

RAESTA 2 - AÑO 2 (2015)

EDITORIAL

Tenemos el agrado de presentarles el segundo número de nuestra Revista Argentina de Estadística Aplicada como un continuum de nuestro compromiso con el enriquecimiento del conocimiento académico y profesional que llevamos adelante desde la Maestría en Generación y Análisis de Información Estadística. Este aporte que intentamos construir es producto del esfuerzo llevado adelante por todos los miembros participantes del comité editorial y académico. En línea con este nuevo número estamos presentando el nuevo portal con el rediseño del soporte, adaptándola a las nuevas pautas elaboradas por el Área de Desarrollo Digital de UNTREFMEDIA

Los artículos que integran este número, fueron previamente evaluados por integrantes de nuestro Comité Académico Editorial y expresan los resultados de la aplicación de herramientas estadísticas para la generación de información y análisis estadístico en áreas temáticas diversas y con distintos niveles de desarrollo, los que, a partir de sus resúmenes y en forma sucinta se describen a continuación.

En el primer artículo, *Aportes al estudio de la opinión pública en las elecciones presidenciales 2007 en Argentina* su autor, Bruno De Santis, egresado de nuestra maestría, muestra la importancia creciente de los estudios de opinión pública en el desarrollo de las estrategias electorales de sus actores y los cambios observados en las percepciones, ex ante y ex post de la ciudadanía, a partir de los resultados de las elecciones presidenciales 2007. Como aporte significativo, se destaca la creciente injerencia de los medios de comunicación en el desarrollo de los procesos electorales de los últimos años.

El segundo artículo, cuya autoría es de Ma. Fernanda Artola e Iván Redini Blumenthal, **Construcción de un índice del nivel socioeconómico del hogar urbano en la República Argentina mediante el análisis de correspondencia múltiple y escalamiento óptimo**, muestra cómo resolvieron la falta de información de ingresos, en un estudio de economía de la salud para segmentar según nivel socioeconómico, a la población de hogares urbanos objeto de estudio. El abordaje del tema lo efectuaron aplicando análisis de correspondencia múltiple y escalamiento óptimo para obtener los ponderadores de las doce variables seleccionadas. Los datos utilizados fueron los de la Encuesta Nacional de Gastos de los Hogares (INDEC-2012/2013). Las variables seleccionadas para calcular el Índice, están referidas al comportamiento de las características de

la vivienda, la condición laboral y educativa del jefe de hogar y del cónyuge relevada en la ENGHO cerrando el proceso de cálculo con la asignación de un valor de quintil cada hogar. Finalmente, se analizan las características de cada uno de los quintiles obtenidos y se indican las fortalezas y limitaciones de la metodología aplicada.

Las **Orientaciones de futuro laboral y educativo de estudiantes secundarios. Análisis multivariado en un diseño muestral complejo** es el tercer artículo presentado cuya autoría corresponde a Rosario Austral, en donde se analizan las orientaciones de futuro laboral y educativo de los estudiantes que a fines de 2008 cursaban el último año de estudio en las escuelas estatales de la Ciudad Autónoma de Buenos Aires. El trabajo de relevamiento de la información de base fue realizado por el Equipo de la Gerencia Operativa de Investigación y Estadística del Ministerio de Educación del GCABA, de la cual la autora forma parte.

Rosario Austral, es egresada de nuestra Maestría y, en su tesis, profundizó el análisis de las dimensiones ó variables del estudio, avanzando además en la exploración de algunos modelos de análisis multivariado sobre los objetivos específicos, utilizando los estimadores más adecuados atendiendo a las características del diseño muestral de tipo complejo aplicado.

En cuanto a la estructura del artículo, luego del marco introductorio, se presenta de manera resumida el modelo analítico utilizado y se describen los principales aspectos metodológicos y estadísticos del trabajo. A partir de allí se exponen los resultados más destacados: la descripción de las orientaciones de futuro laboral y educativo, la modelización multivariada sobre los objetivos de tipo profesional a través de un análisis de regresión logística y algunos análisis que demuestran la importancia de la consideración de la complejidad del diseño muestral en la instancia de análisis de los datos. Por último, se presenta una síntesis de los principales hallazgos.

El siguiente artículo, **Condiciones de socialización, entorno y trayectoria asociados a la reincidencia en el delito** ha sido desarrollado en el marco de un Programa de Investigación del Centro de Estudios Latinoamericano sobre Inseguridad y Violencia de la UNTREF, dirigido por Marcelo Bergman. Dicho Programa se basa en un estudio sobre presos condenados en Argentina, Brasil, Chile, Perú, El Salvador y México. Los autores, Marcelo Bergman, Diego Masello, Christian Arias y Guadalupe Peralta Agüero, han realizado un modelo de análisis multivariado basado en los datos provenientes del relevamiento a 1.033 presos condenados de los sistemas penitenciarios Federal y Bonaerense. El artículo cita como referencia de partida que, en la actualidad, el delito y la punición son analizados desde diferentes enfoques, ya que además de estar en crisis son temas a los cuales se les busca encontrar respuestas y soluciones. Es aceptado por los especialistas en esta problemática que dentro de las cárceles la población es mayoritariamente joven, con bajos niveles de educación y provenientes de clases socioeconómicas medias/bajas y bajas, caracterizadas, entre otras cosas, por los bajos niveles de ingreso.

En este documento se propone que la vinculación explicativa de una conducta delictiva no está dada en forma unívoca por la asociación entre la condición de pobreza del hogar y/o de sus

integrantes y las probabilidades de comisión de delitos y la reincidencia en los mismos, sino que es la consecuencia de la incidencia de otros factores que se relacionan en forma mucho más compleja. Por lo tanto, cabe preguntarse en relación al delito y a la reincidencia en el mismo, ¿qué tan asociados están esos factores a quienes incurren nuevamente en una conducta delictiva y al nivel de violencia al momento de perpetrar un delito?, finalmente, ¿qué factores son estos?

Para responder a estas preguntas se realizó un modelo de análisis multivariado basado en una regresión logística, en el cual se incorporaron variables relacionadas con los entornos o contextos de socialización temprana de los sujetos así como la trayectoria en instituciones como los Institutos de Menores. Los datos utilizados pertenecen a la Encuesta a Población en Reclusión de 2013, en la cual para Argentina se aplicaron más de mil encuestas personales a presos condenados por la justicia federal y ordinaria de la Capital así como por la justicia de la Provincia de Buenos Aires.

Horacio Chitarroni desarrolla: **Un ejemplo de diseño cuasi experimental: uso de la regresión logística binaria en la construcción de un grupo de comparación para la evaluación de impacto de un programa social.** El autor propone el empleo de la regresión logística binaria para la construcción de un grupo de comparación, como testigo para la evaluación del impacto de un programa social. Se basa en una experiencia de aplicación real de tal procedimiento

En la primera parte, aborda brevemente la problemática que plantea la aplicación de diseños puramente experimentales para el caso de la evaluar la implementación de políticas públicas de contenido social y la alternativa de emplear modelos cuasi experimentales con un grupo de comparación construido estadísticamente. También pone en consideración algunos aspectos inherentes a los diseños con doble medición, atendiendo a las dificultades que plantea la frecuente ausencia de una línea de base en el abordaje de Programas Sociales.

La segunda parte se refiere a las características del procedimiento estadístico empleado (la regresión logística binaria) y su utilidad específica para la obtención de grupos de comparación, con las limitaciones e inconvenientes que plantea, las alternativas posibles para sortearlos y los recaudos a adoptar. Finalmente, en la última parte se exponen los resultados provenientes del ejemplo de aplicación de este procedimiento conjuntamente con la interpretación de los mismos.

El último artículo de esta sección, tiene la autoría de Norberto Rodríguez y Julieta Mirensky quienes presentan una **Metodología estadística para la estimación de las superficies sembradas con cultivos extensivos - Método de segmentos aleatorios.** Esta metodología es la que utiliza el Ministerio de Agricultura Ganadería y Pesca (MAGyP) desde la campaña agrícola de los años 2011/2012 para estimar las superficies sembradas con cultivos de tipo extensivos en las principales regiones y jurisdicciones del país, incorporando nuevas tecnologías para el relevamiento y análisis, entre ellas: imágenes satelitales, Sistemas de Información Geográfica (GIS), el GPS, el uso del Índice de Vegetación de Diferencia Normalizada (NDVI).

El método aquí propuesto es el de *observar* –con el significado literal de la palabra- una muestra de *segmentos*, o áreas relativamente pequeñas en forma directa, sin consultar a los dueños de las tierras, a los productores ni a ninguna persona relacionada con las explotaciones que contiene el segmento. La selección original es de puntos aleatorios dentro de *estratos de uso homogéneo del suelo*, que luego se los transforma en segmentos. Es obvio que gran parte de los puntos caerán en lugares que no se pueden acceder con un vehículo y para poder llevar a cabo la observación es necesario *trasladar* el punto hasta el camino más próximo.

Desde el punto de vista de la teoría del muestreo el procedimiento de trasladar origina un sesgo el cual es considerado como un *error no debido al muestreo*. A propósito de esta situación, en el artículo se cita una prueba realizada en un Partido de la Región Pampeana donde se demuestra que el efecto traslado no genera diferencias significativas en las estimaciones de áreas de los cultivos extensivos, objeto de esta medición. Como contraparte de la aceptación del error no muestral, el método garantiza: a) muy alta confiabilidad de los datos por provenir de observaciones “in situ” hechas por expertos, b) no hay error en la medida de las superficies por utilizar tecnología satelital, c) una vez definido el segmento el Sistema de Posicionamiento Global (GPS) permite controlar el operativo y anula el error de ubicación de los segmentos en futuros operativos, d) las muestras son altamente comparables en el tiempo, e) los resultados se obtienen en breve tiempo, en general no más de tres meses, f) reducción notable del presupuesto al no existir revisitas.

En cuanto a su desarrollo, el artículo describe en detalle el proceso de segmentación y de estratificación de acuerdo a criterios de homogeneidad de uso del suelo enunciando las herramientas y tecnologías aplicadas, sin descuidar los aspectos específicos del diseño muestral aplicado y los distintos procedimientos de estimación de parámetros utilizados.

Cierra este número de la Revista una nota sobre el **68º Congreso de la Asociación Mundial de Investigadores de Opinión Pública (WAPOR)** que tuvo lugar, entre los días 16 al 19 de junio de 2015 en el salón del Centro Cultural Borges de la Ciudad de Buenos Aires. El Congreso fue organizado en conjunto entre la institución WAPOR y la Universidad Nacional de Tres de Febrero (UNTREF), siendo el primer evento de esta Institución organizado en el país y contó con una sesión exclusiva, organizado por la UNTREF-MGAyE destinada a la difusión de las actividades de jóvenes investigadores en el área de la Opinión Pública.

ARTÍCULOS

Aportes al estudio de la opinión pública en las elecciones presidenciales 2007 en Argentina.¹

*Bruno De Santis*²

política

medios de comuni

Durante las últimas décadas, se ha producido un importante incremento de la actividad vinculada a la investigación en opinión pública. En este marco, no sólo se han incrementado las demandas de los políticos y candidatos durante las elecciones, sino también los partidos políticos, los líderes de opinión y los medios de comunicación se convirtieron en importantes demandantes de este tipo de estudios. En particular, las investigaciones en opinión pública constituyen actualmente el marco de referencia desde donde los medios de comunicación construyen y segmentan realidades vinculadas con la actividad política.

Puede sostenerse que las encuestas pre-electorales han ejercido una gran influencia en las elecciones, al punto que han cambiado las formas de exposición de un candidato así como también las pretensiones de los mismos durante los períodos previos a las elecciones (Mora y Araujo 2005: Pág. 275). Las posibilidades de un candidato tienen hoy un nivel de previsión que anteriormente no tenían: a través de un sondeo, un candidato puede conocer qué posibilidades concretas tiene de ganar una elección, con lo cuál así puede asegurarse una buena *performance* o bien desestimar su candidatura cuando los resultados se presentan muy adversos. Desde el punto de vista de los candidatos políticos, la fortaleza de las encuestas electorales reside en que tienen la posibilidad de predecir el comportamiento electoral.

Por otro lado, si nos focalizamos en el análisis de las campañas presidenciales, se puede sostener que éstas han tenido un auge mediático durante las últimas décadas. Este fenómeno remite, ante todo, a una experiencia que se vive a nivel mundial y que tuvo como precursores a los países más desarrollados del mundo, siendo siempre Estados Unidos el país con mayor empeño en mediatizar

las campañas electorales y generar debates televisivos. En Argentina, este fenómeno toma su mayor impulso durante la década del noventa, período que coincide con la aparición de diferentes formas de comunicación masiva como la *Televisión por Cable* y, posteriormente, *Internet*. Puede decirse que estos fenómenos modelan la cultura y la formas de comunicación de masas, que a partir de ahora están mucho más diversificadas, perdiéndose esa vieja forma de comunicación masiva donde el público receptor era bastante homogéneo.

Dentro de este contexto, la política necesita adaptarse a estos vertiginosos cambios. Los políticos posmodernos necesitan participar dentro de estas nuevas formas de comunicación virtual que se materializan en debates televisivos, actos de campañas adaptados para la prensa, sitios web y redes sociales para que los futuros votantes se informen diariamente, etc. (Sartori 1997: Pág. 78). Las nuevas formas de comunicación política readaptan también las actividades proselitistas de los partidos políticos, siendo que estos últimos han perdido la función de identificar ideológicamente a los ciudadanos y han pasado a convertirse en instrumentos mucho más flexibles tendientes a formalizar candidaturas individuales (Duran Barba 2006: Pág. 127).

Objetivos del trabajo y enfoque metodológico

El objetivo general de este trabajo es poder explorar el uso creciente las encuestas de *opinión pública y la injerencia de los medios de comunicación en las elecciones presidenciales locales del año 2007 que se realizaron en la República Argentina*. De este objetivo general, se desprenden los siguientes objetivos específicos:

1. Explorar la incidencia de los medios de comunicación en la selección de candidatos.
2. Analizar el comportamiento electoral a partir de variables con mayor nivel de predicción en la orientación del voto.
3. Evaluar el escenario electoral en la etapa posterior a las elecciones presidenciales mencionadas.

En primer lugar, este trabajo recopiló información de prensa (referida exclusivamente a la campaña electoral) proporcionada por los diarios Clarín y La Nación entre los meses de Julio y Octubre del 2007. Se seleccionó dicho período con el objetivo de cuantificar la cantidad de notas periodísticas dedicadas a cada candidato particular. En segundo lugar, se utilizaron dos sondeos realizados por el Centro Estudios en Estadística Aplicada (CINEA). Dichos sondeos fueron elaborados en el marco de una serie de estudios que está desarrollando el Observatorio de Cultura Política de la Ciudad de Buenos Aires y forman parte de un proyecto de investigación y programación científica de la Universidad Nacional de Tres de Febrero. En cuanto a los datos que se obtuvieron de los sondeos, cabe aclarar que los mismos se acotan solamente al ámbito de la Ciudad de Buenos Aires dentro del período electoral señalado. La primera encuesta fue relevada durante el mes de Agosto del 2007 e incluyó una totalidad de 443 casos efectivos y la segunda

durante el mes de Noviembre del mismo año, contemplando un total de 451 casos efectivos. En ambos relevamientos se trabajó con un nivel de confianza de 95,5% y un margen de error de +/- 5%. Ambas encuestas fueron relevadas a través del sistema telefónico de llamadas automáticas IVR.

Presencia de los candidatos en los medios de comunicación

Al finalizar el último tramo de la campaña electoral 2007, prácticamente todas las encuestas daban como ganadora la fórmula presidencial Cristina Fernández de Kirchner y Julio Cobos. Sin la intención de detenerse en las diferencias de cada uno de los candidatos que se presentaron a dichas elecciones, *se observó una asociación entre la presencia de los candidatos en los medios de comunicación seleccionados y el desempeño electoral obtenido durante la elección*. Esta afirmación puede corroborarse a partir de la recopilación de la totalidad de notas periodísticas relevadas durante el mes de Julio a Octubre del 2007 con los diarios anteriormente mencionados. Una vez recopiladas dichas notas, se cuantificó la cobertura periodística referida a cada candidato para luego establecer la comparación con los porcentajes de votos obtenidos que se presentan en el Cuadro I.

Cuadro 1: Cuantificación de notas periodísticas y resultado electoral por candidato

Candidato	Notas periodísticas	Porcentaje de votos
Cristina Kirchner	452	50,4
Elisa Carrió	129	28,4
Roberto Lavagna	129	28,4
Alberto Rodríguez Saá	9	0,2
Otros	9	0,2
Total	709	100

Las inclinaciones en el voto en las elecciones presidenciales 2007

Como primera observación hacia el análisis cuantitativo de las elecciones presidenciales 2007, se detalla en el Cuadro 2 los porcentajes generales en cuanto a la intención de voto de los entrevistados.

Cuadro 2: Intención de voto a presidente

	Intención de voto a Presidente
	Como podrá observarse en el Cuadro 2, al momento en que se realizó la encuesta, Cristina Kirchner no presentaba grandes diferencias con el resto de los candidatos. Si bien en este caso no
Roberto Lavagna	otras jurisdicciones, las diferencias con el resto de las
Elisa Carrió	significativas. De esta manera, la intención de voto en la Ciudad
Alberto Rodríguez Saá	de Buenos Aires tiende a mostrar un importante sesgo si se compara con los resultados
Ricardo López Murphy	electorales a nivel nacional, donde esa tendencia se revierte considerablemente.
En blanco	la otorgado a la injerencia de los medios de comunicación en el
	presente trabajo, es interesante replantear la intención de voto según el diario que los
	entrevistados leen con mayor frecuencia. Esta cuestión se detalla en el Cuadro 3.
Intención de voto (en %)	diario que lee con mayor frecuencia
80	de voto de los
Cristina Kirchner	En el caso del diario La Nación, la preferencia por la candidata
40,1	Carrió quien muestra mayores niveles de intención de voto para ese
	lectores del diario Página 12, la tendencia muestra un vuelco muy
Roberto Lavagna	Kirchner. A pesar que las observaciones no permiten cotejar una
	bio se puede señalar la existencia de una afinidad entre el diario, el
Elisa Carrió	candidato.
100	Intenciones de voto se manejan en base a la probabilidad de que una
.....	candidato o por otro. Ahora bien, dicha situación no implica que los
Total	proyectar un panorama acerca de cómo se desarrollarían los
	cciones. En este sentido, el sondeo utilizado buscó preguntar a los
	o percibían el panorama electoral en ciernes. Es decir, más allá de sus
	preferencias electorales y partidarias (o indiferencia) se les consultó sobre quién podría ser,
	según su propia óptica, el ganador de las futuras elecciones (Cuadro 4).

Cuadro 4: Expectativa de victoria en las elecciones

Porcentaje
66,6
6,1

Expectativa de victoria en elecciones a presidente	
1,8 Cristina Kirchner	
19,8 Elisa Carrió	
19,8 Roberto Lavagna	
Alberto Rodríguez Saá	
Ricardo López Murphy	
No sabe	
Total	

ados consideran que finalmente Cristina Kirchner será la ganadora en es. Cabe señalar que existen pocos entrevistados que consideran a idades de ganar. La diferencia es de más del sesenta por ciento y no e otro candidato sea visto con probabilidades de ganar la elección. En efecto, este último dato sobre las expectativas de victoria en las elecciones refleja *cómo en la está presente el desempeño probable de los candidatos que pelearán las* s allá de las preferencias que tienen respecto a cada candidato en

severar que la difusión de las encuestas previas a las elecciones erencia muy rico para analizar y predecir cuáles son los candidatos más relevantes durante las campañas electorales. Por añadidura, esto implica que los medios de comunicación tienen un enorme poder a la hora de formar opiniones. Si luego de llevar a cabo un oteja que un candidato tiene escasos niveles de conocimiento entre la e no se lo perciba como un candidato con probabilidades de ganar la elección. El conocimiento y la exposición de los candidatos parece ser el punto inicial para que la expectativa de victoria adquiriera valores significativos. En el mismo sentido, también es esperable que una porción importante del electorado no tenga en cuenta a un candidato relativamente desconocido a la hora de emitir su voto ni de considerarlo con chances de ganar. Siguiendo los datos del *Cuadro 4*, éste bien podría ser el caso de Alberto Rodríguez Saá (un candidato que proviene de la provincia de San Luis, con menos de 500.000 habitantes) o de López Murphy, un candidato relegado a la tercera posición en las elecciones presidenciales del 2003. En ambos casos, las expectativas de que ganen las elecciones son muy bajas y, simultáneamente, la intención de voto en ningún caso supera el 5%.

A la luz de los resultados analizados hasta aquí, resulta crucial volver a replantearse por qué tantos entrevistados consideran que Cristina Kirchner será la probable ganadora. Si bien no se trata de un candidato en ejercicio del cargo, muchas suposiciones nos permitirían concluir que su nivel de conocimiento al momento en que se realizó el sondeo era alto y que su futuro caudal de votos seguía el mismo curso. Dicha percepción del electorado se mantiene aun cuando la candidata no estaba representando la intención mayoritaria del voto en la Ciudad de Buenos Aires. Quiere decir que se esperaba observar una gran cantidad de voto no oficialista pero que es consciente que esta fuerza triunfará en las elecciones. Este aspecto puede comprenderse de manera más clara a partir del *Cuadro 5*.

Cuadro 5: Expectativa de victoria en las elecciones según intención de voto

Elisa Carrió	Fuente: UNTREF-CINEA. Observatorio de cultura política de la Ciudad de Buenos Aires. Agosto 2007.
69,2	

Expectativa de victoria en elecciones (en %)	El candidato que lee con mayor frecuencia
Cristina Kirchner	En todos los casos, en una proporción de electores que creen que la candidata oficial ganará las elecciones, de los que votan a la candidata oficial consideran que será ella quien triunfe. Entre los votantes de Roberto Lavagna perciben, en mayor proporción que entre los votantes de Elisa Carrió, que también Cristina Kirchner será quien triunfe en las elecciones. Al compararlo con las expectativas de la población, surge una hipótesis ad-hoc de que el análisis de otras campañas electorales. En efecto, el escrutinio definitivo de la elección de Lavagna ocupó el tercer lugar y, por tanto, podría sugerirse un previo análisis de voto de este candidato al constatar que las expectativas de sus votantes son inferiores a su competidor directo (en este caso, Elisa Carrió).
Roberto Lavagna	
Elisa Carrió	
Total	¿Hasta qué punto los electores creen que la candidata oficial ganará las elecciones? ¿Hasta qué punto los electores creen que la candidata oficial ganará las elecciones? Para complejizar el análisis del cuestionario una pregunta donde se les consultó a los entrevistados sobre sus expectativas de ballotage.

Cuadro 6: Expectativa de ballotage

Habría ballotage en las elecciones presidenciales	En los casos de los entrevistados que piensa que habría <i>ballotage</i> , es mayor aún la posibilidad de que la candidata oficial ganaría sin tener que pasar por la segunda vuelta. Las expectativas electorales toman más contundencia cuando esta gran cantidad de entrevistados desechan la posibilidad de que se dé un escenario con <i>ballotage</i> . Esta dimensión puede analizarse mejor siguiendo el Cuadro 7, observando si la preferencia por un candidato u otro cambia cuando se plantea la posibilidad de que ocurra una segunda vuelta.
Sí, creo que sí	
No, creo que no	

Cuadro 7: Expectativa de ballotage según intención de voto

Habría ballotage en las elecciones presidenciales (en %)	Intención de voto a Presidente de la Nación
Sí, creo que sí	Entre los votantes de Roberto Lavagna son los más confiados que pueda darse una segunda vuelta. De esta manera, encontramos un perfil ambiguo entre los votantes de los tres, el perfil del votante de Roberto Lavagna es el que menos confía en su candidato pero, simultáneamente, confía en que pueda definirse la elección sin necesidad de un escenario de <i>ballotage</i> . Mientras, los votantes de Elisa Carrió se sienten más seguros al creer que la candidata elegida ganará las elecciones, sugiriéndose que podría haber una segunda vuelta.
No, creo que no	
No sabe	
Total	En las elecciones para el mes de Octubre de 2007, se realizó una segunda vuelta para analizar algunas cuestiones sobre el desempeño electoral de los candidatos. Los resultados de las elecciones son poco frecuentes ya que las mismas suelen analizarse hechos consumados y, en cierto sentido, ni los partidos políticos ni los medios de comunicación.

analizar hechos consumados y, en cierto sentido, ni los partidos políticos ni los medios de comunicación.

comunicación suelen hacer mucho uso de ellas. No obstante, son muy fructíferas para que puedan vislumbrarse dimensiones que solo pueden analizarse cuando el período electoral ha finalizado.

El primer aspecto que intenta indagarse en una encuesta pos electoral, es consultarle a los entrevistados acerca de su voto anterior en la elección de referencia (*Cuadro 8*).

Cuadro 8: Voto en las elecciones Octubre 2007

Voto anterior (elecciones presidente de la Nación 2007)	Elisa Carrió	La candidata oficial obtuvo un 23,8% de los votos en la Ciudad de Buenos Aires. Sólo puede decirse que las pequeñas diferencias pueden contemplarse inclusive bajo los márgenes de error establecidos para la muestra aplicada en el estudio. No en tanto, es significativa la diferencia que se observa en el caso de Elisa Carrió: esta candidata obtuvo un total de 37,8 por ciento en las elecciones para Capital Federal y sólo el 26,9% declara haberla votado. En este caso, pueden estar jugando distintos factores. En primer lugar, puede suceder que una gran cantidad de personas manifiesten no recordar a quien han votado y que, en consecuencia, una proporción importante lo haya hecho a favor de Elisa Carrió. En un contexto de despolitización muchos encuestados no registren o les cueste recordar con rapidez el voto pasado. Es importante señalar esto, ya que es muy alta la cantidad de casos que declaran no saber a quién han votado. No obstante ello, no sucede lo mismo con el resto de los candidatos, ya que la recordación del voto se asemeja mucho a los resultados finales.
Cristina Kirchner	Roberto Lavagna	Es atractivo pensar que esta diferencia pueda atribuirse a una posible distorsión del verdadero voto.
Alberto Rodríguez Saá	Ricardo López Murphy	
Otros		
No sabe	17,5	
Total		

Como bien lo menciona Manuel Mora y Araujo (Mora y Araujo 2005: Pág. 302), el ocultamiento del voto trata del candidato que ha sido derrotado en las elecciones. El autor sostiene que es en los casos donde las encuestas pos electorales suelen reflejar una importante diferencia entre las declaraciones del voto anterior y los resultados electorales efectivos. Otra de las cuestiones que parece tener influencia en esta dimensión, estaría ligada a la llamada teoría de “La espiral del silencio” expuesta por la politóloga Elisabeth Noelle-Newmann. Esas diferencias que se suelen observarse entre las declaraciones de los encuestados y los resultados electorales, estaría dada a partir de la existencia de una opinión socialmente predominante que tiende a legitimarse y, por lo tanto, a discriminar o a aislar aquellas opiniones que tienen menor preponderancia. Por lo tanto, aún aquellas personas que piensan inicialmente de manera contraria, tienden a modificar sus expresiones en los casos en donde el entorno social puede ser propenso a rechazar opiniones opuestas y con menor grado de legitimidad (Noelle-Newmann 1995: Pág. 87).

Si bien es difícil poder arribar a conclusiones categóricas, las diferencias entre los resultados electorales y las respuestas de los entrevistados que manifestaron haber votado a Elisa Carrió juega a favor de una opinión en ascenso que no es asimilable a la candidata opositora, sino a la continuidad del oficialismo. Para tener una mejor apreciación al respecto, a continuación se

presenta el Cuadro 9.

Cuadro 9: Evaluación de la gestión Néstor Kirchner

Buena	29,5	Fuente: UNTREF-CINEA. Observatorio de cultura política de la Ciudad de Buenos Aires. Noviembre 2007.
Regular negativa	15,5	
Regular positiva	25,0	
Mala	30,0	
No sabe	100	
Total		

Tomando como intervalo 4 = Máximo imagen positiva a 1 = Máximo imagen negativa, el segundo sondeo arroja un valor de 2.86 puntos.

La observación directa de los valores muestra un aumento positivo de la gestión presidencial anterior. Aún así, ¿puede considerarse un aumento significativo o sólo se trata de un aumento coyuntural que no presta verdadera significancia? Para despejar este interrogante, se decidió realizar una comparación de muestras a través del test de muestras independientes (en el Cuadro 10), operación que, a través del SPSS, permite observar si los cambios en los promedios de las distintas muestras son significativos. Es decir, a través de esta operación puede saberse con exactitud si el promedio en la gestión del ex presidente evidencia un cambio significativo posterior a la elección.

Cuadro 10: Test de muestras Independientes

Grupo	1	2
	1,000	1,000

Apesar de poder argumentar que al mejorar la evaluación de la gestión presidencial anterior, implica que otras variables podrían marcar valores significativos como, por ejemplo, percepción

de la situación macroeconómica, social, la economía familiar, etc. Para englobar de manera sintética estas dimensiones de análisis, se les consultó a los entrevistados acerca de sus expectativas sobre la situación del país en el año venidero, tal como se aprecia en el *Cuadro 11*.

Cuadro 11: Expectativas sobre la situación del país en el año venidero

Mejor que ahora		Fuente: UNTREF-CINEA. Observatorio de cultura política de la Ciudad de Buenos Aires. Noviembre 2007.	
		sobre la situación del país el año que viene	
No sabe		de la encuesta pre electoral. Asimismo, también aumenta levemente la proporción de personas que	
Mejor que ahora		la situación será igual al momento de la encuesta. Finalmente,	
Igual que ahora		la situación futura del país también registran valores más bajos.	
Peor que ahora		ciones en relación a este tema puede ser que la victoria del	
No sabe		percepción generalizada de estabilidad, lo cuál permitiría no sólo	
		expectativas positivas para el año 2008, sino también que la	
		ncial anterior registre un promedio más alto. Simultáneamente,	
		uede estar influyendo en aquella proporción de entrevistados que,	
		aun así perciban que la	
		situación futura del país mejorará. Para profundizar el análisis, en el <i>Cuadro 12</i> se compara la	
		evaluación de la gestión de Néstor Kirchner con la percepción y expectativas para el año siguiente.	

Cuadro 12: Evaluación de la gestión de Néstor Kirchner según expectativas

Evaluación de la gestión de Néstor Kirchner (en %)		sobre la situación del país el año que viene		res. Noviembre 2007.	
Buena					
Regular positiva					
Regular negativa					
Mala					
No sabe					

Total	0.708
-------	-------

Si hay una relación entre los que piensan que la situación del país mejorará y la imagen positiva de la gestión del ex presidente. A su vez, este fenómeno se observa de manera contraria: quienes tienen una evaluación mala o regular negativa del ex presidente son más propensos a contestar que la situación del país para el año siguiente será peor que en el momento que se realizó la encuesta.

Para analizar la asociación entre estas dos variables (imagen presidencial y expectativas de la situación del país para el año 2008) resulta valioso utilizar el coeficiente de asociación *Phi* que permite conocer el grado de asociación que hay entre estas dos variables, partiendo de una hipótesis *ad – hoc* nula que, en este caso, podría ser: *la expectativa de la situación del país baja en la medida que aumenta la imagen positiva del ex presidente*. El coeficiente *Phi* asimila un valor de 0 a 1, siendo 0 el valor que indica asociación nula entre las variables y 1 el valor máximo de asociación entre las variables.

Cuadro 13: Test de Chi cuadrado

Test de Chi Cuadrado	
Chi cuadrado	0.00
Likelihood Ratio	0.00
Asociación Real-por-Lineal	0.00

Test de asociación	
Casos válidos	1083
Phi	0.708
Cramer V	0.83

Conclusión

La realización de este trabajo de investigación permitió demostrar inicialmente que, a través de la información de prensa en la campaña presidencial, el aumento de la cantidad de cobertura periodística sobre un candidato tuvo una significativa influencia sobre el desempeño electoral de cada uno de ellos en el transcurso de las elecciones analizadas. Sin apresurarse a establecer una visión simplista que podría entenderse como mayor cobertura- mayor cantidad de votos, puede en

cambio afirmarse que la cobertura periodística pareció influir indirectamente en las preferencias simbólicas de los votantes.

Para todo trabajo de investigación electoral, la intención de voto parece ser el eje crucial para comprender el escenario electoral. Sin embargo, otras dimensiones que se analizaron permiten demostrar que la utilización de variables más proyectivas da cuenta de un escenario mucho más rico y predecible que la simple medición de la intención de voto. En cuanto al análisis pos electoral, las variables que se analizaron permitieron concluir que el comportamiento electoral se ha vuelto mucho más complejo y las decisiones sobre el voto anterior pueden modificarse una vez finalizado el proceso electoral.

Finalmente, cabe mencionar la contribución del presente trabajo para otros estudios. En efecto, la intención de este trabajo fue contribuir a explorar y comprender el creciente avance de las encuestas de opinión pública en Argentina y el escaso tratamiento de la materia en el ámbito académico. Quedará para estudios posteriores la posibilidad de investigar sobre la evolución que tendrán estas transformaciones en donde los principales protagonistas son los medios de comunicación, los candidatos políticos y la ciudadanía.

¿Puede decirse que estas transformaciones vienen para quedarse? ¿Qué implicaciones tendrán los candidatos en el futuro de acuerdo a la creciente dependencia que tienen frente a los medios de comunicación y las encuestas de opinión pública? Todos estos interrogantes no pueden ser comprendidos a la luz de este trabajo, pero queda entonces el lugar para que investigaciones posteriores puedan encontrar respuestas a estos interrogantes.

Bibliografía

- Giovanni Sartori "Homo Videns: La sociedad Teledirigida". Editorial Taurus. 1997.
- Manuel Mora y Araujo "El poder de la Conversación. Elementos para una teoría de la opinión pública". Ediciones La Crujía. Año 2005.
- Elizabeth Noelle Neumann "La espiral del Silencio". Editorial Paidós. 1995.
- Javier Auyero "La zona gris. Violencia colectiva y política partidaria en Argentina Contemporánea". Siglo XXI Editores. 2007.
- Jürgen Habermas "Historia y crítica de la opinión pública". Editorial Gili. 1981.
- Jaime Durán Barba- Santiago Nieto "Mujer, sexualidad internet y política. Los nuevos electores latinoamericanos". Fondo de Cultura Económica. 2006.
- Federico Rey Lennon- Alejandro Piscitelli Murphy "Pequeño manual de encuestas de opinión pública". Ediciones La Crujía. 2004.
- Roberto Izurieta, Rubén M. Perina y Christopher Arterton "Estrategias de comunicación para gobiernos" Ediciones La Crujía. 2003.
- Gustavo Martínez Pandiani "Marketing Político. Campañas, medios y Estrategias Electorales". Ugerman Editor. 2004.

ARTÍCULOS

Construcción de un índice del nivel socioeconómico del hogar urbano en la República Argentina mediante el análisis de correspondencia múltiple y escalamiento óptimo.

