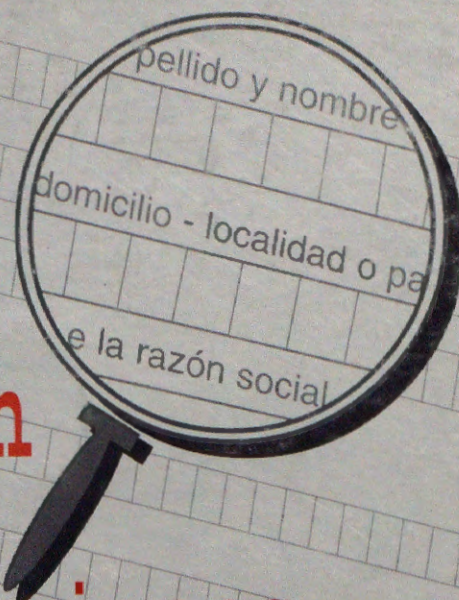


Uso
de técnicas
multivariadas
para la
detección
de posibles
adulteraciones
en encuestas



IDENTIFICACIÓN DE LA EXPLOTACIÓN Y DEL PRODUCTOR

Jefe de Supervisión

Cod.

Supervisor

2. Identificación del productor

Apellido y nombre de productor o denominación de la razón social

Apellido y nombre
domicilio - localidad o paraje
e la razón social

3. Otros datos

Calle / Ruta, N° / Km, Piso, Depto. (solo si difiere de la dirección de la explotación)

Cod. post.

Localidad o paraje más cercano

Cas. post.

Correo o estafeta postal

Teléfono (nueva numeración)

Celular o 3° teléfono (nueva numeración)

Fax o 2° teléfono (nueva numeración)

4.1 Hectáreas

4.2 Parc.



REPÚBLICA ARGENTINA
MINISTERIO DE ECONOMÍA

SECRETARÍA DE PROGRAMACIÓN ECONÓMICA Y REGIONAL

INSTITUTO NACIONAL DE ESTADÍSTICA Y CENSOS

Período de referencia: 1° julio 1998 - 30 junio 1999

5. Explotaciones sin límites

Reliene

35

ESTUDIOS

INDEC

SI/10.35
EJ 8

SLA

C.E.S.

Centro Estadístico de Servicios
DIRECCIÓN DE DIFUSIÓN
I. N. D. E. C.

ISBN 950-896-243-7

31 MAR 2000

Uso
de técnicas
multivariadas
para la
detección
de posibles
adulteraciones
en encuestas



REPÚBLICA ARGENTINA

MINISTERIO DE ECONOMÍA

SECRETARÍA DE PROGRAMACIÓN ECONÓMICA Y REGIONAL

INSTITUTO NACIONAL DE ESTADÍSTICA Y CENSOS

35

ESTUDIOS

INDEC

La serie ESTUDIOS INDEC tiene por objeto dar a conocer los resultados de investigaciones de carácter cuantitativo en los campos sociodemográficos y económicos mediante el aprovechamiento exhaustivo y adecuado de la información estadística disponible.

Con estas publicaciones el INDEC habrá de proporcionar a los usuarios, además de las series habituales, trabajos analíticos con objetivos definidos, así como instrumentos conceptuales y metodológicos que auxilien, orienten y abrevien las investigaciones en los campos mencionados.

La presente publicación resume tareas llevadas a cabo por la Dirección de Metodología Estadística, a cargo de Norma Pizarro de Pereira. El trabajo fue realizado por Adriana Semorile y Martín Donadio.

SIGNOS CONVENCIONALES UTILIZADOS POR EL INDEC

Para la sustitución o complementación de un dato numérico el INDEC utiliza los siguientes signos, según el caso:

- * Dato provisorio
- Dato igual a cero
- 0 Dato igual a cero por redondeo de cifra positiva
- 0 Dato igual a cero por redondeo de cifra negativa
- . Dato no registrado
- ... Dato no disponible a la fecha de presentación de los resultados
- /// Dato que no corresponde presentar debido a la naturaleza de las cosas o del cálculo
- s Dato confidencial por aplicación de las reglas del secreto estadístico
- e Dato estimado por extrapolación, proyección o imputación

© Queda hecho el depósito de la Ley N° 11.723
Buenos Aires, 2000

Responsable de la edición Prof. Leonardo Piatti

PERMITIDA LA REPRODUCCIÓN PARCIAL CON MENCIÓN DE LA FUENTE

PUBLICACIONES DEL INDEC

Las publicaciones editadas por el Instituto Nacional de Estadística y Censos se encuentran a la venta en INDEC, Centro Estadístico de Servicios, Julio A. Roca 615, PB. 1067 Buenos Aires, Argentina.

Por cualquier consulta puede dirigirse personalmente al Centro Estadístico de Servicios, o bien comunicarse a los teléfonos 4349-9650/52/54/62, al fax: 4349-9621, o a través de correo electrónico E-Mail: CES@indec.mecon.ar.

Horario de atención de 9,30 a 16,00.

SI/10.35

SACA

Ey 8

3

S.E.S.

