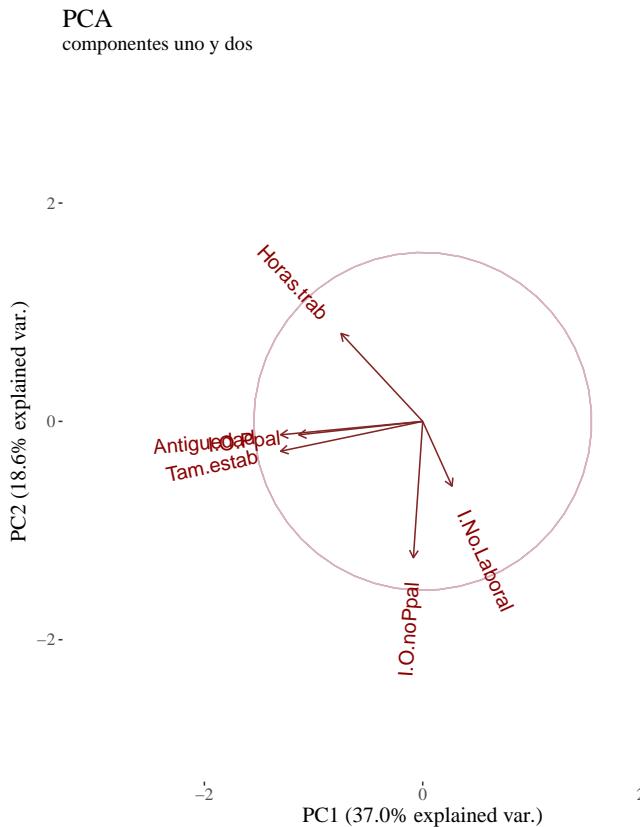


En este gráfico se observa como el primer componente contrasta a los individuos con ingresos no laborales respecto del resto. En este sentido, la primera componente genera una división entre aquellos hogares insertados en el mercado de trabajo, respecto de los hogares que se encuentran fuera del mismo.

Por su parte, el segundo componente contrasta a quienes trabajan muchas horas de aquellos que tienen ingresos por ocupaciones secundarias. Vale recordar que la pregunta de horas trabajadas es exclusiva de la ocupación principal. En este sentido, el segundo componente contrasta a aquellos individuos que tienen una única ocupación de tiempo completo, respecto de quienes tienen varias ocupaciones, y un ingreso de una ocupación secundaria, y de quienes no están insertos en el mercado de trabajo.

El Gráfico 15 muestra el biplot asociado a los componentes descritos arriba.

Gráfico 15



En el Gráfico 15 se ve una asociación negativa entre las horas trabajadas en la ocupación principal y los ingresos de otras fuentes, lo cual es natural porque indica que quien dedica muchas horas a su ocupación principal, no puede tener otro tipo de ocupaciones. Por su parte, aparece una fuerte correlación positiva entre antigüedad, tamaño del establecimiento y los ingresos por ocupación principal. Es decir, quienes tienen trabajos estables, en empresas de mayor tamaño tienen también mayores ingresos. Es destacable esta asociación respecto de las horas trabajadas, que se asocian débilmente a los ingresos. Es decir, lo relevante para los ingresos altos esta más relacionado, en estas primeras dos componentes, a las características del puesto más que a la carga horaria del mismo, aunque este último se asocie negativamente con la posibilidad de tener otro tipo de ingresos.

