Bussi, Javier
Marí, Gonzalo
Méndez Fernanda
Instituto de Investigaciones Teóricas y Aplicadas, Escuela de Estadística

COMPONENTES PRINCIPALES ROBUSTAS: UNA APLICACIÓN A LOCALIDADES DE LA PROVINCIA DE SANTA FE

Resumen

El análisis de componentes principales (ACP) es una técnica muy utilizada dentro de los métodos estadísticos multivariados. El objetivo de este método es representar adecuadamente un conjunto de n observaciones con p variables a través de un número menor de variables construidas como combinaciones lineales de las originales. Se basa en el cálculo de autovalores y autovectores de la matriz de covariancias (o correlaciones). La presencia de valores atípicos en los datos puede distorsionar la matriz de covariancias muestrales. Por este motivo se han propuesto diversas formas de tratar esta dificultad a partir de técnicas robustas.
La inferencia que utiliza métodos *bootstrap* clásicos aplicados a estimadores robustos requiere menos supuestos pero implica un alto costo computacional y una pérdida de robustez ante la presencia de observaciones atípicas. Una alternativa computacionalmente más sencilla y resistente a la presencia de *outliers* es el *Boostrap* Rápido y Robusto (FRB: Fast and Robust Bootstrap).
Se presenta una aplicación del uso del método de componentes principales robusto y la inferencia a partir del método *bootstrap* robusto a datos de indicadores de carencias críticas provenientes del Censo Nacional Población, Hogares y Viviendas 2010.

Abstract

Principal components analysis (PCA) is a widely used technique within multivariate statistical methods. The purpose of this technique is adequately representing a set of n observations and p variables through fewer variables constructed as linear combinations of the original ones. It is based on the calculation of eigenvalues and eigenvectors of the covariance (or correlation) matrix. The presence of outliers in the data can distort the sample covariance matrix. Therefore various ways have been proposed to deal with this difficulty using robust techniques.
The Bootstrap inference applied to classical robust estimators requires fewer assumptions but involves high computational cost and a loss of robustness in the presence of outliers. An alternative computationally simpler and more resistant to the presence of outliers is the Fast and Robust Bootstrap (FRB).
The use of the robust principal components method and the robust bootstrap inference is illustrated on a dataset of indicators of critical needs of communes of the province of Santa Fe from the National Census Population and Housing 2010.

Palabras claves: Componentes Principales; Estimadores MM; Boostrap Rápido y Robusto

1. Introducción

El análisis de componentes principales (ACP) es una técnica muy utilizada dentro de los métodos estadísticos multivariados. El objetivo de este método es representar adecuadamente un conjunto de n observaciones con p variables a través de un número menor de variables construidas como combinaciones lineales de las originales. La técnica se basa en el cálculo de autovalores y autovectores de la matriz de covariancias o de correlaciones de las variables originales. La presencia de valores atípicos en los datos puede distorsionar la matriz de covariancias muestrales. Por este motivo se han propuesto diversas formas de tratar esta dificultad a partir de técnicas robustas. En este trabajo se considera un tipo de ACP basado en estimaciones robustas de la matriz de forma. En particular, se calculan los vectores y valores propios del estimador MM de la matriz de forma. Los estimadores MM están diseñados para ser altamente resistentes a los valores atípicos y altamente eficientes para datos normales.

Principalmente interesa la inferencia del análisis de componentes principales basado en el estimador MM. Tanto en el método clásico como en las versiones robustas, se puede inferir a través de distribuciones normales asintóticas bajo el supuesto de distribuciones elípticas. En el caso de valores atípicos, generalmente este supuesto no se cumple con lo cual se deben utilizar técnicas alternativas, entre las cuales se encuentra el método *bootstrap*.

La inferencia que utilizan los métodos *bootstrap* clásicos aplicados a estimadores robustos requiere menos supuestos pero implica un alto costo computacional y una pérdida de robustez ante la presencia de observaciones atípicas. Una alternativa computacionalmente más sencilla y resistente a la presencia de outliers es el *Boostrap* Rápido y Robusto (FRB: Fast and Robust Bootstrap), que en principio puede ser utilizado para cualquier estimador que sea solución de un sistema de ecuaciones suaves de punto fijo tales como los estimadores MM.

En la sección 2 se explican algunos aspectos teóricos de la ACP sobre la base de las estimaciones de MM y se describe el método *Bootstrap* rápido y robusto.

En la sección 3 se presenta el uso del método de componentes principales robusto y la inferencia a partir de método *bootstrap* robusto a datos de indicadores de carencias críticas provenientes del Censo Nacional Población, Hogares y Viviendas 2010.

La Sección 4 contiene algunas observaciones finales.

2. Metodología

2.1. Análisis de Componentes Principales Clásico

El análisis de componentes principales (ACP) tiene por objetivo representar apropiadamente la información provista por un grupo de n observaciones donde se consideran p variables, reduciendo el número de estas pero resignando una baja cantidad de información.

Esta representación se realiza a través de la creación de r nuevas variables no observables que resultan ser combinaciones lineales de las p variables originales ($r < p$).

Por ejemplo, en casos donde se presentan variables con alta asociación, se puede reducir el número de las mismas, seleccionando sólo algunas de ellas pero que expliquen un alto porcentaje de la variabilidad total original. Es posible representar las observaciones en un espacio de menor dimensión (r) pudiendo identificar variables latentes que expliquen la variabilidad de los datos y al mismo tiempo generando variables no correlacionadas que facilitan la interpretación de los resultados obtenidos.

Suponiendo que se cuenta con n elementos de una población en los que se miden p-variables, los cuales son las observaciones de un vector aleatorio $\boldsymbol{x}$ p-dimensional con media $\boldsymbol{\mu}$ con una matriz de covariancias $\boldsymbol{\Sigma}$ de dimensión $p \times p$. La primera componente principal es la combinación lineal de las variables originales que resulta de la proyección que tiene mayor variabilidad de manera tal que:

$$Var(\boldsymbol{b}_1'\boldsymbol{x}) = max \quad \text{dado que} \quad \|\boldsymbol{b}_1\| = 1 \tag{1}$$

siendo $\boldsymbol{x}'\boldsymbol{b}_1$ la forma que toma la combinación lineal. La segunda componente principal cumple (1) y además que $\boldsymbol{b}_2'\boldsymbol{b}_1 = 0$. De manera similar se computan las componentes principales subsiguientes. El número total de componentes es *p*, los autovalores de la matriz $\boldsymbol{\Sigma}$ son $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ y sus respectivos autovectores son $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_p$. La variancia de cada componente principal está dada por:

$$Var(\boldsymbol{b}_j'\boldsymbol{x}) = \lambda_j$$

El número q de componentes puede ser seleccionado a través del criterio de porcentaje de variancia explicada:

$$\frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

El ACP ha sido extensamente tratado en la bibliografía estadística, una detallada explicación del mismo puede ser consultada en Peña (2002) y Johnson y Wichern (2007).

2.2. Análisis de Componentes Principales Robusto

Los estimadores MM de locación y forma se basan en dos funciones de pérdida que debe cumplir con las siguientes condiciones de regularidad:

(R1) La función ρ debe ser real, doblemente diferenciable y $\rho(0) = 0$.

(R2) La función ρ debe ser estrictamente creciente en $[0, c]$ y constante en $[c, \infty)$ para una constante finita c.

Los estimadores MM de locación, forma y covariancia pueden entonces ser definidos de la siguiente manera:

Sea $\boldsymbol{X}_n = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\} \subset \mathbb{R}^p$ con $n \geq p+1$. Sean ρ_0 y ρ_1 de manera tal que satisfagan (R1) y (R2) y sea $(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}_n)$ el estimador S multivariado, es decir que minimiza $|\boldsymbol{C}|$ sujeto a:

$$\frac{1}{n}\sum_{i=1}^{n} \rho_0\left([(\boldsymbol{x}_i - \boldsymbol{T})^t \boldsymbol{C}^{-1}(\boldsymbol{x}_i - \boldsymbol{T})]^{1/2}\right) = b$$

entre todas las $(\boldsymbol{T}, \boldsymbol{C}) \in \mathbb{R}^p \times \mathrm{DPS}(p)$. Aquí $\mathrm{DPS}(p)$ representa el conjunto de matrices definidas positivas simétricas de orden *p*×*p*. Denotando a $\hat{\sigma}_n$ con $\left|\widetilde{\boldsymbol{\Sigma}}_n\right|^{1/(2p)}$ resulta entonces que el estimador MM de locación y forma $(\widehat{\boldsymbol{\mu}}_n, \widehat{\boldsymbol{\Gamma}}_n)$ minimiza:

$$\frac{1}{n}\sum_{i=1}^{n} \rho_1\left([(\boldsymbol{x_i} - \boldsymbol{T})^t \boldsymbol{G}^{-1}(\boldsymbol{x}_i - \boldsymbol{T})]^{1/2}/\hat{\sigma}_n\right)$$

entre todas las $(\boldsymbol{T}, \boldsymbol{G}) \in \mathbb{R}^p \times DPS(p)$ para las cuales $|\boldsymbol{G}| = 1$. El estimador MM para la covariancia resulta: $\widehat{\boldsymbol{\Sigma}}_n = \hat{\sigma}_n^2 \widehat{\boldsymbol{\Gamma}} n$.