*María Fernanda Artola*¹ / *Iván Redini Blumenthal*²

pobreza

índice socioeconómico

El artículo presenta la construcción de un índice que tiene como finalidad asignar a cada hogar urbano de la República Argentina un nivel socioeconómico en base a los datos provistos por la Encuesta Nacional de Gasto de los Hogares que se realizó en áreas urbanas de todo el país, entre marzo de 2012 y marzo de 2013 (ENGHO 2012/2013).²

Dicho índice se integra en el marco de un estudio que comenzó a desarrollarse en el año 2014 y que tiene como fin estimar la relación de costo-efectividad de la vacuna contra el rotavirus y su impacto en el gasto de bolsillo de los hogares urbanos. La estimación del gasto de bolsillo, requirió efectuar un relevamiento a familias con niños menores de cinco años con gastroenteritis aguda en un grupo de centros de salud. Ante la dificultad de incluir en dicho cuestionario preguntas relativas al nivel de ingreso del hogar, resultó necesario construir un índice socioeconómico. Se aplicó el análisis de correspondencia múltiple (ACM), técnica exploratoria dentro de los métodos multi-variados estadísticos de interdependencia, utilizada para identificar relaciones sistemáticas entre variables donde a priori no hay una hipótesis respecto a la naturaleza de dichas relaciones y se utilizaron los resultados para obtener los pesos óptimos de las variables que componen el índice.

A continuación, se presenta la metodología utilizada, los pasos seguidos para su aplicación y los resultados encontrados. Por último, a modo de discusión, se menciona el alcance de la metodología, limitaciones y las consideraciones para su aplicación. ■

Metodología utilizada

El análisis de correspondencia es una técnica exploratoria de interdependencia desarrollada por Jean-Paul Benzécri (Benzécri, 1973) cuyo objetivo es resumir una gran cantidad de datos en un número reducido de dimensiones, con la menor pérdida de información posible. Esta técnica basada en la descomposición de una matriz en valores singulares, puede presentarse en términos geométricos a través de visualizaciones gráficas así como también de forma analítica construyendo escalas de las categorías de las variables categóricas bajo análisis. Su objetivo es similar al de los métodos factoriales, aunque a diferencia de éstos, se aplica sobre variables categóricas.

Cuando el análisis contempla sólo dos variables o una tabla de doble entrada se denomina análisis de correspondencia simple (ACS) y cuando se aplica a más de dos variables se denomina análisis de correspondencia múltiple (ACM).

El ACM en su versión asimétrica contempla utilizar una matriz binaria asimétrica (tabla disyuntiva). Para construir dicha matriz, se recodifican las variables categóricas como variables binarias y para cada modalidad de cada variable hay sólo dos respuestas posibles: [0 = el encuestado no elige esa modalidad; 1 = el encuestado elige dicha modalidad]. Esta matriz es de dimensión (N X J), donde N es el tamaño de la muestra y J es el número total de modalidades asociadas a las P preguntas (variables). La suma de cada una de las filas es P y la suma de cada una de las columnas es la cantidad de individuos que asume la modalidad respectiva.

En el caso particular analizado, la dimensión de la tabla disyuntiva es de 20.960 X 39 donde las filas representan la cantidad de hogares encuestados en la ENGHO 2012/13 y 39 son el total de las modalidades asociadas a las 12 variables finalmente seleccionadas. Una vez aplicado el ACM a la *tabla disyuntiva completa*, los resultados se presentan en coordenadas estándares para las columnas mientras que para las filas se presentan como coordenadas principales. La inercia indica la calidad de representación y este valor varía entre 0% y 100%, indicando el 100% una perfecta representación.

El ACM en su versión simétrica consiste en construir la tabla simétrica de Burt a partir de la matriz binaria, con todos los cruces de las categorías de las P variables, resultando en una matriz de orden (J X J), siendo de orden 39 en este caso. Cada bloque de la matriz de Burt es una submatriz formada por tablas de contingencia de las variables dos a dos, excepto los bloques que están en la diagonal principal que son las tablas de contingencia de cada variable consigo misma.

De acuerdo a lo indicado por Michael Greenacre, el ACM efectuado mediante una matriz binaria

asimétrica (*tabla disyuntiva completa*) es equivalente a aquel efectuado con la tabla de Burt (Greenacre, 2008).

Por otro lado, el “escalamiento óptimo” comprende un conjunto de métodos multivariados que buscan obtener la máxima discriminación entre las categorías de las variables mediante la construcción de pesos óptimos.

El escalamiento óptimo se vincula con el ACM en su forma asimétrica como simétrica. En el primer caso, a partir de la tabla disyuntiva completa, se puede pensar al ACM como un método de escalamiento óptimo, o en otras palabras, una técnica para la cuantificación de las categorías de las variables y la maximización de la distancia entre las mismas. A los fines de obtener los pesos óptimos, se asignan a las J modalidades dicotómicas de la *tabla disyuntiva completa* los pesos π

i

, π

h

, π

k

,...,

m

y luego se maximiza la varianza de las puntuaciones de las observaciones (combinaciones lineales de las filas):

$$\max \var{ \alpha_i + \beta_h + \gamma_k + \cdots + \pi_m }$$

s.a.

$$\sum_i c_i \alpha_i^2 + \sum_h c_h \beta_h^2 + \sum_k c_k \gamma_k^2 + \cdots + \sum_m c_m \pi_m^2 = 1$$

donde

c

i

,c

j

,

c

k

es la masa μ de la columna i,h,k, respectivamente.

Dado que la varianza máxima coincide con el valor que asume la inercia de la primera dimensión del ACM asimétrico y, además, los pesos óptimos obtenidos a partir de la maximización de la varianza coinciden con las coordenadas estándares, utilizaremos el ACM asimétrico para obtener los pesos óptimos del índice socioeconómico.

Aplicación del método

Como se mencionó previamente, a los fines de estudiar el impacto de la vacuna de rotavirus en los hogares de diferentes estratos socioeconómicos, se encuestaron a los familiares de los niños menores de cinco años con gastroenteritis aguda en las salas de espera de centros de salud en Ciudad de Buenos Aires, San Salvador de Jujuy y la ciudad de Salta durante el primer semestre del año 2014.

Dicha encuesta estuvo compuesta de cinco secciones. Las primeras dos eran relativas a los cuidados hospitalarios, medicamentos y gastos no habituales vinculados a la enfermedad de rotavirus. Las restantes secciones fueron destinadas a relevar información relacionada con la composición del hogar y sus características físicas, situación ocupacional y nivel educativo del jefe del hogar y del cónyuge.

Al respecto, no pudieron ser incluidos todos los aspectos del hogar y del jefe del hogar que incorpora la ENGH 2012/2013 en la encuesta de gastroenteritis aguda ni en el índice socioeconómico debido a la posible sensibilidad de los encuestados, quienes se encontrarían en las salas de espera de los centros de salud.


Tomando en consideración las limitaciones mencionadas, se procedió a buscar diferentes alternativas para clasificar los hogares por estratos de ingresos o socioeconómicos a partir de los datos obtenidos por la encuesta de rotavirus.

Revisando estudios sobre pobreza, se observó que la misma es usualmente abordada bajo dos enfoques: el directo y el indirecto. La aproximación directa implica analizar la pobreza estructural (inercial) relevando aspectos vinculados a la calidad de la vivienda, condición habitacional, el acceso a servicios y a la educación. Por otro lado, el enfoque indirecto efectúa un relevamiento de la pobreza coyuntural (pauperización) vinculada a los niveles de ingresos. Cuando se presentan situaciones de pobreza tanto bajo el enfoque directo como indirecto, se denomina pobreza crónica o total. Sin embargo, pueden existir sub-universos cuando no son coincidentes las aproximaciones.

Clasificar a los hogares de los hogares de los encuestados según su nivel de ingreso, requería identificar los ingresos familiares de los encuestados para luego asignarles un quintil de ingreso en base a los quintiles contruidos por el INDEC para la Encuesta Nacional de Gastos de los Hogares (2012/2013) o la Encuesta Permanente de Hogares. Sin embargo, este camino


resultaba poco recomendable tanto por el posible rechazo por parte de los encuestados a brindar información como por la posible distorsión de la información brindada por un segmento relevante de quienes responden. Una alternativa frecuentemente utilizada para sortear dicha dificultad consiste en solicitar a las personas que indiquen el rango donde se ubica el ingreso total del hogar. Sin embargo, dicha alternativa no resultaba suficiente para resolver las limitaciones mencionadas anteriormente.

Tomando en consideración las limitaciones señaladas en el párrafo anterior respecto a los ingresos, se procedió a clasificar a los hogares por nivel socioeconómico, lo cual implicó construir una variable compleja y multidimensional no observada (latente) que la denominamos nivel socioeconómico.

En cuanto al software utilizado para aplicar el método se emplearon los paquetes estadísticos Foreign, Ca y principalmente FactoMineR del software libre R Project for Statistical Computing y para procesar las bases se utilizó Stata SE 11.1 e IBM SPSS Statistics 19. 

Resultados y discusión

En base a las variables incluidas tanto en la encuesta de gastroenteritis aguda como en la ENGH0 2012/2013, se aplicó el ACM en su forma asimétrica a la ENGH0 2012/2013 ponderado por los expansores muestrales.

Luego de probar distintas combinaciones de variables y analizar los resultados en términos de la inercia relativa, signos de coordenadas estándares, contribuciones y coseno al cuadrado de cada modalidad, el índice finalmente obtenido  estuvo compuesto por las siguientes variables:

1. características del baño
2. características del piso
3. disponibilidad de cloacas
4. disponibilidad de teléfono fijo
5. pavimento en la cuadra
6. tipo de combustible utilizado para cocinar
7. cobertura médica del jefe del hogar
8. nivel de educación del jefe de hogar
9. nivel educativo del cónyuge
10. jefe de hogar de género femenino
11. situación laboral del jefe del hogar
12. hogar nuclear o extendido con niños menores de 14 años

A partir de los resultados obtenidos del análisis de correspondencia múltiple, se observa que la inercia de la primera componente alcanzó 0,2856 explicando el 12,71% de la inercia total mientras que las tres primeras componentes en conjunto explican el 25,91% y a partir de la 17ª dimensión

en conjunto se explica el 79,24% de la inercia. (Ver Cuadro N°1 del Anexo: *Inercias principales*). A pesar de los bajos valores de la inercia explicada tanto por el primer factor como por los tres primeros, el scree plot que se presenta a continuación indica una disminución importante en la pendiente a partir de la tercera dimensión lo que da cuenta que a partir de la cuarta dimensión, cada dimensión adicional aporta muy poco en términos individuales.

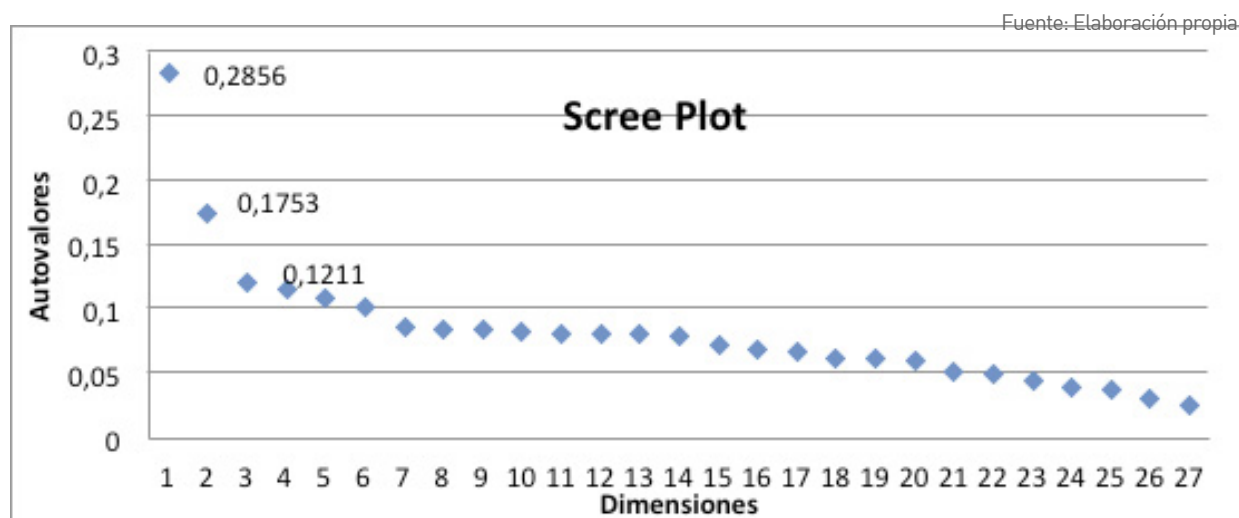


Gráfico 1: Scree Plot

Si analizamos la correlación al cuadrado de cada variable con cada dimensión, surge que las variables que caracterizan al hogar como ser el tipo de pisos, la presencia o no de pavimento en la cuadra de la vivienda, el tipo de desagote en el baño, la disponibilidad de cloacas, teléfono fijo y cobertura explícita del JH son las que están más correlacionadas con el primer eje. Por otro lado, las variables que hacen referencia a la situación laboral del Jefe de Hogar (JH), nivel educativo del cónyuge y el sexo del JH, se correlacionan principalmente con los ejes 2 y 3. Por último, las variables hogar nuclear con niños menores de 14 años y nivel educativo del JH se correlacionan tanto con la primera como con la segunda dimensión.

Lo anterior se puede ver con más detalle al analizar las contribuciones por modalidad y el grado de representación de cada modalidad por parte de cada una de las dimensiones. En términos generales, se observa que la inercia en la primera dimensión está explicada principalmente por las variables que caracterizan al hogar mientras que la inercia de la segunda y tercera dimensión se concentra en las características laborales y nivel educativo del jefe del hogar y del cónyuge (Tabla N°2 , Tabla N°3 y Gráfico N° 1 del Anexo).

Por otro lado, los coeficientes de las coordenadas estándares (pesos óptimos del escalamiento) en la primera dimensión indican que se reducen (incrementan) a medida que las características del hogar y/o del JH y su cónyuge son más favorables (menos favorables). A partir de la segunda y

tercera dimensión no se observa una consistencia entre los signos de las diferentes variables y esto se puede explicar porque los otros ejes dan cuenta del comportamiento de un grupo más reducido de variables inherentes al jefe del hogar y del cónyuge. Esta situación nos impulsó a trabajar únicamente con la primera dimensión, además del hecho de que los pesos obtenidos en el escalado óptimo en su versión asimétrica coinciden con las coordenadas estándares del primer eje principal obtenidas a partir del ACM.

En línea con lo anterior, fue considerado el alfa de Cronbach que permite evaluar la fiabilidad de la prueba y testear dicha fiabilidad si se excluyera un determinado ítem. Esta medida está basada en el promedio de las correlaciones de las variables que forman parte de una misma dimensión teórica (componente principal), obteniendo como resultado 0,77 para la primera dimensión. En general, el valor mínimo aceptable es de 0,6 aunque hay autores que sugieren un valor mínimo de 0,7, lo cual indicaría que en este caso los resultados son aceptables.

A diferencia de la escala original de las variables, la escala óptima para la primera dimensión no sitúa a las categorías de cada variable a distancias iguales. Por ejemplo, en el caso que la vivienda esté ubicada en una calle con pavimento la puntuación es negativa (-0,32), de lo contrario la puntuación es positiva (+0,98). El mismo comportamiento se observa en la variable pisos donde la puntuación es negativa (-0,3117) cuando el piso es de cerámica, madera, baldosa, mosaico, mármol o alfombra y suma (+1,4) puntos si es de cemento, ladrillo fijo y (+2,22) puntos cuando es de ladrillo suelto o de tierra.

A partir de los valores de escala óptimos, se procedió a calcular el puntaje total para cada uno de los hogares encuestados en la ENGH0 2012/2013 y se observó que cuanto más bajo es el índice, más alto es el estrato socioeconómico del hogar, confirmándose el comportamiento individual de las variables bajo análisis. Finalmente, a partir de los puntajes totales, se construyeron los quintiles a nivel hogar como se detalla en la Tabla N°1.

Fuente: elaboración propia en base a ENGH0 2012/2013. Procesado con SPSS.

Quintil socioeconómico	Índice socioeconómico	
	Puntaje Mínimo	Puntaje Máximo
5	-11,06	-5,67
4	-5,67	-3,09
3	-3,09	0,27
2	0,27	5,37
1	5,38	25,9

Tabla 1: Quintiles socioeconómicos

El cuadro N° 4 del Anexo: *Cantidad de hogares por modalidad y por quintil socioeconómico* indica que el comportamiento de cada variable es consistente con cada quintil socioeconómico construido. A continuación, se presentan a modo de ejemplo la forma en la cual varían ciertas variables a lo largo de los quintiles construidos a partir del índice socioeconómico.

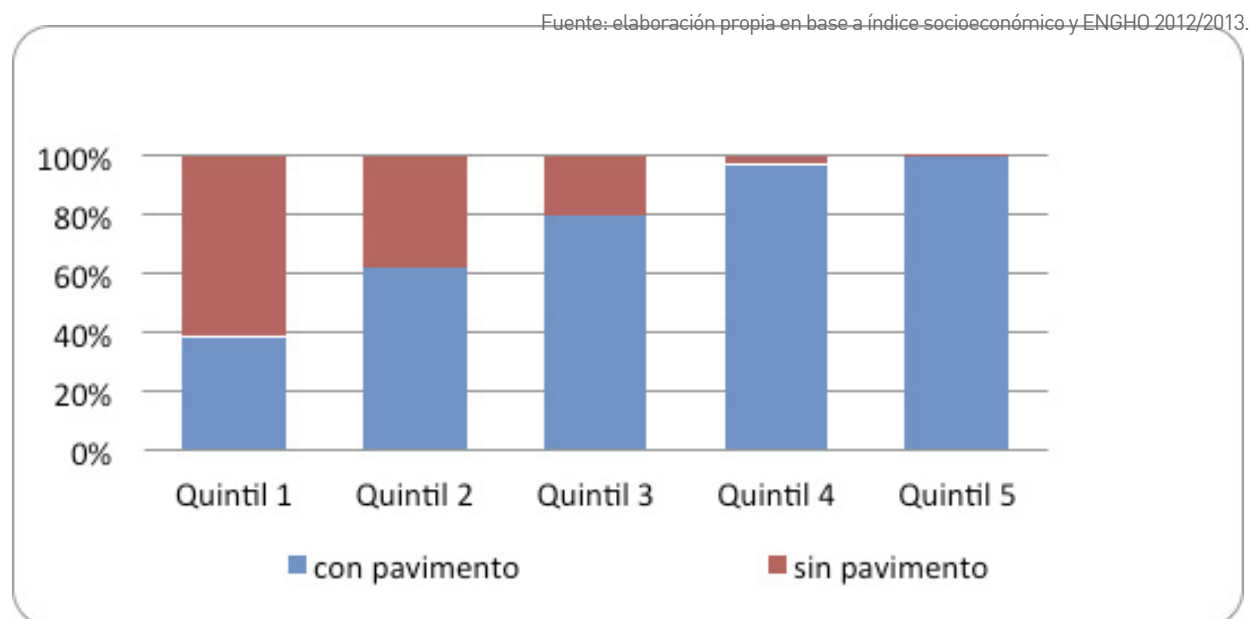


Gráfico 2: Distribución de los hogares por presencia de pavimento y quintil socioeconómico

En el caso de la variable “presencia de pavimento en la cuadra”, mientras que en el primer quintil muy pocos hogares poseen pavimento en su cuadra, en el quinto quintil la mayoría lo tienen, según se pone de manifiesto en el Gráfico N° 2. Por otro lado, el tipo de pisos cambia a medida que uno avanza de quintil, mientras que en el primer quintil el 48% de los hogares tiene cemento o ladrillo fijo, dicho porcentaje se reduce a 19% en el segundo quintil y a menos de 1% en el quinto quintil.

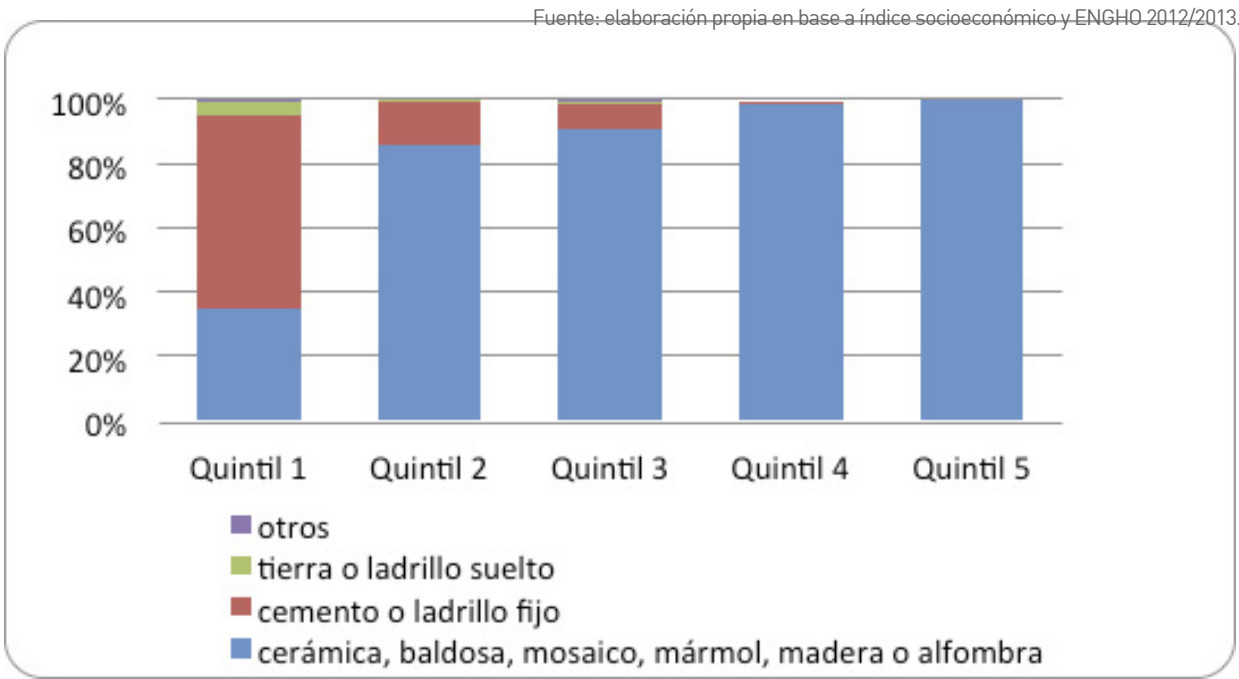


Gráfico 3: Distribución de los hogares según tipo de pisos y nivel socioeconómico

Por otro lado, la cantidad de hogares con alto nivel educativo del jefe del hogar y el cónyuge se incrementa a medida que mejora el nivel socioeconómico. Como se observa en el Gráfico 4, en el primer quintil, el 61% de jefes de hogar no alcanzó a finalizar la secundaria y sólo un 37% la finalizó mientras que, en el quinto quintil, un 66% de los jefes de hogar alcanzaron un nivel educativo de superior/universitario incompleto o más y un 23% de los jefes de hogar finalizaron los estudios secundarios.

Fuente: elaboración propia en base a índice socioeconómico y ENGHO 2012/2013.

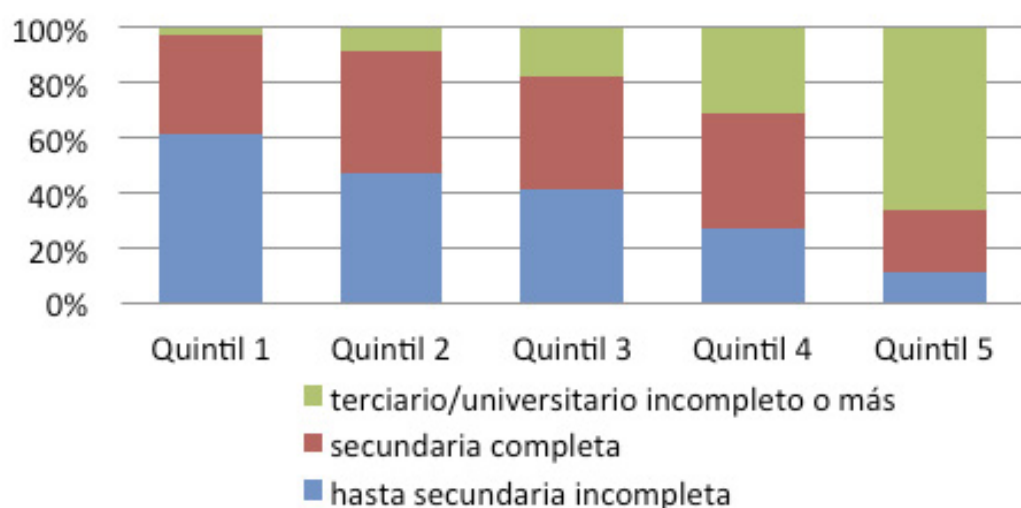


Gráfico 4: Distribución de los hogares según nivel educativo del jefe del hogar y quintil socioeconómico

Similar comportamiento se observa en el caso de la educación del cónyuge según se ilustra en el gráfico N° 5.

Fuente: elaboración propia en base a índice socioeconómico y ENCHO 2012/2013.

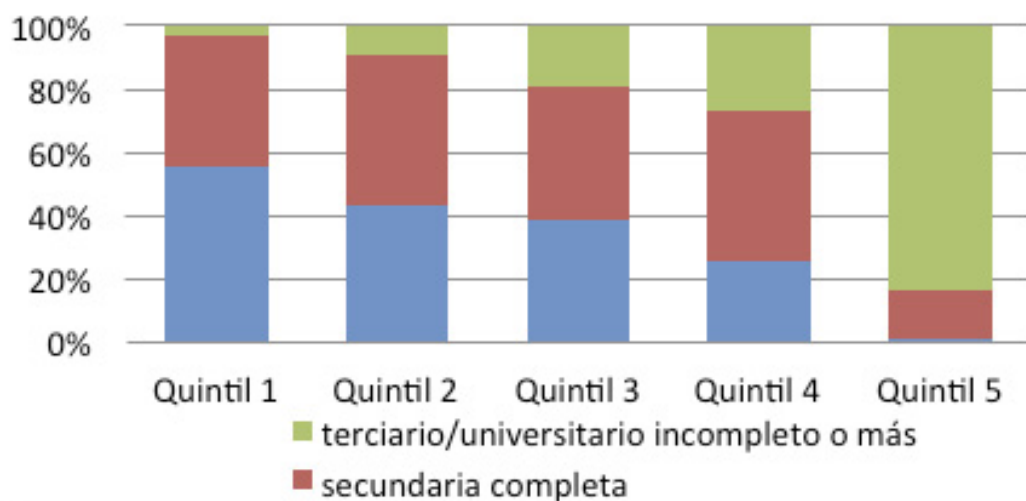


Gráfico 5: Distribución de los hogares según nivel educativo del cónyuge y quintil socioeconómico

En el caso de la situación laboral del jefe del hogar, a pesar de que esta variable contribuye poco a la inercia de la primera dimensión, se puede apreciar que la participación de la relación de dependencia con descuentos para jubilación y obra social sube a lo largo de los quintiles en

detrimento de la relación de dependencia sin descuentos. Por otro lado, los cuenta propistas pierden participación y los patrones incrementan su participación pero no a gran escala. Algo interesante para señalar es que la proporción de desocupados se incrementa a medida que mejora el quintil. A pesar del bajo poder explicativo de esta variable se podría indagar acerca de las características (duración, motivo, etc.) de dicho desempleo en cada uno de los quintiles.

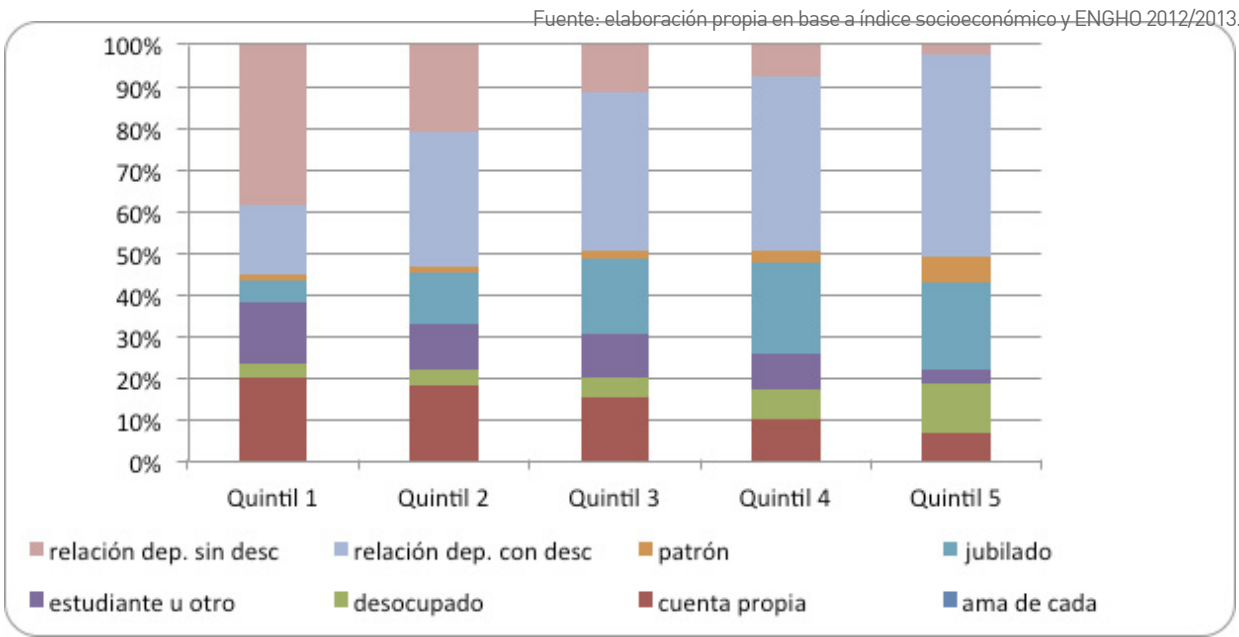


Gráfico 6: Distribución de los hogares según condición laboral del jefe del hogar y quintil socioeconómico

Por otro lado, si se analizan los gastos e ingresos promedios de cada quintil socioeconómico, se observa que a medida que se avanza de quintil, se incrementan tanto el ingreso y gasto en términos totales y per cápita.

Fuente: elaboración propia en base a ENGHO 2012/2013. Procesado con SPSS.

Variable de la ENGHO 2012/2013	Quintiles del índice socioeconómico				
	Quintil 1	Quintil 2	Quintil 3	Quintil 4	Quintil 5
	Valor promedio en \$				
Ingreso total del hogar	\$ 3.860,49	\$ 5.001,20	\$ 5.671,05	\$ 7.212,70	\$ 9.625,54
Ingreso del hogar per cápita	\$ 1.242,97	\$ 1.793,01	\$ 2.253,37	\$ 3.182,73	\$ 4.552,52
Gasto total del hogar	\$ 3.127,27	\$ 3.790,67	\$ 3.912,48	\$ 5.401,02	\$ 6.744,74
Gasto del hogar per cápita	\$ 876,36	\$ 1.367,49	\$ 1.391,63	\$ 1.869,12	\$ 3.504,50

Tabla 2: Quintiles socioeconómicos

Finalmente, al comparar los quintiles del gasto per cápita del hogar que surgen de ENGHO 2012/2013 con los quintiles contruidos a partir del índice socioeconómico, se observa que hay más coincidencias en los quintiles de los extremos que en los centrales. Esto puede deberse a que, a pesar de que ambas variables miden la pobreza desde un enfoque directo, la definición empleada de pobreza no sea exactamente la misma. Por otro lado, este bajo nivel de coincidencia podría estar reflejando una limitación de la metodología empleada producto de haber considerado únicamente la primera dimensión del análisis de correspondencia múltiple.

Fuente: elaboración propia en base a ENGHO 2012/2013. Procesado con SPSS.
Quintiles del gasto per cápita del hogar (expandido)

	Quintil 1	Quintil 2	Quintil 3	Quintil 4	Quintil 5
	Cantidad de hogares				
Quintil 1	1.010.950	620.813	370.118	180.070	57.396
Quintil 2	548.073	611.532	492.533	409.360	177.243
Quintil 3	422.587	488.917	541.465	437.636	352.895
Quintil 4	183.215	322.694	482.870	650.865	595.803
Quintil 5	75.043	195.365	352.542	561.503	1.056.012

Tabla 3: Quintiles socioeconómicos y quintiles de gasto per cápita del hogar

Consideraciones finales

En este artículo se presentó el proceso de construcción de un índice que permite clasificar a los hogares urbanos por nivel socioeconómico. Se pensó a esta variable como latente y dado que las coordenadas estándares del primer eje principal obtenidas a partir del análisis de correspondencia múltiple (ACM) coinciden con los pesos óptimos obtenidos del escalado óptimo, se aplicó el ACM.

Una vez obtenidos y analizados los resultados del ACM, en base a los ponderadores en la primera dimensión, se procedió a calcular el puntaje total para cada uno de los hogares encuestados de la ENGH0 y se construyeron los quintiles. Se observó que a medida que aumenta el puntaje del índice, más bajo es el estrato socioeconómico del hogar.

Entre las limitaciones del índice, cabe señalar que el indicador construido es de carácter estático atento a que se utilizó únicamente la ENGH0 2012/2013. Se podría haber utilizado la Encuesta Permanente de Hogares (EPH) que se publica trimestralmente y luego efectuar un seguimiento de los pesos óptimos a lo largo de determinado período de tiempo. Sin embargo, se optó por utilizar la ENGH0 2012/2013 debido a que abarca mayor cantidad de hogares urbanos y cubre más dimensiones que la EPH.

Por otro lado, el indicador fue construido únicamente a partir de los pesos óptimos de la primera componente, aspecto que debería ser considerado para mejorar la precisión del mismo utilizando algún método que permita incorporar más dimensiones obtenidas a partir del análisis de correspondencia múltiple.

Analizando la literatura, las variables educación del cónyuge, sexo del jefe del hogar y presencia de niños menores a 14 años en hogares nucleares no son comunes al construir índices socioeconómicos o de pobreza. Sin embargo, fueron incluidas debido a que los registros de nacimientos y mortalidad infantil se cruzan generalmente con el nivel de educación de las madres de los niños. Estas variables no contribuyen de manera significativa a la inercia de la primera dimensión, sin embargo, la inclusión de las mismas se justifica con los objetivos del trabajo de gastroenteritis que da marco al estudio.

Por último, cabe destacar que a pesar de las limitaciones mencionadas, la utilidad de este método al diseñar indicadores alternativos multidimensionales de nivel socioeconómico, pudiendo ser considerada para estudios de desarrollo y pobreza. La flexibilidad propia de la técnica utilizada permite explorar variantes en la construcción de los índices y considerar la inclusión de distintas variables, pudiendo considerar a la ENGH0 2012/2013 para tal fin así como también otras encuestas.

