Hachuel, Leticia
Boggio, Gabriela
Cuesta, Cristina
Instituto de Investigaciones Teóricas y Aplicadas en Estadística, Escuela de Estadística

ESTIMACIÓN Y COMPARACIÓN DE LAS TASAS DE DESOCUPACIÓN DE LOS AGLOMERADOS DE LA EPH A TRAVES DE MODELOS DE EFECTOS FIJOS Y ALEATORIOS*

1. INTRODUCCIÓN

El análisis de los datos sobre desocupación provistos por la Encuesta Permanente de Hogares (EPH), a través de los distintos aglomerados urbanos, es de particular interés tanto para obtener estimaciones de las verdaderas tasas de desocupación como para comparar sus valores entre aglomerados. Esta encuesta se realiza en forma sistemática dos veces por año – ondas mayo y octubre- en 28 centros urbanos importantes del país. En cada uno de estos aglomerados la información se releva a través de una muestra de tipo estratificado con etapas múltiples de selección.

Dentro de este contexto, los modelos estadísticos constituyen una herramienta útil para conseguir ambos fines: estimación y comparación. En muestras pequeñas las estimaciones directas frecuentemente muestran mayor variabilidad que los verdaderos valores y en estos casos resulta útil ajustar modelos de efectos aleatorios para obtener mejores estimaciones. También estos modelos resultan adecuados cuando existe poca diferencia entre los valores observados en las áreas. Básicamente el recurso utilizado consiste en emplear la información de todas las áreas para estimar la proporción de un área dada. Los modelos de efectos fijos, en cambio, se utilizan para comparar la probabilidad de un evento en distintas áreas y no son adecuados para obtener estimaciones ya que estas reproducen las observaciones directas.

Este trabajo tiene por objetivo utilizar la modelización estadística de las tasas de desocupación por aglomerados que incluye efectos fijos y aleatorios para la comparación y estimación de dichas tasas. Finalmente se ensaya la posibilidad de predecir las tasas de desocupación en un período futuro, aunque con la restricción de supuestos fuertes que pueden no estar acordes a la realidad.

* Trabajo realizado en el marco del proyecto de investigación “Modelos referidos al fenómeno del desempleo a partir de la Encuesta Permanente de Hogares: Alcance y Limitaciones”. PICT 02-09897. Agencia de Promoción de la Investigación Científica y Tecnológica.

2. METODOLOGÍA

Modelos para mejorar las estimaciones

Los modelos de efectos aleatorios, como recurso para mejorar estimaciones de proporciones muestrales, asumen que las verdaderas proporciones de un evento de interés varían de acuerdo a alguna distribución. De esta manera se emplea la información de todas las áreas para estimar la proporción en un área dada.

Sea π_i la verdadera proporción en el área $i, i = 1,...,I$. Sean y_i variables binomiales con tamaños de muestra n_i y parámetros $\{\pi_i\}$. Las proporciones muestrales $\{p_i = y_i / n_i\}$ son en realidad las estimaciones máximo verosímiles de π_i en el siguiente modelo de efectos fijos saturado

$$\log it(\pi_i) = \alpha + \beta_i, \quad i = 1,...,I \tag{1}$$

sujeto a la restricción $\sum \beta_i = 0$ o $\beta_1 = 0$. Este modelo es saturado ya que tiene I parámetros no redundantes para las I observaciones binomiales, de modo que bajo este modelo resulta $\hat{\pi}_i^{(1)} = p_i$, i=1,...,I.

Como en general las estimaciones muestrales tienen mayor variabilidad que las verdaderas proporciones, Agresti (2000) propone estimadores más consistentes a partir de modelos de efectos aleatorios.

El modelo más simple de efectos aleatorios es el siguiente.

$$\log it(\pi_i) = \alpha + u_i, \quad i = 1,...,I \tag{2}$$

donde se supone que los efectos aleatorios $\{u_i\}$ son independientes con distribución $N(0, \sigma^2)$.

Luego de estimar α y σ se estima $\log it(\pi_i)$ usando $\hat{\alpha} + \hat{u}_i$ donde $\hat{u}_i$ es la predicción del efecto aleatorio basado en los datos observados. A partir de este logit se obtiene la estimación de la probabilidad del evento en cada área: $\hat{\pi}_i^{(2)}$.