La idea es estimar la escala a través de un S-estimador robusto, y luego estimar la locación y la forma utilizando otra función de pérdida que brinde un resultado más eficiente. Las estimaciones de locación y escala mantienen el punto de quiebre de la escala auxiliar.

En este caso se consideran funciones de pérdida pertenecientes a la familia de funciones de Tukey.

$$\rho(x) = \begin{cases} \dfrac{x^2}{2} - \dfrac{x^4}{2c^2} + \dfrac{x^6}{6c^4}, & |x| \leq c \\ \dfrac{c^2}{6}, & |x| \geq c \end{cases}$$

donde el valor $c > 0$ es fijado por el investigador como una constante de ajuste. Los estimadores S se pueden considerar un caso especial de los estimadores MM realizando una particular elección de la función de pérdida. No se considera el punto de quiebre en este trabajo en el ACP debido a ciertas características del espacio paramétrico de los autovalores. La función de influencia del estimador MM de forma no depende del estimador inicial S de escala si este es consistente. Lo mismo ocurre para el estimador MM de locación. La función de influencia de los estimadores MM de covariancia resultan en una combinación de las funciones de influencia de los estimadores S con las funciones de pérdida. Más detalles sobre punto de quiebre y función de influencia en ACP se presentan en Salibián-Barrera et al (2006).

Sea la matriz $\boldsymbol{\Gamma}$ con autovalores $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ y $\upsilon_1, \upsilon_2, .., \upsilon_p$ los correspondientes autovectores. El método de ACP robusto se basa en el estimador MM y consiste en estimar los autovalores y los autovectores a través de los autovalores y autovectores del estimador de forma $\widehat{\boldsymbol{\Gamma}}_n$.

Los estimadores MM deben ser utilizados para mejorar la eficiencia de los estimadores S cuando dicha mejora es necesaria, esto ocurre cuando la dimensión es baja ($p < 15$) como ocurre en la aplicación presentada en este trabajo. Para dimensiones superiores el

estimador S es altamente eficiente por sí mismo. Finalmente, la diferencia en la complejidad computacional entre el paso extra para obtener el estimador MM, es despreciable en comparación con el cálculo inicial del estimador S.

2.3. *Bootstrap* Rápido y Robusto en el Análisis de Componentes Principales

Asumiendo proximidad a la normalidad, métodos inferenciales para el ACP robusto basado en estimadores MM pueden ser derivados a través de los resultados asintóticos. Pero este no siempre resulta ser el caso y el *bootstrap* puede brindar mejores resultados. Sin embargo, el *bootstrap* clásico puede conducir a un alto costo computacional, ya que calcular el estimador MM, en particular el estimador S inicial, puede consumir mucho tiempo, lo cual es particularmente costoso para grandes conjuntos de datos en dimensiones altas. Otra característica es la inestabilidad del procedimiento, ya que la presencia de observaciones atípicas en la muestra original pueden repetirse en grandes números en las muestras *bootstrap*. Por lo tanto, aunque el estimador MM en la muestra original brinde una solución robusta para el ACP, puede fallar en las muestras *bootstrap* con muchas observaciones atípicas. Es decir, el *bootstrap* clásico puede ser menos robusto que el propio estimador MM. Por lo tanto, con el fin de superar el problema asociado con aplicar el *bootstrap* clásico en datos potencialmente contaminados, se presenta el *Bootstrap* Rápido y Robusto (FRB) en el ACP.

Los estimadores MM definidos anteriormente pueden ser escritos como un sistema de ecuaciones de punto fijo de la forma siguiente:

$$\widehat{\boldsymbol{\mu}}_n = \left(\sum_{i=1}^{n} \frac{\rho_1'\left(d_i/\left|\widehat{\boldsymbol{\Sigma}}_n\right|^{1/(2p)}\right)}{d_i}\right)^{-1}\left(\sum_{i=1}^{n} \frac{\rho_1'\left(d_i/\left|\widehat{\boldsymbol{\Sigma}}_n\right|^{1/(2p)}\right)}{d_i}\boldsymbol{x}_i\right)$$

$$\widehat{\boldsymbol{\Gamma}}_n = G\left(\sum_{i=1}^{n} \frac{\rho_1'\left(d_i/\left|\widehat{\boldsymbol{\Sigma}}_n\right|^{1/(2p)}\right)}{d_i}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_n)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_n)^t\right)$$

$$\widetilde{\boldsymbol{\Sigma}}_n = \frac{1}{nb}\left(\sum_{i=1}^{n} p\frac{\rho_0'(\tilde{d}_i)}{\tilde{d}_i}(\boldsymbol{x}_i - \widetilde{\boldsymbol{\mu}}_n)(\boldsymbol{x}_i - \widetilde{\boldsymbol{\mu}}_n)^t + \left(\sum_{i=1}^{n} \widetilde{\omega}_i\right)\widetilde{\boldsymbol{\Sigma}}_n\right)$$

$$\widetilde{\boldsymbol{\mu}}_n = \left(\sum_{i=1}^{n} \frac{\rho_0'(\tilde{d}_i)}{\tilde{d}_i}\right)^{-1}\left(\sum_{i=1}^{n} \frac{\rho_0'(\tilde{d}_i)}{\tilde{d}_i}\boldsymbol{x}_i\right)$$

donde $G(\boldsymbol{A}) = |\boldsymbol{A}|^{-1/p}\boldsymbol{A}$ para matrices $\boldsymbol{A}$ de dimensión pxp y donde

$$d_i = \left[(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_n)^t \widehat{\boldsymbol{\Gamma}}_n^{-1}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_n)\right]^{1/2}$$

$$\tilde{d}_i = \left[(\boldsymbol{x}_i - \widetilde{\boldsymbol{\mu}}_n)^t \widehat{\boldsymbol{\Sigma}}_n^{-1}(\boldsymbol{x}_i - \widetilde{\boldsymbol{\mu}}_n)\right]^{1/2}$$

$$\widetilde{\omega}_i = \rho_0(\tilde{d}_i) - \rho_0'(\tilde{d}_i)\tilde{d}_i$$

Este sistema de ecuaciones permite aplicar el FRB, la idea es utilizar las ecuaciones para calcular aproximaciones a las estimaciones MM en cada muestra *bootstrap*. En particular, dada una muestra *bootstrap*, una forma intuitiva de obtener recómputos aproximados en forma rápida es de la siguiente manera:

$$\widehat{\boldsymbol{\mu}}_n^* = \left(\sum_{i=1}^{n} \frac{\rho_1'\left(d_i^*/\left|\widehat{\boldsymbol{\Sigma}}_n\right|^{1/(2p)}\right)}{d_i^*}\right)^{-1}\left(\sum_{i=1}^{n} \frac{\rho_1'\left(d_i^*/\left|\widehat{\boldsymbol{\Sigma}}_n\right|^{1/(2p)}\right)}{d_i^*}\boldsymbol{x}_i^*\right)$$

$$\widehat{\boldsymbol{\Gamma}}_n^* = G\left(\sum_{i=1}^{n} \frac{\rho_1'\left(d_i^*/\left|\widehat{\boldsymbol{\Sigma}}_n\right|^{1/(2p)}\right)}{d_i^*}(\boldsymbol{x}_i^* - \widehat{\boldsymbol{\mu}}_n)(\boldsymbol{x}_i^* - \widehat{\boldsymbol{\mu}}_n)^t\right)$$

$$\widetilde{\boldsymbol{\Sigma}}_n^* = \frac{1}{nb}\left(\sum_{i=1}^{n} p\frac{\rho_0'(\tilde{d}_i^*)}{\tilde{d}_i^*}(\boldsymbol{x}_i^* - \widetilde{\boldsymbol{\mu}}_n)(\boldsymbol{x}_i^* - \widetilde{\boldsymbol{\mu}}_n)^t + \left(\sum_{i=1}^{n} \widetilde{\omega}_i^*\right)\widetilde{\boldsymbol{\Sigma}}_n\right)$$

$$\widetilde{\boldsymbol{\mu}}_n^* = \left(\sum_{i=1}^{n} \frac{\rho_0'(\tilde{d}_i^*)}{\tilde{d}_i^*}\right)^{-1}\left(\sum_{i=1}^{n} \frac{\rho_0'(\tilde{d}_i^*)}{\tilde{d}_i^*}\boldsymbol{x}_i^*\right)$$

donde

$$d_i^* = \left[(\boldsymbol{x}_i^* - \widehat{\boldsymbol{\mu}}_n)^t\widehat{\boldsymbol{\Gamma}}_n^{-1}(\boldsymbol{x}_i^* - \widehat{\boldsymbol{\mu}}_n)\right]^{1/2}$$

$$\tilde{d}_i^* = \left[(\boldsymbol{x}_i^* - \widetilde{\boldsymbol{\mu}}_n)^t\widehat{\boldsymbol{\Sigma}}_n^{-1}(\boldsymbol{x}_i^* - \widetilde{\boldsymbol{\mu}}_n)\right]^{1/2}$$

$$\widetilde{\omega}_i^* = \rho_0(\tilde{d}_i^*) - \rho_0'(\tilde{d}_i^*)\tilde{d}_i^*$$

