Quaglino, Marta
Merello, Juliana
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística

MÉTODOS MULTIVARIADOS EN ESTUDIOS DE VULNERABILIDAD SOCIAL EN LA PROVINCIA DE SANTA FE

1. Introducción

En las últimas décadas, el uso de información censal para el estudio de la pobreza se ha convertido en un recurso fundamental para orientar la formulación de políticas sociales y facilitar la racionalización y optimización del gasto social. La fuente censal para localizar geográficamente las situaciones de privación así como para su caracterización han representado una motivación para el desarrollo de indicadores de la pobreza de los hogares. La Comisión Económica para América Latina (CEPAL), en los años ´70, propuso la aplicación de la metodología de Necesidades Básicas Insatisfechas (NBI) teniendo como principal objetivo identificar hogares y personas que no alcanzaran a satisfacer un conjunto de necesidades consideradas indispensables según niveles de bienestar aceptados como universales. La adopción de esta forma de medición permitió canalizar la inquietud por explotar la riqueza de la información censal mediante mapas de pobreza con un amplio nivel de desagregación geográfica, al tiempo que la incidencia de la pobreza por NBI se presentó como una alternativa en el caso de fuentes que no indagan ingresos de la población.

Sin embargo, la definición de pobreza se encuentra, en general, limitada a lo cuantitativo y esta definición condiciona cultural y operativamente para pensar de manera más realista y relevante el problema. Es necesario asumir una concepción integral o sistemática de las necesidades humanas donde se reconozca como tales no sólo a aquellas que comúnmente se caracterizan como necesidades básicas u obvias: salud, trabajo, vivienda, educación, alimentación, etc., sino también a un complejo de necesidades no tan obvias tales como: ser protagonista de la propia historia o necesidades que interactúan entre sí.

Pensar en la línea de vulnerabilidad – entendida como “fragilidad”, “indefensión” o “desamparo” – permitiría una aproximación más dinámica de la diversidad de situaciones, las que de una manera u otra son partícipes de algún proceso de “privación”. Vulnerabilidad

implica centrar la problemática, también, en los derechos civiles, políticos y sociales lo que permite a su vez reconceptuar las políticas públicas para moverse en la consideración de las necesidades como derechos. Además permite referirse a aquella diversidad de situaciones que entrarían en el espacio de la exclusión.

Por lo tanto "vulnerabilidad" no es lo mismo que pobreza pero la incluye, implica la posibilidad de padecerla en el futuro a partir de condiciones del presente. Es un proceso multidimensional que confluye en el riesgo de un grupo o individuo de ser dañado ante cambios o permanencia de situaciones externas o internas. Por lo tanto implica dos condiciones: los "vulnerados" es decir los que la padecen y podría aquí también entenderse como pobreza y la de los "vulnerables" en quienes el deterioro de las condiciones de vida no está materializado sino que aparece como una situación de alta probabilidad a partir de sus condiciones de fragilidad.

El nivel de vulnerabilidad depende de varios factores que se relacionan. Por un lado, los de orden social y, por el otro, el de los recursos y estrategias que disponen los individuos, hogares o comunidades. La CEPAL, define vulnerabilidad social como aquella que se relaciona con los grupos socialmente vulnerables, cuya identificación obedece a diferentes criterios: según factor contextual que los hace más propensos a enfrentar circunstancias adversas para su inserción social y desarrollo personal, el ejercicio de conductas que entrañan mayor exposición a eventos dañinos o a la presencia de atributos básicos compartidos (edad, sexo, etnia) que se supone les confiere riesgos o problemas comunes.

La identificación de situaciones de vulnerabilidad, su cuantificación y su localización en el territorio es un análisis imprescindible tanto para el diseño de las nuevas acciones o intervenciones como para evaluar los efectos de las políticas públicas implementadas.

En cuanto a la estimación de la vulnerabilidad, en particular en la Provincia de Santa Fe existen varios indicadores que intentan mostrar su realidad en la región. En la búsqueda de información para la formulación de políticas sociales son necesarias definiciones de nuevas categorías. Al ser este un fenómeno multidimensional, se planteó la necesidad de abordarlo desde una óptica multivariada. El objetivo de este trabajo es presentar algunos resultados parciales obtenidos por las autoras[1] al analizar distintas estrategias metodológicas para obtener grupos diferenciados de radios censales de la Provincia de Santa Fe de

[1] La investigación corresponde a la tesina de la Licenciatura en Estadística realizada por la Lic.Juliana Merello (Estadística del IPEC) bajo la dirección de la Dra.Marta Quaglino

acuerdo al grado de vulnerabilidad social de la población. Los resultados que se presentan corresponden a las etapas iniciales del análisis, en las cuales fue necesario tomar decisiones acerca de los valores atípicos ("outliers") multivariados que aparecían dentro de la información disponible y que distorsionaban los resultados de cualquier estrategia de conglomeración. La información de base en el estudio consistió en datos censales sobre sesenta y tres variables medidas sobre fracción y radios la cual debió ser depurada para cumplir con el objetivo último del estudio, que fue la identificación de radios censales con características similares entre sí, generando finalmente mapas de la provincia y de sus localidades, que mostraran la distribución geográfica de las zonas en distintas condiciones de vulnerabilidad.