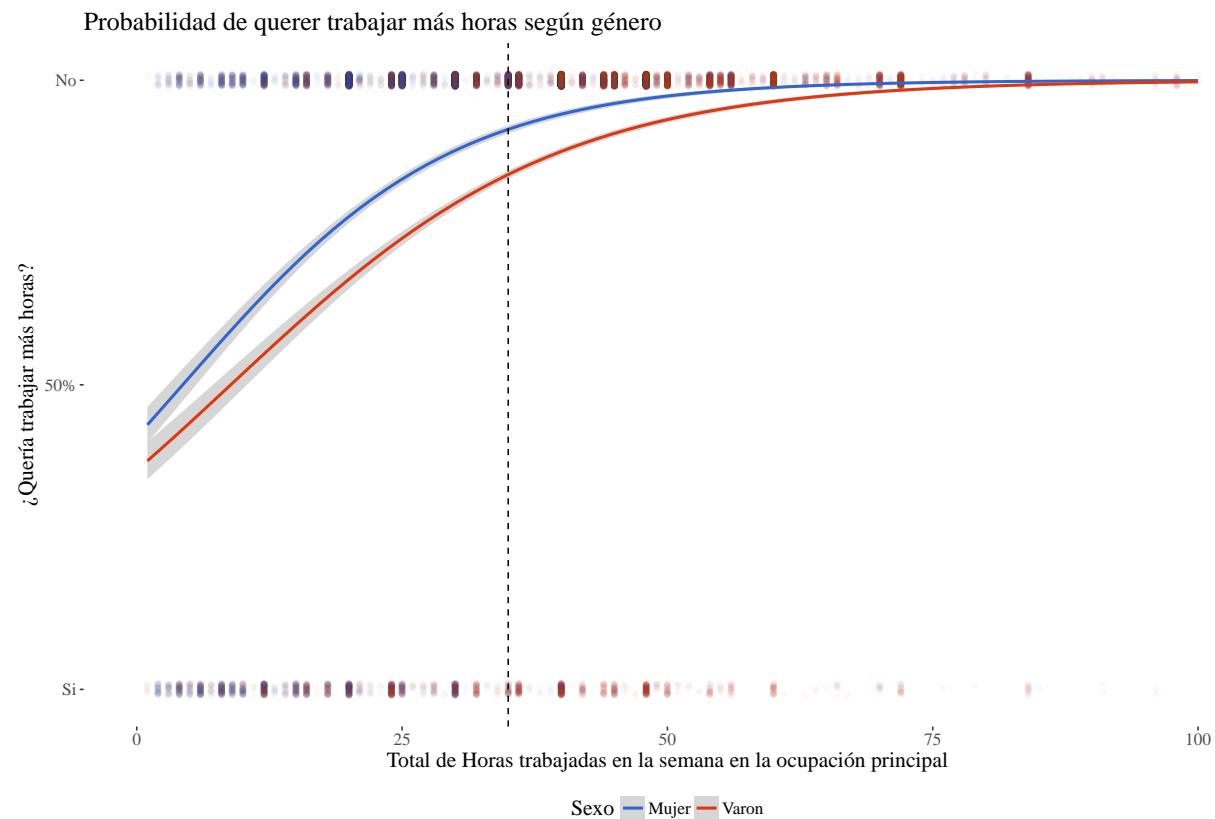
4.2.3. Tiempo de trabajo

Por último, el Gráfico 16 presentan los resultados de ajustar la respuesta respecto a las intenciones de trabajar más cantidad de horas en la semana de referencia respecto a la cantidad de horas efectivamente trabajadas, mediante un modelo logit. Allí se observa como la probabilidad de decir que no se quería trabajar más horas aumenta a medida que el individuo trabajó más horas. En el gráfico se marca con una línea punteada las 35 horas trabajadas, en referencia a la definición de subempleo utilizada en los indicadores de la Encuesta Permanente de Hogares, donde se considera subempleado a aquellos individuos que, trabajando menos de 35 horas semanales, desean trabajar más horas. Por lo tanto, todos aquellos individuos que, ubicados a la izquierda de la línea punteada, respondieron afirmativamente, graficados sobre el valor 'Si', son considerados subocupados. Por su parte, en el gráfico se modeliza la respuesta por separado para hombres y para mujeres, así como los puntos que representan a los individuos también designan el género con el color. Como se observa, la curva de la probabilidad de la respuesta negativa es siempre mayor en las mujeres que en los hombres para toda la distribución de horas trabajadas. También se observa una predominancia del género femenino en los

valores más pequeños del eje x.

La pregunta que motiva esta modelización es respecto a cómo el género del individuo produce en la sociedad una diferenciación en el rol que éste ocupa en el mercado de trabajo. Alejándose de la interpretación estricta del modelo, lo que se puede apreciar es que la mujer ocupa en la sociedad un lugar en el cual debe tener una menor disposición a la inserción y permanencia en el mercado de trabajo respecto del hombre, porque tiene a cargo las tareas no remuneradas al interior de los hogares. Lo interesante del Gráfico 16 es que muestra que esta desigualdad se conserva para todo el rango de horas trabajadas. Si bien es conocido que las mujeres tienen mayores dificultades para ingresar al mercado de trabajo porque comúnmente son las responsables al interior del hogar de las tareas domésticas, también se observa que aquellas mujeres que logran ingresar al mercado de trabajo, siguen haciéndose cargo de dichas tareas y deben tener una mayor disposición horaria para realizarlas. La diferenciación por género en el mercado de trabajo no constituye únicamente una barrera de entrada, sino que continúa una adentro del mercado de trabajo.