Es necesario notar que se siguen utilizando los estimadores $\widehat{\boldsymbol{\mu}}_n$, $\widehat{\boldsymbol{\Gamma}}_n$, $\widetilde{\boldsymbol{\Sigma}}_n$, $\widetilde{\boldsymbol{\mu}}_n$ fijos en el lado derecho de las ecuaciones anteriores. Estas aproximaciones posiblemente subestimen la variabilidad del estimador MM. Para remediar esta situación se aplica una corrección lineal que conduce a una aproximación $\widehat{\boldsymbol{\Gamma}}_n^{R*}$. Ahora, con el fin de obtener muestras *bootstrap* de los autovalores y autovectores de $\widehat{\boldsymbol{\Gamma}}_n$, se propone recalcular estimaciones de forma $\widehat{\boldsymbol{\Gamma}}_n^{R*}$ utilizando el FRB y tomar los autovalores y autovectores de estas estimaciones como las versiones recalculadas de los autovalores y autovectores.

Dada la consistencia de los estimadores para cierta distribución subyacente, la distribución del FRB converge a la misma distribución límite que converge el estimador MM. Esta convergencia puede ser mostrada para el estimador $\widehat{\boldsymbol{\Gamma}}_n$, y en consecuencia la propiedad se demuestra relativamente en forma sencilla para los autovalores y autovectores. Para más detalles sobre esta aproximación lineal se puede consultar en Salibián-Barrera et al. (2006).

Una manera de determinar la variabilidad de las estimaciones MM de los autovalores es a través de la utilización del FRB. Se puede estimar la variancia de los autovalores del estimador MM de forma o se pueden construir intervalos de confianza para los autovalores

de la matriz de forma $\boldsymbol{\Gamma}$. Un método para construir estos intervalos es el método BCa (*Bias Corrected and Accelerated*) para distintos niveles nominales de confianza. Para más detalles sobre este método puede consultarse Davison y Hinkley (1997).

Un gráfico de diagnóstico utilizado para el ACP muestra para cada observación su influencia empírica para el autovalor MM versus su distancia robusta basada en las estimaciones de locación y covariancias MM propuestas por Pison y Van Aelst (2004).

3. Aplicación

Uno de los objetivos que tienen los censos en general, es que los mismos pueden servir como base para la construcción de marcos de muestreo con los cuales desarrollar diseños muestrales para las encuestas que se realizan en los períodos intercensales. Existen dos posibilidades de marcos que tienen su origen en los censos: que el mismo constituya en sí mismo un marco completo de la población, o que a partir del mismo, se pueda construir una muestra maestra de la cual se puedan seleccionar muestras para las encuestas mencionadas. La primera de las opciones, si bien constituye un marco completo de la población, asumiendo que en el censo no se incurrió en problemas de cobertura, conlleva la contra que dado que los censos se realizan cada 10 años, las tareas de actualización del marco así planteado resultan muy difíciles de llevar a cabo.

Como contrapartida, una muestra maestra es un instrumento que permite seleccionar muestras, pero que está compuesta por sólo una parte de la población. La selección de la misma se realiza a partir de diseños muestrales probabilísticos, que permiten inferir a la población de la cual fue seleccionada. La ventaja de este instrumento es que las tareas de actualización son más fáciles de llevar a cabo debido al menor tamaño respecto al marco muestral completo.

Existen en el país antecedentes de muestras maestras. El Instituto Nacional de Estadística y Censos (INDEC) desarrolló luego del Censo Nacional de Población y Vivienda 1991, un Marco de Muestreo Nacional Urbano, que en realidad estaba formado por una muestra representativa de la población objetivo. Luego del Censo Nacional de Población, Hogares y Viviendas 2001, se amplió el marco anterior a zonas rurales además de las urbanas. En el Censo Nacional Población, Hogares y Viviendas 2010, se construyó la Muestra Maestra Urbana de Viviendas de la República Argentina. Estas muestras maestras sirvieron y sirven para la selección de muestras para encuestas a hogares. Pero las mismas no permiten dar estimaciones a niveles de desagregación inferiores a provincia, con la excepción de los aglomerados donde se realiza la EPH.

Por tal motivo, se plantea la necesidad de contar con una muestra maestra de viviendas que sea específica para la provincia de Santa Fe. Si bien es un proyecto preliminar, esta aplicación intenta dar respuesta a una de las primeras dificultades que constituye la estratificación de las localidades de la provincia, las cuales pueden constituir dentro del diseño, las unidades de muestreo de la primera etapa. El objetivo es lograr una estratificación de las localidades. La primera clasificación se refiere al tamaño, considerando la que proviene de la categorización en ciudad, comuna y localidades rurales. Las primeras están formadas por aquellas que tienen una población igual o mayor a 10000 habitantes. El segundo grupo, por aquellas que tienen una población entre 2000 y hasta 9999 habitantes. El último grupo corresponde a la definición de población rural dada por el INDEC que considera como tal a toda localidad con una población inferior a 2000 habitantes.

Dentro de cada uno de los grupos, se pretende estratificar a las localidades en estratos con características socio-demográficas homogéneas medidas a través de una serie de indicadores provenientes de datos censales.

Con el objetivo de ordenar las localidades de la provincia de Santa Fe respecto a sus condiciones socio-económicas de la población que habita en las mismas, se considera la información proveniente del Censo 2010 llevado a cabo por el INDEC. Los datos son indicadores de carencias críticas de las localidades y se encuentran publicados en la página del Instituto Provincial de Estadística y Censos[1] de la provincia mencionada. A continuación se brinda un detalle de los 10 indicadores que fueron utilizados para la aplicación del presente trabajo.

Indicadores (en porcentajes sobre el total correspondiente):

Educativos

JSA: Jefes de hogar sin asistencia escolar

JPI: Jefes de hogar con educación primaria incompleta

JASA: Jefas de hogar sin asistencia escolar

P1419: Población de 14 a 19 años que asiste al nivel de instrucción primario

PNINI: Población de 15 a 19 años que no estudia ni trabaja

Calidad y espacios en la vivienda

HH: Hogares con hacinamiento por cuarto

HPT: Hogares en viviendas con piso de tierra

Servicios básicos en la vivienda

HSA: Hogares en viviendas sin cañería de agua dentro de la vivienda

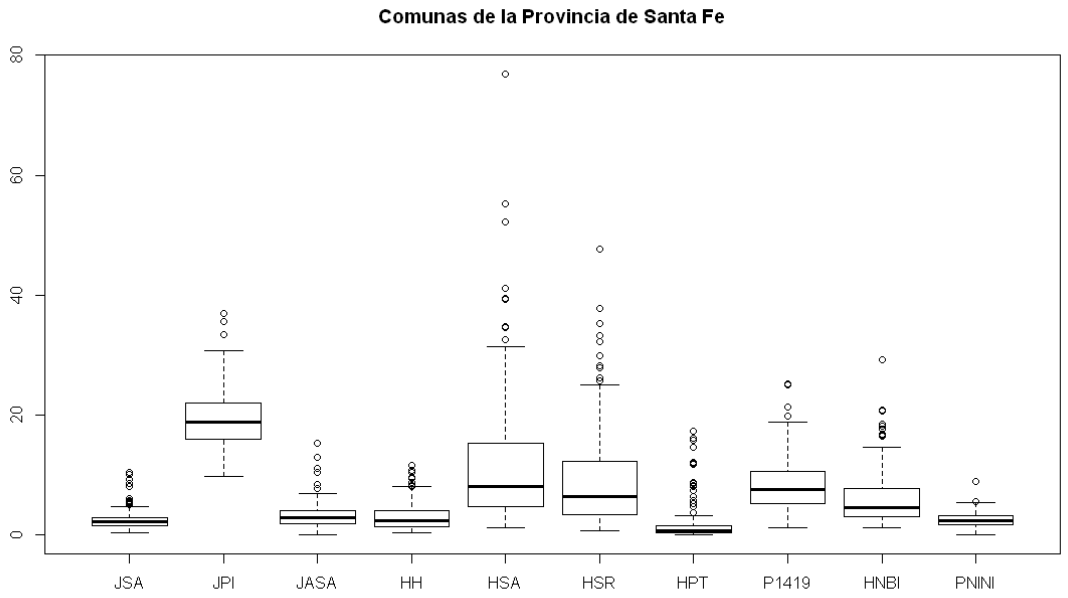
HSR: Hogares en viviendas sin retrete con descarga de agua