2. Material y Métodos

El análisis consideró una matriz de datos de los 3.237 radios censales de la provincia y 63 variables seleccionadas a partir de aquellas que pudieran expresar las diferentes condiciones de vulnerabilidad y sus determinantes, correspondientes al Censo Nacional de Población, Hogares y Vivienda 2001. Se emplea esta fuente de información por su alta representatividad y además porque es la única que presenta desagregación por localidad, fracción y radio. El conjunto de variables se agrupó considerando diferentes temáticas: Vivienda, Datos demográficos, Educación, Ocupación y Salud. En las primeras etapas (no se muestran resultados en este trabajo), se procedió a hacer una selección de variables con el objetivo de eliminar aquellas que brindaran información muy similar. Se realizaron distintos análisis descriptivos para evaluar comparativamente la distribución de las variables teniendo en cuenta las asociaciones entre ellas. Luego se aplicó cluster jerárquico con el método de Ward para agrupar variables y determinar si se formaban algunos grupos en particular en cada área temática y finalmente se interpretaron mapas de las variables en estudio por fracción y radio. Esto facilitó determinar definitivamente que variable incluir y cual excluir del análisis, teniendo en cuenta además, los conocimientos de especialistas en el tema.

Identificación de Valores Atípicos Multivariados.

La detección de valores atípicos es una tarea muy importante en el análisis de datos. Los valores extremos describen el comportamiento de los datos anormales, es decir, datos que se desvían de la variabilidad natural de los datos. A menudo, los valores extremos son de interés primordial, por tanto, es importante identificarlos antes del modelado y análisis. En la aplicación que dio origen a este trabajo, los valores extremos representarían radios censales con condiciones de vida extremadamente desfavorables o favorables.

Una definición exacta de lo que se considera como valor extremo frecuentemente depende de los supuestos ocultos en relación con la estructura de datos y el método de detección aplicado. Sin embargo, algunas definiciones se consideran lo suficientemente generales como para hacer frente a diversos tipos de datos y métodos. Hawkins (1980) define un valor atípico como una observación que se aparta tanto de otras observaciones al punto de despertar la sospecha de que fue generado por un mecanismo diferente. Barnet y Lewis (1994) indican que una observación de la periferia, o errático, es uno que parece apartarse notablemente de los demás miembros de la muestra en la que se produce. Asimismo, Johnson (1992) define un valor atípico como una observación en un conjunto de datos que parece ser incompatible con el resto de ese conjunto de datos.

Los métodos de detección de valores atípicos se han sugerido para numerosas aplicaciones, tales como la detección de fraudes de tarjeta de crédito, ensayos clínicos, análisis de irregularidades en los votos, intrusos en redes, predicción de mal tiempo, sistemas de información geográfica, análisis de rendimiento de los deportistas y otros datos en las tareas de data-mining.

Los métodos de detección de valores atípicos se pueden dividir en univariados y multivariados. Estos últimos forman la mayor parte del cuerpo actual de investigación. En una o dos dimensiones, los valores atípicos que están lo suficientemente alejados de la masa principal de datos son fácilmente identificados con gráficos simples, pero su detección es mas desafiante en dimensiones más altas. Los métodos paramétricos o modelo dependientes, asumen una distribución subyacente conocida de las observaciones e identifican como valores extremos a aquellas observaciones que se apartan de los supuestos del modelo. A menudo son inadecuados para conjuntos de datos de alta dimensionalidad y conjuntos de datos arbitrarios sin previo conocimiento de la distribución de datos subyacente. Dentro de los no-paramétricos están los basados en distancias y otros en técnicas de agrupamiento, donde un grupo de tamaño pequeño puede ser considerado como un cluster de valores atípicos. Los métodos también se pueden particionar entre clasificadores duros y clasificadores blandos. Los primeros particionan los datos en dos conjuntos que no se superponen: los outliers y los no-outliers. Los últimos ofrecen un ranking, asignando a cada dato un factor de clasificación como valor extremo que refleja su grado de alejamiento.

Varios factores afectan la eficiencia de los métodos. En particular, si el modelo normal multivariado es o no adecuado, la dimensión del conjunto de datos, el tipo de los valores extremos, la proporción de valores atípicos en el conjunto de datos y el grado de conta-

minación de los valores atípicos. Estas características hacen que sea recomendable aplicar una batería de métodos en el conjunto de datos a fin de detectar los posibles valores anómalos.

En muchos casos las observaciones multivariadas no puede ser detectada como los valores extremos cuando cada variable se considera de forma independiente. La detección de valores extremos sólo es posible cuando se realiza un análisis multivariado, y las interacciones entre las diferentes variables se comparan dentro de la clase de datos. Un simple ejemplo puede verse en la Figura 1, que presenta puntos de datos que tienen dos medidas en un espacio bidimensional. La observación en la parte inferior izquierda es claramente un valor atípico multivariado, pero no uno univariado. Al considerar cada medida por separado con respecto a la extensión de los valores a lo largo de los ejes X e Y, lo que podemos ver es que caen cerca del centro de las distribuciones univariadas. De esta manera, la prueba para detectar los valores extremos debe tener en cuenta las relaciones entre las dos variables, que en este caso parece anormal.

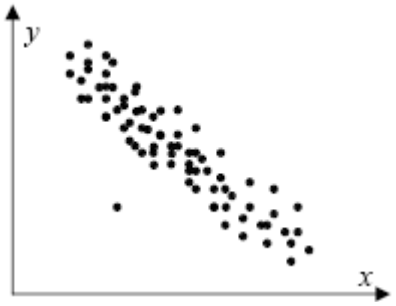


Figura 1: Un espacio bidimensional con una observación extrema

Los conjuntos de datos con múltiples valores extremos o agrupaciones de valores extremos están sujetos a los efectos de enmascaramiento (masking) y empantanamiento (swamping).