Bibliografía

Arakaki, Agustín (2011). La pobreza en Argentina 1974-2006. Construcción y Análisis de la información. Documento de Trabajo N° 15. Centro de Estudios sobre Población, Empleo y Desarrollo. Argentina.

Benzécri JP (1973). L'analyse des données. Tome 1: La taxinomie. Tome 2: L'analyse des correspondances. Dunod, Paris.

Ezzrari, Abdeljaouad y Verme, Paolo. June, 2012. A multiple correspondence analysis approach to the measurement of multidimensional poverty in Morocco 2001 – 2007. Policy Research Working Paper 6087. The World Bank. Middle East and North Africa Region Economic Policy, Poverty and Gender.

Frikkie Booysen; Servaas van der Berg, Ronelle Burger, Michael von Maltitz, Gideon du Rand. Abril 2008. Using an asset index to asses trends in poverty in seven Sub-Saharan African countries. Using an Asset Index to Assess Trends in Poverty in Seven Sub-Saharan African Countries. Elsevier.

Greenacre, Michael. 2008. La práctica del análisis de correspondencias. Julio 2008. Fundación BBVA.

Greenacre, Michael. 2002. The use of correspondence analysis in the exploration of health survey data. Documentos de trabajo 5. Fundación BBVA.

Hoffman Donna, De Leeuw, Jan. 1992. Interpreting Multiple Correspondence Analyses as a Multidimensional Scaling Method. Marketing Letter 3:3. Kluwer Academic Publishers.

Howe, Laura; Hargreaves, James and Huttly, Sharon. January, 2008. Issues in the construction of wealth indices for the measurement of socio-economic position in low income countries. Emerging themes in epidemiology. Bio Med Central.

Instituto Nacional de Estadística y Censos (2003-Mayo). Incidencia de la pobreza y de la indigencia en los Aglomerados Urbanos. ISSN 0327– 7968. Argentina.

Le, Sébastian, Josse, Julie, Husson, Francois. FactoMineR: March 2008. An R Package for Multivariate Analysis. Journal of Statistical Software. Volume 25, Issue 1.

Nenadic, Oleg and Greenacre, Michael. May 2007. Correspondence Analysis in R, with two and three dimensional Graphics: The ca Package. Journal of Statistical Software. Volume 20, Issue 3.

Njong, Aloysius Mom and Ningaye, Paul. August 2008. Characterizing weights in the measurement of multidimensional poverty: An application of data driven approaches to Cameroon. OPHI Working Paper.

Tenenhaus, Michael and Young, Forrest. March 1985. An analysis and synthesis of multiple correspondence analyses, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. Psychometrika. Volume 50, N° 1

Anexo

Fuente: elaboración propia en base a FactoMineR de R Project.

Dimensión	Valor	%	Acumulado%
1	0,2856	12,7141	12,7141
2	0,1753	7,8055	20,5196
3	0,1211	5,3907	25,9103
4	0,1159	5,1614	31,0717
5	0,1087	4,8384	35,9101
6	0,1014	4,5135	40,4236
7	0,0858	3,8206	44,2442
8	0,0852	3,7930	48,0371
9	0,0843	3,7528	51,7900
10	0,0833	3,7098	55,4998
11	0,0821	3,6560	59,1558
12	0,0810	3,6072	62,7630
13	0,0808	3,5953	66,3584
14	0,0796	3,5458	69,9042
15	0,0724	3,2243	73,1285
16	0,0697	3,1036	76,2321
17	0,0677	3,0146	79,2467
18	0,0629	2,8011	82,0478
19	0,0616	2,7420	84,7898
20	0,0600	2,6727	87,4625
21	0,0516	2,2984	89,7609
22	0,0507	2,2552	92,0161
23	0,0448	1,9954	94,0116
24	0,0400	1,7824	95,7940
25	0,0384	1,7113	97,5053
26	0,0306	1,3627	98,8680
27	0,0254	1,1320	100,0000
Total:	2,2461	100	

Tabla N°I: Inercias principales

Variable	Dimensión 1	Dimensión 2	Dimensión 3
Pavimento	0,3161	0,0003	0,0377
Tipo de piso	0,4547	0,0051	0,0045
Tipo de baño	0,3302	0,0071	0,0072
Combustible	0,5170	0,0051	0,0044
Cloacas	0,3290	0,0031	0,0505
Teléfono fijo	0,3148	0,0017	0,0457
Nivel educativo del JH	0,2402	0,2324	0,1839
Cobertura médica explícita del JH	0,3698	0,0023	0,2292
Situación laboral del JH	0,2338	0,5270	0,4109
Nivel educativo del cónyuge	0,1858	0,6226	0,3165
JH de sexo femenino	0,0121	0,4151	0,1587
Hogar con niños menores de 14 años	0,1234	0,2822	0,0036
Inercia promedio	0,2856	0,1753	0,1211

Fuente: elaboración propia en base a FactoMineR de R Project.

Tabla N°II: Correlación al cuadrado de cada variable con cada dimensión

Variable		Coordenadas principales			Coordenadas estándares			Coseno al cuadrado			Contribuciones		
		Dim 1	Dim 2	Dim 3	Dim 1	Dim 2	Dim 3	Dim 1	Dim 2	Dim 3	Dim 1	Dim 2	Dim 3
Pavimento	sin Pavimento	0,9849	0,0300	-0,3402	1,8430	0,0717	-0,9777	0,3162	0,0003	0,0377	6,9570	0,0105	1,9578
	con pavimento	-0,3210	-0,0098	0,1109	-0,6006	-0,0234	0,3186	0,3162	0,0003	0,0377	2,2673	0,0034	0,6380
Tipo de piso	cerámica, baldosa, mosaico,	-0,3117	-0,0187	0,0075	-0,5833	-0,0448	0,0215	0,4437	0,0016	0,0003	2,3263	0,0137	0,0032
	mármol, madera o alfombra	1,4008	0,0704	-0,0394	2,6214	0,1681	-0,1134	0,3891	0,0010	0,0003	9,4759	0,0390	0,0177
	cemento o ladrillo fijo	2,2242	0,5585	-0,3205	4,1621	1,3338	-0,9209	0,0502	0,0032	0,0010	1,4496	0,1489	0,0710
Tipo de baño	tierra o ladrillo suelto										0,0975	0,0433	0,4048
	sin baño	2,3785	1,2419	3,1559	4,4508	2,9661	9,0696	0,0034	0,0009	0,0059	0,7929	0,0221	0,0040
	inodoro con botón/mochila/ cadena y arrastre de agua	-0,1721	-0,0225	0,0080	-0,3221	-0,0538	0,0229	0,3303	0,0057	0,0007			
Combustible	inodoro sin botón/mochila/cadena y arrastre de agua				3,3828	0,4634	-0,2479	0,2438	0,0028	0,0006	6,5565	0,1230	0,0352
	letrina (sin arrastre de agua)	1,8077	0,1940	-0,0863	4,3647	1,1432	-0,6723	0,0768	0,0032	0,0008	2,1876	0,1501	0,0519
	gas de red	2,3324	0,4787	-0,2339	-1,0282	-0,0427	0,0246	0,5054	0,0005	0,0001	5,5156	0,0095	0,0032
Cloacas	tubo o garrafa	-0,5495	-0,0179	0,0086	1,7087	0,0445	-0,0689	0,4795	0,0002	0,0003	8,8833	0,0060	0,0144
	leña o carbón	0,9131	0,0186	-0,0240	5,2051	3,0196	-1,2343	0,0226	0,0047	0,0005	0,6562	0,2209	0,0369
	otro	2,7815	1,2643	-0,4295	0,7939	0,2847	2,2548	0,0011	0,0001	0,0037	0,0311	0,0040	0,2509
Teléfono fijo	no	0,4242	0,1192	0,7846	1,4427	0,1774	-0,8684	0,3326	0,0031	0,0511	6,1803	0,0934	2,2392
	sí	0,7710	0,0743	-0,3022	-0,7986	-0,0982	0,4807	0,3326	0,0031	0,0511	3,4209	0,0517	1,2394
	hasta secundaria incompleta	-0,4268	-0,0411	0,1673	1,2941	-0,1229	0,7570	0,3148	0,0017	0,0457	5,5400	0,0499	1,8954
Nivel educativo del JH	secundaria completa	0,6916	-0,0515	0,2634	-0,8518	0,0809	-0,4983	0,3148	0,0017	0,0457	3,6466	0,0329	1,2476
	terciario/universitario	-0,4552	0,0339	-0,1734	0,8859	1,4643	-1,5251	0,1345	0,2256	0,1690	2,4497	6,6924	7,2600
	incompleto o más	0,0544	-0,4557	0,1895	0,1017	-1,0884	0,5446	0,0018	0,1242	0,0215	0,0322	3,6884	0,9233
Cobertura médica explícita del JH	no	-0,7848	-0,2358	0,5082	-1,4686	-0,5631	1,4605	0,2078	0,0187	0,0871	4,5261	0,6653	4,4759
	sí	1,0588	-0,0827	0,8335	1,9813	-0,1974	2,3954	0,3700	0,0023	0,2293	8,1147	0,0806	11,8607
	Estudiante_u_otro	-0,3493	0,0273	-0,2750	-0,6536	0,0651	-0,7902	0,3700	0,0023	0,2293	2,6769	0,0266	3,9126
Situación laboral del JH	Trabajador_sin_desc	0,3657	0,6005	0,4901	0,6843	1,4342	1,4086	0,0142	0,0383	0,0255	0,3746	1,6457	1,5873
	Cuenta_propia	0,8782	-0,0836	1,0283	1,6434	-0,1997	2,9553	0,1475	0,0013	0,2022	3,6127	0,0533	11,6829
	Desocupado	0,3137	-0,3813	0,3110	0,5871	-0,9106	0,8937	0,0164	0,0243	0,0161	0,4107	0,9880	0,9515
Nivel educativo del cónyuge	Jubilado	-0,4385	0,7756	0,4296	-0,8206	1,8524	1,2346	0,0124	0,0389	0,0119	0,3408	1,7367	0,7714
	Trabajador_con_desc	-0,3563	1,3222	-0,9550	-0,6668	3,1577	-2,7446	0,0237	0,3261	0,1701	0,5824	13,0621	9,8677
	Amas_de_casa	-0,3446	-0,6199	-0,3726	-0,6449	-1,4804	-1,0707	0,0653	0,2112	0,0763	1,2295	6,4786	3,3886
JH de sexo femenino	Patrón	-0,1448	0,3665	-0,9951	-0,2709	0,8753	-2,8598	0,0000	0,0001	0,0005	0,0003	0,0030	0,0316
	hasta secundaria incompleta	-0,5834	-0,9114	0,0023	-1,0917	-2,1768	0,0067	0,0096	0,0234	0,0000	0,2719	1,0811	0,0000
	secundaria completa	0,6131	0,1338	-1,0025	1,1474	0,3195	-2,8810	0,1044	0,0050	0,2791	2,3845	0,1849	15,0342
Hogar con niños menores de 14 años	terciario/universitario	0,1875	-0,8329	0,0364	0,3508	-1,9892	0,1046	0,0117	0,2305	0,0004	0,2557	8,2227	0,0227
	incompleto o más	-0,7389	-1,0062	0,1184	-1,3827	-2,4032	0,3403	0,0986	0,1829	0,0025	2,4371	7,3622	0,1476
	no corresponde	-0,1761	0,8744	0,5015	-0,3296	2,0884	1,4411	0,0190	0,4692	0,1543	0,3443	13,8216	6,5818
Hogar con niños menores de 14 años	no	0,0805	-0,4714	-0,2915	0,1506	-1,1258	-0,8378	0,0121	0,4151	0,1587	0,1231	6,8794	3,8093
	sí	-0,1503	0,8805	0,5445	-0,2813	2,1029	1,5648	0,0121	0,4151	0,1587	0,2300	12,8497	7,1151
	no	-0,3159	0,4777	-0,0540	-0,5911	1,1409	-0,1551	0,1234	0,2822	0,0036	1,6097	5,9975	0,1109
	sí	0,3906	-0,5908	0,0668	0,7310	-1,4110	0,1918	0,1234	0,2822	0,0036	1,9908	7,4173	0,1371

Tabla N°III: Contribuciones y correlaciones de las modalidades de las variables – Tres primeras componentes –

Fuente: elaboración propia en base a R Project.

ACM con FactoMineR -Coordenadas principales-

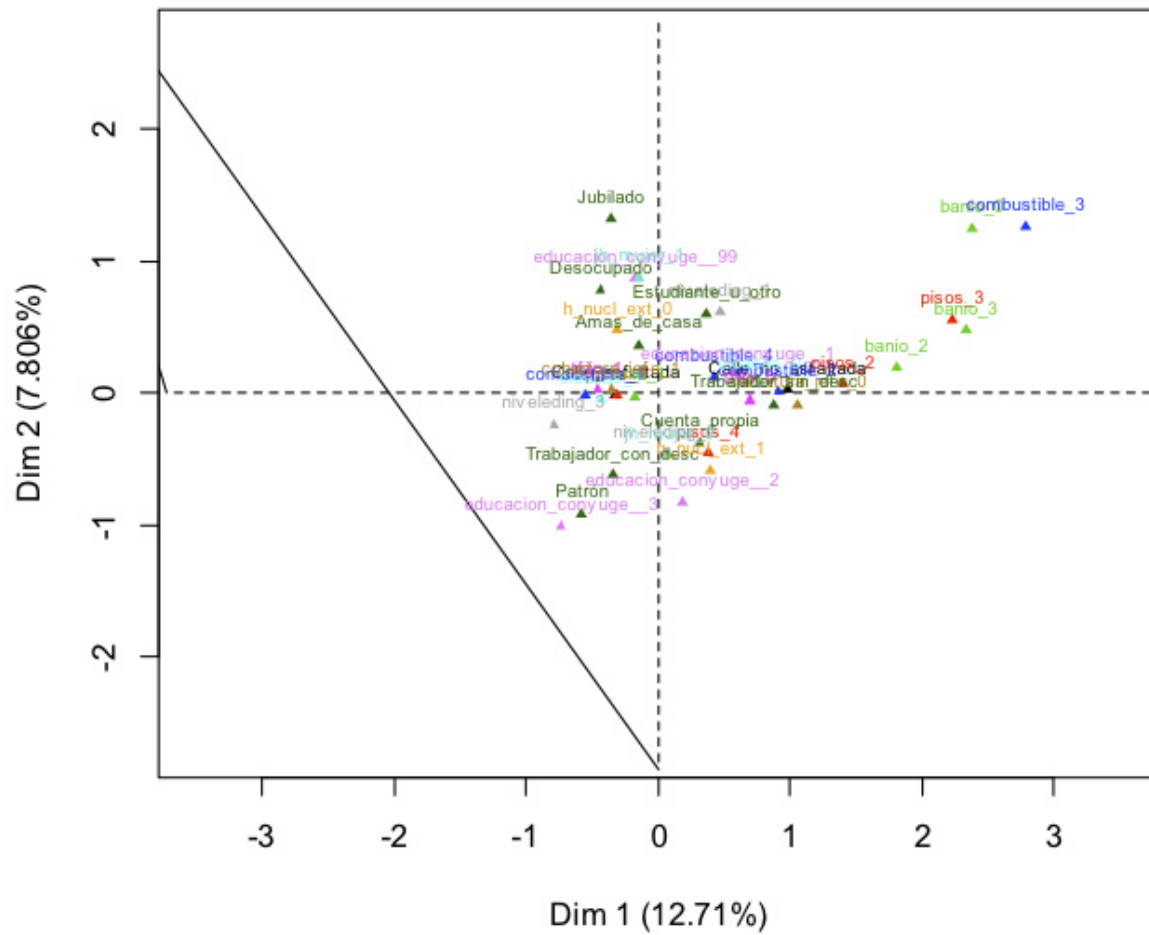


Gráfico I: Coordenadas principales e inercias de los primeros 2 ejes

Pavimento en la cuadra		Tipo de piso en la vivienda		Tipo de arrastre en el baño		Combustible utilizado para cocinar		Cloacas		Teléfono fijo	
Sin pavimento	1,8430	Cerámica, baldosa, mosaico, mármol, madera o alfombra	-0,5833	Sin baño	4,4508	Gas de red	-1,0282	No tiene	1,4427	No tiene	1,2941
Con pavimento	-0,6506	Cemento o ladrillo fijo	2,6214	Inodoro con botón/mochila/cadena y arrastre de agua	-0,3221	Tabo o garrafa	1,7087	Tiene	-0,7986	Tiene	-0,8518
		Tierra o ladrillo suelto	4,1621	Inodoro sin botón/mochila/cadena y arrastre de agua	3,3828	Leña o carbón	5,2051				
		Otros	0,7287	Letrina/ sin arrastre de agua	4,3647	Otro	0,7938				

Nivel educativo del jefe del hogar		Jefe del hogar con cobertura explícita		Situación laboral del JH		Nivel educativo del cónyuge		Jefe del hogar de sexo femenino		Hogar con niños menores de 14 años	
Hasta secundaria incompleta	0,8859	No	1,9813	Estudiante u otro	0,6843	Hasta secundaria incompleta	1,1474	No	0,1506	No	-0,5911
Secundaria completa	0,3017	Si	-0,6536	Trabajador sin descuentos	1,6434	Secundaria completa	0,3508	Si	-0,2813	Si	0,7330
Terciario/universitario o más	-1,4686			Cuenta Propia	0,5871	Terciario/universitario o más	-1,3827				
				Desocupado	-0,8206	No corresponde	-0,3296				
				Jubilado	-0,6668						
				Trabajador con descuentos	-0,6449						
				Ama de casa	-0,2709						
				Patrón	-1,0917						

= Puntaje del índice socioeconómico para el hogar "x"

Cuadro N° I: Representación algebraica del índice: el valor del índice socioeconómico de cada hogar surge del puntaje que adquiere la modalidad seleccionada por los encuestados pertenecientes a determinado hogar. Cada modalidad es excluyente de la otra, por ej. en el caso de pavimento en al cuadra, hay o no hay. Por lo tanto, cada variable suma un puntaje de una sola modalidad.

Fuente: elaboración propia.

Variable	Modalidad	Cantidad de hogares en cada quintil socioeconómico				
		Quintil 1	Quintil 2	Quintil 3	Quintil 4	Quintil 5
Pavimento	con pavimento	860.473	1.391.489	1.788.845	2.168.272	2.233.647
	sin pavimento	1.378.874	847.252	431.316	67.175	6.818
Tipo de pisos	cerámica, baldosa, mosaico, mármol, madera o alfombra	786.216	1.917.274	2.032.229	2.210.810	2.239.307
	cemento o ladrillo fijo	1.342.164	311.762	183.145	15.903	
	tierra o ladrillo suelto	94.296	2.561	15.383		
	otros	16.671	7.144	12.543	8.734	1.158
Tipo de baño	sin baño			6.544		
	inodoro con botón/mochila/cadena y arrastre de agua	1.398.933	2.185.623	2.100.689	2.235.347	2.240.465
	inodoro sin botón/mochila/cadena y arrastre de agua	694.163	46.657	21.034	100	
	letrina /sin arrastre de agua)	146.251	6.461			
Cloacas	no tiene	1.754.235	1.163.539	733.064	248.812	46.665
	tiene	485.112	1.075.202	1.388.659	1.986.635	2.193.800
Combustible utilizado para cocinar	gas de red	180.754	943.618	1.541.152	2.123.842	2.220.766
	tubo o garrafa	2.008.986	1.280.520	683.365	104.534	11.111
	leña o carbón	30.761	438	1.328		
	otro	18.846	14.145	17.655	7.071	8.588
Teléfono fijo	no tiene	1.842.161	1.178.317	831.351	469.832	123.122
	tiene	397.186	1.060.424	1.412.149	1.765.615	2.117.343
Situación laboral	ama de casa	917	430	2.159	881	801
	cuenta propia	456.845	409.940	343.558	234.169	156.451
	desocupado	69.280	83.956	110.583	150.878	265.385
	estudiante u otro	328.645	242.559	236.769	196.378	70.649
	jubilado	121.838	285.107	399.360	484.979	468.907
	patrón	29.065	25.439	42.063	69.346	140.658
	relación dependencia. con descuentos	375.122	725.745	853.331	929.565	1.088.276
	relación dependencia. sin descuentos	857.635	465.565	255.677	169.251	49.338
Presencia de niños menores de 14 años en el hogar	no	672.858	992.108	1.254.209	1.513.610	1.758.516
	sí	1.566.489	1.246.633	989.291	721.837	481.949
Nivel educativo del jefe del hogar	hasta secundaria incompleta	1.367.122	1.052.772	920.529	598.085	246.160
	secundaria completa	817.549	997.646	907.103	941.525	510.740
	terciario/universitario incompleto o más	54.676	188.323	390.968	695.837	1.483.565
Jefe del hogar de género femenino	no	1.620.307	1.545.712	1.505.828	1.524.495	1.096.656
	sí	619.040	693.029	737.672	710.952	1.143.809
Jefe del hogar con cobertura explícita	no	1.512.380	796.530	383.343	83.362	656
	sí	726.967	1.442.211	1.854.959	2.152.085	2.239.809
Nivel educativo del cónyuge	hasta secundaria incompleta	888.297	659.419	533.743	341.771	10.652
	secundaria completa	658.420	737.173	593.567	643.756	159.273
	terciario/universitario incompleto o más	49.467	144.177	265.450	361.815	892.050
	no corresponde	643.163	697.972	850.740	888.105	1.178.490

ARTÍCULOS

Orientaciones de futuro laboral y educativo de estudiantes secundarios. Análisis multivariado en un diseño muestral complejo.¹

Rosario Austral

análisis multivariado

orientaciones de fu

Las trayectorias de vida de las personas se hallan influidas por la estructura de oportunidades del contexto, por el bagaje de recursos heredados o adquiridos, así como por los deseos y las expectativas con respecto a su porvenir. Es por esto que conocer cómo las personas piensan el futuro –en general y en el terreno personal– permite dar cuenta de representaciones que posiblemente graviten en sus cursos de vida (Nurmi, 1991). Independientemente de que la concreción de las aspiraciones se vea luego facilitada o restringida por condiciones objetivas que van mediando entre las expectativas y los logros, las visiones acerca del futuro cobran relevancia en tiempo presente, como autoatribuciones de los horizontes de posibilidad. En otras palabras, se trata de analizar cómo las personas se ven y se piensan a sí mismas, a partir del conocimiento acerca de qué esperan de sí y para sí. En este trabajo, el concepto *orientaciones de futuro* (Gjesme, 1981; Trommsdorff, 1983; Nurmi, 1991) resulta una herramienta conceptual fundamental para el abordaje de la cuestión.

El universo analizado se halla definido en este caso por los estudiantes que, a fines de 2008, se hallaban cursando el último año de estudio en escuelas secundarias de gestión estatal en el ámbito de la Ciudad Autónoma de Buenos Aires. El interés de este recorte analítico radica en la importancia de generar conocimiento acerca de las representaciones de los estudiantes acerca de la formación recibida y las posibilidades futuras, considerando que se trata de un aspecto de importancia en las transiciones laborales y educativas de los futuros egresados.

El trabajo se realizó sobre la base de los datos recabados por el Equipo de Nivel Secundario de la actual Gerencia Operativa de Investigación y Estadística del Ministerio de Educación del GCABA, de la cual la autora forma parte. Fue así como, para la elaboración de la tesis, se profundizó el análisis de las dimensiones comprendidas en las orientaciones de futuro laboral y educativo, avanzándose además en la exploración de algunos modelos de análisis multivariado sobre los objetivos de tipo profesional. Para ello se tomó el recaudo de utilizar los estimadores más adecuados que contemplaran las características del diseño muestral en términos de conglomeración, estratificación y selección sistemática con probabilidades desiguales sin reemplazo.

En cuanto al papel de la escuela, está claro que la obligatoriedad legal y la masificación del acceso a la educación secundaria en el transcurso de los últimos años ha delineado un escenario particular para las trayectorias vitales, laborales y educativas de los jóvenes. En este sentido es que interesó conocer el papel de la escuela en la construcción de los proyectos y aspiraciones de los jóvenes, sin omitir aquellos otros aspectos de tipo contextual, social y biográfico que también podían incidir en los itinerarios vitales.

Cabe la mención de que la producción de los horizontes temporales se halla imbricada con la producción del orden social (Lechner, 2002). En este sentido, no resulta ajena al tema analizado la individualización de las trayectorias vitales (Sennet, 1998) que ha repercutido en términos de ruptura del clásico modelo de "carrera laboral" (Dubet y Martuccelli, 2000; Sennet, 1998; Castel, 1997; Rosanvallon, 1995). En cuanto al contexto local, resulta necesario considerar como telón de fondo la recuperación de un rol activo del Estado en las políticas macroeconómicas con posterioridad a la crisis del 2001 (Novick, 2006), aunque sin llegar a revertir las problemáticas laborales en el segmento juvenil (Miranda y Zelarrayan, 2011; Álvarez y Fernández, 2011; Weller, 2006). Por otra parte, en el plano educativo no puede dejar de mencionarse el establecimiento de la obligatoriedad del nivel secundario (en la CABA establecida en 2002 y, a nivel nacional, en 2006 mediante la Ley 26.206).

Volviendo al foco de este trabajo, la descripción de las orientaciones de futuro abarcó varios aspectos: los objetivos priorizados, los plazos y obstáculos imaginados a futuro y la confianza en las posibilidades de concreción de los objetivos. Dichos aspectos fueron considerados en los planos laboral y educativo, primero separadamente y luego de manera combinada. Asimismo se comparó la influencia relativa de distintos atributos sobre los objetivos de tipo profesional, es decir, aquellos que apuntaban a un desempeño laboral que requería de un período de formación o de capacitación postsecundaria. Es así como se identificaron los perfiles biográficos y sociales asociados a estos planes de futuro, así como las condiciones contextuales y escolares que, en principio, los propiciaban en mayor medida.

En cuanto a la estructura del artículo, luego de este apartado introductorio se presenta de manera resumida el modelo analítico utilizado (sección 2) y se describen los principales aspectos metodológicos y estadísticos del trabajo (sección 3). A partir de allí se exponen los resultados más

destacados: la descripción de las orientaciones de futuro laboral y educativo (sección 4.1), la modelización multivariada sobre los objetivos de tipo profesional a través de un análisis de regresión logística (sección 4.2) y algunos análisis que demuestran la importancia de la consideración de la complejidad del diseño muestral en la instancia de análisis de los datos (sección 4.3). Por último, se presenta una síntesis de los principales hallazgos.

2. El modelo analítico utilizado

El concepto más sustantivo de este trabajo es el de “orientaciones de futuro”. Se trata de un término proveniente de la Psicología Social que se define como la capacidad de anticipación de eventos futuros y que abarca tanto la elaboración cognitiva de planes y proyectos, como el grado de incumbencia, participación y compromiso en el futuro (Gjesme, 1981). Las orientaciones de futuro comprenden dos dimensiones: una “motivacional” referida a los intereses de las personas a futuro y una “cognitiva” que alude a la forma en que las personas piensan acerca del futuro en términos de secuencia temporal y causalidad (Trommsdorff, 1983). La dimensión motivacional comprende tres subdimensiones: a) *el involucramiento* (la medida en que las personas piensan en su futuro), b) *el contenido temático* (los objetivos que se formulan) y c) *el contenido afectivo* (el nivel de optimismo manifestado con respecto a la concreción de ese futuro). La dimensión cognitiva se halla referida a la forma en que se estructuran las orientaciones de futuro y abarca cuatro subdimensiones: a) *la extensión* (los plazos previstos para alcanzar los objetivos), b) *la sensación de control* (identificación de los posibles obstáculos a sortear que, según su naturaleza externa o interna, daría cuenta del nivel de expectativas con respecto a las posibilidades de control de la futura trayectoria) (Nurmi, 1991), c) *la densidad* (cantidad de objetivos propuestos) y d) *el realismo* (grado de concordancia con las posibilidades y modos de concreción de los planes de futuro) (Husman y Lens, 1999; McCabe y Barnett, 2000). 2

La prevalencia empírica y el interés analítico en los objetivos de tipo profesional, obligan a una delimitación conceptual del término “profesión”. De acuerdo con Hargreaves (2000), la misma se asocia a una serie de rasgos como la autonomía, el conocimiento especializado, un período de entrenamiento y prácticas estandarizadas. Es por eso que en este trabajo se consideraron como objetivos profesionales aquellos que consistían en la formulación de trayectos que incluían un período de formación o capacitación luego de la secundaria.

En el Diagrama 1 se presenta el modelo analítico completo: el concepto orientaciones de futuro y cuatro dimensiones o grupos de variables explicativas introducidas en la regresión logística sobre los objetivos profesionales.

Las variables explicativas consideradas en el análisis se agrupan en cuatro dimensiones:

1. Atributos de la oferta educativa local: a) **la ubicación geográfica de las escuelas** (sur/ resto de la CABA) permitió diferenciar los barrios de la zona sur caracterizados por una trama urbana fragmentada, patrones heterogéneos de ocupación del suelo y de distribución poblacional, así

como por mayores niveles de pobreza, desocupación y déficit educativo (Documento interno de la DIE del GCBA, 2008); b) **la modalidad del plan de estudios** (técnico/ bachiller o comercial) posibilitó la diferenciación entre tradiciones formativas e institucionales, con mandatos fundacionales y formatos escolares que podían incidir en las aspiraciones de los estudiantes (Filmus y Moragues, 2003); y c) el turno (diurno/ nocturno) distinguió entre ofertas diurnas y nocturnas que, en general, se muestran asociadas a distintas composiciones socio-educativas de los alumnos.

2. El perfil social del alumnado: a) **el sexo** fue considerado una característica demográfica clave en el marco de profundas transformaciones sociales y culturales que vienen aconteciendo en términos de ampliación de umbrales educativos y de participación laboral de las mujeres; y b) **el origen social educacional** (madre hasta secundario incompleto/ madre con secundario completo o más) fue contemplado a través del nivel educativo materno que, como se ha probado, presenta fuerte gravitación tanto en los horizontes (Bourdieu, 2006; Fieulaine, 2006; Guichard, 1995) como en los logros educacionales (Jorrat, 2008; Hossler y Stage, 1992; Kandel y Lesser, 1969).

3. La biografía educativa y laboral: a) **la trayectoria escolar** (sin/ con sobreedad) fue clave para la distinción entre itinerarios educativos ajustados a los teóricos previstos para la promoción y la terminalidad de la secundaria, y aquellos de una duración mayor que arrojaban indicios de episodios de repetición y/ o abandono escolar; y b) **la experiencia laboral** (sin/ con experiencia) se consideró como un aspecto que podía intervenir fuertemente en las significaciones que componían las imágenes de futuro laboral y educativo de los jóvenes.

4. El contexto institucional: a) **el tamaño de la escuela** (pequeña/ mediana o grande) remite de algún modo a las posibilidades de acompañamiento y seguimiento personalizado de los estudiantes; b) **el nivel de abandono escolar intraanual de la escuela** (bajo/ medio o alto) fue considerado un atributo de importancia como indicador del tipo de ambiente de aprendizaje (Cervini, 2003) y dada su variabilidad institucional (Steinberg y Tófaló, 2013; Steinberg, Cetrángolo y Gatto, 2011; PNUD, 2009); c) **los niveles de valoración estudiantil de los aportes de las orientaciones cursadas, de los aprendizajes y de las experiencias escolares** (bajos o medios/ altos) basados en actitudes, expectativas y percepciones de los propios estudiantes, fueron incluidos en el análisis debido a su importancia en estudios de efectividad escolar (Cervini, 2003).

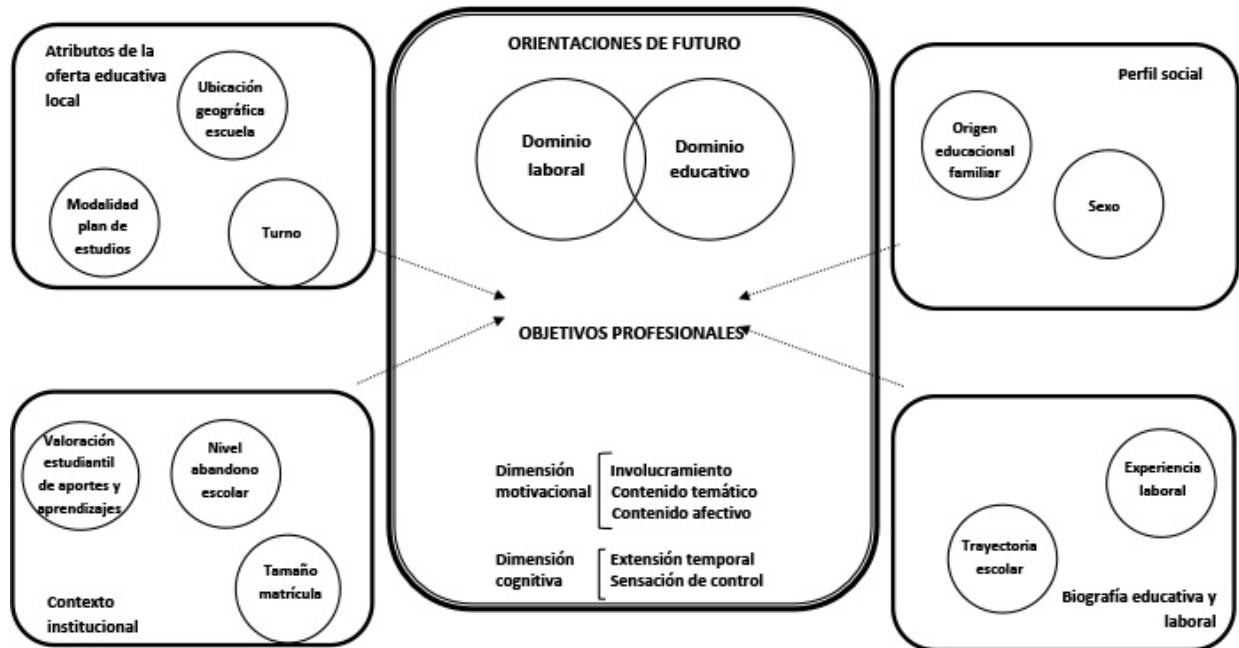


Diagrama 1. Modelo analítico de las orientaciones de futuro laboral y educativo de los estudiantes del último año de la secundaria

3. Aspectos metodológicos y estadísticos

Como se aclaró oportunamente, se utilizó como fuente una encuesta a estudiantes del último año de escuelas secundarias estatales de la CABA (quinto año en las escuelas con orientaciones comerciales y de bachillerato, y sexto en las escuelas técnicas). Se trató de un instrumento autoadministrado con presencia del encuestador frente a cada sección de estudiantes.