HNBI: Hogares con necesidades básicas insatisfechas (NBI)

[1] http://www.santafe.gov.ar/index.php/web/Estructura-de-Gobierno/Ministerios/Economia/Secretaria-de-Planificacion-y-Politica-Economica/Direccion-Provincial-del-Instituto-Provincial-de-Estadistica-y-Censos-de-la-Provincia-de-Santa-Fe/Temas-Especificos/Datos-Estadisticos/Poblacion/Censo-Nacional-Poblacion-Hogares-y-Viviendas-2010/Todos-los-Distritos-de-la-Provincia/Estadisticas/CARENCIAS/Indicadores-de-Carencias-Criticas-segun-Censo-Nacional-de-Poblacion-2010.-Provincia-Santa-Fe

Se utiliza la técnica de componentes principales con el objetivo de obtener a partir de la primera componente, un indicador que dé cuenta de las carencias críticas de las localidades. Se elige esta técnica estadística ya que permite resumir en un indicador agregado las diferentes dimensiones del fenómeno en estudio y permite ordenar las unidades de observación (comuna, localidad, etc.) según sus carencias sociales. En una primera instancia se trabaja a nivel comunas, puede resultar interesante llevar a cabo esta misma aplicación a nivel ciudades y localidades rurales.

Gráfico 1: Indicadores de carencias en las comunas de la Provincia de Santa Fe (en porcentaje)

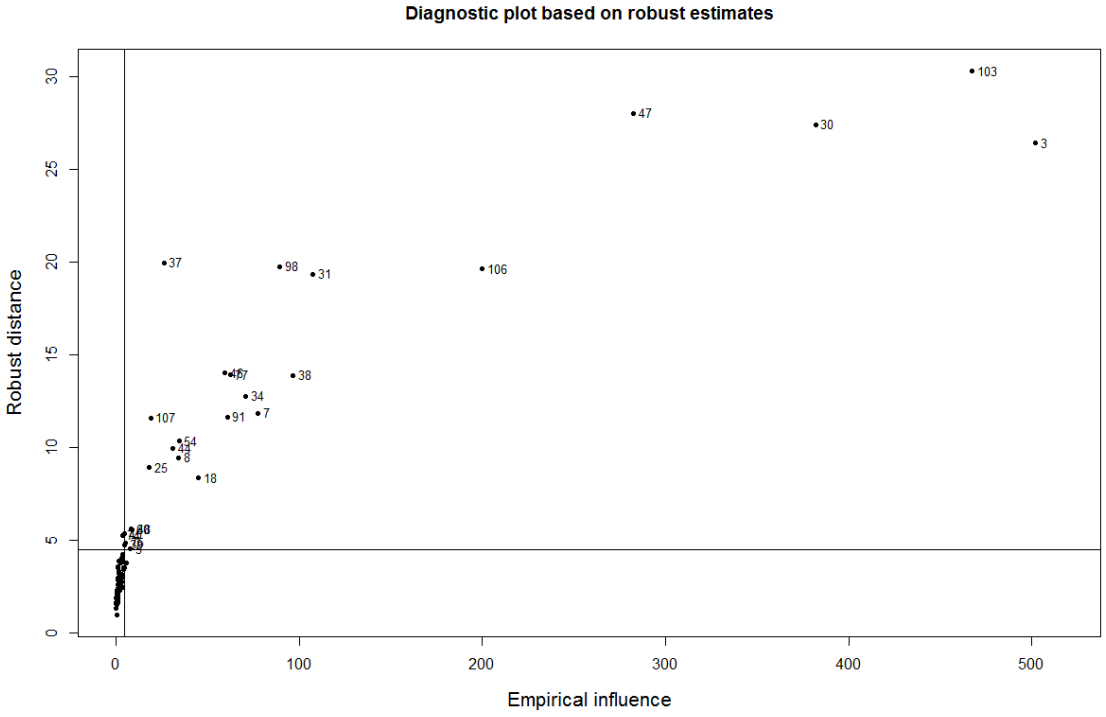


Fuente: Instituto Provincial de Estadística y Censos de la Provincia de Santa Fe.

Se puede observar que los estimadores MM detectan un grupo de 20 observaciones atípicas (con distancias robustas grandes) y que la mayoría de ellas parecen ser puntos de gran influencia para los vectores propios (Gráfico 2). Esto significa que podrían influir en gran medida en un análisis de componentes principales clásico. En cambio, tendrían poca influencia en el análisis de componentes principales robusto porque valores grandes de distancias robustas, por definición, se corresponden con pequeñas ponderaciones en los estimadores MM. Un listado con los códigos que identifican a cada comuna se encuentra en el Anexo.

Luego se procede a realizar un análisis de componentes principales robusto en lugar del clásico con el propósito de asegurar que los estimadores no están influenciados por la presencia de observaciones atípicas y detectar posibles *outliers* en el conjunto de datos.

Gráfico 2: Influencia empírica global de los vectores propios frente a la distancia robusta basado en los estimadores MM



Los estimadores MM de los valores propios de la matriz de forma se presentan en la Tabla 1 junto con los límites de confianza del 95% obtenidos a partir del método *bootstrap* rápido y robusto. Los intervalos de confianza son del tipo BCa (*bias corrected and accelerated*).

Tabla 1: Estimadores MM de los valores propios e intervalos de confianza del 95% basado en el método FRB

Autovalor	**CP1**	**CP2**	**CP3**	**CP4**	**CP5**	**CP6**	**CP7**	**CP8**	**CP9**	**CP10**
Estimador MM	30.0	6.89	5.77	1.87	0.88	0.66	0.60	0.14	0.10	0.09
Límite inferior IC	20.8	4.76	4.46	1.39	0.65	0.50	0.54	0.10	0.07	0.09
Límite superior IC	39.3	8.52	8.25	2.45	1.12	0.85	0.80	0.18	0.14	0.12

A partir de la Tabla 2 y el Gráfico 3, se puede ver que la primera componente principal tiene todas sus coordenadas de signo positivo y puede interpretarse como un promedio ponderado de todas las variables. Luego es un factor que agrega variables de educación, de servicios básicos en la vivienda, de calidad y espacios en la misma. Es decir, proporciona el resumen de tres indicadores de carencias sociales (educación, servicios básicos y espacios en la vivienda) en un sólo índice que se puede utilizar con la finalidad de ordenar a las unidades de observación según sus carencias sociales.

El Indicador de Carencia, obtenido a partir de la primera componente, no se trata de una medición de pobreza ya que no incluye indicadores de ingreso, seguridad social y alimentación (estas variables no están explícitas en el Censo Nacional de Población, Hogares y Viviendas). Su cálculo a nivel comunal puede contribuir con la generación de datos para la toma de decisiones en materia de política social, especialmente para analizar la desigualdad de coberturas sociales que subsisten en el territorio nacional.

Tabla 2: Peso de las componentes principales obtenidas con el método FRB

Variable	**CP1**	**CP2**	**CP3**	**CP4**	**CP5**	**CP6**	**CP7**	**CP8**	**CP9**	**CP10**
JSA	0.07	0.03	0.09	0.21	0.25	0.22	-0.12	0.12	0.84	-0.31
JPI	0.30	-0.43	0.83	-0.01	-0.15	-0.09	0.06	0.01	-0.02	0.01
JASA	0.07	0.01	0.15	0.41	0.66	0.38	-0.21	-0.03	-0.40	0.10
HH	0.12	0.13	-0.03	0.37	-0.14	-0.20	-0.04	0.71	0.06	0.51
HSA	0.71	-0.27	-0.36	-0.30	0.02	0.05	-0.44	0.04	-0.01	0.04
HSR	0.52	0.13	-0.16	0.13	0.08	0.12	0.80	-0.01	-0.04	-0.09
HPT	0.07	0.06	0.01	0.07	0.03	-0.03	0.03	-0.60	0.33	0.72
P1419	0.24	0.81	0.34	-0.35	0.08	-0.08	-0.16	0.01	-0.04	-0.03
HNBI	0.23	0.16	-0.06	0.63	-0.28	-0.39	-0.24	-0.33	-0.10	-0.33
PNINI	0.02	0.14	0.05	0.12	-0.60	0.76	-0.10	-0.03	-0.06	0.04

Se puede apreciar en el Gráfico 3 que todos los pesos se encuentran en el intervalo [-1, 1] ya que corresponden a los coeficientes de los vectores propios normalizados. Se puede concluir que la estimación de la primera componente principal es bastante precisa, puesto que la amplitud de los intervalos de confianza es relativamente pequeña.