Centro Estadístico de Servicios
DIRECCIÓN DE DIFUSIÓN
I. N. D. E. C.

contenido

31 MAR 2000

	página
1. INTRODUCCIÓN	5
2. ENCUESTA NACIONAL DE GASTOS DE LOS HOGARES 1996/97	7
3. ESQUEMA DE CONTROL DE CALIDAD	9
4. APLICACIÓN DE MÉTODOS MULTIVARIADOS	11
4.1 Estudios preliminares	11
4.2 Estudio de componentes principales (ACP) con clasificación jerárquica	12
4.3 Análisis de correspondencias simples	13
4.4 Estudios descriptivos	14
4.4.1 Encuestadores con medianas menores - Box-plots	14
4.4.2 Encuestadores asociados a encuestas atípicas en monto de gasto - Box-plots	15
4.5 Estudios complementarios	15
4.5.1 Aplicación de REGARIMA y ROBREGAR con enfoque de series de tiempo para detectar outliers.	15
4.5.2 Aplicación de modelo lineal generalizado para detectar encuestas con gastos significativamente distintos (menores)	15
5. EJEMPLO DE LA APLICACIÓN DE LAS TÉCNICAS DESCRIPTAS A UNA CIUDAD	15
Cuadro 5.1 ACP-Cantidad de registros F2 y F3	17
Cuadro 5.2 ACP-Monto de gasto F2	19
Cuadro 5.3 Análisis de Correspondencias Simples	20
Cuadro 5.4 Encuestadores con medianas menores	22
Cuadro 5.5 Encuestadores más atípicos en box-plots	23
6. CONCLUSIONES	25
7. REFERENCIAS BIBLIOGRÁFICAS	27

1. Introducción

La precisión de los resultados de las encuestas se verían afectados seriamente si no se efectuara un control de la calidad del trabajo del encuestador que permita, entre otros controles, minimizar la cantidad de encuestas adulteradas.

La adulteración de las encuestas es un hecho difícil de detectar y existen pocos estudios sobre el tema. Habitualmente, se controla a través de trabajos de supervisión y reentrevista incluidos dentro de los programas de control de calidad de las encuestas, pero siempre se verifica sólo una porción del trabajo de cada encuestador.

En el presente trabajo se aborda este problema desde un ámbito diferente: el metodológico. Bajo el supuesto que un encuestador adultera encuestas bajo un patrón de consumo que él considera habitual, la búsqueda de encuestas adulteradas se lleva a cabo empleando técnicas de análisis multivariado de datos. En primer término se aplican métodos factoriales de reducción de dimensiones tales como componentes principales y correspondencias simples, tratando luego de encontrar encuestas parecidas mediante clasificaciones jerárquicas. El empleo de técnicas multivariadas para el control de la calidad fue utilizado con anterioridad por la Dirección de Metodología Estadística en el marco de la Encuesta Permanente de Hogares. En esa oportunidad el objetivo era el control de las tasas de no respuesta. Se utilizó la técnica de clusters para agrupar áreas de acuerdo a los valores de las tasas de no respuesta (rechazo, ausencia y general). Luego se analizó el trabajo de los encuestadores en los que al menos el 50% de sus áreas pertenecían a clusters con valores altos de las tasas.

La eliminación de la adulteración pasa no sólo por la detección de las encuestas adulteradas sino también por la eliminación de las causas "desmoralizadoras" del trabajo del encuestador (trabajo de campo dificultoso o una remuneración inadecuada, entrenamiento deficiente, cuestionarios mal diseñados o con preguntas contradictorias, etcétera) mediante un diseño cuidadoso e inteligente de las muestras y estableciendo amplias oportunidades para la mejora de la calidad de su trabajo.

2. Encuesta Nacional de Gastos de los Hogares 1996/97

Se trató de la primera encuesta con cobertura nacional urbana para localidades de 5.000 y más habitantes. Su propósito fundamental fue obtener las ponderaciones para construir el índice de precios al consumidor.

En la encuesta se investigan como variables temáticas principales, los gastos y los ingresos del hogar, además de características sociodemográficas del hogar y de sus miembros.

Dentro del gasto del hogar se estudian los gastos corrientes y de capital de los hogares, sin incluir los gastos relativos a sus actividades económicas. Los primeros incluyen los gastos de consumo final y los gastos no de consumo (impuestos, transferencias corrientes privadas, contribuciones sociales). Los segundos comprenden las compras de viviendas, objetos valiosos u otros activos que usualmente se engloban bajo el nombre de "gastos de capital".

Dentro de los ingresos se investigan tanto las entradas en dinero como en especie y las variaciones patrimoniales.

Para el relevamiento de la información se utilizaron cuatro cuestionarios:

Cuestionario 1: Características sociodemográficas de los hogares (F1)

Cuestionario 2: Gastos diarios. El hogar registraba el gasto de la semana de referencia en alimentos y en otros bienes y servicios de consumo frecuente (F2).

Cuestionario 3: Gastos varios. Se registraban por recordación los gastos correspondientes a bienes y servicios adquiridos en períodos de tiempo comprendidos entre el mes y el año anterior a la semana de entrevista (F3).

Cuestionario 4: Ingresos. Se consignaban los ingresos que percibieron cada uno de los miembros del hogar, en los seis meses anteriores a la semana de entrevista (F4).

La encuesta se realizó bajo el principio de centralización metodológica del INDEC y descentralización ejecutiva en las Direcciones de Estadística de cada una de las provincias. Cada una de ellas se hizo cargo del relevamiento, codificación, control y depuración de la información ingresada. La captura de la información se realizó en los 11 centros de procesamiento que dependen del INDEC y se transmitía periódicamente a INDEC para su procesamiento definitivo.

Los hogares hallados en cada una de las viviendas seleccionadas estuvieron bajo estudio durante una semana. A esta semana se la denominó "semana de la encuesta". Se relevó información durante 48 semanas, en cada una de las ciudades. La carga de trabajo por encuestador fue de seis viviendas por semana. El hogar era visitado tres veces durante la semana a fin de completar los cuestionarios.

La forma de pago a los encuestadores fue por encuesta. Las efectivamente realizadas tenían un pago mayor que las no respondidas.

3. Esquema de control de calidad

Con el objetivo de controlar la calidad de la información y fundamentalmente disminuir los errores no muestrales (defectos en el trabajo de los encuestadores, declaraciones incorrectas, falta de respuesta, errores de codificación, errores de ingreso, etcétera), la Encuesta estableció un esquema de control en dos etapas: la primera con controles aplicados durante el relevamiento, recepción, análisis e ingreso de las encuestas en el sistema y la segunda donde el Equipo Central, a partir de información acumulada de varias semanas y de todo el país, controló la marcha del operativo sobre la base de la evaluación de distintos indicadores.