Efecto de enmascaramiento. Se dice que un outlier enmascara a un segundo outlier, si el segundo outlier puede ser considerado como un valor extremo sólo por sí mismo, pero no en presencia del primer outlier. Así, después de la eliminación del primer outlier, en una segunda instancia, el otro punto se convierte en un valor atípico. El enmascaramiento se produce cuando un grupo de observaciones extremas sesga las estimaciones de la media y de la covarianza hacia él, y la distancia resultante del valor extremo a la media es pequeña.

Efecto de empantanamiento. Se dice que un outlier empantana una segunda obser-

vación, si ésta última puede ser considerada como un valor extremo sólo bajo la presencia de la primera. En otras palabras, después de la eliminación del primer outlier, la segunda observación se convierte en un no-outlier. El empantanamiento ocurre cuando un grupo de valores extremos sesga las estimaciones de la media y de la covarianza hacia él y lejos de otros valores no periféricos, y la distancia resultante de estos casos a la media es grande, haciéndolos parecer como outliers.

Los métodos estadísticos para la detección de valores atípicos multivariados a menudo indican las observaciones que se encuentran relativamente lejos del centro de la distribución de datos. Se pueden implementar varias medidas de distancia para tal tarea. La distancia de Mahalanobis es un criterio muy conocido que depende de los parámetros estimados de la distribución multivariada. Dadas n observaciones de un conjunto de datos p-dimensionales (n > p), se denota el vector de medias muestral con $\bar{x}_n$ y la matriz de covarianza muestral por S_n, donde $S_n = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)(x_i - \bar{x}_n)^T$. La distancia de Mahalanobis para cada punto i multivariante de datos, i = 1, ..., n, se denota por MSD_i y está dada por:

$$MSD_i = \sqrt{(x_{.i} - \bar{x})^T \, S_n^{-1} (x_{.i} - \bar{x})}$$

Para datos multivariados distribuidos normalmente los valores de la distancia de Mahalanobis tienen aproximadamente una distribución chi-cuadrado con p grados de libertad. En consecuencia, aquellas observaciones con una distancia de Mahalanobis grande se indican como valores atípicos. (por ejemplo, valores superiores al cuantil 97,5%).

Nótese que los efectos de enmascaramiento y empantanamiento juegan un rol importante en la adecuación de la distancia de Mahalanobis como criterio para la detección de valores atípicos. Es decir, los efectos de enmascaramiento podrían disminuir la distancia de Mahalanobis de un valor atípico. Esto puede ocurrir, por ejemplo, cuando un pequeño grupo de outliers atrae a $\boldsymbol{X_n}$ e infla S_n hacia su dirección. Por otra parte, los efectos de empantanamiento podrían aumentar la distancia de Mahalanobis de las observaciones que no son outliers. Por ejemplo, cuando un pequeño grupo de valores atípicos atrae $\bar{X}_n$ e infla S_n lejos del patrón de la mayoría de las observaciones.

Los problemas de enmascaramiento y empantanamiento pueden resolverse usando estimaciones robustas, las cuales por definición están menos afectadas por “outliers”, sien-

do menos probable que influencien los parámetros usados en la MSD. Los puntos que no son atípicos, determinarán completamente la estimación de la forma y posición de los datos. Muchos de los métodos de estimación, incluyendo un método robusto como el de los M-estimadores, fallan si la fracción de "outliers" es mayor que 1/(p+1), donde p es la dimensión del conjunto de datos o número de variables, indicando que en dimensiones grandes, una pequeña cantidad de valores atípicos puede producir estimaciones deficientes. Por lo tanto, las distancias de Mahalanobis deben ser estimadas por un procedimiento robusto a fin de proporcionar medidas fiables para el reconocimiento de los valores extremos.

Entre los estimadores robustos para las medidas de localización y de covariancia, en este trabajo se utiliza el estimador de Mínimo Determinante de Covariancia (MCD) propuesto por Rousseeuw 1998. Dado un conjunto de n datos, los estimadores MCD del vector de medias y de la matriz de covariancias son auqellos basados en una sub-muestra de tamaño h (h ≤ n) que minimiza el determinante de la matriz de covariancias.

$$\text{MCD} = \left(\overline{X}_J^*, S_J^*\right)$$

donde

J = {conjunto de h puntos tales que $\left|S_J^*\right| \le \left|S_K^*\right|$, para cualquier conjunto K de tamaño h}

$$\overline{X}_J^* = \frac{1}{h}\sum_{i\in J} x_i$$

$$\overline{S}_J^* = \frac{1}{h}\sum_{i\in J}\left(x_i - \overline{X}_J^*\right)\left(x_i - \overline{X}_J^*\right)^T$$

El valor h puede pensarse como el mínimo número de puntos no atípicos. El MCD tiene su punto de ruptura más alto en h=⌊(n+p+1)/2⌋, donde $\lfloor\cdot\rfloor$ es la función parte entera. El punto de ruptura representa la mayor fracción de contaminación que un estimador puede tolerar antes de comenzar a comportarse en forma totalmente aberrante. El MCD es calculado con la muestra de tamaño h más "cercana" y, por lo tanto, los puntos más alejados tendrán un efecto pequeño en la estimación MCD de los parámetros de forma y posición.

La distribución de la distancia de Mahalanobis (MSD) cuando se usan medidas robustas de posición y forma, es desconocida. La determinación de los puntos de corte exac-

tos para determinar las distancias que convierten un valor en atípico, no está perfectamente determinado, aunque suelen usarse criterios basados en percentiles.