Gráfico 3: Pesos de la primera componente principal a partir del método FRB

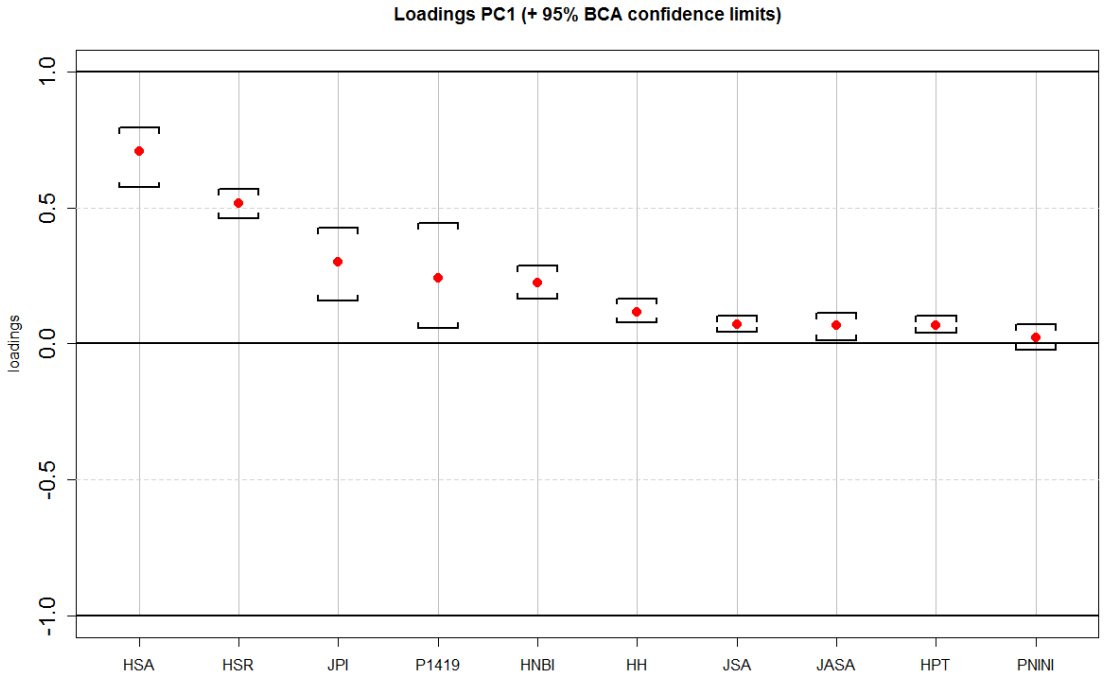
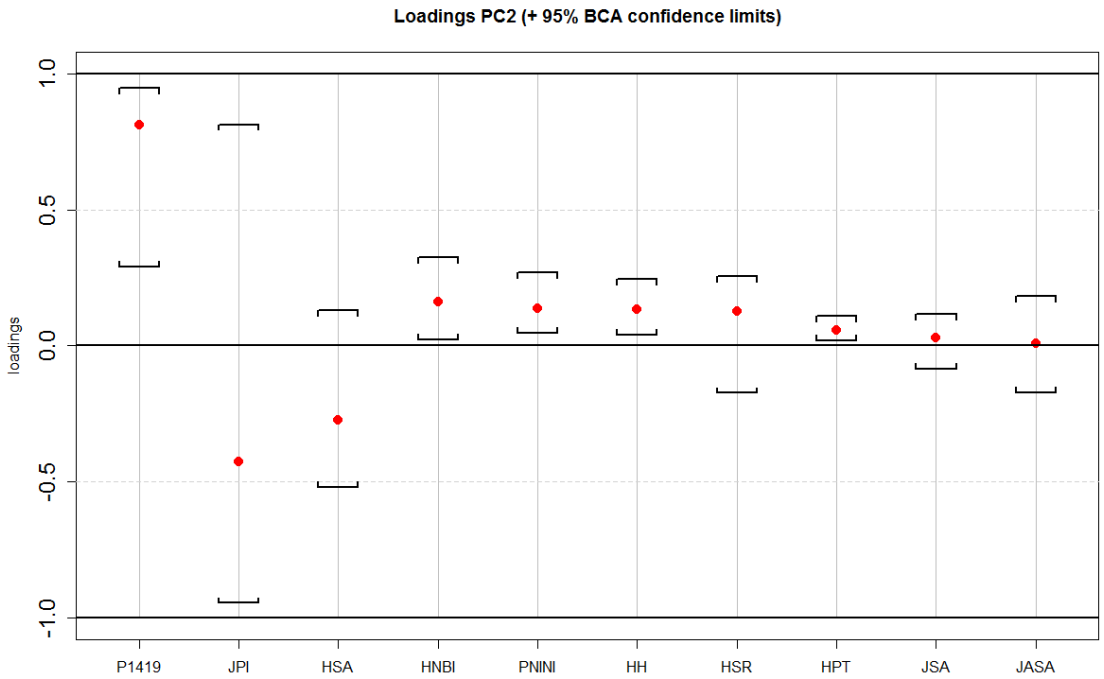
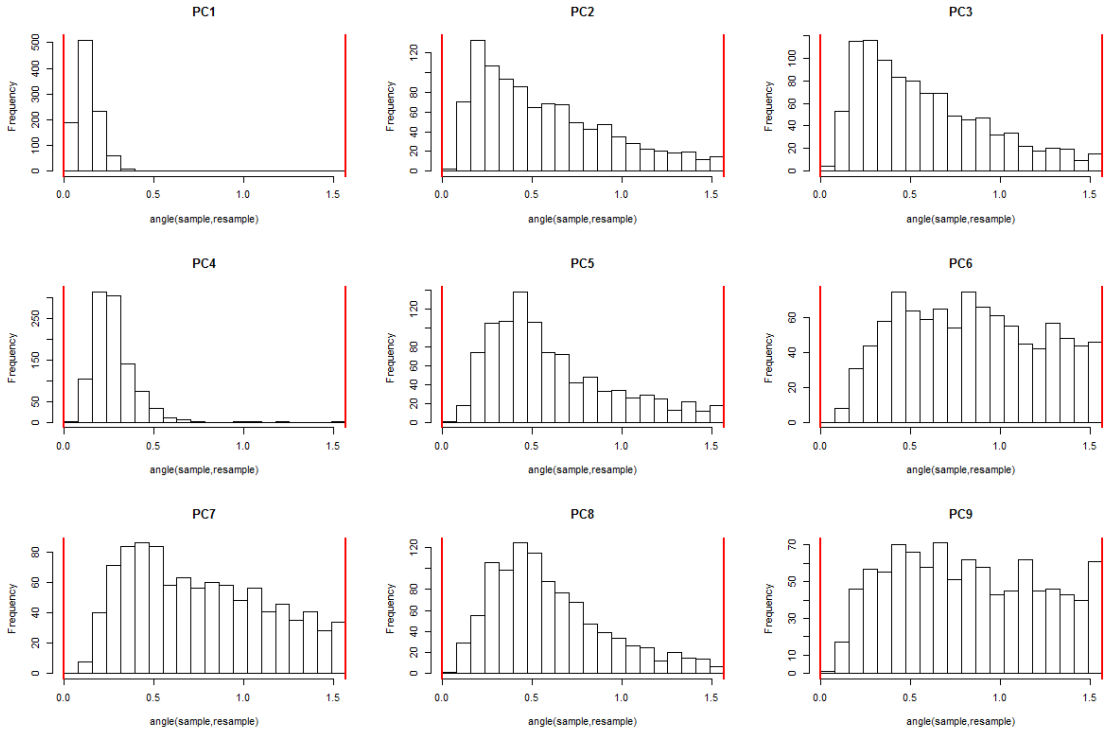


Gráfico 4: Pesos de la segunda componente principal a partir del método FRB



El gráfico 5 muestra, para cada componente principal, los histogramas de los ángulos entre la componente original y la correspondiente componente *bootstrap*. El ángulo entre las dos componentes se expresa por un valor comprendido entre 0 y $\pi/2$. Se indican estos límites por dos líneas verticales en los histogramas. Se puede ver, por ejemplo, que la estimación de la primera componente está más alineada con sus versiones *bootstrap* (la mayoría de los ángulos son cercanos a cero), indicando una baja variabilidad de dicha estimación. Para el resto de las componentes se observa bastante más inestabilidad.