La tarea del Equipo Central se dificultó porque no existía presupuesto suficiente para viajes frecuentes del Equipo a las distintas provincias para efectuar supervisiones directas (reentrevistas). Era necesario encontrar un método que desde el Equipo Central permitiera apuntar a las encuestas sospechosas de anomalías.

Por otra parte, el objetivo era no sólo mejorar la calidad del resultado del trabajo de campo sino también impedir las adulteraciones y el trabajo de baja calidad del encuestador por el efecto disuasivo, que puede tener la detección de las adulteraciones, sobre todo el personal de la encuesta.

La Dirección de la ENGH estableció controles mediante el análisis de las tasas de respuesta, el estudio de las estructuras de gasto, el control de las estructuras poblacionales, el análisis de valores extremos, perfiles provinciales según los registros en gastos diarios y por recordación, evolución periódica del volumen de información en cada provincia por tipo de gastos, etcétera.

La Dirección de Metodología Estadística (DME) ofreció un conjunto de estudios orientados a detectar encuestas con problemas, basado en los datos disponibles en ese momento: montos de gastos por rubro de gastos diarios (F2)¹ y cantidad de registros² por rubro de gastos diarios y gastos por recordación³ (F2 y F3).

Se utilizaron los registros como una aproximación al dato de monto de gasto debido a que estos últimos no se encontraban aún disponibles para los gastos por recordación (F3).

¹ Los gastos diarios incluyen todos los rubros de gastos de alimentos y de no alimentos (lavadero, artículos de tocador, productos medicinales, combustibles, transporte público, correo y teléfono público, textos y útiles escolares, cigarrillos, artículos de limpieza, etcétera)

² Se entiende por registros a cada una de las compras efectuadas por los hogares de cualquier bien o servicio, es decir cada línea de los cuestionarios.

³ Los gastos por recordación se clasificaron según el período de referencia en gastos mensuales, bimestrales, trimestrales, semestrales y anuales. A su vez los gastos mensuales fueron subclasificados en tres categorías: indumentaria, cuotas pagadas y otros gastos mensuales.

4. | A p l i c a c i ó n | d e | m é t o d o s | m u l t i v a r i a d o s |

El objetivo fue tratar de señalar encuestas posiblemente adulteradas. La aplicación de estos métodos se realizó bajo la hipótesis que un encuestador adultera encuestas en forma sistemática, es decir repite un determinado patrón de consumo habitual para sí mismo. Se trata entonces de determinar cuáles son esas encuestas parecidas entre sí y quién es el encuestador característico para investigarlas.

Si los encuestadores que adulteran encuestas, lo hacen sin un patrón subyacente, es muy probable que esas encuestas no sean detectadas por este tipo de método.

Un objetivo secundario fue detectar encuestas que, si bien no estaban adulteradas, eran parecidas debido a la escasez de gastos registrados en ellas y estaban asociadas a un encuestador en particular. Estas encuestas debían revisarse porque podían ser producto del trabajo poco aplicado (a desgano) del encuestador.

Se estudiaron todas las localidades de más de 500.000 habitantes para identificar posibles encuestas adulteradas o con subdeclaraciones mediante la aplicación del método de componentes principales con clasificación jerárquica y del análisis de correspondencias simples con clasificación jerárquica.

4.1. ESTUDIOS PRELIMINARES

En las primeras pruebas, estaba claro el objetivo pero no cómo lograrlo. Se probaron métodos de clasificación (clusters) y métodos de reducción de dimensiones con clasificación posterior (componentes principales con clasificación, correspondencias simples con clasificación posterior).

Se hicieron algunas pruebas aplicando clusters pero en ese caso las encuestas con valores extremos podían afectar la constitución de los grupos oscureciendo la lectura de la clasificación. El método de clusters se aplicó a las provincias de Chubut y Córdoba.

Los métodos de reducción de dimensiones permitieron evitar los valores extremos al efectuar la clasificación jerárquica a partir de la información proveniente de los ejes retenidos.

Otro tema de discusión fue la cantidad de grupos a pedir en la clasificación. En un inicio se pidió una gran cantidad para permitir que el método detectara pequeños grupos de encuestas "parecidas". Se probaron las alternativas de 10, 20 y 30 grupos. En ninguno de estos casos se analizaba cuál era la cantidad óptima de grupos de acuerdo a los índices de nivel de agregación de los nudos de la jerarquía. Este procedimiento se aplicó a las provincias de Buenos Aires y Formosa y al Área Metropolitana.

También se probó otra alternativa que consistía en aplicar componentes principales pidiendo la cantidad de grupos óptima sugerida por los índices del nivel de agregación de los nudos de la jerarquía y repetir el procedimiento de aplicación de componentes principales con clasificación dentro de las clases donde aparecían encuestadores característicos. De esta forma se pensó que, al respetar la cantidad óptima de grupos, no se rompían las estructuras internas de los datos. Pero los resultados obtenidos no fueron satisfactorios en cuanto a detectar encuestas adulteradas, por lo que fue abandonado.

Finalmente se estableció que el procedimiento más conveniente era la aplicación del método de componentes principales con clasificación jerárquica posterior en 20 grupos y la aplicación del análisis de correspondencias simples con clasificación jerárquica posterior en 20 grupos sobre las variables monto de gasto (F2) y cantidad de registros (F2 y F3) que se describen a continuación. Este procedimiento se aplicó para la ciudad de Rosario, el resto de la provincia de Santa Fe, ciudades de Mendoza, La Plata, Córdoba y Tucumán.