El marco muestral fue la base de datos del Relevamiento Anual 2006 del Ministerio de Educación nacional. Se extrajo una muestra de escuelas estatales partiendo de una estratificación que combinó el turno de la oferta educativa, la modalidad del plan de estudios y la ubicación geográfica de los establecimientos. De este modo, la unidad de muestreo fue la escuela-modalidad-turno, siendo cada institución contabilizada tantas veces como modalidades y turnos

h

comprendiera. La asignación de la muestra en los estratos (n_h) fue proporcional a la distribución

h

del universo en los mismos (N_h), y se previeron estimaciones con un 95% de confianza.

Técnicamente se llevó a cabo entonces una selección estratificada de conglomerados de estudiantes con probabilidades desiguales sin reemplazo, ya que al interior de cada estrato se realizó un muestreo sistemático de escuelas-modalidades-turnos con probabilidades proporcionales al tamaño en matrícula. Se encuestó a todos los estudiantes que, al momento del relevamiento, se hallaban presentes en las unidades seleccionadas. Por lo tanto, la probabilidad de inclusión de cada estudiante en la muestra resultó igual a la probabilidad de inclusión de su escuela-modalidad-turno en el correspondiente estrato:

$$\pi_i = n_h \frac{x_i}{\sum_{i=1}^{N_h} x_i}$$

h

i

donde n_h es el tamaño muestral del estrato en cantidad de conglomerados y x_i es la cantidad de alumnos matriculados en el último año de estudio para la escuela-modalidad-turno i .

Se seleccionaron 57 (n) escuelas-modalidades-turnos de un total de 187 (N), realizándose 3.402

encuestas a estudiantes. Solo 10 jóvenes se rehusaron a responder, es decir que, en términos de la relación entre entrevistas concluidas y unidades contactadas (Lohr, 2000), la no respuesta se redujo a una mínima expresión.

El modelo de regresión logística binomial resultó apropiado para el análisis de las orientaciones de futuro de los estudiantes, al aportar pistas acerca de cuáles son los atributos personales, institucionales, sociales y de la oferta educativa que se asocian a la priorización de objetivos de tipo profesional entre los estudiantes. Este tipo de modelo permite explicar una variable binaria que mide la ocurrencia de algún evento particular o la presencia de cierto atributo de interés. En general, esta variable se presenta con valores 0 y 1, y el modelo apunta a explicar su media π que –calculada como la suma de valores dividida por el tamaño muestral– representa la proporción de casos que cumplen con la condición estudiada. La expresión matemática general de la función al incorporar más de una variable explicativa es la siguiente:

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 X_1 + \dots + \beta_k X_k$$

La forma abreviada de la función es:

$$\text{logit}(\pi) = \alpha + \beta_1 X_1 + \dots + \beta_k X_k$$

La expresión antilogarítmica permite obtener el π esperado para cada x:

$$\pi = \frac{e^{\alpha + \beta_1 X_1 + \dots + \beta_k X_k}}{1 + e^{\alpha + \beta_1 X_1 + \dots + \beta_k X_k}}$$

La manera más sencilla de interpretar π es con una aproximación lineal. La línea tangente a la curva de regresión logística tiene pendiente $\pi(1-\pi)$, la cual se minimiza en valores extremos de π y se maximiza en $\pi=0.5$.

Por otra parte, y más allá del tipo de modelización utilizada, se destaca la necesidad de utilizar medidas apropiadas que consideren la complejidad de un diseño muestral estratificado y

s

conglomerado. En primer lugar se usa la corrección de segundo orden Rao-Scott χ^2 a la prueba

Chi cuadrada (χ^2), que corrige la esperanza asintótica (E) y la varianza (V) de la distribución χ^2 utilizando las pruebas de modelos de análisis de varianza de Satterthwaite χ^2 (Lohr, 2000). Se trata de comparar la expresión:

$$X_s^2 = \frac{v X_F^2}{(r-1)(c-1)}$$

con la expresión χ^2 con v grados de libertad, donde:

$$v = 2 \frac{[E(X^2)]^2}{V[X^2]}$$

Como se observa en la primera fórmula, la corrección de segundo orden incluye a la de primer

F

orden χ^2 que consiste en la corrección de la esperanza de χ^2 a partir de la comparación de la

siguiente expresión

$$X_F^2 = \frac{(r-1)(c-1) X^2}{E[X^2]}$$

$(r-1)(c-1)$

con una distribución χ^2 , donde r es el número de filas y c el número de columnas del cruce

de variables en cuestión.

En cuanto a la estimación de varianza, se debió utilizar la aproximación Brewer considerando el muestreo sistemático con probabilidades desiguales sin reemplazo:

$$\hat{\sigma}_{Brewer}^2 = \sum_{i \in s} \check{c}_i^* \hat{e}_i^{*2}$$

donde $\check{c}_i^* = n_h(n_h - 1)^{-1}(1 - \pi_i)$

i

y \hat{e}_i representa los residuos mínimos cuadrados ponderados.

Por último, también se consideró el efecto del diseño muestral sobre la varianza de la estimación, el cual proporciona una medida de la precisión ganada o perdida por el uso del diseño complejo con respecto a una muestra aleatoria simple. De acuerdo con Lohr, el efecto de diseño de una muestra estratificada será generalmente menor a 1, lo cual significa una mejora en la precisión respecto de un muestreo aleatorio simple. En el muestreo por conglomerados, en cambio, el efecto de diseño es generalmente mayor a 1, lo cual indica una pérdida de precisión. Dado que en este caso se trata de una muestra estratificada y conglomerada, “el efecto de diseño total dependerá de si se pierde más precisión con los conglomerados que la ganada con la estratificación” (Lohr, 2000, p.238).

$$ED(plan, \hat{y}) = \frac{V[\hat{y}]}{\left(1 - \frac{n}{N}\right) \frac{S^2}{n}}$$

En la sección siguiente se presentan los principales resultados del análisis de las orientaciones de futuro laboral y educativo de los estudiantes.

4. Resultados destacados

Los resultados se presentan organizados en torno a tres núcleos de análisis: la descripción de las orientaciones de futuro laboral y educativo, el análisis de regresión logística multivariada sobre los objetivos de tipo profesional y algunos desarrollos que demuestran la importancia de la

consideración de la complejidad del diseño muestral en la instancia de análisis de los datos.

4.1 La descripción de las orientaciones de futuro

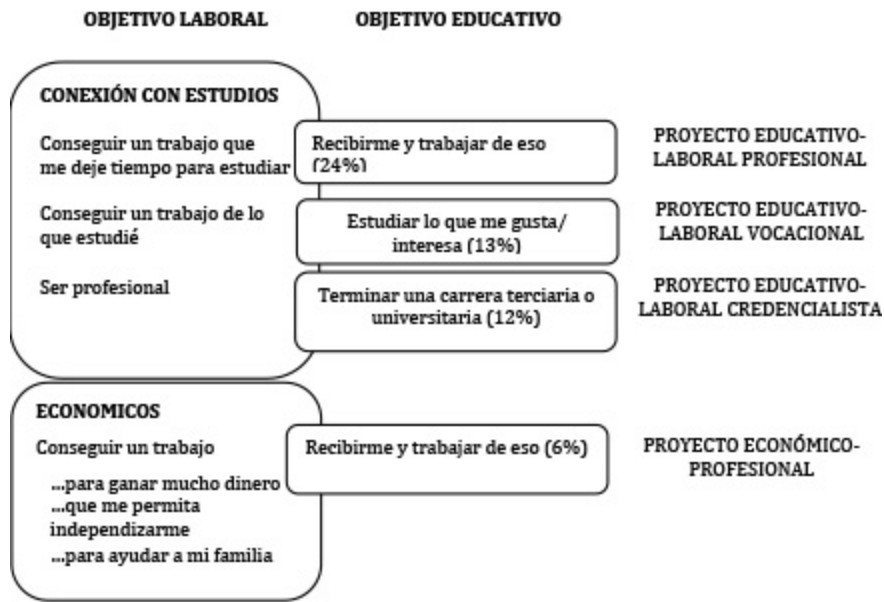
Entre las metas privilegiadas por los estudiantes en cada plano (el laboral y el educativo) prevalecieron los proyectos educativo-laborales (Guichard, 1995; Falco de Jouas, 2003) en los que la educación y el trabajo se hallan fuertemente imbricados. En el plano laboral, cerca de la mitad de los estudiantes priorizó un objetivo que respondía a alguno de los siguientes itinerarios: el de “estudiar para trabajar” con eje en la continuación de estudios con vistas a un desempeño laboral especializado (elegido por el 29%) o el plan de “trabajar para estudiar”, apuntando a una pronta inserción laboral que estableciera condiciones para seguir estudiando (señalado por el 18% de los estudiantes). También al elegir un objetivo en el plano educativo, 1 de cada 3 jóvenes apostó a seguir estudiando para poder trabajar en algo afín a esa formación.

Los estudiantes estuvieron lejos de escatimar realismo al pensar en las dificultades que posiblemente enfrentarían a futuro (6 de cada 10 jóvenes previeron más de un obstáculo). Esto evidencia que la encuesta logró captar lo que los jóvenes consideraban viable o posible más allá de lo “deseable” (Corica, 2010), es decir las expectativas concretas más que las aspiraciones en sentido abstracto (Hauser y Anderson, 1991). Frente a las distintas metas señaladas por los estudiantes, la competencia y la falta de oportunidades figuraron entre las principales amenazas en el plano laboral. En cuanto al futuro educativo, lo que más preocupó fue la escasez de tiempo para estudiar.

Por otra parte, se identificó todo un repertorio de temporalidades en torno a los plazos previstos para la concreción de los objetivos elegidos. Uno de los principales contrastes en la extensión temporal se observó entre las expectativas de inserción laboral a corto plazo y las aspiraciones profesionales de más largo alcance.

Dada la centralidad cobrada por los proyectos educativo-laborales (Guichard, 1995) se profundizó acerca de las conexiones entre los dominios laboral y educativo de las orientaciones de futuro. Fue así como se elaboró una tipología en base a las convergencias halladas entre los objetivos laborales y educativos de los estudiantes, quedando definidos los siguientes tipos (ver el diagrama 2): a) proyectos educativo-laborales profesionales (24% de los estudiantes); b) proyectos educativo-laborales vocacionales (13%); c) proyectos educativo-laborales credencialistas (12%) y d) proyectos económico-profesionales (6%). Más de la mitad de los estudiantes quedó comprendida en alguno de estos enlaces cuyo denominador común era un itinerario futuro de profesionalización, aunque con dinámicas distintas en cuanto a la estructuración temporal y la anticipación de posibles obstáculos.

Diagrama 2 Principales enlaces entre objetivos laborales y educativos



Cabe resaltar como hallazgo de relevancia que los estudiantes imaginaron menos dificultades educativas que laborales. Si bien los modos de “inserción” están lejos de expresar las voluntades o intenciones individuales (Nicole-Drancourt, 1994), muchos relatos juveniles se apoyaron fuertemente en la idea de que la voluntad personal permitiría sortear una estructura de oportunidades limitante (Meo y Dabenigno, 2008). Al respecto –y considerando como telón de fondo los altos índices de fracaso en los inicios de los estudios superiores- surge como reflexión la necesidad de políticas que contribuyan a zanjear estas distancias entre la estructura de oportunidades y los anhelos individuales.

El futuro laboral, en cambio, apareció menos previsible y “controlable” en la mirada de los jóvenes. Únicamente entre algunos estudiantes que aspiraban a una pronta inserción laboral para ayudar a su familia, la meta educativa se visualizó como más dificultosa.

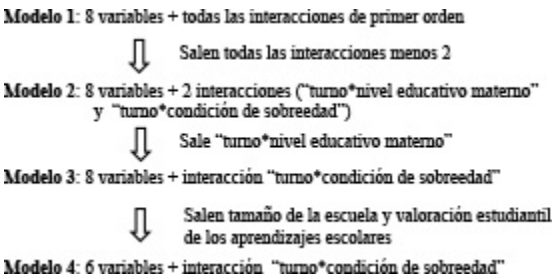
4.2 Los modelos explicativos sobre los objetivos profesionales

La prevalencia empírica y la relevancia teórica de los objetivos de tipo profesional llevaron a emprender el desafío de identificar los distintos aspectos que podían explicar estas preferencias. Para ello se utilizó un modelo de análisis multivariado sobre los objetivos profesionales. Siguiendo el modelo analítico presentado en la sección 2, se consideraron cuatro dimensiones explicativas: contextual, institucional, social y biográfica. Dado que en general los estudiantes habían previsto menos condicionamientos educativos que laborales –quedando sugeridas distintas estructuraciones de los espacios de posibilidad imaginados-, en esta instancia se consideró pertinente analizar por separado cada plano de futuro.

A partir de un conjunto inicial de variables presumiblemente explicativas, la estrategia de análisis permitió –en base a criterios estadísticos- descartar variables, detectar relaciones espurias y

descubrir los efectos puros y combinados de algunas de ellas. En el plano laboral, el análisis bivariado de preselección de variables para el abordaje multivariado, mostró que los objetivos profesionales no expresaban asociación con la zona geográfica, el nivel de abandono escolar de la escuela o la experiencia laboral del joven. Luego, en los cuatro pasos de la regresión logística multivariada, el tamaño de la escuela y el nivel de valoración estudiantil de los aprendizajes escolares también fueron descartadas como variables asociadas a este tipo de metas (los pasos se sintetizan en el diagrama 3).

Diagrama 3. Sucesivos modelos de regresión sobre los objetivos laborales profesionales



El modelo final se presenta en el cuadro 1. La interacción turno*sobreedad resultó fuertemente

7

explicativa en el modelo, con un $\beta=0,725$. Las otras dos variables que resultaron muy asociadas a

5

1

los planes profesionales fueron sexo con $\beta=0,349$ y modalidad del plan de estudios con $\beta=0,309$, significativas al 1%. Aunque en menor medida, también parecieron propiciar este tipo de planes el contexto institucional en términos de valoración estudiantil de los aportes de la orientación

3

4

cursada con un $\beta=-0,242$ y también el nivel educativo materno con un $\beta=0,203$, ambas variables

significativas al 5%. La esquematización del modelo final se presenta en el diagrama 4.

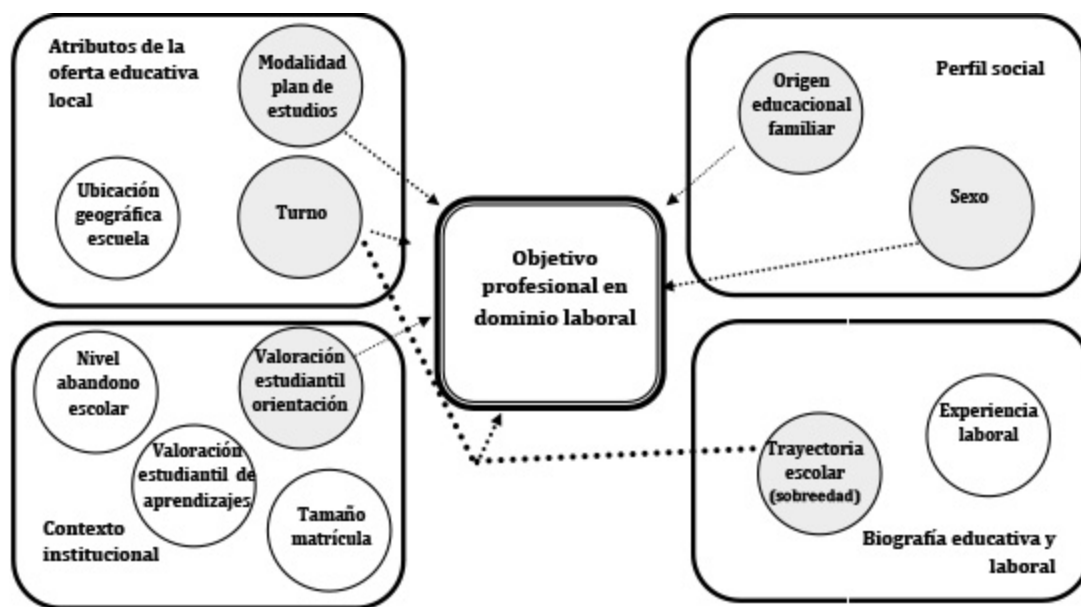
Códigos de significación: **** 0.001 *** 0.01 ** 0.05 * 0.1 " " 1

Fuente: Elaboración propia sobre la base de la encuesta a estudiantes secundarios 2008 (99) E/ ME/ GCABA).

Coeficientes	B	Error estándar	Test Wald			Wald ^{2(a)}	OR Odds ratio (e ^b)	Intervalo de estimación de B (95% de confianza) exp[B± 1.96(S.E.)]		Intervalo de estimación de OR (95% confianza) (e ^b límite intervalo ^c)	
			t	p-valor				inf.	sup.	inf.	sup.
(Intercept)	-1,454	0,125	-11,675	3,90E-14	***						
Modalidad bachiller-comercial	0,309	0,090	3,443	0,001	**	11,854	1,36	0,13	0,48	1,14	1,62
Turno diurno	-0,303	0,154	-1,966	0,057	.	3,865	0,74	-0,60	0,00	0,55	1,00
Alta valoración estudiantil de la orientación del plan de estudios para el trabajo	-0,242	0,097	-2,489	0,017	*	6,195	0,78	-0,43	-0,05	0,65	0,95
Madre con secundario completo o más	0,203	0,077	2,630	0,012	*	6,917	1,22	0,05	0,35	1,05	1,42
Mujer	0,349	0,108	3,227	0,003	**	10,414	1,42	0,14	0,56	1,15	1,75
Sin sobreedad	-0,411	0,170	-2,420	0,020	*	5,856	0,66	-0,74	-0,08	0,48	0,92
Turno * Condición de sobreedad	0,725	0,194	3,736	0,001	***	13,958	2,06	0,34	1,11	1,41	3,02

Cuadro 1. Regresión logística sobre los objetivos laborales profesionales. Estudiantes del último año escuelas secundarias estatales. CABA. Año 2008

Diagrama 4. Modelo explicativo sobre los objetivos laborales profesionales



En cuanto a los signos negativos que presentan las variables turno y sobreedad, estos se deben a la presencia de la interacción entre ambas en el modelo final. Siguiendo a Heeringa, West y Berglund (2010) en el desarrollo de las razones de odds para el análisis de las interacciones en los modelos de regresión logística, se interpretó la incidencia combinada del turno y la sobreedad, ($OR_7 = 2,06$). Se estimaron cuatro modelos logits resultantes de la rotación de las categorías de referencia en las dos variables implicadas en la interacción (ver la columna 4 del cuadro 2). Luego se calcularon las diferencias de los diferentes logits estimados con respecto al logit calculado para la combinación “nocturno-con sobreedad” (columna 5), expresando las mismas como razones de odds (columna 6). Así se constató una diferencia mayor entre los estudiantes sin sobreedad de turnos diurnos ($e^{(\text{logit } 1 - \text{logit } 4)} = 1,034$) respecto de los jóvenes sin sobreedad de los nocturnos ($e^{(\text{logit } 3 - \text{logit } 4)} = 0,916$), lo cual condice con las diferencias en los porcentajes de estudiantes con planes laborales profesionales (28% vs. 16%).

Elaboración propia sobre la base de la encuesta a estudiantes secundarios 2008 (GOlyE/ ME/ GCABA).

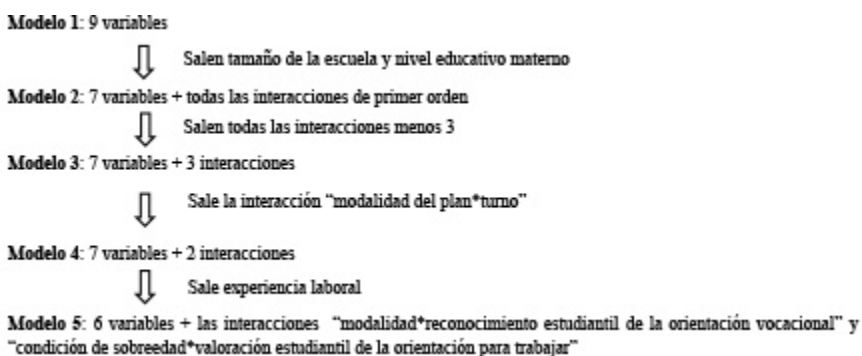
Esquema de covariación (i)	Turno	Condición de sobreedad	Logit _i estimado	Logit _i - logit ₄	$e^{(\text{logit } i - \text{logit } 4)}$
1	Diurno	Sin Sobreedad	0,278	0,033	1,034
2	Diurno	Con Sobreedad	0,215	-0,030	0,970
3	Nocturno	Sin Sobreedad	0,157	-0,088	0,916
4	Nocturno	Con Sobreedad	0,245	0,000	1,000

Cuadro 2 Esquemas de covariación entre turno y condición de sobreedad, logits estimados y razones odds de comparación para la regresión sobre los objetivos laborales profesionales

En cuanto a las regresiones logísticas sobre los objetivos profesionales de estudio, fueron

removidas las siguientes variables: zona geográfica, nivel de abandono escolar de la escuela y valoración estudiantil de los aprendizajes escolares (en el análisis bivariado de preselección de variables), y luego también el tamaño de la escuela, la experiencia laboral y el origen educacional (en los 5 pasos del análisis de regresión que se detallan en el diagrama 5).

Diagrama 5. Sucesivos modelos de regresión sobre los objetivos educativos profesionales



La salida de la regresión final se presenta en el Cuadro 3. Se observó una fuerte incidencia de la variable sexo en los objetivos profesionales, y en segundo orden de importancia, del turno y de dos interacciones: "modalidad* reconocimiento estudiantil de la orientación vocacional" y "sobreedad*valoración estudiantil de la orientación para la inserción laboral". Los resultados

4

indican una fuerte incidencia de la variable sexo en los planes de estudio profesionales, con un ?

= 0,755. Es así como las chances de optar por este tipo de planes se duplican entre las mujeres con respecto a los varones ($\widehat{OR}_4 = 2,13$). Esto abre preguntas acerca de por qué ellas trazaban en mayor medida sus prioridades en torno a esta meta de profesionalización. Las tasas de actividad de los jóvenes presentan diferencias por sexo y pueden aportar pistas en este sentido. De acuerdo con datos de la Encuesta Permanente de Hogares para el total de los aglomerados urbanos, en el año 2008 se hallaban activos el 51% de los varones y sólo el 35% de las mujeres entre 15 y 24 años (Base de datos SITEAL). Podría decirse entonces que las estudiantes mujeres, en contraste con los varones, podían estar ingresando más tardíamente al mercado laboral y pensando en futuros estudios, mientras que quizás ellos se sentían instados a salir más prontamente al mercado laboral. Los datos aportados por la encuesta arrojan algunas pistas en este sentido. En un segundo orden de importancia aparecen las interacciones "modalidad*reconocimiento estudiantil

7

de la orientación vocacional" (? = -0,446 y la "sobreedad*valoración estudiantil de la orientación

para la inserción laboral" ($\beta = -0,448$). Por razones de espacio se obvian en este caso los

desarrollos de las razones de odds para la interpretación de las interacciones en el modelo resultante. El turno de asistencia a las clases es otra condición que apareció con cierta fuerza

2

explicativa ($\beta = 0,300$), dando cuenta de que el cursar en un turno diurno aumenta en alrededor

del 35% las chances de aspiración profesional en el plano educativo ($\widehat{OR}_2 = 1,35$). En el diagrama 6 se sintetizan estos hallazgos.

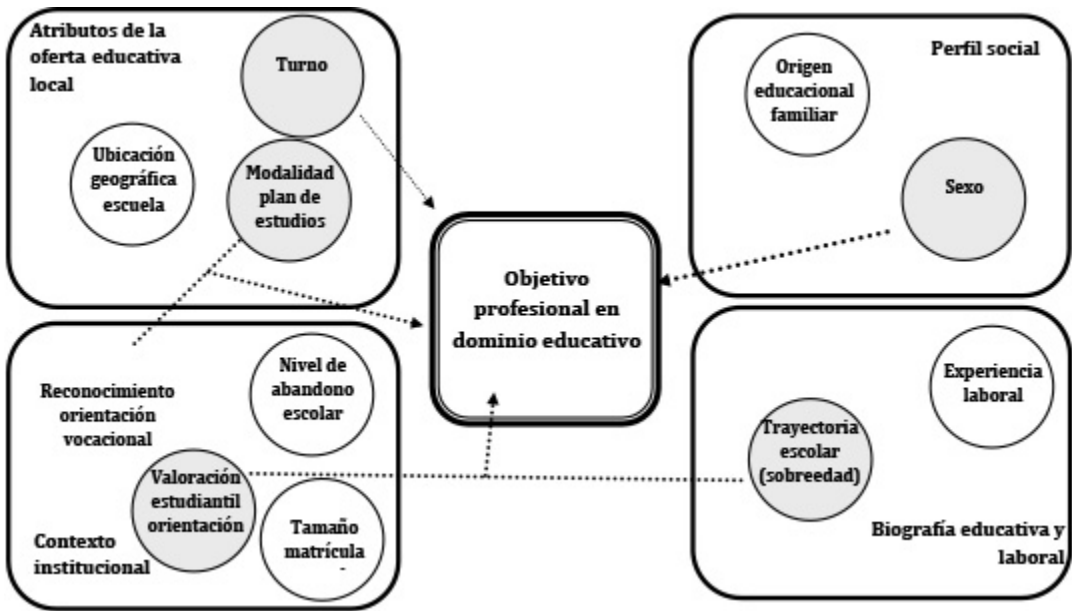
Códigos de significación: **** 0.001 *** 0.01 ** 0.05 * 0.1 " " 1

Fuente: Elaboración propia sobre la base de la encuesta a estudiantes secundarios 2008 (E/ME/SCABA).

Coeficientes	B	Error estándar	Test Wald			Wald ²⁽ⁿ⁾	OR Odds ratio (e^{β})	Intervalo de estimación de B (95% de confianza) $\exp[B \pm 1.96(S.E.)]$		Intervalo de estimación de OR (95% confianza) $(e^{\text{inferior intervalo } \beta}, e^{\text{superior intervalo } \beta})$	
			t	p-valor				inf.	sup.	inf.	sup.
(Intercept)	-1,93	0,113	-17,123	< 2e-16	***						
Modalidad bachiller-comercial	0,301	0,093	3,237	0,003	**	10,478	1,35	0,12	0,48	1,13	1,62
Turno diurno	0,300	0,080	3,746	0,001	***	14,033	1,35	0,14	0,46	1,15	1,58
Alto reconocimiento estudiantil de los aportes de la orientación vocacional	0,120	0,109	1,101	0,278		1,212	1,13	-0,09	0,33	0,91	1,40
Mujer	0,755	0,083	9,094	5.72e-11	***	82,701	2,13	0,59	0,92	1,81	2,50
Sin sobreedad	0,369	0,084	4,374	9.55e-05	***	19,132	1,45	0,20	0,53	1,23	1,71
Alta valoración estudiantil de los aportes de la orientación del plan para trabajar	0,168	0,148	1,132	0,265		1,281	1,18	-0,12	0,46	0,88	1,58
Modalidad * Reconocimiento estudiantil de los aportes de la orientación vocacional	-0,446	0,142	-3,141	0,003	**	9,866	0,64	-0,72	-0,17	0,48	0,85
Condición de sobreedad * Valoración estudiantil de los aportes de la orientación del plan para trabajar	-0,448	0,165	-2,723	0,010	**	7,415	0,64	-0,77	-0,13	0,46	0,88

Cuadro 3. Regresión logística sobre los objetivos educativos profesionales. Estudiantes del último año escuelas secundarias estatales. CABA. Año 2008

Diagrama 6. Modelo explicativo sobre los objetivos educativos profesionales



4.3. El análisis en diseños muestrales complejos

Tal como se explicó en la sección 3, el efecto de diseño sobre la varianza de una estimación se puede medir mediante el cociente de la varianza obtenida en el diseño complejo y la varianza en un “diseño naive” (Kleinbaum, 2002) bajo supuesto de muestreo aleatorio simple con tamaño muestral equivalente. Así, el efecto de diseño permite tener una idea aproximada de la precisión ganada o perdida con el diseño complejo con respecto al muestreo simple al azar. Por razones de espacio solo se presentan en esta sección los resultados referidos a la dimensión laboral.

En este caso se estimó un 25% de estudiantes con objetivos laborales profesionales con un efecto de diseño de 2,3. Al ser dicho efecto mayor a 1, se dedujo que la precisión perdida por haber conglomerado la muestra superó a la ganada por haber estratificado. La comparación se realizó sobre la base de la información presentada en el cuadro 4.

Fuente: Elaboración propia sobre la base de la encuesta a estudiantes secundarios 2008 (GOIyE/ ME/ GCABA).

	p	Efecto de diseño (ED)	Varianza en muestra compleja	Limite superior del intervalo en diseño complejo
			$V [p] = ED V(p)$	$\lim sup_{(complejo)} = p + 1,44 \sqrt{V [p]}$
Bajo supuesto de muestreo aleatorio simple	0,25	1,0	0,00006	0,260
Considerando el diseño complejo	0,25	2,3	0,00013	0,265

Cuadro 4. Estimación de la proporción de estudiantes con objetivos profesionales laborales. Efecto de diseño, varianza y límite superior del intervalo de estimación bajo supuesto de

muestreo aleatorio simple y al considerar el diseño complejo

En la estimación que consideraba la complejidad del diseño se obtuvo una varianza levemente mayor a la calculada bajo supuesto de muestreo aleatorio simple para un tamaño muestral equivalente. Esto se tradujo en un intervalo de estimación más amplio y menor precisión. De ahí que se concluya la importancia de la consideración de la complejidad del diseño muestral en el análisis de resultados. Se trata de pequeñas diferencias que cobran importancia en la contrastación de hipótesis de relaciones y en la modelización.

En el cuadro 5 se observa cómo todos los p-valores obtenidos al realizar las pruebas de hipótesis resultaron más pequeños bajo supuesto de muestreo aleatorio simple que al considerar la estructura compleja del diseño. De esto se desprende que el alejamiento de la independencia estadística se exagera al obviar el peso de la conglomeración de diseño. A modo de ejemplo, al evaluar las relaciones con los objetivos laborales profesionales, se vio que el turno de asistencia a las clases, el tamaño de la escuela, el nivel de valoración estudiantil de los aprendizajes escolares y la condición de sobreedad fueron las variables que, al obviar el diseño en el análisis, rechazaron las hipótesis nulas de independencia estadística con niveles de significancia de 0,1% (***). Al considerar el diseño, en cambio, las significancias se ubicaban en el orden del 5% (*). En cuanto al nivel de valoración estudiantil de los aportes de la orientación cursada para trabajar, se observó una disparidad menos pronunciada al efectuar las pruebas de hipótesis en el plano laboral.

Estos resultados muestran que las diferencias surgidas al considerar y al ignorar el diseño muestral hubieran sido notorias de haberse decidido incorporar a los modelos multivariados sólo aquellas variables que rechazaran las hipótesis nulas con niveles de significación más exigentes (al 1% por ejemplo). Bajo hipótesis de muestreo aleatorio simple hubieran entrado 8 variables, mientras que considerando la complejidad de diseño, hubieran sido 4 las variables introducidas.

Fuente: Elaboración propia sobre la base de la encuesta a estudiantes secundarios 2008 (GOIyE/ ME/ GCABA).

	Bajo supuesto de muestreo aleatorio simple		Considerando el diseño complejo	
	p-valor		p-valor	
Zona	0,094	-	0,283	
Modalidad	0,000	***	0,000	***
Turno	0,000	***	0,039	*
Nivel de abandono escolar	0,134		0,470	
Tamaño de la escuela	0,000	***	0,034	*
Nivel de valoración estudiantil con respecto a los aportes de la orientación cursada para trabajar	0,000	***	0,007	**
Nivel de valoración estudiantil de los aprendizajes escolares	0,000	***	0,018	*
Nivel de valoración estudiantil de actividades de orientación vocacional	-	-	-	-
Nivel educativo materno	0,000	***	0,000	***
Sexo	0,000	***	0,000	***
Experiencia laboral	0,159		0,357	
Sobreedad	0,000	***	0,014	*

Cuadro 5. Pruebas de independencia estadística de las variables explicativas con el objetivo profesional laboral. Comparación entre ignorar y considerar el diseño complejo.

Códigos de significación: "****" 0.001 "***" 0.01 "**" 0.05 "." 0.1 " " 1

Coeficientes	Bajo supuesto de muestreo aleatorio simple					Considerando el diseño complejo				
	B	Error estándar	Test Wald			B	Error estándar	Test Wald		
			t	p-valor				t	p-valor	
(Intercept)	-1,454	0,089	264,640	0,0000	***	-1,454	0,125	-11,675	0,0000	***
Modalidad bachiller-comercial	0,309	0,064	23,445	0,0000	***	0,309	0,090	3,443	0,0014	**
Turno diurno	-0,303	0,092	10,763	0,0010	**	-0,303	0,154	-1,966	0,0566	.
Alta valoración estudiantil de la orientación del plan	-0,242	0,057	18,143	0,0000	***	-0,242	0,097	-2,489	0,0173	*
Madre con secundario completo o más	0,203	0,052	14,928	0,0001	***	0,203	0,077	2,630	0,0123	*
Mujer	0,349	0,052	44,326	0,0000	***	0,349	0,108	3,227	0,0026	**
Sin Sobreedad	-0,411	0,137	9,033	0,0027	**	-0,411	0,170	-2,420	0,0204	*
Turno * Condición de sobreedad	0,725	0,148	23,892	0,0000	***	0,725	0,194	3,736	0,0006	***

Fuente: Elaboración propia sobre la base de la encuesta a estudiantes secundarios 2008 (GOlyE/ ME/ GCABA).

Cuadro 6. Comparación de regresiones sobre los objetivos laborales profesionales al ignorar y al considerar el diseño complejo

Por último, se comparó el modelo final al utilizar los estadísticos apropiados para la consideración de la complejidad de diseño con el modelo “naive” (ver cuadro 6). Si bien los coeficientes Beta estimados resultaron coincidentes en ambas versiones, se observaron diferencias en las pruebas de significancia estadística de los términos de la ecuación de regresión. Al ignorar el diseño se obtuvieron Walds más grandes, obteniéndose p-valores más pequeños que condujeron a rechazos más fuertes de las hipótesis nulas. Por lo tanto, con excepción de la interacción turno*sobreedad, todos los términos de la ecuación de regresión se situaron en distintos “umbrales” de significancia, dependiendo de que el diseño complejo fuera ignorado o considerado.

De todo lo analizado se desprende que la consideración de la complejidad del diseño muestral es un aspecto de suma importancia a considerar en la exploración bivariada y multivariada. Se mostró cómo las diferencias detectadas en la evaluación de las relaciones bivariadas con la variable dependiente de interés pueden dar lugar a distintas preselecciones de variables a incorporar en modelos de mayor complejidad, fundamentalmente cuando las significancias estadísticas van traspasando los “umbrales” más habitualmente utilizados (0,1%, 1% y 5%). Por otra parte, de la comparación de los resultados finales de las regresiones también se deduce que las distorsiones provocadas al ignorar el diseño pueden dar lugar a diferentes recorridos en la búsqueda de modelos más parsimoniosos.