Gráfico 5: Histogramas para los ángulos entre las estimaciones originales y *bootstrap* FRB de las componentes principales



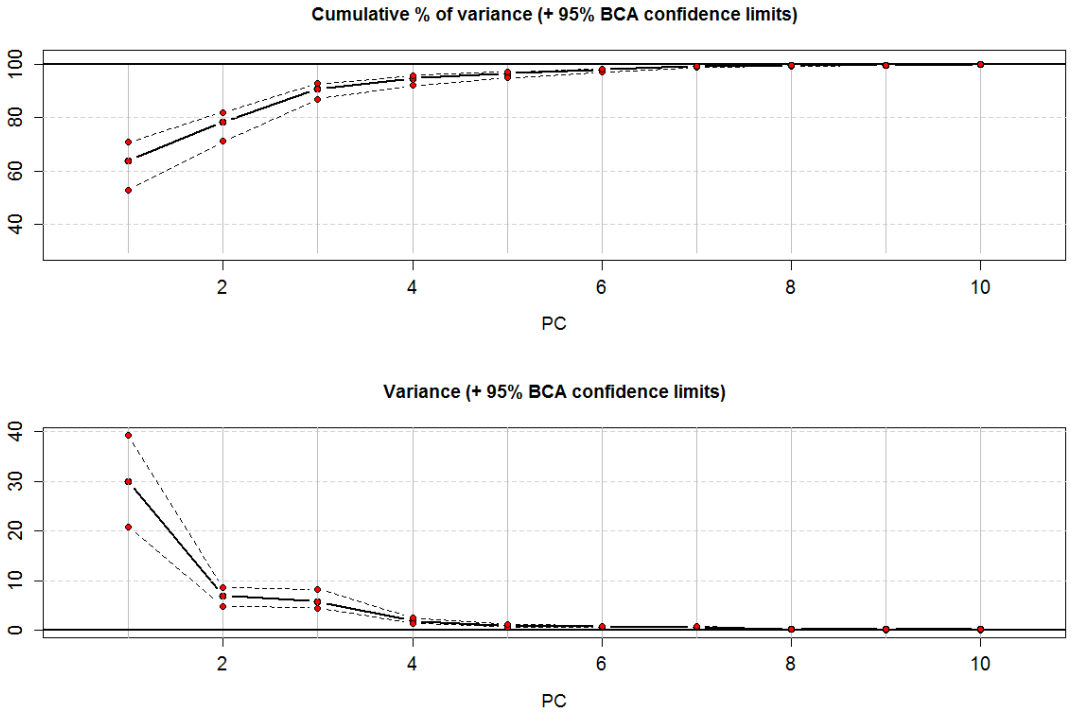
En el gráfico 6 se presenta el porcentaje de variancia explicada por cada componente, junto con los intervalos de confianza basados en el método FRB. Se observa que la primera componente principal explica alrededor del 63% de la variabilidad total; sin embargo el límite inferior del intervalo de confianza se encuentra por debajo del 60%. En general, al seleccionar el número de componentes para retener para su posterior análisis sobre la base de estos porcentajes, puede ser más seguro tener en cuenta los límites inferiores en lugar de los porcentajes estimados.

En este punto es interesante comentar que la primera componente obtenida a partir del método clásico explica una proporción mucho más alta de la variabilidad, el 89,2%. Como ya se mencionó no sería conveniente utilizar la estimación clásica debido a que se detectaron observaciones atípicas y las mismas podrían influenciar las conclusiones (Gráfico 2).

Tabla 3: Porcentaje de variabilidad explicada por cada componente: estimación clásica y método FRB

Componente	CP clásicos	CP robusto	
		Estimación puntual	Intervalo de confianza del 95%
1	89.2	63.8	53.1; 70.8
2	93.4	78.5	71.1; 82.1
3	95.9	90.8	87.0; 92.8
4	97.5	94.7	92.1; 95.8
5	98.3	96.6	94.9; 97.2
6	99.0	98.0	97.2; 98.3
7	99.4	99.3	98.9; 99.5
8	99.8	99.6	99.4; 99.7
9	99.9	99.8	99.7; 99.8
10	100.0	100.0	100.0; 100.0

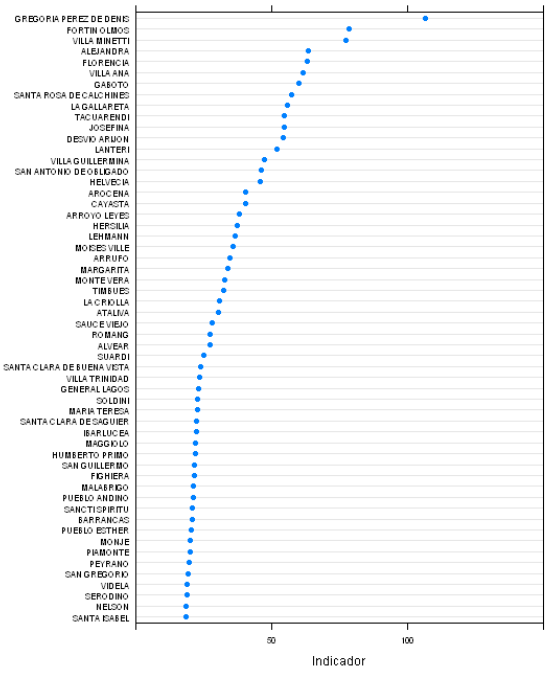
Gráfico 6: Porcentaje de variabilidad explicada por cada componente basado en el método FRB

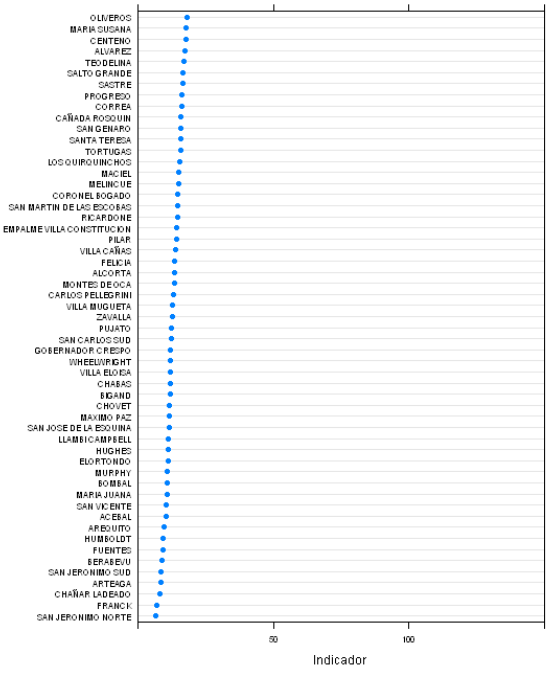


Teniendo en cuenta la primera componente, se construye el Indicador de Carencia para cada una de las comunas, las cuales se presentan en el siguiente gráfico. Puede apreciarse

que las comunas con mayor grado de carencias sociales son: Gregoria Perez de Denis, Fortín Olmos, Villa Minetti, Alejandra, Florencia y Villa Ana.

Gráfico 7: Representación gráfica de las comunas de la Provincia de Santa Fe, según la primera componente





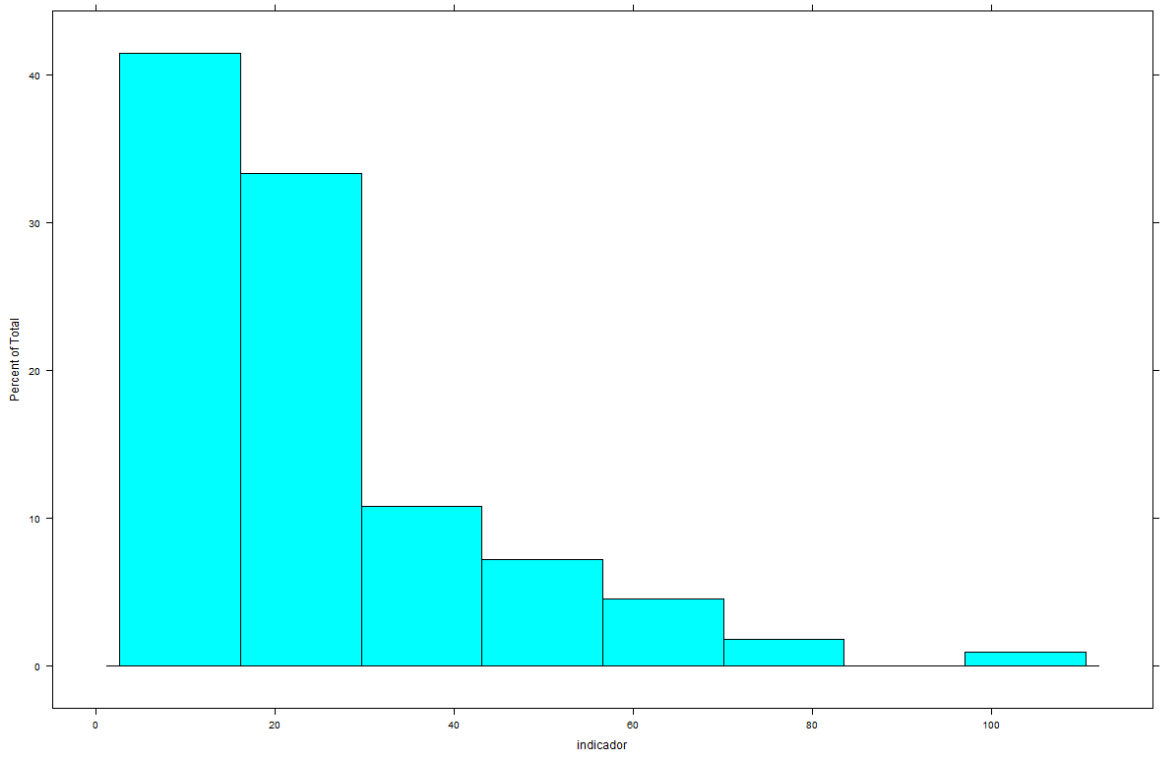
A continuación, el objetivo es construir estratos a partir del Indicador de Carencia. La justificación del uso de esta variable, es que en la mayoría de las encuestas que se realizan en el ámbito de las estadísticas oficiales, la misma se encuentra altamente correlacionada con las variables bajo estudio en dichas encuestas.