Gráfico 16



5. Conclusiones

El presente trabajo se propuso realizar una primera aproximación respecto de las posibilidades y los límites de las técnicas abordadas en el curso Análisis Inteligente de Datos para el estudio de las estadísticas sociales. Para ello, se realizó una serie de ejercicios con diferentes técnicas, en distintos módulos de la Encuesta Permanente de Hogares. Los resultados obtenidos muestran un gran potencial explicativo, si bien no escapan a las dificultades tales como el interrogante respecto del uso de ponderadores de replicación, o la poca correlación presente en ciertos conjuntos de variables. Muchos de los resultados obtenidos padecen de cierta trivialidad, ya que ponen a la luz relaciones previamente conocidas, como es el caso de las condiciones habitacionales en los barrios precarios. Sin embargo, logran aportar cierta ordinalidad en los resultados, lo cual no es de por sí auto evidente. Por su parte, si bien la pregunta que guió el trabajo versaba respecto de las potencias de las

técnicas vistas en las estadísticas sociales y no respecto de los contenidos de dicha disciplina, se realizó una primera incursión respecto del rol de la mujer en el mercado de trabajo, tanto en el Análisis de Componentes Principales respecto del ingreso, como en el modelo logit sobre disponibilidad a la ocupación laboral. En ambos ejercicios el resultado aportó evidencia sobre de la desigualdad de género que existe aún hoy en el mercado de trabajo.

Futuros trabajos

El presente trabajo se planteó ser un punto de partida en un campo de investigación mayor, y por lo tanto las líneas de desarrollos futuros son amplias. En primer lugar, los resultados positivos obtenidos plantean la necesidad de realizar nuevos trabajos, cuyas preguntas guía sean propias del contenido y no de la herramienta. En este sentido, se plantea el desafío de ver la aplicabilidad de los métodos estudiados cuando se tiene una pregunta concreta para responder respecto del mercado de trabajo, las condiciones de vida, u otros posibles análisis. Por su parte, la problemática respecto al uso de los ponderadores tuvo sólo una respuesta parcial y requiere de estudios posteriores que tomen el tema en forma exclusiva para poder obtener respuestas más específicas en el asunto. Por su parte, no todas las técnicas vistas a lo largo del curso fueron utilizadas, por lo que este trabajo puede extenderse para abarcar a las mismas. En particular, son abundantes los trabajos en las estadísticas sociales que requieren de test de contraste de medias de diferentes tipos, así como el análisis de independencia y homogeneidad de las variables. Estos métodos, a diferencia de el Análisis de Correspondencia Múltiple y del Análisis de Componentes Principales, son utilizados de forma habitual en este campo de estudio.

Como desarrollos concretos se plantea a futuro realizar nuevas aperturas por quintiles de ingreso, así como el uso de bases de datos de otros períodos para observar posibles cambios estructurales en la población. Siguiendo la línea respecto del rol de la mujer en el mercado de trabajo, la modelización realizada en este trabajo puede ser complementada con un análisis del módulo *Organización del hogar* de la EPH, donde se pregunta respecto a quién realiza las tareas domésticas en el hogar.

Bibliografía

- Cuadras, C M. 2014. *Métodos de análisis multivariante*. doi:10.1017/CBO9781107415324.004.
- Gelman, Andrew. 2007. «Struggles with Survey Weighting and Regression Modeling». *Statistical Science* 22 (2): 153-64. doi:10.1214/088342306000000691.
- Hastie, Trevor, Robert Tibshirani, y Jerome Friedman. 2009. *The Elements of Statistical Learning*. Vol. 1. doi:10.1007/b94608.
- INDEC. 2016. «Encuesta Permanente de Hogares. Diseño de registro 4to trimestre 2016», 17.
- . 2017. «Indicadores de coyuntura. Fichas técnicas».
- Le Roux, Brigitte, y Henry Rouanet. 2010. *Quantitative applications in the social sciences*. Vols. 07-163.
- Lê, Sébastien, Julie Josse, y François Husson. 2008. «{FactoMineR}: A Package for Multivariate Analysis». *Journal of Statistical Software* 25 (1): 1-18. doi:10.18637/jss.v025.i01.
- Loretan, Mico. 1997. «Generating market risk scenarios using principal components analysis: methodological and practical considerations». *Federal reserve board*, 23-60. <https://www.bis.org/publ/ecsc07b.pdf>.
- Peña, Daniel. 2002. *Análisis de datos multivariantes*. December. doi:8448136101.
- Pfeffermann, Danny. 2011. «Modeling Sampling Data of when». *International Statistical Review* 61 (2): 317-37.
- R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.r-project.org/>.
- Solon, Gary, Steven J Haider, y Jeffrey Wooldridge. 2015. «What Are We Weighting For?» *Journal of Human Resources* 50 (2): 301-16. doi:10.3386/w18859.
- Vu, Vince. 2015. «ggbiplots». *GitHub repository*. <https://github.com/vqv/ggbiplots>.
- Wickham, Hadley. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <http://ggplot2.org>.
- Yang, Libin. 2015. «An Application of Principal Component Analysis to Stock Portfolio Management». Tesis doctoral.
- Young, Rebekah. s.f. «Survey Sample Weights».