4.2. ESTUDIO DE LAS COMPONENTES PRINCIPALES CON CLASIFICACIÓN JERÁRQUICA

El estudio se orienta a la detección de grupos de encuestas con patrones comunes y permite la identificación de los encuestadores característicos de esos grupos. El porcentaje de encuestas realizadas por un encuestador dentro de un grupo dado, es el criterio usado para decidir si tal encuestador es característico del grupo o no.

Es importante señalar que los grupos obtenidos no se refieren obligadamente a encuestas adulteradas sino a encuestas similares entre sí, ya sea en la cantidad de registros por rubro y período de referencia o en el monto de gasto para cada rubro. Las encuestas pueden ser parecidas porque tienen características comunes que son reales. La revisión de las encuestas o la confrontación con el resultado de otros estudios permite determinar si se trata de encuestas adulteradas o reales.

El método de componentes principales trabaja sobre variables cuantitativas y permite reducir la dimensión del problema, mientras que la clasificación posterior posibilita construir grupos a partir del conjunto de las encuestas de acuerdo a una noción de distancia entre las mismas. Así, las encuestas parecidas en la cantidad de registros para cada tipo de gasto forman parte del mismo grupo. Los encuestadores se consideran como variables ilustrativas y es posible individualizar en cada grupo a un encuestador que resulta muy característico.

El análisis en componentes principales se utiliza para describir una matriz de valores numéricos donde las columnas son las variables y las filas los individuos. En este caso las columnas son los tipos de gastos y las filas las encuestas:

Encuestas	Tipo de gastos				
	Pan y cereales	Azúcar y dulces			
1					Cigarrillos
2					
....					
n					

y en cada celda se detalla la cantidad de registros o el monto de gasto -según corresponda- relevado por cada encuestador en cada tipo de gasto.

Un individuo (encuesta) puede representarse por un punto en el espacio de las variables (tipos de gastos) y una variable (tipo de gasto) como un punto en el espacio de los individuos (encuestas).

Dos individuos (encuestas) son considerados tanto más semejantes cuanto mayor sea el número de valores similares que presentan sobre el conjunto de las variables. La comparación entre dos individuos se realiza mediante el cálculo de la distancia euclídea entre ellos. La relación entre dos variables (tipos de gastos) se estudia mediante el coeficiente de correlación.

Es posible tratar a algunas variables como elementos suplementarios, proyectándolas sobre los ejes determinados por las otras variables llamadas activas. En este caso se consideraron como variables suplementarias el código del encuestador y la semana de la encuesta.

Sobre las encuestas, ahora caracterizadas por sus 10 primeras coordenadas factoriales, se efectuó una clasificación jerárquica⁴. El criterio de agregación utilizado es el criterio de Ward.

El resumen final de los resultados listaba, en orden de importancia, sólo los grupos que se consideraban interesantes en cuanto a cuán característico era el encuestador señalado en él. Se detallaba el código del encuestador, el tamaño del grupo, el porcentaje del grupo realizado por el encuestador (en porcentaje de encuestas) y el tipo de registros o de gastos que caracterizaban al grupo. Mediante archivos adjuntos se informaban los datos que permitían individualizar las encuestas integrantes de cada grupo.

Los grupos pequeños (con pocas encuestas), en donde aparecían encuestadores característicos que habían relevado un porcentaje considerable de las encuestas del grupo, eran los indicados para efectuar una revisión especial.

Los grupos numerosos (con muchas encuestas) caracterizados por falta de información indicaban que podía tratarse de hogares que efectuaron sus compras correspondientes a los gastos diarios en una semana distinta a la de referencia. En el caso de estratos bajos podía tratarse de hogares con muy escaso consumo. Pero si se trataba de grupos donde faltaban compras de períodos de referencia mayores y, además, los hogares pertenecían a estratos altos correspondía efectuar una revisión minuciosa de las encuestas involucradas.

4.3. ANÁLISIS DE CORRESPONDENCIAS SIMPLES

El análisis de correspondencias simples permite estudiar los grupos que se forman a partir de otras relaciones entre las variables, además de las lineales (estudiadas mediante el análisis de componentes principales). Así aparecen más frecuentemente en la caracterización de los grupos algunas variables que presentan relaciones lineales débiles con otras, tales como Boletos, Comidas afuera, Material escolar, Cigarrillos, Medicamentos.

Cuando aparecen encuestadores asociados a grupos caracterizados por algún gasto tal como Boletos, Comidas afuera, Combustibles, etcétera, puede tratarse de encuestadores que habitualmente captan esos gastos porque preguntan por ellos más frecuentemente que el resto o bien puede tratarse de datos adulterados.

⁴ En una clasificación jerárquica, se agrupan en primer término los individuos que resultan más cercanos y en cada paso siguiente se agrupan los grupos ya formados que resultan más parecidos. Este proceso se repite tantas veces como sea necesario hasta que todos los individuos están incluidos en un único grupo.

Recordando que se trata de un análisis descriptivo cuyos resultados son indicativos para un estudio posterior, la aplicación del análisis de correspondencias simples se efectuó considerando la misma tabla descripta en el análisis de componentes principales.

Las variables elegidas son: el tipo de gastos y las encuestas. En rigor esta última no es una variable categórica, la interpretación se fuerza con el fin de utilizar las ventajas de la métrica Chi-cuadrado para evaluar otras relaciones además de las lineales.

La misma tabla puede expresarse en términos de sus frecuencias relativas por fila y por columna. Habitualmente se denominan perfiles fila (la suma por fila es 100) y perfiles columna (la suma por columna es 100).

Como en otros métodos multivariados, también aquí es posible representar en el análisis ciertas variables que no inciden en la determinación de los ejes pero que pueden ayudar a la interpretación, se trata de las variables "ilustrativas". En este caso se tomaron como tales el código del encuestador y la semana en que se realizó la encuesta.

Sobre las encuestas, ahora caracterizadas por sus 10 primeras coordenadas factoriales, se efectuó una clasificación jerárquica. El criterio de agregación utilizado es el criterio de Ward.