3. Resultados

Luego de las etapas de depuración de información e identificación de grupos de variables con información similar, se eligieron 36 variables de las 63 originales, para describir la situación de vulnerabilidad en la provincia. La mayoría de las mediciones son en porcentajes respecto del total del radio censal, en los casos que no corresponden a porcentajes, se indica en el nombre de la variable (cantidad o índice). Ellas se muestran agrupadas por área temática. Entre paréntesis el código utilizado para su referencia).

Ocupación:

Población ocupada (PP14)
Ocupados en el sector público (PP16)
Jefes de hogar desocupados (PP64)

Educación

Jefes de hogar varones con nivel primario incompleto (E1)
Jefes de hogar mujeres con nivel primario incompleto (E2)
Personas que asistieron con nivel primario completo (E3)
Personas que asistieron con nivel secundario completo (E4)
Personas que asistieron con nivel superior y terciario completo (E5c)
Personas de 10 a 14 años que no asisten a un establecimiento educativo (E6)
Personas mayores de 5 años que asistieron a un establecimiento educativo (E7)
Personas analfabetas (E8)
Personas entre 5 y 14 años que asisten a un establecimiento educativo (E9)
Personas entre 15 y 18 años que no asisten a un establecimiento educativo. (E10b)
Personas entre 20 y 25 años que asisten a EGB 3 o polimodal (E11)
Personas entre 20 y 25 años que asisten a nivel universitario y superior no universitario (E14)

Vivienda y saneamiento básico

Viviendas con suministro de agua de red potable (PV4)
Viviendas con agua dentro de la vivienda (PV5)
Viviendas que poseen baño con inodoro (PV7)
Viviendas que poseen pozo ciego y cámara séptica (PV8)
Viviendas que poseen cloaca (PV9)

Viviendas que poseen sólo pozo ciego (PV10)
Viviendas con piso de tierra o ladrillo suelto (PV11)
Hogares que poseen gas (PV15)
Hogares con propietarios de la vivienda y el terreno (PV18a)
Hogares con propietarios de la vivienda (PV18b)
Hogares con privación de recursos corrientes (PV21)
Hogares con privación patrimonial (PV22)
Hogares con hacinamiento (PV27)

Salud

Población con cobertura de salud (PP50)
Cantidad de hijos por mujer (PP56)
Porcentaje de hijos nacidos muertos (PP58)

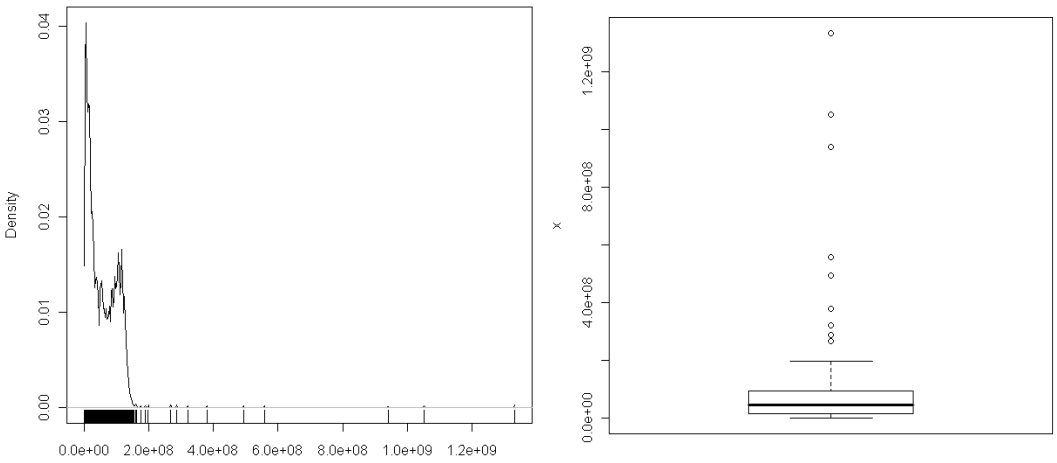
Demografía

Índice de masculinidad (D1)
Índice de dependencia potencial total (D2)
Niños de 0 a 14 años (D3)
Personas mayores de 64 años (D4)
Índice de recambio (PP63)

Se investigó acerca de la presencia de radios censales con valores atípicos, siguiendo un enfoque multivariado, utilizando los métodos descriptos en la sección 2. El objetivo fue separar para un análisis individualizado a aquellos con valores extremos, suponiendo que se encontrarían realidades muy heterogéneas en toda la extensión geográfica de la provincia. Entre los métodos para detectar valores atípicos multivariados se eligió aquel que considera la distancia de Mahalanobis, que resume la información de las variables en estudio y que establece una medida de orden con una distribución conocida. Luego se determinó un valor de corte sobre dicha distribución considerando como valores extremos los puntos con valores mayores al de un percentil observado.

Las variables de la base de datos en su mayoría no se ajustan a una distribución normal al ser consideradas marginalmente, lo cual es condición suficiente para asegurar que la distribución conjunta de ellas no es normal multivariada. Se prefirió por lo tanto calcular la distancia utilizando estimadores robustos, en particular, los MDC (se utilizó la rutina covMDC de R), las cuales se muestran en el Gráfico 1, identificando algunas observaciones con valores muy altos.