Una alternativa, dentro del enfoque de modelos de efectos aleatorios consiste en utilizar información complementaria propia de cada área para mejorar aún más las estimaciones. Sea q_i la proporción del evento en un período anterior. Esta nueva información puede usarse para ajustar el modelo:

$$\log it(\pi_i) = \log it(q_i) + \alpha + u_i, \quad i = 1,...,I \tag{3}$$

donde las $\{q_i\}$ son conocidas y los $\{u_i\}$ se siguen suponiendo independientes con distribución $N(0, \sigma^2)$. El término conocido del predictor lineal, $\log it(q_i)$, se denomina "offset". Si se reordenan los términos del modelo, se obtiene:

$$\log\left(\frac{\pi_i /(1-\pi_i)}{q_i /(1-q_i)}\right) = \alpha + u_i, \quad i = 1,...,I$$

por lo que $\alpha + \mu_i$ representa entonces el logaritmo de la razón de odds, para la i-ésima área, de presentar el evento en un momento, relativo a haberlo presentado en un momento anterior.

Nuevamente, luego de estimar α y σ se estima $\log it(\pi_i)$ y a partir de este logit se obtiene la estimación de la probabilidad del evento en cada área: $\pi_i^{(3)}$.

Estas estimaciones obtenidas a partir de los modelos (2) y (3) mejoran sustancialmente las estimaciones directas especialmente en los casos en que los tamaños de muestra son pequeños o cuando σ es pequeño.

Modelos para la evaluación del efecto aglomerado

Los modelos de efectos fijos reproducen las proporciones muestrales y por lo tanto no son adecuados para mejorar las estimaciones. Sin embargo estos modelos son adecuados para la comprobación de existencia de diferencias entre los aglomerados.

A partir del modelo (1)

$$\log it(\pi_i) = \alpha + \beta_i, \quad i = 1,...,I$$

y considerando a los $\{\beta_i\}$ como los efectos correspondientes a los diferentes niveles del un factor fijo, área o aglomerado, se puede determinar si hay diferencias en las chances de presentar el evento de interés (vs. no presentarlo) entre las distintas áreas.

También se puede postular el modelo de efectos fijos con información previa

$$\log it(\pi_i) = \log it(q_i) + \alpha + \beta_i, \quad i = 1,...,I \qquad (4)$$

o equivalentemente:

$$\log\left(\frac{\pi_i/(1-\pi_i)}{q_i/(1-q_i)}\right) = \alpha + \beta_i, \quad i = 1,...,I$$

donde $\{q_i\}$ son las proporciones conocidas del evento en un período anterior. Este modelo (4) permite detectar si las chances de presentar un evento en comparación con la chance de presentarlo en un período anterior son iguales para todas las áreas.

Predicciones

El modelo (4) antes presentado utiliza información previa acerca del evento en estudio, por lo que resulta natural pensar en su utilización para realizar predicciones. Es decir, el modelo mencionado permitiría predecir la probabilidad del evento de interés por aglomerado donde, en esta oportunidad, q_i es la última estimación disponible de la proporción del evento por área.

Sin embargo, dichas predicciones estarían sujetas al cumplimiento de los siguientes supuestos:

1. Las razones de odds de presencia del evento entre dos períodos se mantienen constantes a través del tiempo.

2. Los efectos de cada área sobre el logit de la probabilidad del evento de interés se mantienen constantes a través del tiempo.

3. APLICACIÓN

La EPH provee la tasa de desocupación correspondiente a 28 aglomerados del país. Considerando el aglomerado Gran Buenos Aires desagregado en Partidos del Conurbano y Ciudad de Buenos Aires se cuenta con un total de 29 aglomerados. En el Anexo se presentan las tasas provistas por el INDEC para los 29 aglomerados en las ondas mayo 2002 y octubre 2002.

En el Gráfico 1 se pueden comparar visualmente las tasas de desocupación en las ondas mayo 2002 y octubre 2002 para los distintos aglomerados. Se observa que consistentemente la desocupación fue mayor en mayo excepto en el aglomerado Gran Resistencia donde la tasa de desocupación subió en octubre . Para comparar ambas tasas se construyen las razones de odds que miden la chance de estar desocupado (versus estar ocupado) en octubre del 2002 respecto de la misma chance en mayo del mismo año. En virtud de los resultados ya expuestos dichas razones de odds toman valores menores que uno en todos los aglomerados, salvo en Gran Resistencia (Gráfico 2).