En este artículo se presentaron algunos resultados de una tesis que analizó las orientaciones de futuro laboral y educativo de los estudiantes que se hallaban cursando el último año en escuelas secundarias estatales de la Ciudad Autónoma de Buenos Aires. La fuente de datos fue una encuesta realizada por el Ministerio de Educación de dicha jurisdicción a fines de 2008 a los alumnos de una muestra de escuelas-modalidades-turnos.

Las generalizadas expectativas de continuidad educativa postsecundaria y adquisición de otras titulaciones en los itinerarios imaginados por los estudiantes se sitúan en un escenario socio-histórico signado por la obligatoriedad del nivel secundario y la progresiva devaluación de sus credenciales en el mercado de trabajo. En primer lugar, resultaron destacados los *proyectos educativo-laborales* (Guichard, 1995) y los jóvenes se mostraron bastante realistas a la hora de anticipar posibles obstáculos. La competencia y la falta de oportunidades estuvieron entre las dificultades más mencionadas al pensar en un objetivo laboral. En el plano educativo, en cambio, se visualizaron menos dificultades, aunque quedando resaltado el temor a no contar con suficiente tiempo para estudiar.

En cuanto a los resultados sustantivos que se desprenden del análisis multivariado, puede decirse que la diferencia detectada entre varones y mujeres en cuanto a la centralidad de los planes profesionales, resultó decisiva tanto en el plano laboral como en el educativo. Fueron las mujeres quienes más fuertemente priorizaron las metas profesionales, lo cual condice con las particularidades halladas en Argentina en cuanto a la gran presencia femenina en la educación superior y su mayor movilidad educacional con respecto a los varones (Jorrat, 2010). En ese sentido, se vio cómo en el plano de las expectativas de los actores aparecía reflejado y delineado un fenómeno social de mayor alcance. La modalidad del plan de estudios también emergió como un aspecto clave en los horizontes de futuro, observándose un hiato entre la formación bachiller o comercial y la técnica. Es así como entre los estudiantes de escuelas técnicas, los planes profesionales fueron menos frecuentes, prevaleciendo entre ellos otras expectativas de inserción laboral más directa desde la secundaria (no es que estos jóvenes descartaran metas profesionales, sino que depositaban más expectativas en oportunidades laborales más inmediatas).

Por otra parte, el origen educacional adquirió fuerza explicativa en el plano laboral -no así en el educativo- lo cual evidencia cómo en la planificación del futuro laboral se expresan con más fuerza aquellos condicionamientos socioeconómicos que orientan a los jóvenes de origen social más bajo hacia metas de inserción laboral y obtención de ingresos antes que de profesionalización. En contrapartida, en el plano educativo los horizontes profesionales tuvieron una difusión amplia e independiente del origen social. La visualización de la educación como terreno de mayor despliegue de posibilidades futuras fue un resultado reiterado. Si bien el desencanto de no hallar correspondencia entre la titulación y las recompensas materiales y simbólicas (Tenti, 2000), podía instar a los jóvenes de origen social más bajo a una ampliación de sus aspiraciones educativas con vistas a mejores oportunidades, es claro que ciertas condiciones contextuales -como una oferta variada, gratuita y sin restricciones formales de acceso-

habilitaban este tipo de opciones a los estudiantes a punto de finalizar la secundaria, sin distinción de orígenes sociales.

Otro eje de análisis se vinculó con el tratamiento estadístico de los datos. En este sentido, se ha argumentado acerca de la importancia de considerar la complejidad del diseño muestral en la instancia de análisis de datos, fundamentalmente cuando se trata de evaluar la existencia de relaciones entre variables, tanto en los análisis bivariados como multivariados. Para mostrar la importancia de considerar aquellos criterios puestos en juego al diseñar y seleccionar la muestra, se compararon los resultados que contemplaban el diseño muestral complejo con otros obtenidos en "diseños naive" (Kleinbaum, 2002) que operaban bajo supuesto de selección aleatoria simple. Se planteó entonces que los efectos de ignorar la estructura de correlaciones adquieren importancia al comparar resultados en las pruebas de hipótesis y conlleva riesgos epistemológicos de enunciación de falsas relaciones.

Bibliografía

- AGRESTI, A. y FINLAY, B. (1997). *Statistical Methods for the Social Sciences* (tercera edición). New Jersey: Prentice Hall.
- ALVAREZ, M. y FERNÁNDEZ, A. (2011). Movilidad ocupacional de los jóvenes en la Argentina durante la postconvertibilidad. Ponencia presentada en 10° Congreso Nacional de Estudios del Trabajo. Buenos Aires. 3 al 5 de agosto de 2011.
- BERGER, Y. (2003). *A Simple Variance Estimator for Unequal Probability Sampling Without Replacement*. Southampton: Social Statistics Research Centre.
- BOURDIEU, P. (2006). *Argelia 60: Estructuras económicas y estructuras temporales*. Buenos Aires: Siglo XXI Editores.
- CASTEL, R. (1997). *La metamorfosis de la cuestión social*. Buenos Aires: Paidós.
- CERVINI, R. (2003). "Relaciones entre composición estudiantil, proceso escolar y el logro en matemáticas en la educación secundaria en Argentina". En *Revista Electrónica de Investigación Educativa*, Vol. 5(1), 1-27.
- CORICA, M. (2010). *Lo posible y lo deseable. Expectativas laborales de jóvenes de la escuela secundaria*. Tesis de maestría publicada, Facultad Latinoamericana de Ciencias Sociales, Buenos Aires.
- DABENIGNO, V., AUSTRAL, R., GOLDENSTEIN JALIF, Y., IÑIGO, L. y SKOUMAL, G. (2007). *Imágenes de futuros laborales. Horizontes sociales y personales de jóvenes escolarizados en el nivel medio de la Ciudad de Buenos Aires*. Ponencia presentada en IV Congreso Nacional y II Internacional de Investigación Educativa. Comahue. 18 al 20 de abril de 2007.
- DIRECCIÓN DE INVESTIGACIÓN Y ESTADÍSTICA (2008). *Diagnóstico socioeducativo de la zona sur. Primer informe de avance (documento interno)*. Buenos Aires: Dirección de Investigación y Estadística/ ME/ GCABA.
- DUBET, F. y MARTUCELLI, D. (2000). *¿En qué sociedad vivimos?* Buenos Aires: Losada.
- FALCO DE JOUAS, R. (2003) *El proyecto educativo-laboral. Un desafío a la orientación universitaria*. Ponencia presentada en Congreso Latinoamericano de Educación Superior en el

Siglo XXI. San Luis. 18, 19 y 20 de septiembre de 2003.

FIEULAIN, N. (2006). *Perspective temporelle, situations de précarité et santé: une approche psychosociale du temps*. Tesis de doctorado. Marseille: Université de Provence.

FILMUS, D. y MORAGUES, M. (2003). "¿Para qué universalizar la educación media?" en *Educación media para todos. Los desafíos de la democratización del acceso*. Tenti Fanfani E. (Comp.). Buenos Aires: Grupo Editor Altamira.

GJESME, T. (1981). "Is there any future in achievement motivation?" en *Motivation and Emotion*, Vol. 2, 115-138.

GUICHARD, J. (1995). *La escuela y las representaciones de futuro de los adolescentes*. Barcelona: Laertes.

HARGREAVES, A. (2000). "Four ages of professionalism and professional learning". En *Teachers and Teaching: History and Practice*, 6(2):151-182.

HAUSER, R. y ANDERSON, D. (1991). "Post-high school plans and aspirations of black and white high school seniors: 1976-86". En *Sociology of Education*, Vol. 64(October), 263-277.

HEERINGA, S., WEST, B. y BERGLUND, P. (2010) *Applied Survey Data Analysis*. Boca Raton, London, New York: Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences.

HOSSLER, D. y STAGE, F. (1992). "Family and High School Experience Influences on the Postsecondary Educational Plans of Ninth-Grade Students". En *American Educational Research Journal*, Vol. 29(2), 425-451.

HUSMAN, J. y Lens, W. (1999). "The role of the future in student motivation". En *Educational Psychologist*, Vol. 34(2), 113-125.

JORRAT, R. (2010). "Logros educativos y movilidad educativa intergeneracional en Argentina". En *Desarrollo Económico*, 49(196): 573-604.

KANDEL, D. y LESSER, G. (1969). "Parental and Peer Influences on Educational Plans of Adolescents". En *American Sociological Review*, Vol. 34(2), 213-223.

KISILEVSKY, M. (2002). *Condiciones sociales y pedagógicas de ingreso a la educación superior en la Argentina. Dos estudios sobre el acceso a la educación superior en la Argentina*. Buenos Aires: IIPE - UNESCO.

KLEINBAUM, D. y Klein, M. (2002). *Logistic Regression. A Self-Learning Text*. New York: Springer-Verlag.

LECHNER, N. (2002). *Las sombras del mañana. La dimensión subjetiva de la política*. Santiago de Chile: LOM Ediciones.

LOHR, S. (2000). *Muestreo: diseño y análisis*. México DF: Thompson Learning.

LONGO, M. (2007). *Anticiparse en el trabajo: el rol del futuro en las trayectorias profesionales de los jóvenes*. Ponencia presentada en 8vo Congreso Nacional de Estudios del Trabajo. Buenos Aires. 8 al 10 de agosto de 2007.

LUMLEY, T. (2011). *Survey: analysis of complex survey samples*. R package version 3.26.

LUMLEY, T. (2010). *Complex Surveys. A Guide to Analysis Using R*. New Jersey: Wiley.

McCABE, K. y Barnett D. (2000). "First Comes Work, Then Comes Marriage: Future Orientation Among African American Young Adolescents". En *Family Relations*, Vol. 49(1), 63-70.

MEO, A. y DABENIGNO, V. (2008). "Los adolescentes y sus visiones de futuro: una primera aproximación a las expectativas educativas en sectores populares de la ciudad de Buenos Aires" en *Cambios epocales y transformaciones en el sistema de educación superior: la universidad*

argentina y los nuevos desafíos. Iriarte, A. (Ed.). Buenos Aires: Teseo.

MIRANDA, A. y ZELARAYAN, J. (2011). La situación de los jóvenes en el mercado de trabajo en la Argentina postconvertibilidad. Ponencia presentada en 10º Congreso Nacional de Estudios del Trabajo. Buenos Aires. 3 al 5 de agosto de 2011.

NICOLE-DRANCOURT, C. (1994). "Mesurer l'insertion professionnelle". En *Revue de sociologie française*, Vol. 35(1), 37-68.

NOVICK, M. . (2006). "¿Emerge un nuevo modelo económico y social? El caso argentino 2003-2006". En *Revista Latinoamericana de Estudios del Trabajo*, Vol. Año 11(18), 53-78.

NURMI, J.E. (1991) "How do adolescents see their future? A review of the development of future orientation and planning". En *Developmental Review*, Vol. 11, 1-59.

PROGRAMA NACIONAL UNIDAS PARA EL DESARROLLO (2009). Abandono escolar y políticas de inclusión en la educación secundaria. Buenos Aires: PNUD.

R DEVELOPMENT CORE TEAM (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna.

ROSANVALLON, P. (1995). La nueva cuestión social. Repensar el Estado providencia. Buenos Aires: Ediciones Manantial.

SENNET, R. (1998). La corrosión del carácter. Las consecuencias personales del trabajo en el nuevo capitalismo. Barcelona: Anagrama.

STEINBERG, C. y TÓFALO, A. (2013). "Aportes para examinar las desigualdades educativas en los grandes centros urbanos. El uso del coeficiente de Gini para analizar la distribución del abandono escolar entre las escuelas secundarias de la Ciudad Autónoma de Buenos Aires". En *Revista Latinoamericana de Metodología de las Ciencias Sociales*, Vol. 3 (1).

STEINBERG, C., CETRÁNGOLO, O. y GATTO, F. (2011). Desigualdades territoriales en la Argentina. Insumos para el planeamiento estratégico del sector educativo. Santiago de Chile: CEPAL.

TENTI FANFANI, E. (2000). Culturas juveniles y cultura escolar. Ponencia presentada en Seminario "Escola Jovem: um novo olhar sobre o ensino médio". Brasília. 7 al 9 de junio de 2000.

TROMMSDORFF, G. (1983). "Future orientation and socialization". En *International Journal of Psychology*, Vol. 18, 381-406.

WELLER, J. (2006). "Inserción laboral de jóvenes: expectativas, demanda laboral y trayectorias" en *Estrategias educativas y formativas para la inserción social y productiva*. Jacinto, De Ibarrola, Girardo, Mochi (Coords.). Uruguay: OIT/ Cinterfor, UNESCO, IIPE, RedEtis.

ARTÍCULOS

“Condiciones de socialización, entorno y trayectoria asociados a la reincidencia en el delito”.¹

Una aproximación explicativa utilizando un modelo multivariado de análisis

*Marcelo Bergman*² / *Diego Masello*² / *Christian Arias*⁴ / *Guadalupe Peralta Agüero*⁵

Introducción

Los sistemas penitenciarios Bonaerense y Federal son los más importantes de la República Argentina. En la actualidad, dentro de los mismos se encuentran alrededor de 38.000 presos (procesados o condenados), que representan al 60% del total de los 63.000 presos existentes en la Argentina. El presente trabajo analiza en profundidad algunas relaciones entre ciertas características de socialización de las personas condenadas y algunos aspectos importantes de la conducta delictiva. La base empírica del análisis está basada en una muestra probabilística de 1.033 casos seleccionados aleatoriamente a partir del sorteo de las unidades penitenciarias en una primera etapa y luego con el sorteo de los internos alojados dentro de las mismas.

Es sabido entre los expertos que estudian temas de criminología que dentro de las cárceles la población es mayoritariamente joven, con bajos niveles de educación y provenientes de clases socioeconómicas medias/bajas y bajas, caracterizadas, entre otras cosas, por los bajos niveles de ingreso que percibe.

Teniendo en cuenta estas características se podría pensar que la vinculación explicativa de la conducta delictiva estaría dada por la asociación entre la condición de pobreza de un hogar y/o de sus integrantes y las probabilidades de comisión de delitos. Sin embargo, este trabajo propone una relación mucho más compleja, donde las condiciones materiales básicas, si bien son un punto de partida para mejorar las condiciones de vida de la población en general y para reducir el delito,

necesitan de otros factores complementarios que tendrían mucha incidencia para explicar por qué una persona puede tener más o menos probabilidades de incurrir en una conducta delictiva.

En este sentido, prestando atención a otros elementos como los contextos de socialización temprana, la trayectoria en instituciones de rehabilitación –como los institutos de menores–, el hecho de tener padres que hayan pasado por la experiencia carcelaria; cabe preguntarse en relación a la reincidencia en el delito ¿qué tan asociados están esos factores a quienes incurren nuevamente en una conducta delictiva y al nivel de violencia al momento de perpetrar un delito?, finalmente, ¿qué factores son estos? y ¿en qué orden de importancia se manifiestan?

En este punto, es importante destacar que el presente estudio no ofrece una explicación acabada del impacto sobre la tasa de reincidencia total, ya que debido a la naturaleza de la fuente de los datos (encuestas a población carcelaria) únicamente es posible realizar estimaciones en torno a la población de reincidentes, comparando por la población de no reincidentes hasta el momento, lo que no implica que jamás vayan a serlo. Sin embargo, los datos son ilustrativos de las características de la población reincidente y permiten establecer vinculaciones entre los factores asociados a ella.

Finalmente, cabe señalar que este trabajo tuvo como meta la elaboración de algunas respuestas para estos interrogantes, de modo de elucidar una parte de la complejidad que representa la acción humana y, de este modo, contribuir a diagnósticos más completos sobre qué dimensiones se pone en juego detrás de la acción delictiva.

Metodología

Para elaborar una posible respuesta a este conjunto de interrogantes se llevó a cabo un modelo de regresión logística, en el cual se incorporaron variables relacionadas con los entornos o contextos de socialización temprana y la trayectoria en instituciones como los institutos de menores. Dichas variables se incorporaron como elementos independientes. Los datos utilizados pertenecen a la Encuesta a Población en Reclusión de 2013, en el cual para Argentina se aplicaron más de mil encuestas personales a presos condenados por la justicia federal y ordinaria de la Capital así como por la justicia de la Provincia de Buenos Aires.

La regresión logística se desarrolló utilizando el modelo por pasos (FowardWald), lo que permitió observar el comportamiento de las distintas variables durante el proceso.

Inicialmente se incorporaron numerosas variables al modelo:

- Edad en años cumplidos
- Nivel de violencia en el hogar cuando era niño/adolescente
- Entorno delictivo (si tuvieron familiares cercanos presos)
- Relación del entorno familiar con drogas y alcohol

- Nivel Educativo Alcanzado del entrevistado (NEA)
- Trayectoria en institutos de menores
- Familiaridad con armas (haber tenido/usado armas de fuego)
- Entorno barrial (bandas en el barrio, peleas en el barrio, accionar de los vecinos, accionar de la policía)

La variable dependiente en el modelo fue “si había reincidido en el delito” (Si/No)

Luego de una iteración de cuatro pasos, el modelo quedó ajustado con las siguientes variables:

Variable	Indicadores	Categorías	Preguntas
Entorno delictivo	1 Familiares presos	ū Con entorno delictivo (estuvo detenido algún familiar o amigo cercano)	-¿Quién de sus familiares estuvo preso alguna vez?
	2 Mejores amigos presos	ū Sin entorno delictivo (sin detenidos o presos dentro de su entorno)	-Antes de que Usted cumpliera la mayoría de edad y sin decirme quien ni sus nombres ¿sabe si alguno de sus mejores amigos cometían delitos, aunque fuera uno, o varios muy de vez en cuando?
Familiaridad con armas	3 Tenencia de armas en sus manos alguna vez en su vida	ū Tuvo arma en sus manos ū No tuvo armas en sus manos.	-¿Alguna vez tuvo usted en sus manos un arma de fuego?
Nivel Educativo Alcanzado (NEA)	4 Máximo nivel educativo alcanzado	ū Nivel bajo (hasta primaria completa) ū Nivel alto (secundaria incompleta y más)	-¿Cuál fue el último grado alcanzado antes de ingresar a la cárcel por primera vez?
Relación del entorno familiar con drogas y alcohol	5 Ingesta de drogas por parte de algunos de los adultos con los que convivía.	ū Familiaridad con drogas y alcohol. (al menos uno de los adultos con los que convivió consumía drogas y/o alcohol frecuentemente)	-¿Alguno de sus padres o de los adultos con los que vivía cuando era chico(a) tomaba(n) alcohol frecuentemente? /
	6 Ingesta de alcohol por parte de algunos de los adultos con los que convivía.	ū Sin familiaridad con drogas y alcohol (ninguno en su hogar consumía drogas y/o alcohol)	-¿Alguno de sus padres o de los adultos con los que vivía cuando era chico(a) consumía(n) drogas?

Variable Dependiente

Reincidencia en el delito	Reincidencia en el delito	1Si	¿Y estuvo preso en una cárcel de adultos?
		2No	

Marco conceptual del análisis

Dentro de los estudios sobre el delito se han analizado en varios países las relaciones entre ciertas características del entorno social y las conductas delictivas, arribándose a conclusiones importantes sobre la influencia de los entornos sociales delictivos (con padres, familiares y amigos presos) en las conductas de los sujetos que viven dentro de dichos entornos. (Rosenberg, 2009, pp. 18-19; Oliver Robertson, 2007, 2012)

Siguiendo esta línea conceptual, en este trabajo se ha llevado adelante un análisis de los contextos de socialización temprana de los presos condenados para observar posibles relaciones de influencia en el ciclo de la delincuencia. Si bien el estudio carece de un grupo de control para tener un mejor escenario para la elaboración de hipótesis, el ser una población tan homogénea robustece las diferencias encontradas.

Complementariamente, no debe pensarse que los contextos de socialización se originan aisladamente, o sea, en forma endógena dentro del hogar sin conexiones claras con el contexto social más amplio. Por el contrario, todas estas variables se relacionan y funcionan concomitantemente en espacios geográficos que permiten su reproducción y potenciamiento. Por estos motivos, también se incorporaron numerosas variables, tanto relacionadas con el propio entorno familiar de los presos como con los barrios donde ellos vivían.

Antes de pasar a los resultados del modelo multivariado conviene aclarar algunas cuestiones relativas a las características de la población analizada. En primer lugar, como se mencionó anteriormente, se trata de una población altamente homogénea puesto que todos son presos condenados, lo que supone un igualamiento a priori en algunos atributos que permiten de mejor manera apreciar la robustez de la influencia de las variables independientes. Por ejemplo, si las mismas relaciones respecto a la influencia del contexto de socialización se hicieran tomando a la población general, las diferencias significativas se apreciarían con mayor facilidad. Es decir, estamos tratando de discriminar propiedades dentro de un conjunto que se presenta con un grado de homogeneidad elevado.

En segundo lugar, además de su condición de presos condenados y, quizás por dicha condición, esta población es muy homogénea también en cuanto a la extracción socioeconómica a la que pertenece así como a su inserción dentro de la estructura socio-productiva. Por lo tanto, la dimensión económica está fuertemente controlada por esta homogeneidad de la propia población analizada.

Por ejemplo, la población de condenados observada en este trabajo muestra que casi la totalidad ha trabajado alguna vez en su vida (92%), asimismo, muchas de estas personas (70%) continuaba trabajando hasta seis meses antes de ser detenido por el delito cometido. La mayoría tenía un trabajo de baja calificación, contando más de un 30% de ellos con ingresos iguales o menores a la línea de pobreza. Del mismo modo, más de dos tercios tenían una antigüedad de menos de un año

en dicho trabajo, lo que evidencia una relación laboral relativamente nueva y bastante cambiante, ya que una proporción superior al 25% declaró a estos trabajos como inestables.

Influencia de los contextos

Lo primero que conviene revisar es la fortaleza de las relaciones que se analizarán en el modelo multivariado, ya que una premisa para avanzar en este tipo de análisis más complejo es que las relaciones originales presenten asociación entre las variables.

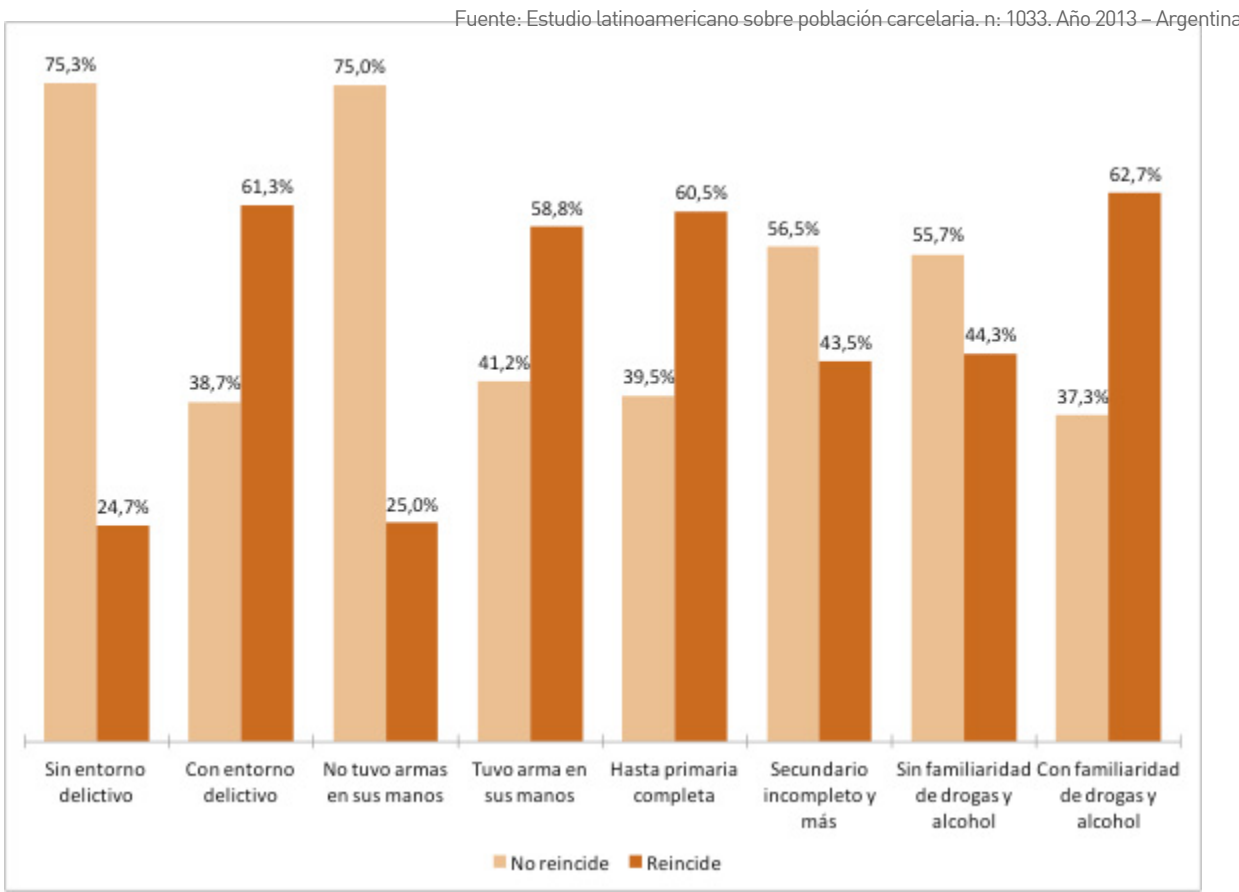


Gráfico 1. Recibió condena anteriormente según entorno delictivo, familiaridad con las armas, nivel educativo y familiaridad con las drogas

El gráfico N° 1 muestra la relación bivariada entre la reincidencia de los presos condenados (observando si habían pasado anteriormente por la cárcel) con algunas variables independientes. Como se puede observar, en todos los casos se evidencian diferencias porcentuales importantes, especialmente en cuanto a haber crecido dentro de un entorno delictivo y en la familiaridad con las armas en el hogar donde se socializaron.

De esta forma, el 61% de los que crecieron dentro de un entorno delictivo observaron una conducta de reincidencia en el delito, mientras que estos valores se reducen al 25% entre aquellos que no crecieron dentro de este tipo de entornos, evidenciándose una diferencia porcentual de 36 puntos, lo que podría significar una asociación entre las variables.

En el caso de la incidencia de la familiaridad en el uso de armas, se observa que el 59% de aquellos que tuvieron o utilizaron armas de fuego han tenido una conducta reincidente, mientras esta proporción disminuye al 25% entre aquellos que no las tuvieron o usaron. Diferencia de 34 puntos porcentuales.

Complementariamente, se le han aplicado a estas relaciones la prueba de χ^2 para comprobar la asociación así como los coeficientes Phi y V de Cramer para revisar la fortaleza de las mismas. Respecto a la prueba de χ^2 los resultados arrojaron valores de 108,46 para entorno delictivo y 81,38 para familiaridad con las armas con significaciones tendientes a cero, lo que implicaría intervalos de confianza superiores al 99% para ambas variables. En cuanto a la aplicación de los coeficientes:

Coeficientes entorno delictivo			
		Valor	Sig.
Para variables nominales	Phi	,325	,000
	Cramer's V	,325	,000
N casos válidos		1025	

Coeficientes familiaridad	
Para variables nominales	
N de casos válidos	

Se evidencia que los valores reflejan un nivel de fortaleza aceptable para la asociación de estos fenómenos, con un nivel de significación tendiente a cero.

Por lo tanto, en las relaciones bivariadas se comprueba la asociación entre el entorno delictivo, la familiaridad con la tenencia y/o uso de armas y la conducta de reincidencia en el delito. De todos modos, esta asociación es una condición necesaria pero no suficiente para pensar en una aproximación explicativa sobre la reincidencia delictiva. Para ello es necesario involucrar otro conjunto de atributos y comprobar qué pasa con la influencia de estos aspectos bajo estas nuevas circunstancias.

Modelo multivariado de análisis

Para avanzar en la características de la reincidencia de los presos condenados, se desarrolló un procedimiento para ajustar un Modelo de Regresión Logística Binaria (RL) que vincula la

prevalencia en reincidencia en el delito (reincidente – no reincidente) de la población bajo estudio con el conjunto total de las variables independientes identificadas anteriormente. La idea inicial estuvo puesta en identificar, en primer lugar, cuáles de todas estas variables tenían un aporte teórico significativo dentro del modelo y cuáles no.

Para ello se introdujeron ocho variables independientes que se suponían relevantes para la explicación de la probabilidad de presentar una situación de reincidencia o no. Luego, el modelo fue ajustando en cuatro pasos la resolución teniendo en cuenta sólo cuatro variables:

		Variables en la ecuación final					
		B	S.E.	Wald	df	Sig.	Exp(B)
Paso final del modelo	Entorno Delictivo	1,263	,170	55,108	1	,000	3,538
	Familiaridad con las drogas	,508	,145	12,306	1	,000	1,662
	Nivel Educativo	-,726	,142	26,214	1	,000	,484
	Familiaridad con las armas	1,198	,186	41,329	1	,000	3,314
	Constant	-2,366	,217	118,559	1	,000	,094

Como se puede observar hay diferentes indicadores, algunos relacionados directamente con algunas características del hogar de los presos condenados cuando eran niños (entorno delictivo, consumo habitual de drogas dentro del hogar) y otras referidas a ciertas características y/o conductas subjetivas como el nivel educativo de las personas y si ha tenido o utilizado armas de fuego.

Inicialmente, se observa que el puntaje Wald (condición para ingresar dentro del modelo) presenta sus valores más elevados para el entorno delictivo y la familiaridad con el uso de armas.

Para el desarrollo del modelo se eligió el método por pasos hacia adelante (forward wald). En dicho procedimiento se van ingresando y eliminando variables en cada uno de los pasos hasta llegar a lo que sería una solución óptima en función de la iteraciones realizadas. Esta técnica de selección de variables es realizada en función del set inicial de atributos incluidos. De este modo, se partió de un modelo inicial, en el que se incluyeron las variables mencionadas anteriormente, que se consideraba que podían incidir y luego, en función de los pasos recorridos, se evalúa estadísticamente cuáles son las variables que menos participan en el modelo y se procede a la eliminación de las mismas. A partir de esta primera selección, se vuelve a aplicar la misma técnica, esta vez solo con las variables que sí inciden estadísticamente. Se hace nuevamente una selección y así sucesivamente hasta que se considera que el modelo obtenido es el que “mejor se ajusta” a las condiciones requeridas.

En este caso el modelo se ha desarrollado en cuatro pasos, siendo este último el que mejor se ajusta a la explicación de asociación con las categorías de la variable dependiente “Reincidencia en el delito”. A continuación se desarrollan los estadísticos del modelo final:

Pruebas omnibus sobre los coeficientes del modelo				
		Chi-square	df	Sig.
Paso 1	Step	113,089	1	,000
	Block	113,089	1	,000
	Model	113,089	1	,000
Paso 2	Step	39,602	1	,000
	Block	152,692	2	,000
	Model	152,692	2	,000
Paso 3	Step	32,692	1	,000
	Block	185,384	3	,000
	Model	185,384	3	,000
Paso 4	Step	12,389	1	,000
	Block	197,773	4	,000
	Model	197,773	4	,000

Resumen del modelo				
Paso	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square	
1	1298,744 ^a	,105	,140	
2	1259,142 ^a	,139	,186	
3	1226,450 ^a	,166	,222	
4	1214,061 ^a	,176	,235	

Como puede apreciarse, en cada uno de los pasos, con la introducción de un nuevo indicador, se robustece el coeficiente de χ^2 , asociándose a un adecuado nivel de significación tendiente a cero. Asimismo, el coeficiente R de Nagelkerke aumenta en cada paso hasta el valor de 0,235 que podría interpretarse como que el modelo explicaría aproximadamente un 20% de la varianza de la variable dependiente. Si se toma en cuenta que dicha variable es la condición de reincidencia, que es una variable sumamente compleja, dicho valor explicativo cobra importancia.

Ahora bien, de las variables ingresadas, las que finalmente quedaron dentro del modelo que refleja estos coeficientes de explicación fueron: Entorno Delictivo, Familiaridad con el uso de armas, Nivel educativo alcanzado, Familiaridad con drogas.

Bondad de Ajuste del modelo prueba de Hosmer y Lemeshow

Por otro lado, se utilizó el Test de Hosmer y Lemeshow para controlar el ajuste de las variables al modelo, para lo cual se plantean las siguientes hipótesis:

0

- H_0 : El modelo ajustado es significativo, se ajusta correctamente.

1

- H_1 : El modelo ajustado no es significativo, no se ajusta correctamente.

Según se observa la prueba de Hosmer and Lemeshow Test la significación es $> 0,05$, por lo tanto

0

se acepta la H_0 (nula), y se concluye que el modelo ajustado es significativo. Es decir, las variables introducidas se ajustan al modelo.

Hosmer and Lemeshow Test			
Paso	Chi-square	df	Sig.
1	,000	0	.
2	,051	2	,975
3	1,918	4	,751
4	4,108	6	,662

Interpretación de los parámetros Estimados del Modelo – ODDS Ratio

Para determinar la significación del coeficiente de regresión se utilizó el estadístico de Wald y el cociente odd (OR = Odd ratio). El estadístico de Wald sigue una distribución χ^2 ; en este caso para

todas las variables introducidas $p = 0.001$, lo cual es significativo, en las cuatro variables introducidas se rechazó H_0 como se observó en la tabla anterior.

Además de los coeficientes se obtiene información de Exp (B), que corresponde al Odds-Ratio asociado a cada factor. El ODDS Ratio cambia cuando la i -ésima variable explicativa regresora se incrementa en una unidad,

Si:

□ > 0 significa que el ODDS RATIO se incrementa.

□ < 0 significa que el ODDS RATIO decrece.

? $= 0$ significa que el factor es igual a uno, lo cual hace que ODDS RATIO no varía.

Cuando el coeficiente B de la variable es positivo se obtiene un odds ratio mayor que 1 y corresponde por tanto a un factor de riesgo o chance. Por el contrario, si B es negativo el odds ratio será menor que 1 y se trata de un factor de protección.

Según se desprende de la lectura de los resultados finales de las variables que finalmente quedaron dentro de la ecuación, la probabilidad de reincidir para cada una de ellas es de:

Exp (B)	Variable	Cuanto aumenta las chances si las demás se mantienen constantes
3,538	Entorno delictivo	La estimación de la chance de los que vivieron dentro de un entorno delictivo es 3 veces más elevada de aquellos que no vivieron en dicho entorno
3,314	Familiaridad con uso de armas	La estimación de la chance de los que se familiarizaron con el uso de armas de niños en el hogar es 3 veces más elevada de aquellos que no lo hicieron
0,484	Nivel educativo	La estimación de la chance de los que alcanzaron un bajo nivel educativo es cerca de 0,5 veces más elevada de aquellos que lograron mayores niveles de instrucción
1,662	Familiaridad con drogas	La estimación de la chance de los que vivieron dentro de un hogar donde se consumía alcohol y/o drogas es casi 2 veces más elevada de aquellos que no vivieron dichas situaciones

Finalmente, es importante aclarar que para utilizar el modelo es necesario que las variables independientes involucradas no estén correlacionadas entre ellas. Si la correlación entre dos variables es alta, entonces los resultados del modelo de regresión logística son poco confiables.