Gráfico 1: Distribución de las distancias cuadradas de Mahalanobis (frecuencias y box-plot)

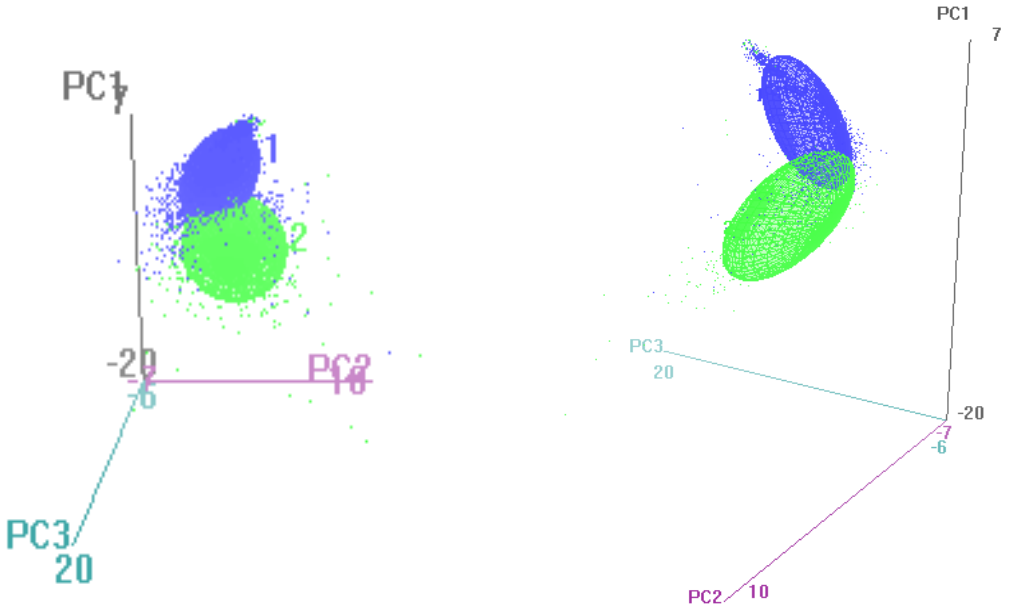


Dado que la forma de la distribución no sugería alguna distribución conocida para ajustar a la distribución de la variable, se utilizó el criterio de separar los 10 puntos que exceden a p= Q3+ 1,5 * I(Q1,Q3) siendo I(Q1,Q3) el rango intercuartílico, según se muestra en el Box-Plot (p=212.637.778).

Con el objetivo de analizar el conjunto de datos restantes, en la búsqueda de otros valores extremos, se calcularon los estimadores robustos para vector de medias y matriz de covariancias con los datos de la base reducida y con ellos se re-calcularon las distancias de Mahalanobis. Nuevamente la distribución de frecuencias no mostró semejanza con algún modelo conocido y se decidió utilizar como punto de corte el 3º cuartil de dicha función (p=93.847.784). Los radios con $D^2_{robusta}$>93.847.784 se identificaron como valores extremos.

Para poder analizar las características de estos dos subgrupos de radios censales identificados (según magnitud de $D^2_{robusta}$) se recurrió a la técnica de Componentes Principales aplicada sobre la matriz de correlaciones. El gráfico 2 muestra su proyección sobre las tres primeras componentes principales, las que representan un 59% dela variabilidad total de los datos. Los grupos están bien diferenciados, especialmente sobre la primera componente principal (CP1). El grupo de valores no extremos (Grupo 1) es el que toma valores altos de CP1 y el grupo de valores extremos (Grupo 2) obtiene los valores más bajos. Sobre CP2 y CP3 no se observan diferencias muy marcadas. Los distintos ángulos mostrados en el gráfico 2 evidencian esta característica.

Gráfico 2: Proyección de radios censales con valores atípicos (G2) y no atípicos (G1) sobre las tres primeras componentes principales



Según las correlaciones de las Componente Principales con las variables originales (Tabla 1 del Anexo) las variables que más influyen en la formación de la primera Componente son las relacionadas a nivel educativo, cobertura de salud, presencia de servicios y agua dentro de la vivienda. Las variables que la influyen positivamente son las que denotan condiciones favorables en estos aspectos, mientras que las que tienen coeficientes negativos son variables que representan condiciones desfavorables. La segunda Componente Principal representa en mayor medida la información de población ocupada y hogares sin privación de recursos corrientes, mientras que la tercera Componente Principal está influenciada por la presencia de las variables vivienda sólo con pozo ciego y con pozo ciego y cámara séptica. Una característica a destacar es que la CP1 y la CP3 se relacionan con aspectos estructurales de la realidad social, mientras que la CP2 está asociada a condiciones coyunturales, que pueden variar más rápidamente en el tiempo.

Finalmente se identificaron tres grupos de radios censales: 2420 con valores no atípicos (80% de las población de radios censales), 807 con valores extremos (19,9%) y 10 con valores muy extremos (0.1%). La Tabla 1 muestra su composición respecto de población urbana o rural.

Tabla 1: Porcentaje de radios censales según sean Rurales, Mixtos o Urbanos

Grupo	Rural	Mixto	Urbano	Total	Cantidad de radios
1 – Valores no extremos	5,7	1,5	92,8	100	2.420
2 – Valores extremos	56,2	6,4	37,4	100	807
3 – Valores muy extremos	60	10	30	100	10

Se representaron los radios censales según estos tres agrupamientos en mapas de la provincia de Santa Fe y por separado en ciudades de Rosario y Santa Fe. Los radios censales del grupo de valores extremos se ubican en el centro y norte de la provincia y en la periferia de dichas ciudades. Los del grupo de valores muy extremos en su mayoría se encuentran dispersos por el sur de la provincia y en la ciudad de Rosario (Gráficos 3, 4 y 5).

Gráfico 3: Radios censales de Santa Fe según valores extremos o no extremos.