Gráfico 1. Tasas de desocupación en los 29 aglomerados de la EPH en mayo 2002 y octubre 2002

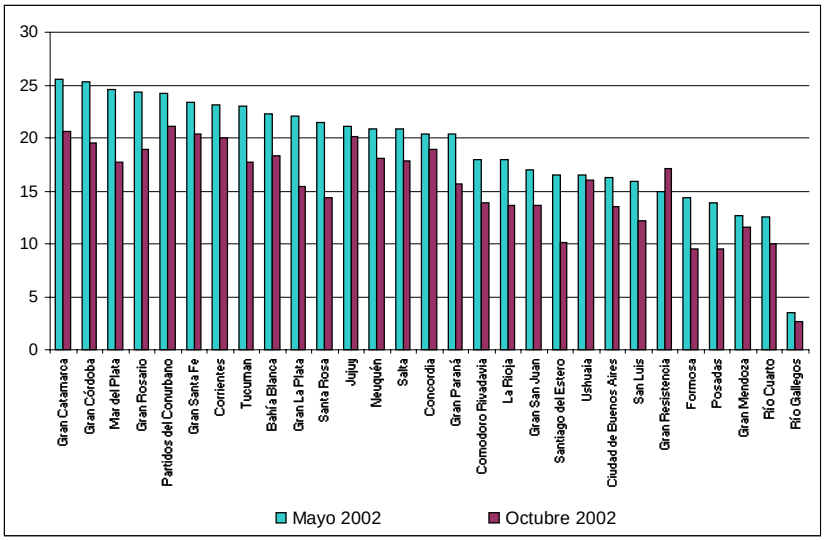
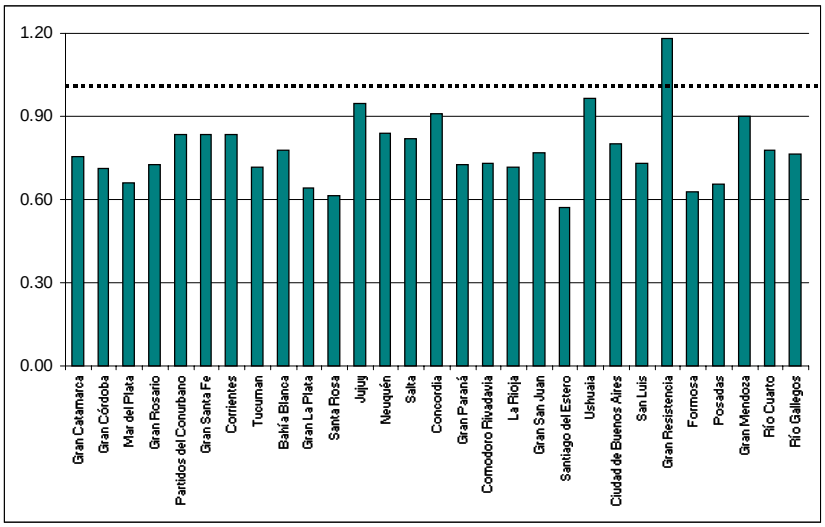


Gráfico 2. Razones de odds de desocupación entre las ondas de octubre y mayo de 2002 para los 29 aglomerados



Con esta información se pueden obtener nuevas estimaciones de las tasas de desocupación mediante el ajuste de los modelos de efectos aleatorios descriptos y comparalas a partir de los modelos de efectos fijos.

Modelos para estimación

Con el fin de suavizar las estimaciones directas de las tasas de desocupación provistas por la EPH se ajusta en primer lugar el modelo (2),

$$\log it(\pi_i) = \alpha + u_i, \quad i = 1,...,I$$

donde π_i es la probabilidad de desocupación en octubre de 2002 en el i-ésimo aglomerado y $\{u_i\}$ son los respectivos efectos aleatorios independientes con distribución $N(0,\sigma^2)$.

A partir del ajuste de dicho modelo se obtiene una estimación de la variabilidad entre los logaritmos de las chances de desocupación entre los aglomerados que resulta ser $\hat{\sigma}^2$=0.1892 (p=0.0006). Esta significación estadística indica que hay heterogeneidad entre las chances de estar desocupado en los distintos aglomerados en la onda octubre de 2002.