Uno de los elementos que se deben tener en cuenta a la hora de estratificar una población es la distribución de la variable bajo estudio. Existen en la literatura un conjunto de métodos de estratificación que son válidos para distribuciones simétricas, tales como Dalenius y Hodges (1959), y Ekman (1959).

El problema de estos métodos es que no brindan soluciones precisas cuando la distribución de la variable bajo estudio es asimétrica. Existen métodos que son apropiados para este tipo de distribuciones, entre los cuales existen métodos aproximados, como por ejemplo Gunning y Horgan (2004), que resultan ser menos precisos que los óptimos pero de una aplicación más simple.

Debido a que la distribución de la variable Indicador de Carencia es asimétrica (Gráfico 8), se va a considerar el método de Gunning y Horgan (2004). Esta estrategia recurre a una progresión geométrica para determinar los límites de los estratos, bajo el supuesto de obtener coeficientes de variación (CV) aproximadamente iguales en cada uno de los estratos. La misma se basa en una observación de Cochran (1961).

Gráfico 8: Histograma del indicador de carencia para las comunas de la Provincia de Santa Fe



El objetivo al estratificar una población es subdividirla en intervalos, siendo los límites de éstos $k_0 < k_1 < \cdots < k_L$. Dado que no se cuenta con la variable bajo estudio, *y,* se utiliza una variable auxiliar conocida *x* que esté correlacionada con la variable de interés.

Estos límites se obtienen teniendo como objetivo que los coeficientes de variación $CV_h = S_{xh}^2/\bar{x}_h$ sean iguales para todos los estratos $h = 1, \dots, L$, o sea,

$$\frac{S_{x1}}{\bar{x}_1} = \frac{S_{x2}}{\bar{x}_2} = \cdots = \frac{S_{xL}}{\bar{x}_L}$$

Bajo el supuesto que la distribución dentro de cada estrato es aproximadamente uniforme, tenemos que

$$\bar{x}_h \approx \frac{k_h + k_{h-1}}{2}$$

$$S_{xh} \approx \frac{1}{\sqrt{12}}(k_h - k_{h-1})$$

Luego, una aproximación a los coeficientes de variación viene dada por

$$CV_h = \frac{(k_h - k_{h-1})/\sqrt{12}}{(k_h - k_{h-1})/2}$$

con lo cual para obtener coeficientes de variación iguales se debe cumplir

$$\frac{k_{h+1} - k_h}{k_{h+1} + k_h} = \frac{k_h - k_{h-1}}{k_h + k_{h-1}}$$

Esta relación recurrente se reduce a

$$k_h^2 = k_{h+1}k_{h-1}$$

siendo los límites de los estratos los términos de una progresión geométrica

$$k_h = ar^h \qquad (h = 0{,}1, \dots, L)$$

Luego $a = k_0$, el valor mínimo de la variable, y $ar^L = k_L$, el máximo de la variable, con lo cual la razón constante puede calcularse como $r = (k_L/k_0)^{1/L}$.

Para la presente aplicación se consideró la construcción de 5 estratos, los cuales van desde el estrato 1 (muy baja carencia) hasta el estrato 5 (muy alta carencia). A continuación se presentan para cada uno de los nodos en los cuales se divide la provincia que se encuentra

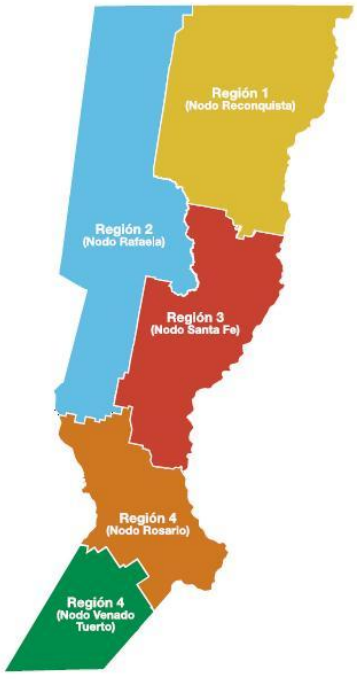
descripto en el Plan Estratégico Provincial Santa Fe[2] (Gráfico 9), el porcentaje de comunas que pertenecen a cada uno de los estratos mencionados.

Tabla 4: Clasificación de las comunas de la Provincia de Santa Fe de acuerdo al nodo y al estrato de carencia

Nodo	**Estrato según nivel de carencia**					
	1	**2**	**3**	**4**	**5**	**Total**
Venado Tuerto	6	8	3			17
Rosario	6	26	8			40
Santa Fe	4	9	5	7		25
Rafaela	2	2	7	4	2	17
Reconquista			3	5	4	12
Total	**18**	**45**	**26**	**16**	**6**	**111**

Se puede observar que a medida que nos movemos entre Nodos de sur a norte, tienden a aparecer una mayor cantidad de comunas en aquellos estratos que denotan una mayor carencia en los indicadores considerados.

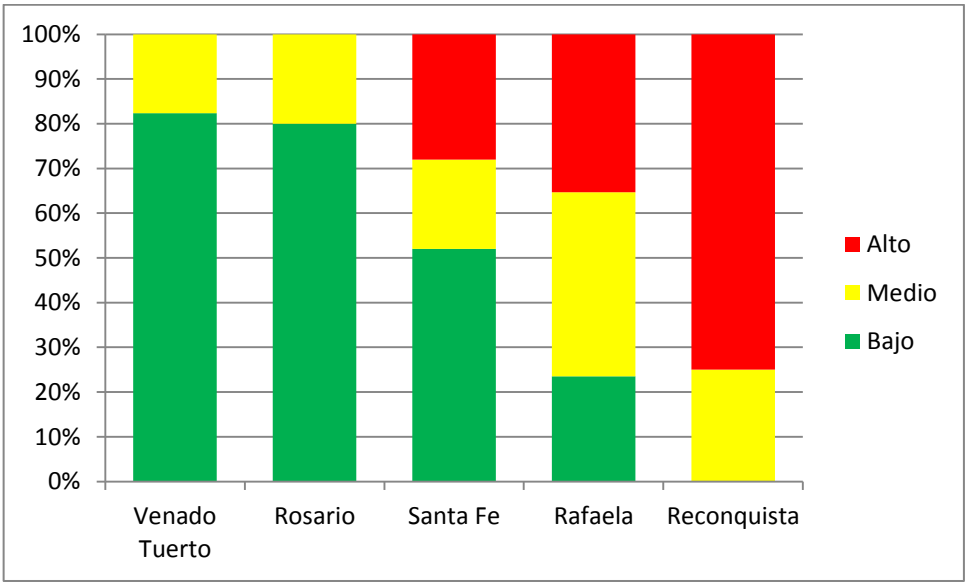
Gráfico 9: Ubicación de los Nodos de la provincia de Santa Fe de acuerdo al Plan Estratégico Provincial



[2] http://www.santafe.gov.ar/index.php/web/guia/gobernador?cat=plan_estrategico

Considerando los estratos 1 y 2 como baja carencia, el estrato 3 como carencia media, y los estratos 4 y 5 como alta carencia, se muestra en el Gráfico 10, la distribución de las comunas de acuerdo al grupo al cual pertenecen en cada uno de los nodos mencionados. Como se puede observar, a medida que se va de sur a norte en la provincia, las comunas tienden a tener mayores carencias.