Primero, debe evitarse que en el modelo de regresión planteado pueda producirse el fenómeno de la colinealidad, porque daría lugar a soluciones inestables. Se habla de colinealidad cuando dos o más variables independientes que se introducen en el modelo de regresión están altamente correlacionadas entre sí.

Por lo tanto, a partir del análisis de las correlaciones que se presentan en la tabla siguiente, queda evidenciado que no existe colinealidad entre las variables introducidas como independientes.

Correlation Matrix							
		Constant	entornodelictivoVI	Constant	entornodelictivoVI	familiaridadarmasVI	neabis
Step 1	Constant	1,000	-.881				
	entornodelictivoVI	-.881	1,000				
Step 2	Constant			1,000	-.532	-.654	
	entornodelictivoVI			-.532	1,000	-.189	
	familiaridadarmasVI			-.654	-.189	1,000	
Step 3	Constant			1,000	-.504	-.591	-.256
	entornodelictivoVI			-.504	1,000	-.180	-.043
	neabis			-.256	-.043	-.138	1,000
	familiaridadarmasVI			-.591	-.180	1,000	-.138
Step 4	Constant			1,000	-.467	-.571	-.272
	entornodelictivoVI			-.467	1,000	-.182	-.051
	familiaridadarmasVI			-.261	-.065	1,000	-.110
	neabis			-.272	-.051	-.138	1,000
	familiaridadarmasVI			-.571	-.182	1,000	-.138

A continuación, en la última tabla, se muestran los casos observados de reincidentes. En esta tabla se cruzan los valores estimados y observados de esta situación calculando el porcentaje de coincidencias; el mismo se aproxima a un 66,9%, lo que representa una buena proporción de imputación de casos en el total. Del mismo modo, cabe señalar que el modelo tiene una mejor capacidad predictiva tanto para aquellos casos que son reincidentes en el delito (83,5%) así como para aquellos que no lo son (49,5%).

Tabla de clasificación

Observados			Predecidos		
			Reincidencia en el delito		Porcentaje de ajuste
			0 No reincide	1 Reincide	
Paso 4	Reincidencia en el delito	0 No reincide	246	251	49,5
		1 Reincide	86	436	83,5
	Porcentaje general				66,9

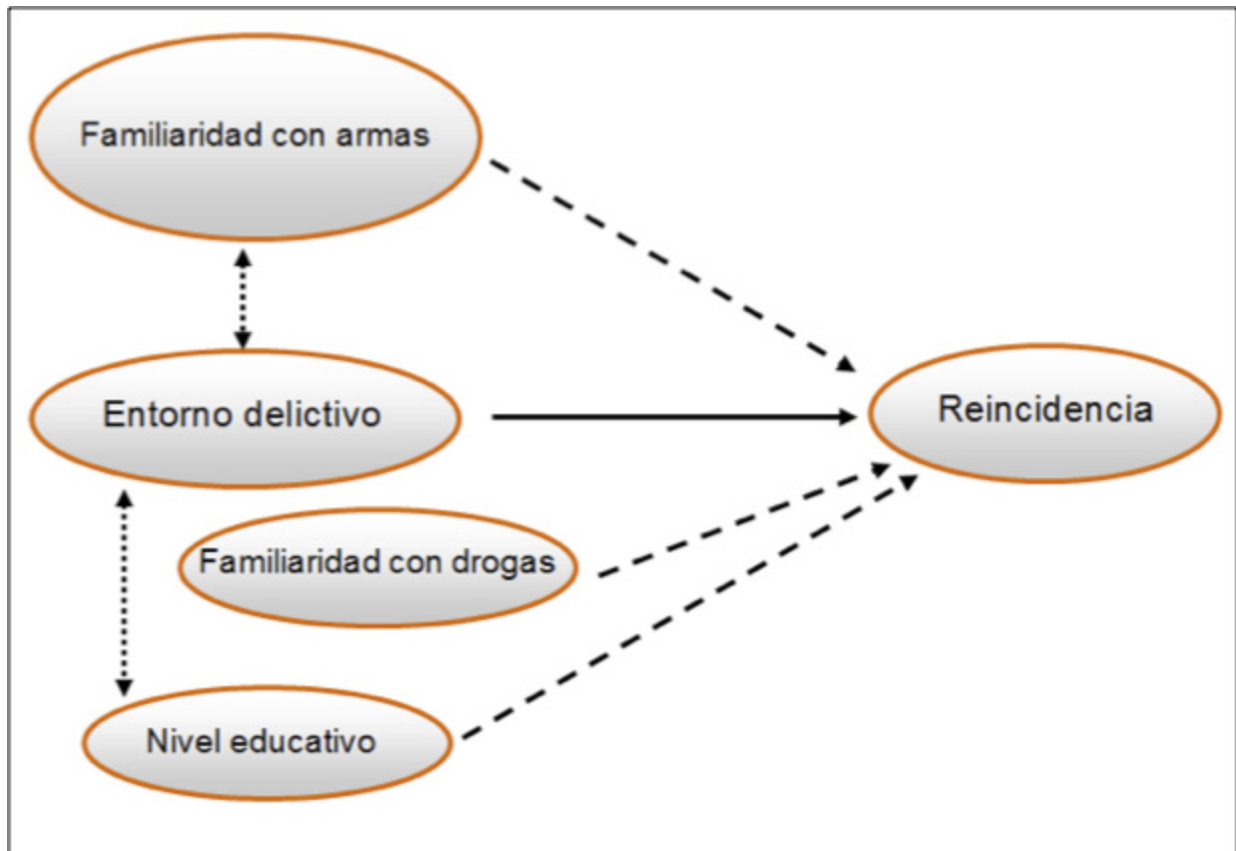
Conclusiones

Con el modelo obtenido se alcanza un porcentaje de coincidencia entre la estimación de probabilidad de reincidir en el delito y el valor observado de un 66,9%, con un buen nivel de estimación tanto para la categoría de reincidentes como para los que no lo son.

En términos generales, crecer en un hogar donde se observa cierta familiaridad (entendida como un consumo frecuente) con el alcohol y las drogas, una familiaridad con el uso de armas de fuego y, especialmente, que familiares y/o amigos hayan pasado por la cárcel, están asociados en la conformación de un esquema de socialización que tiene impacto positivo con la probabilidad de tener conductas de reincidencia delictiva.

A estos aspectos que son propios de la forma en que se estructura el hogar hay que adicionarle las características y logros educativos de los sujetos que viven en dichos hogares así como las circunstancias del entorno más amplio como el barrio donde crece y las características de los vecinos con los entabla relaciones.

En términos gráficos, para tener una mejor comprensión del proceso de relaciones, se presenta el siguiente esquema:



Con el esquema queremos significar de un modo sintético la complejidad de las relaciones que se establecen entre las variables independientes y el fenómeno a explicar: la probabilidad de reincidencia en el delito.

El aspecto que mayor fuerza explicativa tiene es la presencia del entorno delictivo en el hogar donde el sujeto vivió cuando era niño y, aunque no se establezca una causalidad directa y unívoca, la presencia de familiares y/o amigos con pasado carcelario influyen fuertemente en la probabilidad de reincidencia en el delito.

Finalmente, debe tenerse en cuenta que una variable puede tener valor predictivo aunque no sea parte del mecanismo causal que produce el fenómeno en estudio. Por lo tanto, la aplicación de este modelo se ha concentrado en estimar la contribución de los distintos factores mencionados sobre la reincidencia. Además, es recomendable tener especial precaución con los términos “relación”, “correlación” o “causalidad”. Que dos factores estén relacionados no implica de ninguna manera que uno sea causa del otro; en particular porque todas forman en su conjunto una compleja relación.

Bibliografía

Agnew, Robert, (1992), Foundation for a general strain theory of crime and delinquency, en Criminology vol. 30, pp. 47-88.

Bergman, Marcelo; Masello, Diego; Arias, Christian, (2014), Delito, marginalidad y desempeño institucional en la Argentina: resultados de la encuesta de presos condenados, Universidad Nacional de Tres de Febrero, Buenos Aires, Argentina.

Camacho Rosales, Juan, (2001), Estadística con SPSS para Windows, Alfaomega Grupo Editor, México D.F.

García Ferrando, Manuel, (1995), Socioestadística. Introducción a la estadística en sociología, Alianza Editorial, Madrid.

Robertson, Oliver, (2012), Convictos colaterales: niños y niñas de progenitores presos, en Publicaciones sobre los refugiados y los derechos humanos, Quaker United Nations Office, Ginebra, Suiza.

Robertson, Oliver, (2007), El impacto que el encarcelamiento de un(a) progenitor(a) tiene sobre sus hijos, en Serie: Mujeres en la cárcel e hijos de madres encarceladas, Quaker United Nations Office, Ginebra, Suiza.

Rosemberg, Jennifer, (2008), La niñez también necesita de su papá: hijos e hijas de padres encarcelados, en Publicaciones sobre los refugiados y los derechos humanos, Quaker United Nations Office, Ginebra, Suiza.

ARTÍCULOS

Un ejemplo de diseño cuasi experimental: uso de la regresión logística binaria en la construcción de un grupo de comparación para la evaluación de impacto de un programa social.

Una aproximación explicativa utilizando un modelo multivariado de análisis

Horacio Chitarroni ¹

grupo de comparación

diseño experiment

Este artículo trata acerca del empleo de la regresión logística binaria para la construcción de un grupo de comparación útil para la evaluación de impacto de un programa social. Se basa en una experiencia de aplicación real de tal procedimiento llevada a cabo por el autor.

En la primera parte se aborda brevemente la problemática que plantea la implementación de diseños puramente experimentales en el caso de la evaluación de políticas públicas de contenido social y la alternativa de emplear modelos cuasi experimentales con un grupo de comparación construido estadísticamente. También se ponen en consideración algunas cuestiones inherentes a los diseños con doble medición, al tiempo que se abordan las dificultades que plantea la frecuente ausencia de una línea de base en el caso de los programas sociales. Asimismo, se explicitan los requisitos que debieran cumplimentar los grupos de comparación contruidos mediante modelación estadística.

La segunda parte se refiere a las características del procedimiento estadístico empleado (la regresión logística binaria) y su utilidad específica para la obtención de grupos de comparación, con las limitaciones e inconvenientes que plantea, las alternativas posibles para sortearlos y los

recaudos a adoptar.

En la tercera parte se exponen los resultados provenientes de un ejemplo de aplicación conjuntamente con la interpretación de los mismos. Por fin, se enuncian unas breves conclusiones.

1. Las evaluaciones de impacto y los diseños experimentales

Tal como lo señala Baker (2000: 2) “En general se considera que los diseños experimentales, conocidos también como aleatorización, son las metodologías de evaluación más sólidas”.

La situación ideal del diseño experimental, tal como la describen habitualmente los textos, es aquella en que el grupo de control es una muestra extraída de la misma población de la cual proviene el grupo de tratamiento. Es más: la idea es que ambos grupos resulten de la partición al azar de una misma muestra, antes de que comience a operar el estímulo (Heirinch, Maffioli y Vázquez, 2010). Cuando ello es así, el azar nos otorga “garantías” de que ambos grupos no diferirán significativamente en nada. Están controlados los factores conocidos y también los desconocidos por el investigador. Tal como lo ha dicho Blalock (1971: 29):

En la práctica, confiamos en que las leyes de probabilidad produzcan distribuciones similares de todos los factores personales que intervienen en el experimento (...) La principal ventaja de la aleatorización es que toma en cuenta numerosos factores en forma simultánea sin que nos veamos obligados a saber cuáles son.

En ocasiones, también se recurre al sistema denominado *matching* o apareamiento: en este caso pueden formarse, a partir de la muestra original, pares de casos que se asemejen en los valores de un conjunto de variables que se consideren cruciales en la investigación. Por ejemplo, sujetos de igual edad, sexo, nivel educativo y situación ocupacional. Hecho esto, cada miembro de cada par así conformado sería adjudicado en forma aleatoria al grupo experimental o al grupo de control. El propósito es que cada integrante del grupo experimental tenga un gemelo en el grupo de control. Este procedimiento es, sin embargo, trabajoso y está limitado a muestras de tamaño reducido. Por lo demás, no garantiza igualdad en todos los posibles factores intervinientes, sino tan solo en aquellos que han sido tenidos en cuenta en el pareo.

Obviamente, la primera situación mencionada es poco frecuente en la realidad. Podría lograrse en la prueba de un tratamiento nuevo, si se dispusiera de un buen número de pacientes con una similar dolencia y se los dividiera aleatoriamente: algunos recibirían un tratamiento tradicional (o bien ningún tratamiento) y otros la nueva droga. Claro que esto no podría hacerse legítimamente si conllevara algún tipo de riesgos para los pacientes. En el caso de las políticas sociales se agregan dificultades adicionales, tal como lo señala Baker (2000: 2):

Aunque los diseños experimentales se consideran el método óptimo para estimar el

impacto de un proyecto, en la práctica conllevan varios problemas. Primero, la aleatorización podría ser poco ética debido a la negación de beneficios o servicios a miembros de la población de por sí calificados para el estudio. Como un ejemplo extremo se podría citar la negación de tratamiento médico que podría salvar la vida de algunos miembros de la población.

Segundo, puede ser políticamente difícil proporcionar una intervención a un grupo y no a otro. Tercero, el alcance del programa podría significar que no hubiera grupos sin tratamiento, como en el caso de un proyecto o cambio de política de amplio alcance.

Una alternativa sería apelar a algún criterio justificable de priorización: los más necesitados o vulnerables –en el supuesto de que ello pudiera ranquearse– o bien los que reunieran más de las condiciones de admisibilidad si éstas fueran varias. Pero si usáramos un criterio de este segundo tipo, ya quedaría vulnerada la igualdad inicial: habría un sesgo pues los que entraran más tarde serían “menos algo” que los priorizados.

Supongamos que se eligiera a quienes aunque reúnen las condiciones deciden por su cuenta no postularse: aquí habría un sesgo de autoselección. Algo tendrán de diferente –una condición desconocida y no visible– quienes se autoexcluyen. Tampoco esto funcionaría aceptablemente.

Eventualmente sería algo más admisible tomar a aquellos que fueron rechazados por alguna razón de orden administrativo, tal que no pudiera suponerse que establece diferencias, es decir, no asociable a otras condiciones. Ello no resulta sin embargo tan fácil: si alguien no tiene –por ejemplo– documentación en regla porque es extranjero o hijo de extranjeros, ya eso mismo establece diferencias. Se asocia a condiciones que podrían ser objeto de la evaluación o interactúa con ellas. Un niño carente de documentación no podría recibir la AUF², pero tampoco podría ir a la escuela.

Algunas de estas alternativas son, sin embargo, empleadas a veces en diseños cuasi experimentales, menos rigurosos, a falta de otras mejores.

Medición antes/después e interacciones

En algunas ocasiones en que no se puede garantizar la igualdad inicial por procedimientos de azar, es posible en cambio acudir a una medición previa de la variable dependiente en ambos grupos. Esta primera medición permite tener la certeza de que los grupos no difieren significativamente en dicha variable².

Luego de introducido el estímulo sobre uno de los grupos, es posible realizar una segunda medición de la variable dependiente en ambos. No es preciso que cada grupo mantenga el valor inicial: puede haber un “efecto maduración”, consistente en cambios operados debido a la influencia de otros factores o al simple transcurso del tiempo.

En estos casos se emplea el método denominado “diferencia de diferencias”: se establece la diferencia entre la medición inicial y la final en cada uno de los grupos. Tal vez ambos cambiaron, pero si la variable independiente produjo los cambios, el grupo que recibió tratamiento debiera haberla variado más: mostraría, pues, una diferencia mayor entre ambos momentos de medición. Y esta “diferencia de diferencias” se atribuiría a la acción del estímulo.

Podría ocurrir, sin embargo, que haya un efecto de la primera medición que se combine con el estímulo, reforzándolo o atenuándolo, en una suerte de efecto interacción. Discriminar estas interacciones requiere la formación de cuatro grupos aleatorizados o igualados, si el factor experimental es dicotómico, en tanto que el número aumenta a $2 \cdot N$ si hay N valores del factor experimental (Campbell y Stanley, 1995).

1º (E)	2º (E')	3º (C)	4º (C')
Ambas mediciones con estímulo	Sólo estímulo y medición final	Ambas mediciones sin estímulo	Sólo medición final sin estímulo

Así, el primer grupo experimental (E) recibiría el estímulo y sería medido en ambas oportunidades, en tanto que el segundo grupo experimental (E') solo sería sometido a la medición final. Otro tanto se haría con los grupos de control (C y C'). La comparación de la variación en los cuatro grupos.

Diseños ex post

Pero hay un segundo problema. Muy frecuentemente la formulación de las políticas no contempla desde su origen una instancia de evaluación: la necesidad de hacerlo surge a posteriori, cuando ya llevan cierto tiempo de ejecución. Cuando las acciones cuyo impacto se desea evaluar ya se han estado ejecutando y han causado –o no– sus efectos.

En estas evaluaciones *ex post*, sería preciso desplazarse hacia atrás en la búsqueda del grupo de control. Y la medición inicial está irremediablemente ausente, puesto que el fenómeno cuyo impacto deseamos medir o comprobar ya ha ocurrido.

¿Qué se puede hacer en estas situaciones en que las condiciones ideales del experimento están ausentes o no son reproducibles? No queda otra alternativa que procurar aproximaciones a ellas.

Cuando estas son las condiciones contamos, inevitablemente, con una única medición llevada a cabo después de cierto tiempo “bajo programa”. Desconocemos, en rigor, la situación inicial de los beneficiarios: no hay una línea de base. Lo cual no sería tan grave si contáramos con un grupo de

control proveniente de la misma población que los beneficiarios: rezagados que todavía no recibieron ningún beneficio pero que se postularon junto con los actuales beneficiarios. Y a condición de que el criterio de priorización hubiese sido aleatorio, porque entonces, cabría suponer la igualdad inicial, que haría innecesaria la primera medición. Pero supongamos que no lo tenemos: por ejemplo porque todos entraron ya al programa.

Es en estos casos cuando resulta posible apelar a la conformación de un grupo de comparación *ex post* y proveniente de una población que no es la de los beneficiarios, mediante un procedimiento estadístico. Se trata de establecer la “probabilidad de participar” o *propensity score* (Rosembaun y Rubin, 1983; 1985; Caliendo y Copeinig, 2005; Heirinch, Maffioli y Vázquez, 2010). El propósito estriba en contar con un grupo de no participantes asimilable a los participantes, al punto que se pueda asumir la igualdad inicial. El supuesto que hará falta asumir es que si ambos grupos hubieran sido objeto de una medición previa no hubieran presentado diferencias apreciables, por pertenecer ambos a la misma población. Por lo tanto, si tales diferencias se manifiestan en la medición *ex post*, podremos atribuir las acciones del programa.

En rigor, se trata de un suceso cuasi experimental de la situación ideal, pero de muy frecuente uso en las ciencias sociales (y especialmente en la evaluación de políticas) puesto que las condiciones del diseño experimental puro raramente son reproducibles.

Entre las técnicas de diseño cuasi experimental en general se considera que las técnicas de comparación pareada son la alternativa subóptima al diseño experimental. Gran parte de la literatura sobre metodologías de evaluación se centra en el uso de este tipo de evaluaciones, lo que indica el frecuente uso de las comparaciones pareadas y los numerosos desafíos que plantea el contar con grupos de comparación poco adecuados. En los últimos años se han producido significativos avances en las técnicas de correspondencia de puntuación de la propensión (...) Se pueden emplear métodos cuasi experimentales (no aleatorios) para realizar una evaluación cuando es imposible crear grupos de tratamiento y de comparación a través de un diseño experimental. Estas técnicas generan grupos de comparación que se asemejan al grupo de tratamiento, al menos en las características observadas (...) Cuando se usan estas técnicas, los grupos de tratamiento y de comparación por lo general se seleccionan después de la intervención usando métodos no aleatorios. Por lo tanto, se deben aplicar controles estadísticos para abordar las diferencias entre los grupos de tratamiento y de comparación y emplear técnicas de pareo sofisticadas para crear un grupo de comparación que sea lo más similar posible al grupo de tratamiento. En algunos casos también se selecciona un grupo de comparación antes del tratamiento, aunque la selección no es aleatoria (Baker, 2000: 3 a 5).

Los requisitos

La literatura referida a la selección de grupos de comparación basados en la “propensión a participar”, usualmente plantea requisitos bastante estrictos. En rigor los datos de beneficiarios y

no beneficiarios debieran provenir de la misma fuente: una encuesta común. O, al menos, una encuesta aplicada a los beneficiarios paralela en el tiempo a otra encuesta domiciliaria realizada en la misma área geográfica y con el mismo instrumento de recolección.

Inclusive se ha enfatizado la conveniencia de que los encuestadores sean los mismos o al menos pueda garantizarse que recibieron una instrucción análoga antes de abocarse al trabajo de campo. Todo ello, encaminado a asegurar la fiabilidad y comparabilidad de los datos, evitando la introducción de sesgos. O al menos, garantizando que de haber sesgos, estos no se manifestarían de modo disímil entre uno y otro grupo.

¿Qué tipo de sesgos podrían producirse?: por ejemplo, que los sectores que residen en cierto tipo de barrios estén subrepresentados en la muestra. Las villas de emergencia, por caso, son lugares de difícil y riesgoso acceso para encuestadores, si ellos no cuentan con algún contacto interno que franquee el ingreso. Y generalmente la proporción de hogares residentes en villas en las encuestas de hogares (por caso la EP⁴) son inferiores a las que revelan los censos. Ocurre otro tanto, en el extremo opuesto, con la población que habita barrios cerrados, a los que no pueden acceder fácilmente los encuestadores.

Sin embargo, si ese sesgo de selección afectara por igual a ambas muestras (GT y GC), aunque no se lo pudiese evitar no afectaría en demasía la comparación. Por ejemplo, que en ambos grupos estuvieran subrepresentados los sujetos pertenecientes a los estratos más altos: si eso sucediera en igual medida en los hogares del grupo de tratamiento y del grupo de comparación, no ocasionaría dificultades de gran importancia⁵.

Por ejemplo, en la Ciudad Autónoma de Buenos Aires se realizaba un relevamiento de los beneficiarios del programa Ciudadanía Porte⁶ –seleccionados al azar del padrón– con el mismo instrumento y en el momento en que estaba en campo la Encuesta Anual de Hogares de la Ciudad. Y con los mismos equipos de la Dirección de Estadísticas y Censos. Y resultaba posible “pegar” una encuesta a la otra ya que ambas contendrán los mismos campos.

No obstante, sería muy importante que la encuesta general identifique mediante una pregunta a los beneficiarios del programa a evaluar, de manera de no correr el riesgo de incluirlos indebidamente –por desconocer que lo son– en el grupo de comparación, que resultaría así severamente contaminado. Pues algunos miembros del mismo podrían experimentar cambios imputables al programa, estrechando las “diferencias”. Eso conduciría a subestimar los efectos de la exposición al programa.

En el caso de la encuesta de hogares de la Ciudad Autónoma de Buenos Aires sucedía efectivamente así. El programa estaba lo suficientemente extendido como para que la encuesta captara beneficiarios en su muestra. En la encuesta de 2011, por caso, había en la muestra 357 casos de hogares beneficiarios: casi un 6% del total. Estos hogares debieran ser excluidos del grupo de control.

2. El procedimiento

Una vez que se cuenta con una base en la que los beneficiarios del programa “cohabitan” con el resto de la población, la tarea será conformar el grupo de comparación proveniente de esta última mediante la estimación de la “probabilidad de participar” (del programa).

En la base hay hogares provenientes del padrón de beneficiarios del programa –los que efectivamente “participan”– y hay hogares relevados por la encuesta general que no son beneficiarios (habremos identificado y excluido deliberadamente a los hogares beneficiarios relevados por la encuesta general, si es que los hubiera).

Tendremos una variable dependiente dicotómica o binaria cuyos valores serán conocidos: ser o no ser beneficiario. La idea es emplear una herramienta estadística que permita predecir esa condición, basándose en un conjunto de características observables de los hogares, sobre las que la base de datos también cuente con información relevada de manera análoga. La regresión logística resulta un procedimiento adecuado a tales fines (Austin, 2011).

La regresión logística binaria

La regresión logística binaria es un instrumento estadístico de análisis multivariado, de uso tanto explicativo como predictivo. Resulta útil su empleo cuando se tiene una variable dependiente dicotómica (un atributo cuya ausencia o presencia hemos puntuado con los valores cero y uno, respectivamente) y un conjunto de variables predictoras o independientes, que pueden ser cuantitativas (en cuyo caso se las denomina como *covariables* o *covariadas*) o categóricas (Hosmer y Lemeshow, 2000). En este último caso, se requiere que sean transformadas en variables *dummy*, es decir variables simuladas.

Los propósitos del análisis consisten en:

- a) Predecir la probabilidad de que a alguien le ocurra cierto “evento”. Por ejemplo: estar desempleado =1 o no estarlo = 0; ser pobre = 1 o no pobre = 0; ser beneficiario de un programa social =1 o no serlo = 0).
- b) Determinar qué variables pesan más para aumentar o disminuir la probabilidad de que a alguien le suceda el evento en cuestión.

La asignación de probabilidad de ocurrencia del evento a un cierto sujeto, así como la determinación del peso de cada una de las variables dependientes en esta probabilidad, se basan en las características que presentan los sujetos a los que, efectivamente, les ocurren o no dichos sucesos.

- Por ejemplo, la regresión logística tomará en cuenta los valores que asumen en una serie de variables (edad, sexo, nivel educativo, posición en el hogar, origen migratorio, etc.) los sujetos que son efectivamente beneficiarios ($B = 1$) y los que no lo son ($B = 0$).
- En base a ello, predecirá a cada uno de los sujetos –independientemente de su estado real y actual– una determinada probabilidad de ser desocupado (es decir, de tener valor 1 en la variable dependiente) en base a un conjunto de características que podemos suponer que dan cuenta de dicha condición.
- Por ejemplo, si se tratara de un programa que transfiere ingresos a familias vulnerables y con niños. Si alguien es una jefa de hogar joven, con tres hijos pequeños, de baja educación y carente de un empleo estable cubierto por la seguridad social (aunque no sea beneficiaria) el modelo le predecirá una alta probabilidad de serlo (puesto que la proporción de beneficiarios en el grupo así definidos es alta), generando una nueva variable con esa probabilidad estimada.
- Y procederá a clasificarlo como “beneficiario esperado” en otra nueva variable, que será el resultado de la predicción.
- Y además, sopesará cuál es el peso de cada una de estas variables independientes en el aumento o la disminución de esa probabilidad.
- Por ejemplo, cuando aumenta la educación disminuirá en algo la probabilidad de ser beneficiario. En cambio, cuando el sexo pase de 0 = varón a 1 = mujer, aumentará en algo la probabilidad de serlo porque la proporción de beneficiarios es mayor entre las jóvenes mujeres es mayor que entre los jóvenes varones.
- El modelo, obviamente, estima los coeficientes de tales cambios.
- Cuanto más coincidan los estados pronosticados con los estados reales de los sujetos, mejor ajustará el modelo.

En el caso que nos ocupa, se trata de que encontremos un conjunto de variables independientes o predictoras disponibles en la base que permitan identificar adecuadamente a los hogares que efectivamente sean beneficiarios de un programa o política pública, otorgándoles una alta probabilidad de serlo.

Hemos de probar con diferentes variables hasta dar con un modelo que ajuste adecuadamente. Es decir, que prediga razonablemente la condición de los hogares beneficiarios. Que los estados pronosticados coincidan lo más posible con los reales.

Por cierto que cualquier modelo estadístico comete errores. El mejor modelo posible seguramente no identificara como tales a algunos pocos hogares beneficiarios que puedan tener un perfil atípico: por ejemplo, aun en un programa de transferencias de ingresos como Ciudadanía Porteña o en la Asignación Universal por Hijo habrá entre los beneficiarios hogares con perfiles estructurales de clase media, que sin embargo no tienen miembros ocupados en empleos formales o bien tienen bajos ingresos (o al menos sus ingresos no son comprobables). Estos hogares “engañan” al modelo y serán “rechazados” por el mismo: les pronosticará una baja probabilidad de participar y los clasificará como “no beneficiarios”. A esto se le denomina error de exclusión o error de tipo I: el modelo rechaza una hipótesis verdadera. Para que se pueda confiar en que el modelo ajusta adecuadamente, este error de exclusión debiera ser pequeño: involucrar solo a una proporción baja de los realmente beneficiarios.

Pero al revés, también el modelo debiera “rechazar” a los hogares no participantes, reconociéndolos como tales en base a sus características diferentes a las de los beneficiarios. Sin embargo, es probable –y en nuestro caso deseable– que haya entre los hogares no beneficiarios algunos que tengan unas características que los hagan parecer beneficiarios. El modelo los tomaría equivocadamente por tales y les adjudicaría una alta “probabilidad de participar”. En este caso, el estado pronosticado no coincidiría con el real. Esto es lo que se denomina el error de inclusión o error de tipo II: el modelo acepta una hipótesis falsa.

Sea bienvenido este segundo error, porque al cometerlo y mediante su predicción equivocada, el modelo habrá dado con un conjunto de hogares no beneficiarios de los que podremos suponer, en principio, que se asemejan razonablemente a los beneficiarios. ¡Buenos candidatos a integrar un grupo de comparación!

La probabilidad en la regresión logística

La función logística refleja la probabilidad del evento (por ejemplo: ser participante de un programa, ser desempleado, ser graduado universitario, etc.), expresada como *Odds ratio* o chances (Hair et al, 1999):

$$\text{Prob participante} / \text{prob no participante} = \frac{P}{1-P} = e^Z$$

Donde:

E = base del logaritmo natural (2,718)

$$1 \quad 1 \quad 2 \quad 2 \quad \dots \quad n \quad n$$

$$Z = a + b_1 X_1 + b_2 X_2 + \dots + b_n X_n$$

Como vemos, el exponente Z es una ecuación de regresión lineal en la que:

$Z = \text{Log. PP/PNP}$ (*Prob participante/prob no participante*)

a = constante del modelo (la ordenada al origen de la regresión, es decir el valor de la variable dependiente cuando todas las variables independientes asumen el valor cero)

X = variables independientes

b = pesos de cada variable independiente, que pueden ser positivos o negativos (cuando X varía en una unidad, el logaritmo del cociente PD/PND aumenta o disminuye en b unidades)

Finalmente, la probabilidad de ser participante = 1 se obtiene:

$$\text{Prob}(p) = E^Z / (1 + E^Z)$$

Esta última es una probabilidad acotada entre 0 y 1.

Para establecer los valores de la variable dependiente en términos de una probabilidad cuyos límites son cero y uno, la regresión logística supone, en lugar de una relación lineal entre las variables –como lo es el ajuste a una recta de mínimos cuadrados ordinarios– el ajuste a una curva en forma de S. En este modelo, cuando la variable independiente asume valores bajos, la probabilidad se aproxima a cero, en tanto que así como crecen los valores de la variable independiente, esta probabilidad se incrementa a lo largo de la curva pero, a partir de cierto valor la pendiente comienza a decrecer y se estabiliza, de modo que la probabilidad se aproxima a la unidad sin excederla (Hair et al, 1999).

En vez de los mínimos cuadrados, en la regresión logística los valores de las estimaciones se establecen en base al denominado supuesto de *máxima verosimilitud*.

¿Modelos excluyentes o inclusivos?

La regresión logística, al clasificar los casos, adopta por defecto el procedimiento del punto de corte situado en 0,50. Si la probabilidad de participar estimada supera ese valor, entonces el caso es clasificado como “participante”. Pero esto es de algún modo arbitrario.

Supongamos que con ese valor de corte, el modelo es capaz de reconocer como participantes solamente al 60% de los hogares que lo son. Y a su vez, solamente se engaña clasificando un 2% de “falsos participantes” candidatos a grupo de comparación. Sería un modelo en exceso restrictivo, que incurre en muy poco error de inclusión y en demasiado error de exclusión.

Tendremos, pues, que regularlo. Si bajamos a 0,30, el modelo resultante clasifica correctamente a más del 80% de los participantes (redujo su error de exclusión) y en cambio “deja entrar” el triple de no participantes: alrededor de 6%. Es decir, aumenta su error de inclusión y se vuelve más permisivo, con lo que funciona mucho mejor a los efectos de nuestros requerimientos.

Restricciones

Deben tenerse en cuenta algunos preceptos para no “hacer trampas”.

Las variables sobre las que se esperan impactos del programa no deben formar parte del conjunto de variables independientes incluidas en el modelo (Caliendo y Copeinig, 2005). ¿Por qué? Pues porque si el programa tiene efectivamente impactos, el grupo de tratamiento y el de comparación no debiera esperarse que se asemejen en los valores de estas variables, sino que diverjan.

Por caso, si un programa de transferencias condicionadas incluye la obligación de escolarizar a los adolescentes, entonces esperaríamos que en los hogares beneficiarios la tasa de asistencia de los adolescentes fuera mayor que en el grupo de comparación. Si incluimos esta variable en el modelo estaríamos seleccionando hogares que tienden a ocultar el impacto del programa.

Otro tanto ocurriría con eventuales efectos no deseados. Cierta literatura señala que los PTC tienen impactos negativos sobre la tasa de actividad, especialmente en el caso de las mujeres. De manera que la tasa de actividad de los hogares (el porcentaje de miembros activos sobre el total de miembros, o mejor de miembros adultos) tampoco debiera emplearse como variable predictora. Porque contribuiría a seleccionar en el grupo de comparación hogares con similares tasas de actividad, con lo que podría disimular un eventual efecto reductor de la propensión a trabajar entre los beneficiarios, si es que realmente lo hubiera.

En otros términos, las variables independientes del modelo debieran ser “neutras” en cuanto a los efectos esperados del programa. En cambio, sí es posible y recomendable incluir variables relacionadas con los criterios de admisión, tales como la presencia de niños y adolescentes o la de trabajadores informales.

Y por cierto otras variables relacionadas con cierto “perfil” del hogar: el tamaño, la relación entre menores y total de miembros, el nivel educativo de los adultos.

¿Qué sucedería con los ingresos familiares? En principio, se diría que si ambas encuestas los han indagado de igual manera, podrían incluirse como predictores a condición de que, en el caso de los hogares realmente beneficiarios, pudieran restarse de los ingresos totales los que corresponden a la transferencia de la prestación del programa. Pero cabría una objeción: si existe el supuesto de que el ingreso de un PTC podría sustituir ingresos de otras proveniencias y alentar el retiro del mercado laboral de trabajadores con ingresos bajos, al restar los ingresos del PTC estaríamos subestimando los que realmente podría tener el hogar sin esa transferencia, con lo que sesgaríamos hacia abajo la selección del grupo de comparación.