A partir de $\hat{\sigma}^2 = 0.1892$, $\hat{\alpha} = -1.6533$ y de los valores predichos $\hat{u}_i$, se estiman las probabilidades de desocupación en cada aglomerado de acuerdo a:

$$\hat{\pi}_i = \frac{e^{\hat{\alpha}_i + \hat{u}_i}}{1 + e^{\hat{\alpha}_i + \hat{u}_i}} .$$

Estas probabilidades estimadas intentan mejorar las estimaciones directas, sin embargo, la variabilidad hallada (que es moderadamente alta) y los tamaños de muestra de cada aglomerado indican que este modelo puede no mejorar sustancialmente las estimaciones directas. Por este motivo se postula el modelo de efectos aleatorios con offset, que incluye información particular de cada aglomerado y por lo tanto es de esperar que capte mejor las características propias de cada área.

El ajuste del modelo de efectos aleatorios con offset (3) usa como información auxiliar las tasas de desocupación correspondientes a mayo de 2002 simbolizadas por {q_i}:

$$\log\left(\frac{\pi_i/(1-\pi_i)}{q_i/(1-q_i)}\right) = \alpha + u_i, \quad i = 1,...,I$$

En este caso la variabilidad entre aglomerados, $\hat{\sigma}^2 = 0.005425$, no resulta significativa (p=0.9787), es decir, los aglomerados son homogéneos en términos de las razones de odds de estar desocupado en octubre de 2002 respecto a mayo del mismo año. El hecho de haber obtenido una variabilidad pequeña indica que el modelo es considerablemente bueno para suavizar las estimaciones directas.

Nuevamente a partir de $\hat{\sigma}^2 = 0.005425$, $\hat{\alpha} = -0.2297$, $\log it(q_i)$ y de los efectos aleatorios $\hat{u}_i$, se estiman las probabilidades de desocupación según:

$$\hat{\pi}_i = \frac{e^{\log it(q_i) + \hat{\alpha}_i + \hat{u}_i}}{1 + e^{\log it(q_i) + \hat{\alpha}_i + \hat{u}_i}} .$$

Es importante recalcar que en este último modelo se explican las razones de odds de desocupación entre una onda y la anterior mientras que en el modelo (2) se explican los odds o chances de desocupación en una determinada onda.

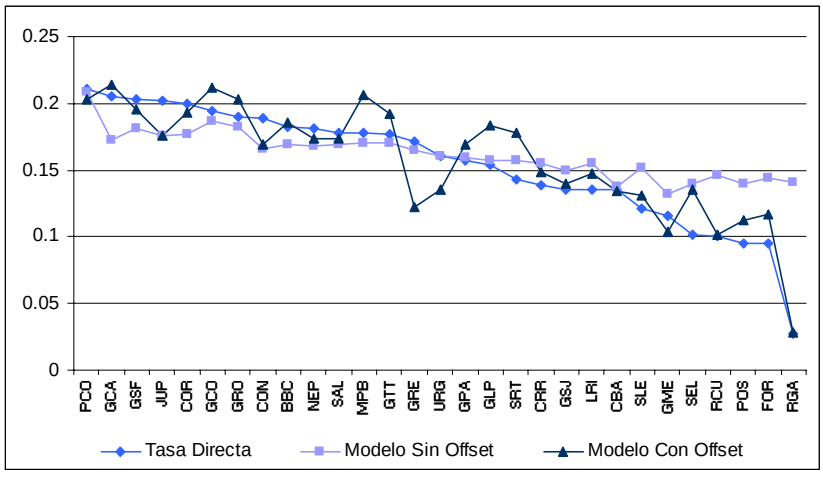
En la Tabla 1 se muestran criterios de bondad de ajuste para los modelos (2) y (3). El modelo (3) presenta menores valores de AIC y BIC, la cual sugiere su elección reafirmado por el hecho de haber obtenido una muy baja variabilidad en el modelo (3) respecto al modelo (2).

Tabla 1: Medidas de bondad de ajuste de los modelos de efectos aleatorios con o sin offset

Modelo	AIC	BIC
ef. aleat. Sin offset (2)	193.0	195.7
ef. aleat. Con offset (3)	157.5	160.3

En el Gráfico 3 se comparan las estimaciones obtenidas con el modelo 2 y el modelo 3 con las estimaciones directas. En el mismo los aglomerados están ordenados en orden decreciente según la tasa de desocupación directa. Las estimaciones a partir del modelo de efectos aleatorios sin offset presenta valores más "estables" o similares para todos los aglomerados, mientras que las estimaciones obtenidas con el modelo de efectos aleatorios con offset si bien es más irregular logra "captar" mejor valores extremos.