En el resumen final de los resultados se caracterizaba a cada encuestador por los tipos de gastos a los que se encontraba asociado en forma frecuente y por aquellos a los que no se encontraba asociado.

En otras palabras, era posible apreciar qué gastos captaba y qué gastos no captaba cada encuestador describiendo una suerte de perfil de cada uno de ellos. Este perfil permitió caracterizarlo, junto con la información proveniente de otros varios métodos de control, de manera de analizar más profundamente el resultado de su trabajo y decidir diferentes caminos a seguir tales como correcciones de su trabajo en campo, instrucción sobre ciertos puntos para evitar errores, efectuar supervisiones especiales, etcétera.

4.4. ESTUDIOS DESCRIPTIVOS

Como forma de complementar los estudios anteriores y facilitar la interpretación de sus resultados, se realizaron análisis descriptivos de la información.

4.4.1 Encuestadores con medianas menores - Box-Plots

Se investigaron las variables cantidad de registros y monto de gasto por tipo de gasto a través de todas las encuestas realizadas por cada encuestador y se estudiaron las diferencias existentes entre la cantidad de registros (o el monto de gasto) por encuestador para cada uno de los gastos. Se informaba para cada gasto quiénes eran los encuestadores para los cuales la mediana de la distribución de la cantidad de registros (o el monto de gasto) por encuesta resultaba menor al resto de los encuestadores de acuerdo a los diagramas de box-plots construidos⁵.

⁵ La implementación de la muestra en campo garantizaba que cada encuestador relevaba encuestas de todos los estratos socioeconómicos.

4.4.2 Encuestadores asociados a encuestas atípicas en monto de gasto - Box-Plots

En este caso se analizaban las encuestas en conjunto por tipo de gasto, sin calcular promedios por encuestador y se identificaban los encuestadores que habían realizado las encuestas con valores extremos. En todos los casos los extremos hallados eran extremos superiores, es decir que el monto de gasto de las encuestas realizadas por los encuestadores detallados era mayor que el resto.

Podían existir varias razones que explicaran esta situación, por ejemplo que los encuestadores hayan relevado encuestas pertenecientes a estratos altos (habitualmente con gastos más elevados) que se tratara de errores de registración o que se tratara de encuestadores que captaban mayor cantidad de información.

4.5. ESTUDIOS COMPLEMENTARIOS

4.5.1 Aplicación de REGARIMA Y ROBREGAR con enfoque de series de tiempo para detectar outliers

Se trató de un estudio especial aplicado a los datos del Área Metropolitana que tuvo por objeto detectar valores atípicos considerando la variable tiempo, en este caso semanas.

Se aplicaron en primer término los métodos REGARIMA y ROBREGAR usando los datos como series temporales de frecuencia 4 (debido a las 4 semanas en el mes). Se estudiaron cada uno de los rubros de gastos.

Luego se individualizaron las semanas que se presentaban con mayor frecuencia a través de los diferentes rubros de gastos. Con posterioridad se listaron las encuestas de esas semanas para ubicar a las posibles causantes de los valores atípicos.

4.5.2 Aplicación de Modelo Lineal Generalizado para detectar encuestadores con gastos significativamente distintos (menores)

Se aplicó el modelo lineal generalizado para cada uno de los rubros de gastos comparando los promedios por encuestador. Se individualizaban los encuestadores cuyos promedios eran significativamente menores a los promedios de los restantes encuestadores.

5. EJEMPLO DE LA APLICACION DE LAS TÉCNICAS DESCRIPTAS A UNA CIUDAD

A continuación se describe el análisis efectuado sobre las encuestas de una ciudad mediante la aplicación de las técnicas multivariadas mencionadas.

En primer término se detallan los resultados de la aplicación del análisis de componentes principales. Se reproducen párrafos del resumen final y sus cuadros respectivos.

Análisis de la información

Se han listado en orden de importancia los grupos de encuestas para los cuales fue posible detectar un encuestador característico. En primer lugar se detallan los encuestadores que resultan más característicos y en último lugar los que resultan menos característicos. Se señala el código del encuestador, el tamaño del grupo, el porcentaje del grupo realizado por el encuestador (porcentaje de encuestas) y el rubro de gastos que caracterizan al grupo.

Analizando ambos cuadros aparecen grupos pequeños (con pocas encuestas) con encuestadores característicos que han relevado un porcentaje considerable de las encuestas del grupo, que hacen recomendable una revisión especial. Entre ellos, es posible incluir el grupo 9/20, el grupo 10/20, el grupo 11/20 y el grupo 18/20 determinados con la cantidad de registros del F2 y F3. Cuando la variable es monto de gasto del F2 es posible incluir el grupo 10/20.

Estudiando los resultados del análisis a través de la cantidad de registros del F2 y F3 y del monto de gasto del F2, es factible ver que existe un grupo numeroso que se caracteriza por la falta de información: el grupo 3/20 (113 encuestas en registros y 152 en monto de gasto) en donde aparece el encuestador D como característico. Este encuestador ya fue señalado en el análisis descriptivo por la misma razón. Al analizar el grupo a través de los registros del F2 y F3 es posible ver que también se caracteriza por la falta de gastos de períodos de referencia mayores a la semana, con lo cual podría pensarse que se trata de encuestas de estratos bajos.