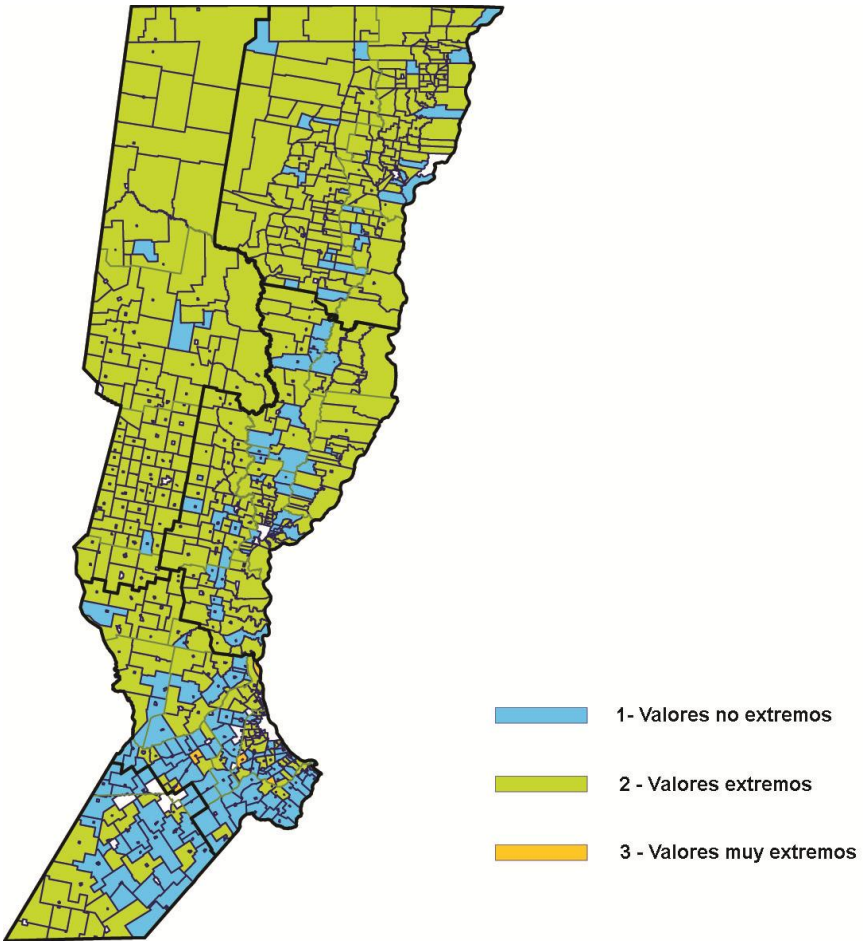


Gráfico 4: Radios censales de Rosario según valores extremos o no extremos.

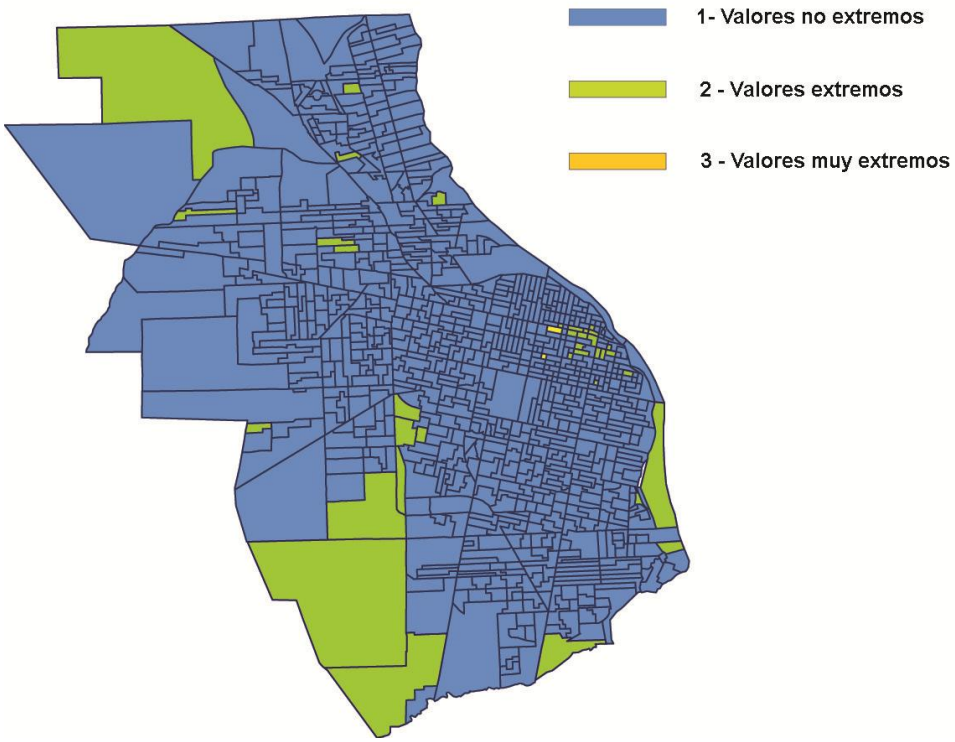
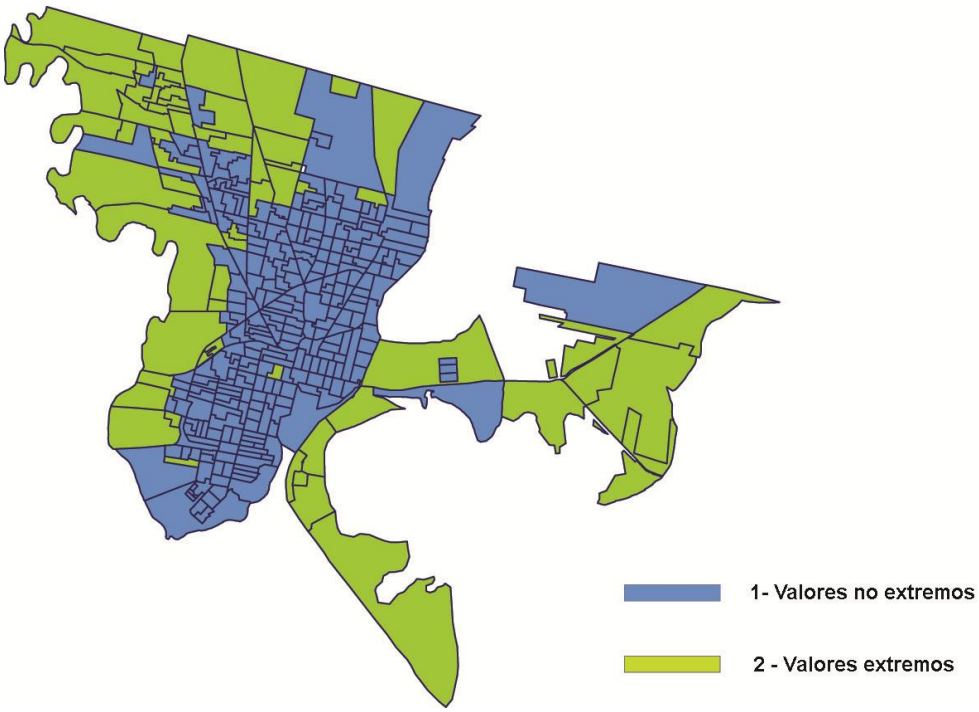