Otra posible razón estriba en el carácter frecuentemente volátil de los ingresos monetarios de los hogares que suelen ser beneficiarios de programas sociales, como también de aquellos hogares que, sin serlo, se les asemejan. Estos hogares suelen tener miembros ocupados en puestos de trabajo del sector informal, frecuentemente inestables. Incluso la intensidad horaria de sus ocupaciones puede ser variable en lapsos relativamente cortos de tiempo. Por eso, parece más razonable nutrir el modelo con variables observables de carácter más estructural y permanente.

¿Cómo proceder?

El subconjunto de hogares que integran el error de inclusión –los “falsos participantes” – serán pues la cantera que permita conformar el grupo de comparación.

¿Cuál es el procedimiento a emplear para seleccionar el grupo definitivo? El procedimiento que sugiere la literatura (Rosenbaum y Rubin, 1985; Lazo y Philipp, 2003; Caliendo y Copeinig, 2005) es el del “vecino más próximo”. E indica que para cada caso integrante del grupo de tratamiento (excluidos previamente los atípicos a los que el modelo estadístico no pudo reconocer como tales y les otorgó baja propensión a participar) hemos de encontrar, dentro del subconjunto del “error de inclusión” o “falsos participantes” entre tres y cinco casos con una propensión a participar cercana.

Las distancias entre las propensiones (que se expresan siempre como probabilidades fijadas entre cero y uno) suelen medirse en términos cuadráticos. Es decir como diferencia elevada al cuadrado entre ambas probabilidades. Y será aceptable incluir en el grupo de comparación los tres o los cinco casos cuya distancia máxima con su correspondiente del grupo de tratamiento no supere 0,01 (Lazo y Philipp, 2003).

La idea es que por cada caso en el grupo beneficiario se cuente con tres o con cinco como controles. Dado que procuramos “aparear” beneficiarios con los más parecidos entre los no beneficiarios, es aceptable que un mismo caso del grupo de comparación pueda, simultáneamente, servir de equivalente para varios beneficiarios con los que guarde una

diferencia reducida en la puntuación de la propensión a participar. Vale decir, no es necesario que el apareamiento sea uno a uno.

Las limitaciones

Por cierto que estas recomendaciones no siempre son fáciles de cumplimentar. La mayor de las dificultades es que el error de inclusión sea muy pequeño. Hay pocos hogares parecidos a los beneficiarios y que no lo sean. Esto ocurre especialmente cuando un programa muestra una cobertura muy extendida sobre la población meta.

Por ejemplo, no sería muy fácil encontrar hogares con menores a cargo de trabajadores informales y que no reciban la AUH. O también podría suceder que los que halláramos guardaran alguna diferencia esencial con los beneficiarios: por ejemplo podrían ser migrantes recientes cuyos hijos carecen de documento de identidad.

En parte, esto podría solucionarse mediante la admisión de unas diferencias más amplias al hacer el *matching*. O bien tornando mas lato el criterio del numero de “gemelos” en el grupo de comparación (Lazo y Philipp, 2003).

Puede ocurrir que algunos miembros del grupo de tratamiento directamente no encuentren gemelos en el grupo de comparación dentro del rango de puntuación aceptado en la diferencia de la propensión a participar. En este caso debieran ser excluidos (Caliendo y Copeinig, 2005). Y ello podría sesgar los resultados: supongamos por un momento que los impactos del programa fuesen menores –o mayores– en cierto tipo de hogares “atípicos”, que no obstante participar del mismo arrojen valores extremos (muy altos o muy bajos) en la puntuación de la propensión. Su exclusión conduciría a ocultar estos efectos diferenciales y con ello a sobrestimar o subestimar el impacto, según el caso.

Cómo calcular las diferencias

En el esquema clásico del diseño experimental, se supone que las diferencias debieran calcularse entre casos pareados.

Como habría en el grupo de comparación más de un caso gemelo por cada beneficiario (entre tres y cinco) los valores de estos casos en la/s variables/s testeadas debieran ser promediados y luego se calcularía, para cada una de estas variables, la diferencia entre el valor de cada caso integrante del grupo de tratamiento y el promedio de los valores de sus casos testigos o gemelos en el grupo de comparación.

Y luego, se promediarían las diferencias así obtenidas: ese promedio sería la diferencia entre el grupo de tratamiento y el grupo de comparación.

Este procedimiento no ofrecería dificultad si la variable o variables a testear (las de impacto del

programa] fuesen variables cuantitativas: el número de controles realizados durante el embarazo o la cantidad de visitas al médico en el primer año de vida de un niño.

Pero más usualmente son categóricas. La asistencia o no a un establecimiento educativo o el hecho de estar o no estar inserto en el mercado de trabajo, por ejemplo.

Esto tendría una solución relativamente sencilla si estas variables se convierten en dummy con valor 1 para la respuesta afirmativa (asiste a la escuela, está inserto en la actividad económica, etc.) y valor 0 para la respuesta negativa (no asiste, no trabaja, etc.). Estos valores podrían promediarse para los casos gemelos del grupo de comparación, aunque esos promedios no serían ya cero y uno. Supongamos que un caso cualquiera del grupo de tratamiento tuviera tres equivalentes por puntuación próxima en el grupo de comparación. En el primero el jefe de hogar trabaja y recibe puntuación 1, mientras que en los tres hogares apareados hay dos que trabajan y uno que no: el promedio sería 0,666. Y la diferencia 0,333. La interpretación intuitiva de diferencias así obtenidas resultaría no obstante un tanto confusa y oscura.

Sin embargo, hay otra dificultad mayor cuando los elementos seleccionados no son individuos sino colectivos –como es el caso de los hogares– y cuando los presuntos impactos del programa son diversos y pueden manifestarse sobre distintos miembros de estos colectivos.

Al tratar de calcular las diferencias de hogar a hogar, en el caso de la asistencia escolar, que se calcula por edades diferentes (5 años, 6 a 12 años, 13 a 15 años, etc.), comprobaríamos que no todas las familias tienen niños situados en los mismos tramos de edad. Por ejemplo, en un hogar beneficiario puede haber un niño en edad de escolaridad primaria y un adolescente en edad de asistir al secundario. Pero en los hogares más próximos en la puntuación de la propensión a participar, no necesariamente habrá niños y adolescentes. Puede ser que haya unos u otros. Y la inexistencia de unos u otros hará imposible el cálculo de las diferencias en todos los tramos previstos.

Otro tanto sucedería con la inserción laboral de los varones adultos: podría no haberlos, por caso en los hogares monoparentales.

Si fuera necesario prever, al hacer el *matching*, formar subcategorías de hogares con conformaciones específicas antes de seleccionar los más próximos en las puntuaciones, la construcción del grupo de comparación se tornaría sumamente dificultosa. Y más probablemente imposible.

Una alternativa

Una alternativa por la que puede optarse para solucionar estos inconvenientes cuando los impactos del programa que deben verificarse son variados y afectan a diferentes miembros de los hogares involucrados, estriba en el cálculo global de las diferencias entre el grupo de tratamiento

y el grupo de comparación y no caso a caso.

Si sencillamente se adopta como grupo de comparación el conjunto de los hogares que integran el error de inclusión y que cuentan con un hogar próximo en su “propensión a participar” entre los beneficiarios, desaparece la necesidad de correspondencia uno a uno entre niños de cada tramo de edad, por ejemplo. Simplemente hemos de computar, para cada grupo etario, la tasa global de asistencia de cada tramo de edad en un grupo y otro. Y las diferencias se computan de ese modo. Otro tanto se haría con las tasas de actividad o las de empleo masculinas, femeninas, etc.

Claro que, eliminado el *matching* caso a caso y simplemente ateniéndonos a un grupo de comparación conformado por el error de inclusión, surge naturalmente la necesidad de hacer algunos controles que nos permitan asumir el supuesto de la igualdad entre ambos grupos.

Los recaudos

Un primer recaudo consistirá en examinar el rango de las puntuaciones de “propensión a participar” tanto en los beneficiarios como en los que no lo son pero el modelo identificó como tales, así como sus dispersiones.

Si las puntuaciones mínimas y máximas fuesen muy diferentes, podrían eliminarse los casos extremos en ambos grupos con el propósito de atenuar estas diferencias. Inclusive, es posible emplear como grupos de impacto solo los subconjuntos de casos que guarden razonable proximidad. Esta posibilidad de usar solo los casos que se ubiquen en una zona de convergencia en cuanto a las puntuaciones, ha sido plenamente justificada por la literatura (Jalan y Ravallion, 1998).

Ello, no obstante, en el caso del grupo de tratamiento, reproduciría el riesgo de encubrir efectos diferenciales sobre casos extremos. Esto no ocurriría en el grupo de comparación, cuyos extremos podrían recortarse sin riesgo para asimilarlo al grupo de tratamiento.

Luego, se compararían las medias de las puntuaciones de uno y otro grupo con el test de t de Student para la diferencia de medias de muestras independientes: el propósito sería, en este caso, que pudiésemos conservar la hipótesis de nulidad que afirmaría que ambas muestras provienen de poblaciones con medias iguales.

Asimismo, los controles podrían replicarse para un conjunto de variables estructurales no vinculadas con los impactos esperados del programa. Por ejemplo, tamaño de los hogares, cantidad de menores, edades o niveles educativos de los jefes de hogar y o sus cónyuges, etc. Se esperarían que ambos grupos no difirieran de manera significativa en este tipo de características. Y ello también podría ser testado con la prueba de t de Student para la diferencia de medias de muestras independientes (y su extensión al caso de las proporciones cuando se trata de variables categóricas). Nuevamente, el propósito de estos controles sería poder admitir la hipótesis de

nulidad.

¿Doble ciego?

En un experimento –cualquiera que fuese– lo ideal es la aplicación del procedimiento denominado de “doble ciego”.

Esto es muy usual por ejemplo en los experimentos médicos. En un experimento de “doble ciego”, ni los individuos participantes que integran los grupos a comparar ni los investigadores que implementan el procedimiento saben quién pertenece al grupo de comparación (y recibe placebos) y quién al grupo experimental. Solamente después de haberse analizado todos los datos, y concluido el experimento, los investigadores conocen qué individuos pertenecen a cada grupo.

De esta forma se evita el riesgo de la introducción de sesgos en la evaluación de los resultados imputables a preconcepciones por parte del que evalúa los mismos. Se busca con ello evitar que las expectativas del investigador influyan sobre el resultado observado.

Es preciso señalar que cuando empleamos el método de construcción estadística del grupo de comparación no es posible poner en práctica el experimento “doble ciego”, en tanto el propósito del modelo estadístico es reconocer a los *verdaderos* participantes: esa es la variable dependiente que se procura predecir.

Y esperamos que el modelo atribuya equivocadamente esa condición a no participantes que se les parezcan lo suficiente, asignándoles una “propensión a participar” próxima a la de los miembros del grupo de tratamiento.

Estas posiciones no se pueden invertir. Si ignoráramos *cual es cual*, podríamos construir un modelo que fuera capaz de reconocer a no participantes. Por ejemplo, sujetos en su mayoría sistemáticamente más ricos que los participantes. Y luego les apareara a algunos de estos últimos, atípicos, que se les asemejaran. Esta inversión, obviamente, sesgaría totalmente los resultados.

3. Un ejemplo de aplicación

A continuación se exponen resultados de la aplicación del procedimiento que hemos estado considerando para la selección del grupo de comparación en la evaluación de un programa de transferencias de ingresos condicionadas (PTC) que, en el marco de esta exposición, denominaremos Plan Solidario.

Los cuadros que se muestran –las salidas resultantes de la aplicación del modelo– son los resultantes de una de las pruebas previas. En la evaluación final se empleó un modelo muy semejante pero no igual.

El Plan Solidario consiste en un programa de transferencias de ingresos condicionadas (PTC) a hogares de bajos ingresos. Conlleva por lo tanto contraprestaciones –mejor, “corresponsabilidades” (Mazzola, 2012) – por parte de los hogares beneficiarios que incluyan niños y adolescentes: de asistencia a un establecimiento educativo en las edades obligatorias y de cumplimiento del plan de vacunación vigente en el caso de los niños.

Puesto que los ingresos de los hogares no son comprobables cuando sus miembros (o al menos algunos de ellos) no son trabajadores asalariados registrados en la seguridad social –y esta es la situación más frecuente entre los hogares de bajos ingresos– se aplica un *proxy means* test para la admisión de los hogares postulantes al programa que, en caso de dudas se complementa con procedimientos de análisis cualitativo mediante visitas y entrevistas.

Una vez admitidos al programa los hogares reciben una prestación monetaria que depende de su composición y tamaño. Pasado cierto tiempo desde la iniciación de las acciones del Plan Solidario se trata de evaluar sus eventuales impactos. Seguramente los gestores de la política se preguntarán si ella ha arrojado los resultados esperados: los hogares mejoraron su nivel de vida, los niños y adolescentes asisten más a la escuela, se hacen más controles médicos preventivos y se vacunan más, etc.

Asimismo –y puesto que los PTC son frecuente objeto de críticas– también habrá de verse si la recepción de un ingreso complementario no ha desalentado las búsquedas laborales e incentivado a los miembros de los hogares a trabajar menos, lo que sería un efecto no deseado de la política. Para todo ello, es preciso contar con un grupo de comparación compuesto por hogares de características estructurales semejantes a las de los que reciben el programa (“elegibles”, digamos), pero que no hayan sido beneficiarios del mismo.

Las similitudes, sin embargo, no deben fincar en aspectos sobre los que se espera que el programa tenga efectos. No tendremos que seleccionar hogares donde los niños asistan en la misma proporción a la escuela, por caso, puesto que esperamos que el programa establezca una diferencia en este punto. Si así lo hiciésemos estaríamos “forzando” a los hogares del grupo de comparación a tener tasas de asistencia escolar similares a las del grupo de tratamiento y eso tendería a encubrir el impacto del programa, jugándonos en contra. Tampoco los ingresos serán una variable de selección, porque justamente la política que estamos evaluando consiste en proporcionar ingresos adicionales.

Aunque en este caso se podría emplear el ingreso descontando la prestación del programa, tampoco resulta del todo aconsejable. Pues en los sectores más vulnerables los ingresos suelen ser variables y poco estables, porque también las ocupaciones lo son.

Es mejor, pues, apelar –más que a los ingresos presentes– a la capacidad potencial de obtenerlos. Así como las necesidades de los hogares para la subsistencia. Así, los niveles educativos de los adultos, el tamaño y conformación de los hogares, la disponibilidad de fuerza de trabajo y el tipo de inserción laboral de los miembros en edad de trabajar, sí que pueden ser variables de selección. En este último caso, porque permitiría discriminar “negativamente” a los hogares con miembros insertos en forma estable en el mercado de trabajo formal, que en rigor resultan menos “elegibles” y por lo tanto no abundan entre los beneficiarios.

En el ejemplo que aquí se presenta, las variables seleccionadas como independientes fueron:

- Tenencia precaria de la vivienda (Tempre) ⁹
- Cantidad de menores de 18 años en el hogar (Men18_1)
- Jefe del hogar con baja educación (jebajed_1) ¹⁰: variable dummy
- Jefe de hogar trabajador protegido (jeprot_1) ¹⁰: variable dummy
- Tasa de dependencia económica del hogar (tasdep) ¹¹
- Tasa de cobertura de salud del hogar (tasalud) ¹²
- Total de miembros del hogar (pobtot)
- Hacinamiento en la vivienda (hacinam) ¹²: variable dummy

A continuación la tabla de clasificación resultante:

Tabla de clasificación*

Observado			Pronosticado	
			Beneficiario del Plan Solidario	
			0	Sí
Paso 1	Beneficiario del Plan Solidario	0	5496	894
	Sí		9	288
	Porcentaje global			
				86,5

a. El valor de corte es ,050

Como puede apreciarse, el modelo reconoce adecuadamente al 97% de los hogares realmente beneficiarios. Tiene un error de tipo I o de exclusión muy bajo (apenas 9 de los 297 hogares beneficiarios presentes en la muestra fueron erróneamente clasificados como no beneficiarios). Pero en cambio, el error de admisión es afortunadamente mayor: puesto que 86% de los hogares no participantes son correctamente clasificados, hay 14% (894 hogares) que son “confundidos” con los participantes; candidatos a integrar el grupo de comparación.

El ajuste del modelo se revela, pues, altamente satisfactorio. Veamos ahora la significación de cada una de las variables predictoras:

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1 ^a	TENPRE_1	,151	,034	19,219	1	,000	1,163
	MEN18_1	,428	,095	20,117	1	,000	1,534
	JBAJED_1	,757	,151	25,020	1	,000	2,133
	JEPROT_1	-,398	,194	4,217	1	,040	,672
	TASDEP	1,046	,263	15,844	1	,000	2,847
	TASALUD	-2,339	,191	149,840	1	,000	,096
	POBTOT	,120	,065	3,399	1	,065	1,127
	HACINAM	,241	,182	1,759	1	,085	1,273
	Constante	-3,601	,279	166,441	1	,000	,027

a. Variable(s) introducida(s) en el paso 1: TENPRE_1, MEN18_1, JBAJED_1, JEPROT_1, TASDEP, TASALUD, POBTOT, HACINAM.

Todas ellas se muestran significativas a través del estadístico de Wald: prueba de significación cuya hipótesis nula es que los coeficientes B de la regresión logística son iguales a cero (Hosmer y Lemeshow, 2000). En todos los casos es posible rechazar esta hipótesis nula con una probabilidad de error siempre menor a 10% (8,5% en el caso del hacinamiento).

Los coeficientes B tienen, asimismo, los signos teóricamente esperados: por caso, la tenencia precaria de la vivienda, la cantidad de menores presentes en el hogar, el tamaño del hogar o la presencia de un jefe con baja educación aumentan la probabilidad de participar del programa. En

tanto que la presencia de trabajadores protegidos o de personas con cobertura de salud la disminuyen.

En la siguiente tabla (prueba ómnibus sobre los coeficientes del modelo) se muestra una prueba Chi Cuadrado que evalúa la hipótesis nula de que los coeficientes B de todas las variables introducidas, a excepción de la constante, son cero. Este estadístico está obtenido por diferencia entre los valores del $-2LL$ antes y luego de la introducción de las variables independientes. La significación de Chi cuadrado (menor a 0,001) nos permite rechazar tal hipótesis nula sin vacilación alguna.

Pruebas omnibus sobre los coeficientes del modelo			
	Chi cuadrado	gl	Sig.
Paso	815,743	8	,000
Paso 1 Bloque	815,743	8	,000
Modelo	815,743	8	,000

Las R cuadrado de Cox y Snell y Nagelkerke son interpretables como proporción de varianza explicada. Los valores obtenidos no fueron altos, como lo muestran las salidas. Nuevamente, la razón es que se espera que en este empleo del modelo logístico haya un error clasificatorio, que nos permita obtener “no participantes” parecidos a los “participantes”. Así las variables independientes estarán lejos de agotar las fuentes de variación de la variable dependiente (“participación/no participación”).

Resumen del modelo			
Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	1614,694 ^a	,115	,377

a. La estimación ha finalizado en el número de iteración 7 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Prueba de Hosmer y Lemeshow			
Paso	Chi cuadrado	Gl	Sig.
1	23,950	8	,002

Finalmente, se testea también la significatividad global del modelo con la prueba de Hosmer y Lemeshow. Se trata de una prueba de significación estadística cuya hipótesis de nulidad supone que los estados observados no difieren de las predicciones. Esta hipótesis de nulidad, en este caso pudo ser rechazada con una probabilidad de error de tipo I (rechazar una hipótesis nula verdadera) muy baja: 2%. Cuando, *prima facie*, se trataría de un indicador desfavorable para evaluar un modelo: se diría que entonces clasifica mal, incurre en errores clasificatorios. Pero en este uso del modelo, dichos errores son –ya se lo ha dicho– deseables. Esperamos que el modelo “se confunda” y tome a algunos “no beneficiarios” por “beneficiarios”, estimándoles una “probabilidad

de participar” (*propensity score*) cercana a los primeros. De no haber este error de admisión, no tendríamos grupo de comparación. Es decir, en este caso se trata de un error deseable. En cambio, hemos de esperar que los verdaderos beneficiarios sean reconocidos con poco error de exclusión: así sucede, felizmente, con el 97% de ellos.

Conclusiones

La evaluación del impacto de las políticas responde a la necesidad de comprobar si los objetivos previstos en las mismas se cumplen. Este breve artículo abordó el uso de la regresión logística binaria para solucionar algunas dificultades propias de la evaluación de políticas y programas sociales, ante la imposibilidad de emplear diseños experimentales puros y sin contar con línea de base.

Sin ignorar los reparos y recaudos que deben tenerse en cuenta en el empleo de grupos de comparación contruidos mediante métodos estadísticos, este procedimiento se revela como una solución subóptima y un fértil recurso en la evaluación de impacto de las políticas públicas.

No obstante, y teniendo en cuenta la creciente necesidad de dar cuenta del resultado de las acciones del Estado que demandan recursos provenientes del conjunto de los contribuyentes, se remarca la importancia de que la planificación de las políticas prevea instancias de evaluación futura y contemple mediciones de línea de base que incluyan, siempre que sea posible, población no inserta en el programa a evaluar o cuyo ingreso al mismo se vea diferido. Ello, con el objetivo de poder emplear el procedimiento de la “diferencia de diferencias” sin el requisito de asumir el exigente supuesto de la equivalencia inicial en ausencia de la medición previa al inicio de las acciones del programa que, como hemos visto, exige la medición *ex post*.

Pero toda vez que ello no siempre es posible, se espera que el presente artículo sea una contribución a la tarea de quienes deben asumir la tarea de evaluar las políticas públicas.

Notas

Austin, P. (2011). “An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies”. *Multivariate Behavioral Research*, 46:3, 399-424, DOI: 10.1080/00273171.2011.568786.

Baker, J. (2000). *Evaluación de impacto de los proyectos de desarrollo en la pobreza. Manual para profesionales*. Washington D. C.: Banco Mundial.

Caliendo, M. y Copeinig, S. (2005). *Some Practical Guidance for the Implementation of Propensity Score Matching*. Forschungsinstitut zur Zukunft der Arbeit Institute for the Study of Labor. Discussion Paper No. 1588. Bonn, Germany.

Hair, J., Anderson, R., Tatham, R y Black, W. (1999). *Análisis Multivariante*. Madrid: Prentice Hall.

Heinrich, C., Maffioli, A. y Vázquez, G. (2010). *A Primer for Applying Propensity-Score Matching Impact-Evaluation Guidelines*. Technical Notes. Inter-American Development Bank.

Hosmer, D. W. y Lemeshow, S. (2000). *Applied Logistic Regression*. New York: John Wiley & Sons, INC.

Jalan, J. y Ravallion, M (1998). "Income Gains from Workfare: Estimates for Argentina's TRABAJAR Program Using Matching Methods." Washington, D.C. Development Research Group, Banco Mundial.

Lazo, T. y Philipp, E. (2003). "Uso de la regresión logística para la construcción de un grupo de control" en Lago Martínez, S., Gómez Rojas, G. Y Mauro, M. (coordinadoras): *En torno de las metodologías: abordajes cualitativos y cuantitativos*. Buenos Aires: Editorial Proa XXI.

Mazzola, R. (2012). *Nuevo Paradigma. La Asignación Universal por Hijo en la Argentina*. Buenos Aires: CEDEP (Centro de Estudios y Desarrollo de Políticas)/Prometeo.

Rosenbaum, P. y Rubin, D. (1985). "The central role of propensity score in observational studies for causal effects". *Biométrica* N° 70. Great Britain.

ARTÍCULOS

Metodología estadística para la estimación de las superficies sembradas con cultivos extensivos - método de segmentos aleatorios.

Norberto V. Rodríguez/ Julieta Mirensky

estimador

tecnología satelita

El presente trabajo corresponde a la metodología del diseño de una muestra con el objetivo básico de estimar las superficies sembradas con cultivos de tipo extensivos en las principales provincias, regiones y jurisdicciones del país, que ha comenzado a aplicarse en el ámbito del Ministerio de Agricultura Ganadería y Pesca (MAGyP) desde la campaña agrícola de los años 2011/2012.

En relación al *tamaño de los granos* del cultivo y de la *época en que se siembra*, en el país se distinguen dos tipos de campañas agrícolas denominadas:

- **Fina:** corresponde a los cultivos extensivos invernales sembrados en Mayo/Julio y cosechados en Noviembre/Enero, como trigo, cebada, centeno, avena y colza.
- **Gruesa:** cultivos extensivos que comienzan a sembrarse en Septiembre y se cosechan hacia Enero, los principales son soja, maíz, sorgo y girasol.

Es de fundamental importancia para la planificación en el sector agrícola disponer de estimaciones anuales de las áreas sembradas con los cultivos para ambas campañas.

Históricamente, la información estadística del MAGyP se basó en un **Método Subjetivo** que parte de *informantes calificados* con la incorporación de controles y validaciones con datos provenientes de fuentes comerciales y productivas (acopiadores, cooperativas, productores, distribuidores de

agroquímicos y semillas, etc.).

Los informantes calificados proveen gran variedad de datos necesarios, pero en lo que se refiere a la estimación de las hectáreas cultivadas tienden a producir desfasajes acumulativos en el tiempo y la no captación suficientemente rápida de los cambios regionales. Esto motiva la necesidad de disponer de información *objetiva*, es decir que no esté fundamentada en opiniones personales.

En consecuencia se desarrolló el método que se presenta en su faceta estadística, que se denominó de *Segmentos Aleatorios*, basado en un muestreo probabilístico de áreas, para la estimación de las superficies sembradas con cultivos extensivos en diferentes zonas agrícolas del país.

2. Incorporaciones tecnológicas

2.1. Antecedente de las encuestas por muestreo en el sector

Hasta hace unos 15 años se realizaban encuestas agropecuarias basadas en *entrevistas directas* (face to face) a productores, esta metodología tuvo que ser dejada de lado debido a una creciente *tasa de no respuesta* debido a motivos, tales como:

- a. Aumento del número de productores no residentes en la explotación, que se los debe buscar en domicilios ubicados en alguna localidad no siempre cercana, lo cual requiere de encuestadores con dotes de detective.
- b. Surgimiento del método de producción *“Pool de siembra”*, caracterizado por un sistema empresarial transitorio que asume el control de la producción agropecuaria mediante el arrendamiento de grandes extensiones de tierra, el uso de equipos propios o contratados para siembra, fumigación, cosecha y transporte. En general son difíciles de localizar y usualmente se niegan a suministrar información.
- c. Aparición de robos e inseguridad, que produce el efecto de tranquera cerrada y para ingresar a la explotación se necesita gestionar un permiso previo.
- d. En el caso que el encuestador logre contactar al productor, es frecuente la negación a responder el cuestionario o que se brinden respuestas incorrectas (generalmente asociado a conflictos impositivos).

La consecuencia de los problemas citados era un alto costo operativo para muy escasa respuesta confiable, incremento de las visitas a un mismo productor y demoras en la recolección y procesamiento de los datos.

En el mediano plazo no se espera el mejoramiento de las condiciones ya señaladas, por ello se

necesita un diseño de muestra y un operativo de campo que no dependa de “entrevistas a productores”.

2.2. Incorporaciones tecnológicas del método

El nuevo método de segmentos aleatorios contempla el uso de nuevas tecnologías:

- Datos básicos obtenidos mediante la observación directa de la cobertura del suelo de cada lote, sin consultar a los productores ni a los dueños de las tierras.
- Uso de imágenes de alta y mediana resolución originadas por los satélites Landsat y Spot, el sensor Modis ubicado en los satélites Terra y Aqua y el programa Google Earth.
- Sistemas de posicionamiento global (GPS), para determinar coordenadas que permiten ubicar puntos de referencias sobre la superficie terrestre y sobre las imágenes satelitales.
- Software de información geográfica (ArgGIS, Quantum GIS).
- Programas “SPSS Statistical” y “Minitab” para el procesamiento, análisis y obtención de resultados.
- Estrictos protocolos de trabajo para organizar la información y elaborar bases de datos.

3. Objetivos y condicionamientos

3.1. Objetivo general del método

- ***Estimar las superficies sembradas con cultivos de tipo extensivos especificados, a nivel de Partidos*** -en la Provincia de Buenos Aires- y ***Departamentos*** -en las restantes provincias-. Estos constituyen el nivel máximo de desagregación, la adición de niveles conducen a regiones o incluso a provincias completas.

3.2. Objetivos específicos

- Obtener estimaciones anuales que sirvan como base de ajuste de múltiples estimaciones subjetivas.
- Generar un Sistema de Información Geográfica (SIG) que permita:
 - El soporte de los sucesivos operativos a campo
 - La fuente de consulta para problemas de gestión y planificación
 - La generación de Información Agrícola en forma de series temporales

- Establecer un proceso de mejora continua de la metodología, incluyendo el aumento progresivo de la superficie en estudio con la incorporación de nuevos Partidos y Departamentos del país.

3.3. Condicionamientos del diseño

El diseño de la muestra tuvo que superar varias condiciones restrictivas; se citan:

a. Suministrar estimaciones para los niveles de Partidos/Departamentos en los que se desarrolla actividad agrícola extensiva (se excluye La Patagonia y zonas cordilleranas). Estos son un número importante y se necesita una muestra de gran tamaño.

b. Información básica obtenida mediante visualización directa de la cobertura del suelo por parte de un observador, con el soporte de imágenes satelitales. Este procedimiento implica la pérdida de datos que no son observables, como intención futura de siembra, régimen de tenencia, etc.

c. Estimar los errores debidos al muestreo, a efectos de verificar que las superficies con los principales cultivos se ubiquen dentro de niveles aceptables. Por tanto, se precisa una muestra probabilística.

d. Estratificar dentro de cada Partido/Departamento creando zonas de uso homogéneo del suelo, con el objetivo de disminuir el error debido al muestreo. La estratificación en el sector agrícola es una tarea compleja que requiere utilizar tecnología satelital.

e. Definir áreas relativamente pequeñas de superficie terrestre, que son las unidades de observación que se les adjudicó la denominación de *segmentos*. Seleccionar sobre imágenes y mapas no presenta dificultad, pero al observador en campo le resultará difícil sino imposible acceder a dichas áreas en caso de que se ubiquen en lugares sin caminos públicos.

f. Releva el 100% de cada segmento, pueden contener además de los cultivos extensivos, otros cultivos (producción hortícola, frutícola, etc.) y desperdicio agrícola (urbanizaciones, montes, caminos, inundaciones transitorias, etc.).

g. Reducir el tiempo destinado a trabajo de campo y procesamiento de datos, idealmente a no más de tres meses.

h. Repetir dos veces por año el operativo -coincidente con las campañas agrícolas- dentro de un presupuesto restringido a los recursos del MAGyP (personal, equipamiento informático y vehículos disponibles).

Nótese que la condición 1 actúa en sentido opuesto a las condiciones 7 y 8.

La condición 2 conduce a tener que aceptar un sesgo estadístico, ya que para poder efectivizar las

observaciones de los segmentos, se requiere alterar la aleatoriedad de la selección originalmente asignada.

4. Proceso Estadístico de elaboración del diseño de la muestra

4.1. Esquema del diseño de la muestra

Desde el punto de vista de la teoría del muestreo, se trata de un *diseño de muestra probabilístico de áreas, en una sola etapa, con estratificación de las unidades de muestreo de acuerdo al uso de la tierra y selección simple al azar de estas unidades dentro de cada estrato*.

En esta aplicación la unidad de muestreo es el *segmento*, que una vez definido pasa a integrar un sistema de visitas periódicas repetitivas en el tiempo, para ser observados de acuerdo a las necesidades de información, es decir que se le adjudica el carácter de *permanente*.

El procedimiento en cuanto al diseño de muestra requirió cumplir la siguiente serie de pasos.

4.2. Eliminación de superficies con probabilidad nula de ser cultivadas

La tarea inicial consistió en excluir aquellas superficies con probabilidad nula (actual o futura) de producción agrícola. Se lleva a cabo por Partido/Departamento Provincial, mediante un análisis de imágenes satelitales Lansat y del programa Google Earth, también se tuvo en cuenta el conocimiento de los delegados del MAGyP. Las áreas excluidas corresponden a:

- Ciudades y poblaciones
- Cuerpos de agua permanente (lagunas y ríos) y bajos inundables
- Sierras y pendientes pronunciadas
- Franjas de playas, dunas, bosques

Estas superficies configuran el *Descarte*, obviamente no revisten de interés para el operativo y solo importan a efectos de su exclusión del marco de selección.

4.3. Estratificación

Se define como *Estrato* a aquellas zonas dentro de cada Partido/Departamento que se caracterizan por un uso agrícola homogéneo del suelo. Los estratos presentan formas irregulares y no tiene porqué formar superficies continuas ni seguir límites de jurisdicciones; pueden cortar las explotaciones agropecuarias con líneas precisadas sobre las imágenes, aunque no observables sobre la superficie.

El objetivo fundamental de la estratificación es aumentar la precisión de las estimaciones que se obtengan por la muestra (estimaciones con menor error debido al muestreo), pero no reviste

ningún interés suministrar información a nivel de estrato.

La estratificación es una tarea realizada por especialistas en la interpretación de imágenes, en este caso provenientes de los satélites Landsat y Spot y del sensor MODIS (Espectroradiómetro de Imágenes de Resolución Moderada) ubicado en los satélites Terra -órbita en dirección Norte-Sur cruzando el Ecuador por la mañana- y Aqua -órbita de Sur a Norte cruzando el Ecuador por la tarde- y además se utiliza el programa Google Earth para determinar y medir estas zonas.

Desde un punto de vista general se diferencian 4 estratos básicos:

- **Estrato A:** Zonas con un alto porcentaje de aptitud agrícola
- **Estrato B:** Zonas mixtas con un mayor porcentaje de aptitud ganadera
- **Estrato C:** Zonas sólo aptas para uso ganadero
- **Estrato D:** Tierras con probabilidad nula de utilización agrícola o pecuaria.

No obstante, en algunas provincias y zonas, de acuerdo a las necesidades de estimación propias de cada una, se pueden adicionar otros estratos de interés.

La estratificación se actualiza continuamente, ya que el uso de la tierra es susceptible a modificaciones en el tiempo.