Cuadro 5.1 Análisis de componentes principales
Cantidad de registros F2 y F3

Cantidad de registros F2 y F3			
Encuestador	Cantidad de encuestas del grupo	% del grupo realizado por el encuestador	Registros característicos de las encuestas pertenecientes al grupo
D	37 Grupo 10/20	73%	Corraos y teléfonos
D	34 Grupo 11/20	62%	Trimestrales Indumentaria Mensuales resto
D	113 Grupo 3/20	41%	Por la falta de: Boletos Artículos de higiene y tocador Mensuales resto Bebidas Azúcar y dulces Otros alimentos Cigarrillos Indumentaria Artículos de limpieza Correos y teléfonos Semestrales Comidas afuera, etcétera
B	42 Grupo 5/20	50%	Cigarrillos Bebidas Boletos
E	14 Grupo 9/20	71%	Anuales Cuotas Semestrales
E	51 Grupo 1/20	39%	Boletos
A	45 Grupo 16/20	44%	Combustibles Pan y cereales Frutas y verduras Carnes Lácteos y huevos Bimestrales Mensuales resto
A	69 Grupo 13/20	38%	Boletos Frutas y verduras Bimestrales Lácteos y huevos Pan y cereales Carnes

Cuadro 5.1 (continuación)

Cantidad de registros F2 y F3			
Encuestador	Cantidad de encuestas del grupo	% del grupo realizado por el encuestador	Registros característicos de las encuestas pertenecientes al grupo
A	50 Grupo 15/20	40%	Artículos de limpieza Artículos de higiene y tocador Otros alimentos Aceites y grasas Frutas y verduras Pan y cereales Azúcar y dulces Boletos Bimestrales Carnes Lácteos y huevos Mensuales resto
A	23 Grupo 12/20	48%	Medicamentos
C	9 Grupo 18/20	56%	Comidas afuera Mensuales resto Semestrales Indumentaria Otros alimentos Trimestrales Material escolar Medicamentos Azúcar y dulces Artículos de limpieza, etcétera

Cuadro 5.2 Análisis de componentes principales
Monto de Gasto F2

Monto de gasto F2			
Encuestador	Cantidad de encuestas del grupo	% del grupo realizado por el encuestador	Registros característicos de las encuestas pertenecientes al grupo
D	91	47%	Carnes Lácteos y huevos
	Grupo 2/20		
B	69	38%	Cigarrillos
	Grupo 14/20		
A	20	55%	Correos y teléfonos
	Grupo 10/20		
A	45	42%	Artículos de limpieza Artículos de higiene y tocador Frutas y verduras Aceites y grasas Carnes Lácteos y huevos Pan y cereales Bebidas Azúcar y dulces Otros alimentos
	Grupo 16/20		
D	152	32%	Por la falta de:
	Grupo 3/20		Carnes Lácteos y huevos Pan y cereales Frutas y verduras Aceites y grasas Bebidas Boletos Cigarrillos, etcétera
A	31	45%	Combustibles Pan y cereales Carnes Frutas y verduras Lácteos y huevos Otros alimentos
	Grupo 18/20		
B	40	37%	Cigarrillos Bebidas Carnes Pan y cereales Lácteos y huevos Boletos
	Grupo 15/20		

A continuación se detallan los resultados de la aplicación del análisis de correspondencias simples. Como en el caso anterior se reproducen algunos párrafos del resumen final y el cuadro anexo.

Análisis de la información

Con esta información es posible realizar varias lecturas. Por una parte puede apreciarse que los encuestadores D y E se caracterizan por la **presencia de pocos** tipos de gastos y en cambio por la **ausencia de muchos** tipos de gastos. Estos datos son coincidentes con los resultados del análisis descriptivo.

En cambio los encuestadores B y C se caracterizan por la **presencia de muchos** tipos de gastos y por **ausencia de una cantidad similar o menor**, de la misma forma en que aparecen señalados en los estudios de outliers superiores.

Por otro lado es posible describir una suerte de perfil de cada encuestador a través de los tipos de gastos que le son característicos (que capta más frecuentemente) y los que no lo son (que no capta frecuentemente). Por ejemplo un análisis del encuestador D permitiría arriesgar el siguiente perfil: capta frecuentemente gastos en Combustibles e Indumentaria y en cambio, no capta Material escolar, Artículos de higiene y tocador, Comidas afuera, Medicamentos y gastos de período de referencia mayor.

Cuadro 5.3 Análisis de correspondencias simples

Encuestador	Caracterizado por la	Tipo de gastos	
		según cantidad de registros	según monto de gastos
A	Presencia de	Medicamentos	Comidas afuera
			Otros alimentos
			Combustibles
	Ausencia de	Indumentaria	Artículos de higiene y tocador
		Mensuales	Artículos de limpieza
		Pan y cereales	Azúcar y dulces
		Bebidas	Boletos
		Cigarrillos	Cigarrillos
			Material escolar
B	Presencia de	Bebidas	Cigarrillos
		Artículos de higiene y tocador	Bebidas
		Azúcar y dulces	Carnes
		Artículos de limpieza	Lácteos y huevos
		Otros alimentos	Comidas afuera
		Aceites y grasas	Lavandería
		Cigarrillos	Medicamentos
	Ausencia de	Frutas y verduras	Correo y teléfono
		Carnes	
		Bimestrales	
		Indumentaria	
		Mensuales resto	
		Combustibles	
		Semestrales	
		Trimestrales	

Cuadro 5.3 (continuación)

Encuestador	Caracterizado por la	Tipo de gastos	
		según cantidad de registros	según monto de gastos
C	Presencia de	Semestrales	Artículos de higiene y tocador
		Otros alimentos	Artículos de limpieza
		Material escolar	Azúcar y dulces
		Azúcar y dulces	Medicamentos
		Comidas afuera	
		Boletos	
	Ausencia de	Pan y cereales	Carnes
		Combustibles	Frutas y verduras
		Frutas y verduras	Pan y cereales
			Lácteos y huevos
			Combustibles
D	Presencia de	Indumentaria	Combustibles
		Pan y cereales	
		Frutas y verduras	
	Ausencia de	Anuales	Material escolar
		Cuotas	Artículos de higiene y tocador
		Otros alimentos	Comidas afuera
		Material escolar	Medicamentos
		Azúcar y dulces	
		Comidas afuera	
E	Presencia de	Anuales	Boletos
		Cuotas	Frutas y verduras
	Ausencia de	Indumentaria	Combustibles
		Mensuales	Medicamentos
		Pan y cereales	
		Medicamentos	
		Artículos de higiene y tocador	
		Artículos de limpieza	
		Otros alimentos	
		Cigarrillos	

Por último, se incluyen dos cuadros con el resumen de los resultados de los estudios descriptivos efectuados sobre la información de la misma ciudad. El primero de ellos detalla los encuestadores con medianas menores y el segundo, los encuestadores que resultaron más atípicos en el estudio de Box-Plots.