Gráfico 5: Radios censales de ciudad de Santa fe según valores extremos o no extremos.



Para completar la descripción de estos grupos se analizaron las medianas de cada variable, excluyendo de este análisis al conjunto de 10 radios con valores muy extremos. Estos además poseen características heterogéneas entre sí.

Tabla 4: Medianas de las variables bajo estudio para cada subgrupo

Variable	Mediana Grupo 1	Mediana Grupo 2
D1 Índice de masculinidad	91,48	109,09
D2 Índice de dependencia potencial total	58,37	69,27
D3 Porcentaje de niños de 0 a 14 años	21,05	34,02
D4 Porcentaje de personas mayores de 64 años	6,37	4,38
E1 Porcentaje de jefes de hogar varones con nivel primario incompleto	12,67	30
E10b Porcentaje de personas entre 15 y 18 años que no asisten a un establecimiento educativo	16,67	42,86
E11 Porcentaje de personas de 20 y 25 años que asisten a EGB 3 o polimodal	3,22	1,92
E14 Porcentaje de personas entre 20 y 25 años que asisten a nivel universitario y superior no universitario	27,75	5
E2 Porcentaje de jefes de hogar mujeres con nivel primario incompleto	22,71	33,33
E3 Porcentaje de personas que asistieron con nivel primario completo	34,25	44,46
E4 Porcentaje de personas que asistieron con nivel secundario completo	21,48	7,97
E5c Porcentaje de personas que asistieron con nivel superior y terciario completo	9,76	2,49
E6 Porcentaje de personas de 10 a 14 años que no asisten a un establecimiento educativo	0,87	2,70
E7 Porcentaje de personas mayores de 5 años que asistieron a un establecimiento educativo	68,95	63,33
E8 Porcentaje de personas analfabetas	1,18	5
E9 Porcentaje de personas entre 5 y 14 años que asisten a un establecimiento educativo	98,72	96,09
PP14 Porcentaje de población ocupada	76,39	81,25
PP16 Porcentaje de ocupados en el sector público	18,47	12,5
PP50 Porcentaje de población con cobertura de salud	68,72	39,65
PP56 Cantidad de hijos por mujer	1,78	2,60
PP58: porcentaje de hijos nacidos muertos	3,88	3,86
PP63 Índice de recambio	72,05	17,51
PP64 Porcentaje de jefes de hogar desocupados	9,28	7,94
PV10 Porcentaje de viviendas que poseen sólo pozo ciego	12,35	29,55
PV11 Porcentaje de viviendas con piso de tierra o ladrillo suelto	0	5,33

PV15 Porcentaje de hogares que poseen gas	94,23	0
PV18a Porcentaje de hogares propietarios de la vivienda y el terreno	75,36	49,37
PV18b Porcentaje de hogares propietarios de la vivienda	0,36	1,78
PV21 Porcentaje de hogares con privación corriente	15,09	12,42
PV22 Porcentaje de hogares con privación patrimonial	3,87	20,92
PV27 Porcentaje de hogares con hacinamiento.	2,03	11,61
PV4 Porcentaje de viviendas con suministro de agua de red potable	99,58	1,54
PV5 Porcentaje de viviendas con agua dentro de la vivienda	96,6	59,29
PV7 Porcentaje de viviendas que poseen baño con inodoro	98,81	80,71
PV8 Porcentaje de viviendas que poseen pozo ciego y cámara séptica	14,21	23,53
PV9 Porcentaje de viviendas que poseen cloaca	56,42	0

Las medianas del grupo de radios censales con valores extremos se mantuvieron superiores a las medianas del grupo de valores no extremos en las variables que son desfavorables en educación, malas condiciones de vivienda, índice de masculinidad, porcentaje de niños y de población ocupada. mientras que en las variables que representan buena educación, condiciones favorables en la vivienda, porcentaje de población ocupada en el sector público, porcentaje de población con cobertura de salud, y porcentaje de población mayor a 64 años las medianas se mantuvieron superiores en el grupo de valores no extremos con respecto a las del grupo de valores extremos. La única variable para la cual no hubo diferencia en la mediana entre ambos grupos es porcentaje de hijos nacidos muertos (PP58).