4.4. Procedimiento para estratificar

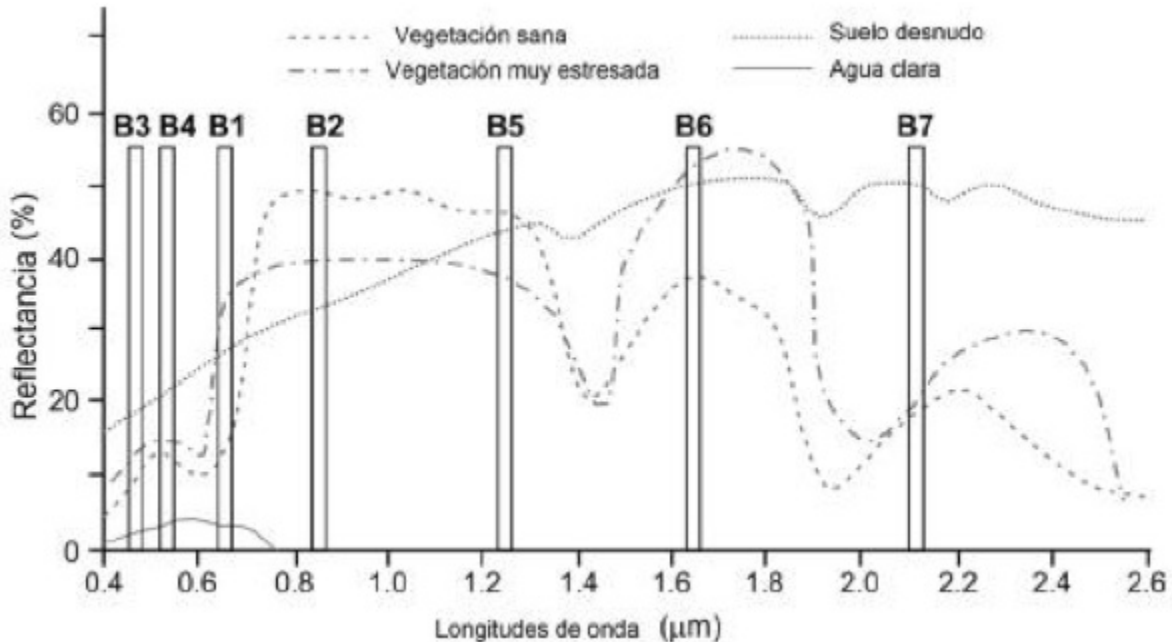
Entre 2010 y 2012 la estratificación se realizó en forma visual por expertos en interpretación de imágenes del Landsat; éste es un método subjetivo, ya que distintas personas pueden conformar los estratos de forma diferente. Por otra parte toda imagen se relaciona con una fecha y un cambio de la fecha puede modificar los resultados.

Advertido de esto en el año 2012 se incorporó el uso del Índice de Vegetación de Diferencia Normalizada (**NDVI**, por sus siglas en inglés), el cual permite estimar el estado de desarrollo de la vegetación a partir de datos espectrales proporcionados por el sensor MODIS. Terra-MODIS y Aqua-MODIS que cubren la superficie de la Tierra en aproximadamente dos días, adquiriendo datos en 36 bandas espectrales de las cuales las Bandas 1 y 2 son las que interesan para la medición de coberturas vegetales.

El Índice aprovecha el particular comportamiento radiométrico de la vegetación sana en las diferentes bandas espectrales, especialmente el visible y el infrarrojo cercano. En el visible se dispone de la banda 1 correspondiente al rojo, donde los pigmentos de la hoja absorben la mayor parte de la energía que reciben del sol y en consecuencia la reflectancia es baja. La banda 2 está ubicada en el infrarrojo cercano y el comportamiento es inverso al anterior, ya que es escasa la energía que absorben los pigmentos de la hoja y la reflectancia es alta.

El siguiente gráfico presenta la ubicación de las 7 primeras bandas del sensor MODIS,

identificadas desde la B1, hasta la B7 en relación de la longitud de onda. Se marcó con un círculo verde el comportamiento normal de la vegetación sana en las bandas 1 y 2, que son las que particularmente interesan.



La banda 1 (rojo) mide la reflectancia en la longitud de onda entre 0,620 μm a 0,670 μm y la banda 2 (infrarojo) en la longitud de onda entre 0,841 μm a 0,876 μm. En el gráfico se observa que la vegetación sana tiene un porcentaje de reflectancia mayor en la banda 2 que en la banda 1. Este contraste es usado para calcular el NDVI para cada pixel, el cual se define como el cociente:

$$NDVI = \frac{\text{Infrarojo cercano} - \text{Rojo}}{\text{Infrarojo cercano} + \text{Rojo}} = \frac{\text{Banda 2} - \text{Banda 1}}{\text{Banda 2} + \text{Banda 1}}$$

El NDVI expresa la relación entre la energía absorbida y emitida, útil para medir la cantidad, salud y vigor de la vegetación. Las variables banda 2 y banda 1 se definen por las medidas de reflexión espectral que adquieren en esas regiones del espectro electromagnético. El índice toma valores en el intervalo [-1, +1], lo que lo convierte en un indicador sumamente útil.

El procedimiento que determina el estado de la vegetación en un pixel consiste en tomar el valor más alto del NDVI durante periodos consecutivos de 16 días. Esto genera una serie de puntos con los cuales se construye una curva multimodal, variable según el tipo de cobertura. En el siguiente ejemplo se presenta la curva generada por un mismo pixel a través de dos años, en los cuales en su superficie se sembró primero trigo, luego soja y finalmente maíz.



El pixel del ejemplo sería clasificado como “*pixel agrícola*”. En caso de que se presente una línea más regular cercana y paralela a la ordenada sería un “*pixel agrícola*”.

El paso siguiente es la delimitación de los estratos: en la imagen satelital se superpone una grilla hexagonal de 2.500 hectáreas de superficie, donde cada hexágono estará compuesto por una cantidad de píxeles -agrícolas y no agrícolas-.

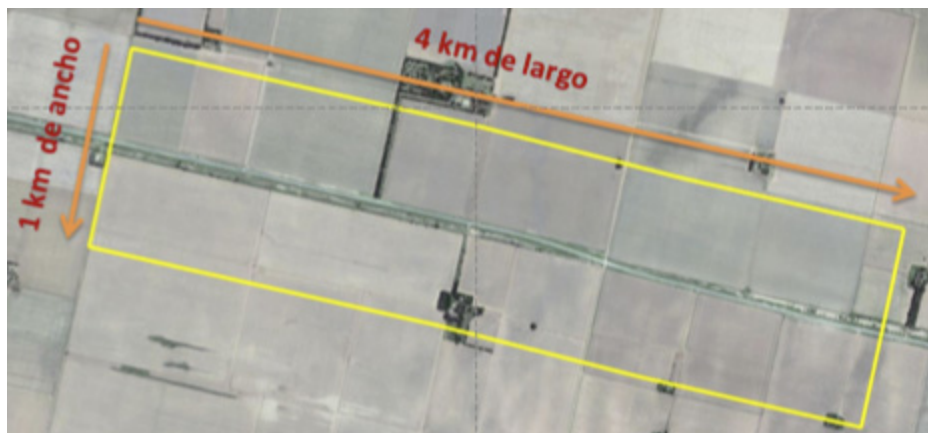
Se calcula para cada hexágono la proporción de píxeles destinados a la actividad agrícola. Si el porcentaje de píxeles agrícolas del hexágono es:

- $P\% > 70\%$ es un hexágono agrícola: A= Agrícola
- $30\% \leq P\% \leq 70\%$ es un hexágono mixto: B= Mixto
- $P\% < 30\%$ es un hexágono ganadero: C= Ganadero

Se agrupan todos los hexágonos de un mismo tipo, y por interpretación de las imágenes resultantes se forman dentro de cada Partido/Departamento, los estratos de uso internamente homogéneo del suelo.

4.5. El segmento como unidad de muestreo

Se mencionó que la Unidad de Muestreo del diseño es el **Segmento** que se define dentro de cada estrato como una superficie con forma de polígono, generalmente rectangular, generada a partir de una ruta o camino identificable, de 4 Km de largo y 500 metros a cada lado. Por tanto, el área total de cada segmento se aproxima a las 400 hectáreas.



Se realiza un esfuerzo para lograr segmentos de igual tamaño a los efectos de cumplir dos objetivos: 1) evitar tener que incorporar la “superficie del segmento” como una variable en un estimador y 2) facilitar la comparabilidad entre segmentos.

En la mayoría de las provincias se logró tener segmentos iguales, las excepciones fueron algunos Departamentos con alto nivel de Descarte que hicieron necesario definir segmentos de tamaño variable; en estos casos se trabaja con *estimadores por razón a la superficie del segmento* en lugar de *simple expansión*.

En cada segmento elegido:

- Se observa en forma directa el 100% de su superficie. Consecuentemente, salvo errores de interpretación, el segmento constituye la *verdad terrestre* o *verdad de campo*.
- Se identifican las *Unidades de Uso de Suelo (UUS)* que son las diferentes coberturas que contiene, sean o no agrícolas.

4.6. Tamaño de la muestra, por Partido/Departamento

El método para estimar las superficies sembradas parte de seleccionar una muestra aleatoria de puntos. Como se necesita suministrar estimaciones para cada Partido/Departamento, la cantidad de puntos -futuros segmentos- se determinan independientemente para cada estrato de estas jurisdicciones.

El tamaño de la muestra es una aproximación que se realiza teniendo en cuenta los criterios estadísticos conjuntamente con el conocimiento de: a) el tipo de Estrato, b) la superficie del Estrato, c) información previamente disponible de la zona y d) comportamiento de los parámetros de los principales cultivos.

El principal procedimiento de estimación del tamaño de muestra utiliza un modelo de regresión logarítmica donde la variable explicativa es la superficie del estrato y la explicada la cantidad de

segmentos. Es decir:

h 10

$$n = A + B \times \log (\text{Superficie Estrato } h)$$

h

Donde: n : tamaño de la muestra del h-ésimo estrato

Otra alternativa también utilizada para estimar el tamaño de la muestra es partir de un Coeficiente de Variación prefijado y estimaciones del error relativo, según la expresión:

$$n_h = \frac{z^2 \cdot (CV_h)^2}{(ER_h)^2}$$

Donde

$$CV_h = \frac{\hat{\sigma}_{Y_h}}{Y_h} * 100 \quad \text{Coeficiente de variación}$$

z = valor de la variable de una distribución normal estandarizada de probabilidad

$$CV_h = \frac{\hat{\sigma}_{Y_h}}{Y_h} * 100 \quad \text{Error Relativo}$$

La suma de puntos elegidos dentro de cada estrato de un Partido/Departamento es el tamaño de la muestra de esta jurisdicción y por agregación del total provincial.

4.7. Selección de los Segmentos

Se seleccionan **puntos aleatorios** sobre la superficie de cada combinación Estrato y Partido/Departamento, con excepción de la zona de Descarte. Es decir que la unidad estadística original es un punto sobre la superficie terrestre. Hasta aquí el diseño de muestra es estrictamente aleatorio.

El inconveniente de este procedimiento es que los puntos caen en cualquier parte de la superficie, frecuentemente dentro de explotaciones y lotes no alcanzables a través de un camino. Esta situación hace necesario reubicarlos a la posición accesible más cercana, evitando producir sesgos significativos por pérdida de aleatoriedad. La solución consiste en trasladar el punto hasta el camino más cercano, tal que pueda arribarse con un vehículo y sin necesidad de contar con la autorización previa del propietario, productor u ocupante de las tierras.

En la práctica se utiliza un Sistema de Información Geográfico (GIS) y se superpone la capa de imágenes satelitales con la de caminos y a partir del punto se crea una circunferencia a la que se incrementa progresivamente el radio, hasta producir el *primer contacto* con un camino identificable. La intercepción "Circunferencia-Camino" define el *punto elegido*.

Dado que no es posible estimar áreas en base a puntos, se procede a transformar el punto en una superficie siguiendo los siguientes pasos:

1. Creación de un segmento lineal a partir del punto elegido sobre el camino: para ello se establecen dos nuevos puntos opuestos y equidistantes, a dos kilómetros del mismo. Se unen para dar origen al segmento lineal de 4 kilómetros de largo.
2. Transformación del segmento lineal en un polígono mediante el trazado de dos líneas paralelas a 500 metros a cada lado del camino.

Surge así una superficie de forma rectangular -excepto por la aparición de alguna curva o algún descarte- de 4 Km de largo por 1 Km de ancho, que salvo pequeños ajustes comprende un área de 400 Ha.

Se considera que desde un camino una distancia 500 metros usualmente es visible para el observador, no obstante deberá encontrar la forma de acercarse a los lotes más alejados.

Una consecuencia del mecanismo de traslado es que dos segmentos podrían solaparse en un mismo camino. En casos como el descrito se separan dejando 2 km. libres entre la finalización de uno de y el comienzo del otro.

Otra ocurrencia asociada al proceso es la necesidad de alterar la forma rectangular de los segmentos ante la presencia de descarte (manteniendo las 400 Ha.). Ejemplo de ello son



Cada segmento cumple con los siguientes requisitos:

- a. Pertenece en su totalidad a un mismo Estrato de un Partido/ Departamento.
- b. Posee un tamaño de 400 hectáreas, con una tolerancia de hasta $\pm 5\%$ (entre 380 y 420 hectáreas). No obstante se tuvieron en cuenta excepciones a esta regla en algunos Departamentos que presentaron estratificaciones complejas.
- c. Ostenta el carácter permanente: Como no existe sesgo por cansancio de la unidad de respuesta ya que no se entrevistan personas, una vez delimitado cada segmento en la cartografía digital no será modificado en los sucesivos operativos, excepto alguna circunstancia que lo justifique. De esta forma la información de cada segmento es una serie de tiempo, de comparabilidad directa.
- d. Se identifican mediante un código de 9 dígitos que define:
 - Provincia = primeros dos dígitos
 - Partido o Departamento = siguientes tres dígitos
 - Estrato = letra A, B o C
 - Orden = últimos tres dígitos

De esta manera no hay posibilidad de que existan dos segmentos con igual numeración. El código facilita la determinación, comparación y la repetición de operativos semejantes en el tiempo.

Previo a la asignación del carácter “permanente” de los segmentos, personal técnico del Ministerio realiza una salida a campo para evaluar “in situ” la factibilidad de acceso y el recorrido de los segmentos. Si surgen anomalías, por ej., caminos cerrados, se informa a Sede Central para que evalúe la situación y actúe en consecuencia.

4.8. El sesgo debido al traslado

Trasladar el punto de su ubicación aleatoria original hasta el camino más cercano se presenta como la única posibilidad de lograr la verdad de campo, ya que la alternativa de consultar el productor no es factible y tampoco lo es estimar a partir de clasificaciones supervisadas o no supervisadas de imágenes satelitales -en Argentina la observación del uso del suelo en las imágenes presenta un alto grado de confusión y los resultados no son aceptables-.

La duda que se plantea es si el corrimiento genera un sesgo que pueda ser significativo. Con el

objetivo de tener una respuesta sobre este Interrogante se implementaron varias pruebas, la principal fue una investigación llevada a cabo en el Partido 25 de Mayo de la Provincia de Buenos Aires que consistió en delimitar pares de segmentos, uno sobre el punto original (segmento original), y el otro sobre el punto trasladado (segmento trasladado).

Mediante acuerdos con asociaciones de productores se logró a modo de excepción los permisos para entrar al interior de los campos y observar el contenido de los lotes de los segmentos originales.

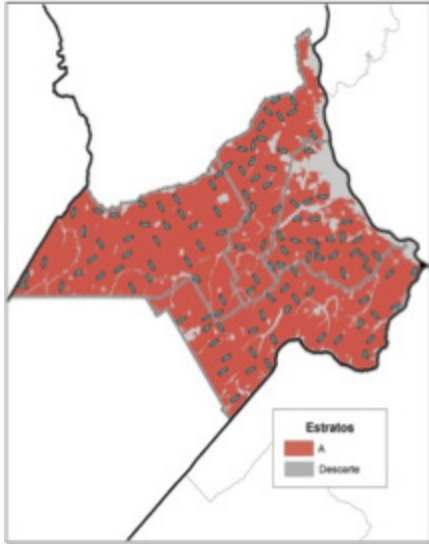
Se llevaron a cabo dos estimaciones por separado, una utilizando el conjunto de segmentos trasladados y otra con el conjunto de segmentos originales.

Los resultados fueron dos estimaciones independientes para cada cultivo del Partido de 25 de Mayo, las que no presentaron diferencias considerables en los cultivos importantes, aunque sí en los cultivos menos extensivos y localizados que no son de interés para el presente estudio. Es decir que como una primera conclusión se considera que en las Provincias de la Región Pampeana el efecto traslado no es significativo en cultivos extensivos, especialmente porque hay suficientes caminos y los traslados son distancias cortas realizadas siempre dentro de un mismo estrato. En provincias extrapampeanas no se han llevado a cabo este tipo de estudios.

4.9. Ejemplo de segmentos ubicados en la Delegación Casilda

El siguiente mapa muestra a los segmentos ubicados en la Delegación Casilda de la Provincia de Santa Fe, la cual comprende los departamentos de: Caseros, Constitución, Rosario y San Lorenzo.

Esta región de la Provincia de Santa Fe y abarca 1.058.626 Ha. Es una zona que se caracteriza por un uso de suelo muy homogéneo, por tanto sólo fue definido un solo estrato agrícola de 934.803 Ha., más un descarte de 123.823 Ha. Se puede observar que se ubicaron en total 114 segmentos.

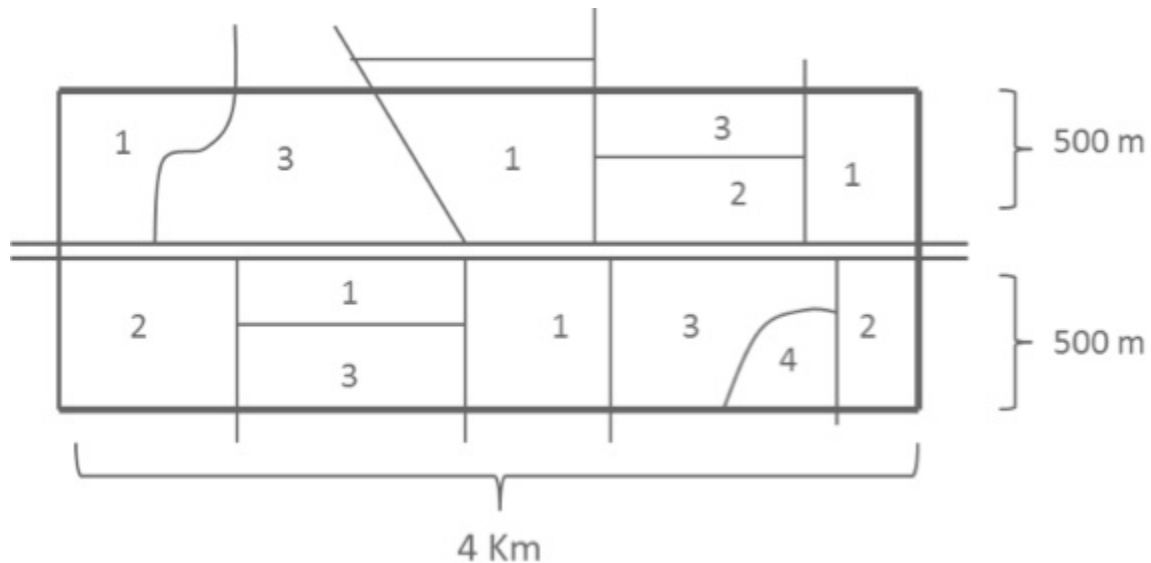


5. Unidades del Uso del Suelo

Las *Unidades de Uso del Suelo (UUS)* son las diferentes coberturas que se le presentan al observador a medida que recorre cada uno de los segmentos, pueden ser agropecuarias o no agropecuarias y constituyen la *verdad de campo*.

Cada segmento se integra de una cantidad variable de UUS.

Se supone el siguiente ejemplo de un segmento compuesto con solo cuatro tipos de UUS:



1 = Trigo

2 = Cebada Cervejera

3 = Avena
4= Caserío

De los datos del cultivo y la superficie de cada lote, surge el siguiente cuadro:

Cobertura	Superficie en hectáreas
1 – Trigo	125
2 - Cebada Cervecera	90
3 – Avena	130
4 - Desperdicio (caseríos, camino, otros)	55
TOTAL	400

No obstante que los usos no agrícolas identificables son previamente descartados del marco, dentro de cada segmento siempre aparecen diferentes tipos de *Desperdicios*, frecuentemente transitorios que no fueron eliminados, tales como: rutas y caminos, viviendas rurales, parques, arboledas, vías férreas, canales, aeródromos, cementerios, instalaciones deportivas, esteros, médanos, floraciones rocosas, pendientes pronunciadas, cañadas, anegamientos temporales, etc. Este conjunto no reviste interés, no obstante se lo estima a partir de una variable que se define como la diferencia entre la superficie total del segmento y la superficie agrícola relevada.

Las variables del estudio que interesan son los *usos agropecuarios* del suelo. En total se consideran 33 usos:

Algodón	Alpiste	Arroz	Arveja Seca	Avena	Barbecho Convencional
Barbecho Químico	Campo Natural	Cártamo	Cebada	Centeno	Colza
Garbanzo	Girasol de Primera	Girasol de Segunda	Lenteja	Lino	Maíz de Primera
Maíz de Segunda	Maní	Mijo	Otros Cultivos	Pasturas Permanentes	Poroto
Potrero	Rastrojo de fina	Rastrojo de gruesa	Soja de Primera	Soja de Segunda	Sorgo Granífero 1º
Sorgo Granífero 2º	Trigo	No Relevado			

La variable superficie “No Relevada” corresponde a toda UUS que pese a haber sido delimitada en el segmento, por imposibilidad de acceso o incluso por omisión, no se le identificó la cobertura. En la práctica no reviste significación.

6. El Operativo a Campo

6.1. Tareas de los Delegados

La obtención de la información básica de cada segmento es responsabilidad de los Delegados del MAGyP con asiento en el interior del país, estos cumplen el rol de *observador*. A cada Delegado se le suministra el material necesario para cumplir adecuadamente su tarea:

1. Archivos SIG por Partidos/Departamento, de Segmentos, de UUS, de cartografía base, de rutas y de ubicación de principales ciudades y pueblos.
2. Imágenes satelitales del Lansat 8.
3. Archivos GPS con las coordenadas de inicio y fin de cada segmento.
4. Planillas de campo varias, útiles en el relevamiento.

El observador con el GPS marca los waypoints de todas las UUS que va encontrando, a ambos lados de la línea de recorrido del segmento, como también las UUS ubicadas lejos de la línea de recorrido pero que están dentro del segmento.

Terminadas estas tareas, remite la información mediante FTP (Protocolo de transferencia de archivos) a la Oficina Central de Estimaciones.

6.2. Tareas en la Oficina Central de Estimaciones

Una vez recibida la información, se realiza un *análisis de consistencia y coherencia*, y si aparecen dudas al comparar lo relevado con lo visualizado en la imagen satelital se consulta con el Delegado responsable.

Se procesa la información y se elaboran las *Planillas Resumen de la Muestra*, de la cual surge para cada Partido/Departamento el cruzamiento de los segmentos como filas y las diferentes coberturas como columnas. A partir de las mismas, se ejecuta el proceso de expansión y por el programa SPSS se obtienen las estimaciones de las superficies de cada tipo de cobertura.

Se consideran dos tipos de estimadores con sus respectivas variancias:

- *Estimadores por simple expansión*, para Partidos/Departamentos donde los segmentos son de igual tamaño (400 Ha).
- *Estimadores por razón separada*, para Partidos/Departamentos donde los segmentos son de diferente tamaño.

7. Generación de los estimadores

7.1. Definiciones básicas de la teoría del muestreo

El objetivo primario de toda *investigación por muestreo es estimar parámetros* -valores cuantitativos desconocidos- de una *población* en estudio. Para ello se utiliza el procedimiento de encuestar u “*observar*” a una *muestra* de las unidades que conforman la población. Si la selección se realiza por algún procedimiento aleatorio aceptado por la teoría del muestreo, la muestra es probabilística.

En el presente estudio los parámetros corresponden a las superficies en hectáreas de cada una de las diferentes coberturas del suelo en cada Partido/Departamento y en una campaña especificada.

Por su parte los *estimadores* se definen como las expresiones matemáticas construidas a partir de los datos de la muestra. Como su nombre lo indica tienen como objetivo la estimación de los parámetros.

Toda estimación originada por el método del muestreo tiene dos tipos de errores, los *no debidos* al proceso de muestreo y los *debidos* al hecho de observar parcialmente la población.

- Los no debidos al muestreo, son difíciles de medir y en consecuencia rara vez se los llega a conocer; en general dependen de la forma de captación de los datos, luego se pueden reducir con un buen análisis de *consistencia y coherencia* de la información básica. En la presente investigación es de suponer que este tipo de error no es significativo debido a que los datos se obtienen por observación directa por personal experto y conocedor del sector, con estrictos protocolos, control por GPS y con áreas consistidos y medidas con imágenes satelitales. No obstante en la selección aparece un corrimiento del segmento original hacia caminos y es causa de un surgimiento de error no debido al muestreo.
- Los debidos al proceso de muestreo que no se pueden anular, pero en el caso de una muestra probabilística se los puede medir. Los dos principales indicadores son el *Error Estándar* (E.Std) y el *Coefficiente de Variación* (CV).

Ambas medidas están afectadas por el tamaño de la muestra y en consecuencia, salvo excepción, serán mayores a nivel de Partido/Departamento que a nivel de una región. Estos indicadores de la variabilidad cumplen con la condición general de “cuando más pequeño es mejor”.

- El **Error Estándar** es una medida de la variabilidad dada en cifras absolutas, en la presente investigación expresa el grado de precisión con que la estimación de cada cultivo se aproxima a la verdadera cantidad de hectáreas sembradas.
- El **Coefficiente de Variación** es una cifra relativa expresada en porcentaje. Esta dada por el cociente entre las estimaciones del error estándar y la superficie estimada de cada cobertura.

CV =

Error Estándar
Es

x 100

El CV asocia la confiabilidad que necesita la investigación con cada parámetro. En la presente se utiliza la siguiente regla:

Estimación del CV	Confiabilidad de la estimación
$CV < 5,0 \%$	Excelente
$5,0 \% \leq CV < 10,0 \%$	Muy Buena
$10,0 \% \leq CV < 15,0 \%$	Buena
$15,0 \% \leq CV < 20,0 \%$	Poco confiable pero aceptable
$20,0 \% \leq CV < 25,0 \%$	Escasamente confiable
$CV > 25,0 \%$	No confiable

Cuando el CV supera el 25%, se interpreta que las superficies sembradas presentan mucha variabilidad y la estimación debe ser tomada cuidadosamente.

7.2. Estimadores por simple expansión de cada tipo de cobertura

7.2.1. Cantidad de hectáreas con una cobertura determinada en un segmento de un estrato

Dentro de un determinado segmento cada una de las diferentes coberturas de la simboliza se la simboliza.

$$h_i =$$

y

$$\sum_{j=1}^{m_{hi}}$$

hi

y

(1)

Dónde

hij

y : Cantidad de hectáreas con una determinada cobertura que posee el j-ésimo lote, del i-ésimo segmento, del estrato h de un Partido/Departamento.

hi

m : Total de lotes con una determinada cobertura que posee el i-ésimo segmento, del estrato h de un Partido/Departamento.

hi

Sup.Med : Cantidad de hectáreas medidas (reales) que posee el i-ésimo segmento del estrato h, de un Partido/Departamento.

hi

y : Cantidad de hectáreas con una determinada cobertura que posee el i-ésimo segmento, del

estrato h, de un Partido/Departamento.

La superficie de los segmentos tienen una variación tolerable en hectáreas de:

h_i

380 Ha. \leq Sup.Med. \leq 420 Ha.

Por tal motivo, la formula (1) adiciona el factor de ajuste proporcional dado por:

$$F^{hi} =$$

h_i

F : Salvo excepciones arroja un valor muy cercano a la unidad, al multiplicar cada cobertura del segmento y hace que estas se refieran a una superficie de exactamente 400 Ha.

La población de segmentos que contiene cada estrato, se calcula como el cociente entre la superficie total del estrato sobre 400 hectáreas.

$$N_h =$$

N_h

Dónde: Sup.Est : Cantidad de hectáreas que posee el estrato h .

7.2.2. Estimador por simple expansión del total de hectáreas de una cobertura en un estrato h .

Está dado por:

$$\hat{y}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi} = \frac{Sup.Est_h}{n_h} \sum_{i=1}^{n_h} \frac{1}{Sup.Med_{hi}} \sum_{j=1}^{m_{hi}} y_{hij} \quad (2)$$

Dónde:

N_h

n_h

: Estimador por simple expansión del total de hectáreas de una determinada cobertura sembrada

en el estrato h

y_h

n_h : Cantidad de segmentos seleccionados que integran la muestra del estrato h

7.2.3. Estimador por simple expansión del total de hectáreas de una cobertura de un Partido/Departamento

Se obtiene como una suma de los estratos especificados en el Partido/Departamento

$$\hat{y} = \sum_{h=1}^L \hat{y}_h \quad (3)$$

El mismo estimador expresado según la simbología dada por Horvitz y Thompson.

$$\hat{y} = \sum_{h=1}^L \sum_{i=1}^{n_h} \frac{y_{hi}}{\pi_{hi}} \quad (4)$$

y_{hi}

Dónde π_{hi} es la Probabilidad de que el i-ésimo segmento de la población del estrato h, integre la

y_h

y_h

muestra en una selección simple al azar de “n ” unidades sin reemplazo de la población N .

$$\pi_{hi} = \frac{n_h \text{ Sup. Med}_{hi}}{\text{Sup. Est}_h}$$

$$h_i \quad h_i$$

En el procesamiento se considera $w = 1/h_i$ como el factor de expansión.

7.2.4. Estimadores por Simple Expansión del Error Estándar y el Coeficiente de variación del Total de Hectáreas de una cobertura de un Partido/Departamento

El estimador por simple expansión de la variancia del estimador de un total de una cobertura para un Partido/Departamento, viene dada

$$\hat{\sigma}_y^2 = \sum_{h=1}^L \frac{N_h^2(1-f_h)}{(n_h-1) \cdot n_h} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_{hi})^2 \quad (5)$$

El error estándar del estimador de un total de una cobertura para un Partido/Departamento

$$\hat{\sigma}_y = \sqrt{\hat{\sigma}_y^2} \quad (6)$$

Y el estimador del coeficiente de variación de un total de una cobertura para un Partido/Departamento

$$CV = \frac{\hat{\sigma}_y}{\bar{y}} \times 100 \quad (7)$$

7.3. Estimadores por razón de cada tipo de cobertura

7.3.1. Estimadores por razón del total de hectáreas con una cobertura en un Estrato

Cuando los segmentos seleccionados presentan variaciones en su superficie, el tamaño del segmento se convierte en una variable aleatoria que usualmente esta correlacionada positivamente con cada tipo de cobertura.

El estimador por razón utiliza ambas variables, la superficie en hectáreas de la cobertura en particular y la superficie en hectáreas del segmento. Toma la forma.

$$\hat{y}_{(R)h} = \frac{\sum_{i=1}^{n_h} y_{hi}}{\sum_{i=1}^{n_h} x_{hi}} \cdot X_h = r_h \cdot X_h \quad (8)$$

Donde:

\hat{y}_{hi}

$(R)h$

: Estimador por razón separada del estrato h

h_i

y : Cantidad de hectáreas con una determinada cobertura en el i -ésimo segmento seleccionado del estrato h

h_i

x : Cantidad de hectáreas del i -ésimo segmento seleccionado del estrato h

h

r : Razón entre la superficie de una determinada cobertura y la cantidad de hectáreas del total de segmento seleccionados en la muestra en el estrato h .

h

X : Superficie total de hectáreas del estrato h . Este es un parámetro poblacional previamente conocido ya que corresponde a la superficie satelital medida del estrato.

7.3.2. Estimadores por razón del total de hectáreas de una cobertura en un Partido o Departamento.

Es la suma a través de los estratos

$$\hat{y}_{(R)} = \sum_{h=1}^L \hat{y}_{(R)h} \quad (9)$$

Donde:

□

(R)

: Estimador por razón separada para un Departamento/Partido

7.3.3. Estimador por razón separada del error estándar y del coeficiente de variación del total de una cobertura en un Partido o Departamento

El estimador de la variancia del estimador de razón es:

$$\hat{\sigma}_{\hat{y}_{(R)}}^2 = \sum_{h=1}^{h=L} \frac{N_h^2(1-f_h)}{n_h} \cdot \frac{\sum_{i=1}^{n_h} (y_{hi} - R_h x_{hi})^2}{n_h - 1} \quad (10)$$

El estimador del error estándar de razón

$$\hat{\sigma}_{\hat{y}_{(R)}} = \sqrt{\hat{\sigma}_{\hat{y}_{(R)}}^2} \quad (11)$$

El estimador del coeficiente de variación

$$CV_{(R)} = \frac{\hat{\sigma}_{(R)}\hat{y}}{\hat{y}_{(R)}} 100 \quad (12)$$

8. Estimadores por intervalos de confianza

Las fórmulas anteriores corresponden a estimadores puntuales, en el estudio también se elaboran para cada Partido/Departamento estimadores por intervalos e confianza de las diferentes coberturas. Los que responden a la expresión general:

$$\hat{y} \pm z_{\left(1-\frac{\alpha}{2}\right)} \hat{\sigma}_y \quad (13)$$

Donde

\hat{y} : Estimador puntual de una cobertura (por simple expansión o por razón) para un Partido/Departamento

$\hat{\sigma}_y$: Estimador puntual del error estándar de una cobertura (por simple expansión o razón) para un Partido/Departamento.

z: Valor de la variable de la distribución normal estandarizada correspondiente a una confianza

(0,95)

en probabilidad de “ $1-\alpha = 0,90$ ”. En este caso en particular $z = 1,645$.

9. Presentación de la información

Como resultado de la muestra se presentan cuadros para cada Partido/Departamento con las diferentes coberturas en las campañas de cosecha fina y de cosecha gruesa, con las siguientes estimaciones sobre la superficie en hectáreas sembradas y otras coberturas,

- Estimación de la superficie sembrada de cada cobertura considerada a nivel de Partido o Departamento.
- Estimación de los errores debidos al muestreo de los estimadores considerados en el punto anterior
- Estimación del coeficiente de variación porcentual de cada cobertura considerada a nivel de Partido o Departamento.
- Estimación por niveles de confianza del 90%. Estos son elaborados a nivel de Partido y de

departamento utilizando la distribución de probabilidad “t” de Student o la distribución Normal.

Bibliografía

Ministerio de Agricultura Ganadería Pesca de la República Argentina, Dirección de Información Agropecuaria y Forestal – Año 2013 - Método de segmentos aleatorios, Metodología para la estimación de la superficie sembrada con cultivos extensivos – Versiones 2 y 3 -

FAO Organización de las Naciones Unidas para la Agricultura y la Alimentación – Encuestas Agrícolas con múltiples marcos de muestreo – Año 1996 - Volumen 1.

Cochran, Willams .G. – Año 1963 - Sampling Techniques – 3 ed. – Wiley.

Tzitziki Janik García-Mora, Jean-François Mas - Evaluación de imágenes del Censor Modis - Vol. 63, N° 1 – Año 2011 - Boletín de la Sociedad Geológica Mexicana.

Scheaffer, Mendenhall, Ott – Año 1987 - Elementos de muestreo – Grupo Editorial Iberoamérica.