Gráfico 10: Distribución de comunas por grado de carencia en cada nodo de la provincia de Santa Fe



4. Conclusiones

En este trabajo se presentó un análisis de componentes principales robusto a partir de los estimadores MM. Por otro lado, se consideró el método *Bootstrap* Rápido y Robusto para estimar intervalos de confianza para la proporción de variancia explicada de las componentes, y para calcular los límites de confianza de las cargas de las componentes principales robustas. Se desarrolló una aplicación de estos métodos a datos correspondientes a indicadores de carencias de comunas de la provincia de Santa Fe provenientes del Censo Nacional de Población, Hogares y Viviendas 2010, con el objetivo de lograr una estratificación de las mismas para un futuro marco de muestreo. A partir de la primera componente robusta se construyó un Índice de carencias, a partir del cual se estratificó a las comunas en cinco estratos a partir del método geométrico, el cual es apropiado para poblaciones asimétricas. Se observó que el estrato al cual pertenecen las comunas está relacionado con el nodo donde se encuentran ubicadas, encontrándose que a medida que uno se mueve de sur a norte en la provincia, las comunas tienden a tener mayores carencias.

REFERENCIAS BIBLIOGRÁFICAS

Ekman, G. (1959). An Approximation Useful in Univariate Stratification. *Annals of Mathematical Statistics*, 30, 219-229.

Cochan, W.G. (1961). Comparison of methods for determining stratum boundaries. *Bulletin of the International Statistical Institute*, 32, 345-358.

Davison, A. C.; Hinkley, D. V. (1997).Bootstrap Methods and their Application, *Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.*

Efron, B., Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Efron, B. (1979). Bootstrap Methods: another look at the jackknife. *Annals of Statistics*, 7, 1-26.

Gunning, P. and Horgan, J. M. (2004). A simple algorithm for stratifying skewed populations. *Survey Methodology*, 30, 159-166

Hampel, F. (1968).*Contributions to the theory of robust estimation*.PhD. Thesis, University of California, Berkeley.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley & Sons.

Hastie, T. T. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (Second ed.). New York: Springer-Verlag.

Huber, P., Ronchetti, E. (2009). *Robust Statistics (Second ed.).* Hoboken, New Jersey: John Wiley & Sons, Inc.

Johnson, R.; Wichern, D. (2007). *Applied Multivariate Statistical Analysis, 6th edition*. Pearson Education Limited.

Jureckova, J., Picek, J. (2006). *Robust Statistical Methods with R.* Boca Raton, Florida: Chapman & Hall/CRC. John Wiley and Sons, Ltd.

Maronna, R.A.; Martin, R.D.; Yohai, V.J. (2006), *Robust Statistics: Theory and Methods*. John Wiley and Sons.

Peña, D. (2002). *Análisis de Datos Multivariantes*. McGraw-Hill/Interamericana de España.

Pison, G.; Van Aelst, S. (2004). Diagnostic Plots for Robust Multivariate Methods, *Journal of Computational and Graphical Statistics, 13, 310-329.*

Salibian-Barrera, M. (2000). *Contributions to the theory of robust inference*. Ph.D. thesis, Dept. Statist., Univ. British Columbia,Vancouver.

Salibian-Barrera, M., Van Aelst S., Willems G. (2006). PCA Based on Multivariate MM-Estimators with Fast and Robust Bootstrap. *Journal of the American Statistical Association,* 101, 1198-1211.

Salibian-Barrera, M., Zamar, R. H. (2002). Bootstrapping robust estimates of regression.*The Annals of Statistics*, 30, 556-582.

Van Aelst, S., Willems, G. (2005).Multivariate regression S-estimators for robust estimation and inference.*StatisticaSinica*, 15, 981-1001.

Van Aelst, S., Willems, G. (2013). Fast and robust bootstrap for multivariate inference: The R package FRB. *Journal of Statistical Software*, 53 (3),1–32.URL: http://www.jstatsoft. org/v53/i03/.

ANEXO

Tabla A.1: Códigos de las comunas de la Provincia de Santa Fe

Código	Comuna	Código	Comuna	Código	Comuna
1	Acebal	38	Helvecia	75	Romang
2	Alcorta	39	Hersilia	76	Salto Grande
3	Alejandra	40	Hughes	77	San Antonio de Obligado
4	Alvarez	41	Humberto Primo	78	San Carlos Sud
5	Alvear	42	Humboldt	79	San Genaro
6	Arequito	43	Ibarlucea	80	San Gregorio
7	Arocena	44	Josefina	81	San Guillermo
8	Arroyo Leyes	45	La Criolla	82	San Jerónimo Norte
9	Arrufo	46	La Gallareta	83	San Jerónimo Sud
10	Arteaga	47	Lanteri	84	San José de la Esquina
11	Ataliva	48	Lehmann	85	San Martín de las Escobas
12	Barrancas	49	Llambi Campbell	86	San Vicente
13	Berabevu	50	Los Quirquinchos	87	Sancti Spiritu
14	Bigand	51	Maciel	88	Santa Clara de Buena Vista
15	Bombal	52	Maggiolo	89	Santa Clara de Saguier
16	Cañada Rosquín	53	Malabrigo	90	Santa Isabel
17	Carlos Pellegrini	54	Margarita	91	Santa Rosa de Calchaquíes
18	Cayastá	55	María Juana	92	Santa Teresa
19	Centeno	56	María Susana	93	Sastre
20	Chabás	57	María Teresa	94	Sauce Viejo
21	Chañar Ladeado	58	Máximo Paz	95	Serodino
22	Chovet	59	Melincué	96	Soldini
23	Coronel Bogado	60	MoisesVille	97	Suardi
24	Correa	61	Monje	98	Tacuarendi
25	Desvío Arijoçón	62	Monte Vera	99	Teodelina
26	Elortondo	63	Montes de Oca	100	Timbués
27	Empalme Villa Constitución	64	Murphy	101	Tortugas
28	Felicia	65	Nelson	102	Videla
29	Fighiera	66	Oliveros	103	Villa Ana
30	Florencia	67	Peyrano	104	Villa Cañás
31	Fortín Olmos	68	Piamonte	105	Villa Eloisa
32	Franck	69	Pilar	106	Villa Guillermina
33	Fuentes	70	Progreso	107	Villa Minetti
34	Gaboto	71	Pueblo Andino	108	Villa Mugueta
35	General Lagos	72	Pueblo Esther	109	Villa Trinidad
36	Gdor Crespo	73	Pujato	110	Wheelwright
37	Gregoria Pérez de Denis	74	Ricardone	111	Zavalla

Tabla A.2: Ordenación creciente de las comunas de la Provincia de Santa Fe, según el Indicador de Carencia obtenido a partir de la primera componente

Estrato*	Comuna	Estrato	Comuna	Estrato	Comuna
1	San Jerónimo Norte	2	San Martín de las Escobas	3	María Teresa
1	Franck	2	Coronel Bogado	3	Soldini
1	Chañar Ladeado	2	Melincué	3	General Lagos
1	Arteaga	2	Maciel	3	Villa Trinidad
1	San Jerónimo Sud	2	Los Quirquinchos	3	Santa Clara de Buena Vista
1	Berabevu	2	Tortugas	3	Suardi
1	Fuentes	2	Santa Teresa	3	Alvear
1	Humboldt	2	San Genaro	3	Romang
1	Arequito	2	Cañada Rosquín	3	Sauce Viejo
1	Acebal	2	Correa	3	Ataliva
1	San Vicente	2	Progreso	3	La Criolla
1	María Juana	2	Sastre	3	Timbués
1	Bombal	2	Salto Grande	3	Monte Vera
1	Murphy	2	Teodelina	3	Margarita
1	Elortondo	2	Alvarez	3	Arrufo
1	Hughes	2	Centeno	4	Moises Ville
1	Llambi Campbell	2	María Susana	4	Lehmann
1	San José de la Esquina	2	Oliveros	4	Hersilia
2	Máximo Paz	2	Santa Isabel	4	Arroyo Leyes
2	Chovet	2	Nelson	4	Cayastá
2	Bigand	2	Serodino	4	Arocena
2	Chabas	2	Videla	4	Helvecia
2	Villa Eloisa	2	San Gregorio	4	San Antonio de Obligado
2	Wheelrwright	2	Peyrano	4	Villa Guillermina
2	Gobernador Crespo	2	Piamonte	4	Lanteri
2	San Carlos Sud	2	Monje	4	Desvío Arijón
2	Pujato	3	Pueblo Esther	4	Josefina
2	Zavalla	3	Barrancas	4	Tacuarendí
2	Villa Mugueta	3	Sancti Sprititu	4	La Gallareta
2	Carlos Pellegirni	3	Pueblo Andino	4	Santa Rosa de Calchines
2	Montes de Oca	3	Malabrigo	4	Gaboto
2	Alcorta	3	Fighiera	5	Villa Ana
2	Felicia	3	San Guillermo	5	Florencia
2	Villa Cañas	3	Humberto Primo	5	Alejandra
2	Pilar	3	Maggiolo	5	Villa Minetti
2	Empalme Villa Constitución	3	Ibarlucea	5	Fortín Olmos
2	Ricardone	3	Santa Clara de Saguier	5	Gregoria Perez de Denis

* Estrato 1: muy baja carencia, Estrato 2: baja carencia, Estrato 3: carencia media, Estrato 4: alta carencia y Estrato 5: muy alta carencia.