Cuadro 5.4 - Encuestadores con medianas menores

Tipo de gastos	Encuestadores con medianas menores
Pan y cereales	E
Carnes	—
Aceites y grasas	D
Lácteos y huevos	E
Frutas y verduras	D, E
Azúcar y dulces	D
Otros alimentos	D
Bebidas	D
Comidas en restaurante, bares, etcétera	D
Artículos de limpieza	D
Lavandería	—
Productos medicinales y farmacéuticos	—
Combustible	E
Boletos, servicios, transportes	D
Correo, teléfono	—
Material escolar	—
Cigarrillos, tabaco	E, D
Artículos higiene y tocador	D
Indumentaria	—
Cuotas	D
Otros gastos mensuales	B
Gastos bimestrales	—
Gastos trimestrales	—
Gastos mensuales	A
Gastos anuales	—

Nota: se señala en **negrita** a aquellos encuestadores cuyo promedio de registros (monto de gasto) también fue detectado como significativamente diferente (menor) en cada tipo de gasto según el modelo lineal generalizado.

A partir de los resultados de estos estudios la Dirección de la Encuesta obtuvo un perfil del trabajo de cada encuestador en la provincia, disponiendo un operativo de supervisión que realizó el Equipo Central.

Se investigó el trabajo del encuestador D señalado en los estudios. Se analizaron los cuestionarios correspondientes a los hogares del grupo 3/20 del estudio de componentes principales y se comprobó que presentaban perfiles de consumo llamativamente parecidos y que la gran mayoría de los Cuestionarios 2 no habían sido completados por el informante, sino por el encuestador. Procedieron a supervisar en campo esas encuestas, comprobando que el encuestador no realizaba las encuestas en las viviendas seleccionadas y alteraba el régimen establecido de visitas al

Cuadro 5.5 - Encuestadores más atípicos en Box-Plots

Tipo de gastos	Más atípicos en Box-Plots
Pan y cereales	—
Carnes	C, B
Aceites y grasas	B, C, E
Lácteos y huevos	—
Frutas y verduras	—
Azúcar y dulces	B, A, E, C
Otros alimentos	C, B, D
Bebidas	A, B, C, E
Comidas en restaurantes, bares, etcétera	—
Artículos de limpieza	C
Lavandería	—
Productos medicinales y farmacéuticos	A, B, C, E
Combustible	A, B, C
Boletos, servicios, transportes	—
Correo, teléfono	—
Material escolar	—
Cigarrillos, tabaco	B
Artículos higiene y tocador	B, C, A
Indumentaria	C, D
Cuotas	A, E
Otros gastos mensuales	C, B
Gastos bimestrales	—
Gastos trimestrales	—
Gastos semestrales	C, B, D
Gastos anuales	—

Nota: se advierte que en todos los casos los outliers hallados son superiores y no inferiores, es decir que la cantidad de registros de las encuestas realizadas por los encuestadores detallados es **mayor** que el resto.

Los encuestadores citados más frecuentemente como **outliers superiores** son: **B y C**.

hogar, haciendo sólo una entrevista de apertura para obtener los datos del Cuestionario 1 de características sociodemográficas, completando el resto de los cuestionarios con información falsa.

También investigaron el trabajo del encuestador A, caracterizado por la presencia de muchos grupos de gastos alimenticios y la ausencia de gastos por recordación como indumentaria y otros gastos mensuales. Analizaron las encuestas de los grupos en los que el mismo era representativo y comprobaron similares irregularidades que para el encuestador D. Por ende, supervisaron en campo con idénticos resultados.

6. Conclusiones

La adulteración de información puede afectar seriamente la exactitud de los resultados de una encuesta. Por este motivo, es sumamente importante el desarrollo de estrategias orientadas a identificar encuestas adulteradas, para luego proceder a su corrección o eliminación.

En este estudio de carácter exploratorio, se investigó la posibilidad de emplear técnicas de análisis multivariado de datos como una herramienta útil en la detección de grupos de encuestas adulteradas. El objetivo concreto era señalar encuestadores con algún patrón "sospechoso" en sus encuestas. Luego, llevando a cabo una revisión especial, se determinaba si tales sospechas tenían o no fundamento.

La metodología propuesta se aplicó en el contexto de la ENGH. A pesar de no contar con experiencia previa, lo cual dificultó en un principio la tarea, los resultados fueron en general positivos. Se comprobó que algunos de los encuestadores señalados a través de los estudios realizados, no efectuaban las visitas a los hogares como debían, sino que llenaban los cuestionarios con información inventada por ellos. En estos casos se procedió a eliminar la totalidad de las encuestas realizadas por tales encuestadores, evitando de esta manera un potencial sesgo en las estimaciones.

Por último es importante remarcar el alcance de estos estudios. De la misma forma en que se detectaron anomalías en la ciudad de los ejemplos, hubo otra ciudad, donde las "sospechas" señaladas en el informe pudieron ser explicadas por causas normales (no adulteración). Es así que para una correcta lectura de los resultados es imprescindible comprender el carácter tentativo de estos estudios. Bajo ninguna circunstancia los resultados obtenidos en ellos deben ser considerados como verdaderos, sino como indicativos de una probabilidad que debe ser corroborada por medio de análisis más detallados y una supervisión especial.