4. Comentarios Finales

La identificación de situaciones de vulnerabilidad, su cuantificación y su localización en el territorio es un análisis imprescindible tanto para el diseño de las nuevas acciones o intervenciones, como para evaluar los efectos de las políticas públicas implementadas. Las mediciones de la situación social basadas en los censos tienen la particularidad de proporcionar datos que abarcan la totalidad del territorio pudiéndose aplicar en áreas no cubiertas por otras fuentes, lo cual las convierte en un instrumento adecuado para el análisis de áreas específicas y para la toma de decisiones a nivel de los gobiernos locales.

Fijado el objetivo de obtener grupos diferenciados de radios censales de la Provincia de Santa Fe de acuerdo al grado de vulnerabilidad social de la población, fue necesario realizar un proceso de selección de variables censales entre aquellas que pudieran expresar

las diferentes condiciones de vulnerabilidad y sus determinantes y además una identificación de los radios censales con valores extremos, previo al empleo de algoritmos de agrupamiento.

En este trabajo se presentó una síntesis de la metodología utilizada en esta etapa de identificación de valores atípicos respetando el concepto multivariado de la observación. Su empleo sobre la información acerca de vulnerabilidad social identificó tres subgrupos: uno de valores muy extremos (10 radios censales), otro de valores extremos (con 807 radios), y un grupo mayoritario (2.420 radios) de valores no extremos. El grupo de valores no extremos estaba formado casi en su totalidad por radios urbanos (92,8%), mientras que los grupos de valores extremos y muy extremos contenían una proporción mayor de radios rurales (56,2 % y 60%, respectivamente). Se representaron los radios censales según estos tres grupos en los mapas de la provincia de Santa Fe y en las ciudades de Rosario y Santa Fe, observando que los radios censales del grupo de valores extremos se ubicaban en el centro y norte de la provincia y en la periferia de dichas ciudades y los del grupo de valores muy extremos en su mayoría se encontraban dispersos por el sur de la provincia y en la ciudad de Rosario.

Estos resultados permitieron a posteriori, aplicar algoritmos de Análisis Cluster en cada subgrupo, para detectar finalmente conglomerados de radios censales con características similares entre sí, a los efectos de dar respuesta al objetivo primordial planteado.

REFERENCIAS BIBLIOGRÁFICAS

Ben-Gal, I. E. (2005) "*Outlier Detection"*. The Data Mining and Knowledge Discovery Handbook.

Filzmoser, P.; Maronna, R.; Werner, W. (2008) "*Outlier Identification in High Dimensions"*. Computational Statistics & Data Análisis, Vol. 52, Nº 3

Filzmoser, P. (2005) "*Identification of Multivariate Outliers: A Performance Study"*. Austrian Journal of statistics. Vol 34, Nº 2.

Hardin, J.; Rocke, D. M. (2004) "*Outlier detection in the multiple clustersetting using the minimum covariance determinant Estimator".* Computational Statistics & Data Analysis n 44.

Hodge, V.J.; Austin, J. (2004) "*A survey of outlier detection methodologies".* Artificial Intelligence Review, Nº 22.

Klawonn, F.; Rehm, F. (2006) "*Clustering techniques for outlier detection"*. Wang J (ed) Encyclopedia of Data Warehousing and Mining, Idea Group, Hershey.

Pizarro, R. (2001) *"La vulnerabilidad social y sus desafíos: una mirada desde América Latina"*. CEPAL, Serie Estudios Estadísticos y Prospectivos, Nº 6.

ANEXO

Tabla 1: Correlaciones de las variables con las tres primeras componentes principales

Variable	CP1	CP2	CP3
D1	-0,638	0,328	-0,055
D2	-0,530	-0,101	0,067
D3	-0,832	-0,318	0,006
D4	0,070	0,412	-0,085
E1	-0,860	0,077	-0,051
E10B	-0,772	0,025	-0,116
E11	-0,066	-0,358	-0,078
E14	0,853	0,175	0,293
E2	-0,495	-0,127	-0,240
E3	-0,691	-0,012	-0,513
E4	0,875	-0,110	0,100
E5C	0,773	0,227	0,362
E6	-0,395	0,061	0,143
E7	0,471	0,352	-0,487
E8	-0,716	0,112	0,393
E9	0,352	-0,170	-0,283
PP14	0,126	0,855	-0,156
PP16	0,304	-0,361	0,233
PP50	0,876	0,284	-0,065
PP56	-0,882	-0,097	0,106
PP58	-0,102	-0,079	-0,015
PP63	0,755	0,325	0,116
PP64	-0,218	-0,830	0,149
PV10	-0,464	0,009	-0,629
PV11	-0,641	0,177	0,512
PV15	0,763	-0,193	0,213
PV18A	0,581	-0,349	-0,253
PV18B	-0,466	-0,239	0,404
PV21	0,020	-0,682	-0,314
PV22	-0,799	0,324	0,097
PV27	-0,728	-0,311	0,320
PV4	0,618	-0,512	0,176
PV5	0,857	-0,203	-0,317
PV7	0,753	-0,302	-0,409
PV8	-0,225	-0,035	-0,631
PV9	0,775	-0,040	0,444
λ_i	14,10	3,87	3,24
% de variancia explicada	39,16	10,74	9,07