7. Referencias bibliográficas

- ANDERSON T.W. (1984), *An introduction to Multivariate Statistical Analysis*. 2nd ed. John Willey and Sons, New York.
- CRESPI, L.P. (1945). "The Cheater Problem in Polling", *Public Opinion Quarterly*, Winter, pp. 431-445.
- CRIVISQUI, E. (1993), *Análisis factorial de correspondencias. Un instrumento de investigación en ciencias sociales*, Universidad Católica de Asunción.
- CUADRAS C.M. (1981), *Métodos de Análisis Multivariante*, Eunibar, Barcelona.
- ESCOFIER, B. y PAGES, J. (1992), *Análisis factoriales simples y múltiples. Objetivos, métodos e interpretación*, Universidad del País Vasco, Bilbao.
- INDEC (1997), *Documento interno elaborado por la Dirección de Metodología Estadística dirigido a la Encuesta Permanente de Hogares*.
- LEBART, L., MORINEAU, A. y FENELON, J.P. (1985), *Tratamiento estadístico de datos. Métodos y Programas*, Marcombo Boixareu Editores, Barcelona-México.
- MORRISON, D.F. (1990), *Multivariate Statistical Methods*, 3rd. ed., Mc Graw Hill Series in probability and statistics, New York.

Publicaciones anteriores

1. (1984) La pobreza en la Argentina: indicadores de necesidades básicas insatisfechas a partir de los datos del Censo Nacional de Población y Vivienda 1980.
2. (1984) Los Censos de Población del '80: taller de análisis y evaluación.
3. (1985) La juventud de la Argentina.
4. (1986) Proyección de la población 1970-2025: urbana y rural y económicamente activa por sexo y grupo de edad.
5. (1986) Sistema integrado de estadísticas continuas sobre la infancia.
6. (1987) Exportaciones industriales: perfil y comportamiento de las empresas exportadoras de manufacturas.
7. (1987) Encuesta industrial de electrónica.
8. (1987) Los Censos del '90: características económicas de la población.
9. (1987) Economía no registrada.
10. (1988) Tablas de mortalidad 1980-1981: total y jurisdicciones.
11. (1988) Encuesta de gastos e ingresos de los hogares.
12. (1988, 1989) Trabajadores por cuenta propia: Encuesta del Gran Buenos Aires.
13. (1989) La pobreza en el conurbano bonaerense.
14. (1989) La pobreza en la ciudad de Posadas.
15. (1989) Estimaciones y proyecciones de población 1959-2025 (versión revisada).
16. (1989) La pobreza en Santiago del Estero y La Banda.
17. (1989) La pobreza en la ciudad de General Roca.
18. (1989) La pobreza en la ciudad de Neuquén.
19. (1990) Encuesta de hogares: Errores de muestreo y efectos de diseño.
20. (1992) Cantidades Consumidas: Encuesta de gasto de los hogares.
21. (1992) El consumo y el ingreso de los hogares frente a una caída de la tasa de inflación. EGH 1985/1986.
22. (1993) Encuesta de gastos e ingresos de los hogares: San Salvador de Jujuy. Setiembre 1981. Agosto 1982.
23. (1994) Estimaciones y proyecciones de población 1950-2050. Total del país.
24. (1996) Perfil de los hogares y de la población con Necesidades Básicas Insatisfechas (NBI).

25. (1996) El cultivo de la jojoba.
26. (1997) Horticultura bajo cubierta.
27. (1997) Hogares particulares con menores: perfil sociodemográfico.
28. (1998) Hogares particulares con adultos mayores: perfil sociodemográfico.
29. (1997) La migración internacional en la Argentina: sus características e impacto.
30. (1998) Hogares particulares con jóvenes: perfil sociodemográfico.
31. (1998) Encuesta sobre la conducta tecnológica de las empresas industriales argentinas.
32. (1999) Un modelo de predicción del Estimador Mensual Industrial (EMI).
33. (1999) Ajuste estacional de series elaboradas por INDEC.
34. (1999) La producción industrial argentina en los años '90.

ESTE VOLUMEN
USO DE TÉCNICAS MULTIVARIADAS PARA LA DETECCIÓN DE
POSIBLES ADULTERACIONES EN ENCUESTAS
SERIE ESTUDIOS N° 35
CON UNA TIRADA DE 800 EJEMPLARES
SE TERMINÓ DE IMPRIMIR
EN EL MES DE FEBRERO DE 2000
EN EL DEPARTAMENTO PUBLICACIONES DEL INDEC,
AV. PRESIDENTE JULIO A. ROCA 609
1067 BUENOS AIRES
REPÚBLICA ARGENTINA

0102 Llenado del inicio Fin Fecha

0101 Ubicación Código Pdo. Frac. Radio Seq.

0102 Código de 0203 Cod. 0205 Cod. Censista

0203 Cod. Supervisor

ARTICULO 10: Las informaciones que se suministran a los organismos que integran el Sistema Estadístico Nacional, en cumplimiento de la presente ley serán estrictamente secretas, y solo se utilizarán con fines estadísticos. Los datos deberán ser suministrados y publicados, exclusivamente en compilaciones de conjunto, de modo que no pueda ser

I. IDENTIFICACIÓN DE LA EXPLOTACIÓN Y

1101 Calle / Ruta, Nº / Km

Cód. Localidad o paraje más cercano

2. Identificación del productor y del

1201 Apellido y nombre del productor o denominación de la razón social

3. Otros datos

1702 Calle / Ruta, Nº / Km, Piso, Depto. (solo si difiere de la dirección de la

Cód. Localidad o paraje más

1703 Correo o estafeta postal

Cas. Fax o 2° teléfono (nueva)

1704 Teléfono (nueva)

1705 Celular o 3° teléfono (nueva)

5. Explotaciones sin

1706

1501 Hectáreas

2 Parc. Periodo de referencia: 1° julio

Relone