Gráfico 3. Probabilidades de desocupación estimadas para octubre de 2002 por aglomerado, a partir de los modelos de efectos aleatorios sin y con offset



Modelos para evaluar la existencia del efecto aglomerado

Los modelos postulados para la comparación de las tasas de desocupación entre aglomerados consideran al aglomerado como un efecto fijo. En primer lugar se ajusta el modelo (1),

$$\log it(\pi_i) = \alpha + \beta_i, \quad i = 1,...,I$$

donde π_i es la probabilidad de desocupación en octubre de 2002 y β_i el efecto fijo correspondiente al i-ésimo aglomerado.

El efecto correspondiente a los aglomerados resulta significativo (p<.0001), lo cual indica que la chance de estar desocupado en octubre de 2002 (vs. no estarlo) no es igual en todos los aglomerados. Así por ejemplo, la chance de estar desocupado en los Partidos del conurbano es 0.27 mientras que la chance de estar desocupado en Rio Gallegos es 0.03. Estas diferencias se visualizan en el Gráfico 1 donde se observa claramente que la tasa de desocupación no es pareja en todos los aglomerados.

Se ajusta luego el modelo (4), de efectos fijos con offset

$$\log\left(\frac{\pi_i/(1-\pi_i)}{q_i/(1-q_i)}\right) = \alpha + \beta_i, \quad i = 1,...,I$$

donde π_i es la probabilidad de desocupación en octubre de 2002, β_i el efecto fijo correspondiente al i-ésimo aglomerado y $\{q_i\}$ son las proporciones muestrales del evento en mayo 2002.

En este caso, el efecto aglomerado no resulta significativo (p=0.9833), es decir la razón de odds de estar desempleado en octubre de 2002 respecto a mayo de 2002 es igual en todos los aglomerados. Considerando que no hay efecto aglomerado, un estimación de dicha razón de odds es $e^{\hat{\alpha}} = e^{-0.2633} = 0.7685$ (ver Gráfico 2). Es decir, la chance de estar desocupado en octubre de 2002 se redujo en aproximadamente un 23% respecto a mayo de ese año en todos los aglomerados.

Predicciones

Se propone utilizar la metodología de estimación a través de modelos de efectos aleatorios con offset para intentar predecir la probabilidad de desocupación por aglomerado para la onda mayo de 2003 utilizando la información de octubre de 2002

Es decir, se utiliza como información auxiliar $\{q_i\}$ a las estimaciones directas de la proporción de desocupados en octubre de 2002 por aglomerado.

Estas predicciones serán adecuadas sólo bajo los siguientes supuestos:

1. Las razones de odds de desocupación entre mayo y octubre de 2002 se mantienen constantes en octubre de 2002 a mayo de 2003.

2. Los efectos de cada aglomerado sobre el logit de la probabilidad del evento de interés se mantienen constantes de octubre de 2002 a mayo de 2003.

Es importante destacar que estos supuestos pueden no ser sustentables con la realidad económica Argentina.

En el Gráfico 4 se muestra la evolución de las tasas de desocupación a través del tiempo. En el mismo se observa la tasa de desocupación muestral en los períodos mayo 2002 y octubre 2002 y la tasa predicha para mayo 2003 a partir del modelo (3).

El 31 de Julio de 2003 el INDEC dió a conocer las tasas de desocupación obtenidas por la EPH correspondiente a la Onda mayo de 2003. En el Gráfico 5 se comparan estas tasas de desocupación con las predichas a partir del modelo (3). La mayor diferencia se observa en Gran Resistencia (donde se produjo una sobreestimación). El aglomerado Gran Santa Fe no pudo ser evaluado debido a que su relevamiento se postergó por la catástrofe hídrica acontecida en dicha ciudad.

El promedio de las diferencias absolutas entre las tasas observadas en mayo de 2003 y las predichas para ese período es 2.75%. En 19 (65%) de los aglomerados las predicciones estan subestimadas respecto al valor observado en mayo de 2003. En el restante 35% las predicciones están sobreestimadas. Los aglomerados en que se observa una mayor diferencia entre el valor predicho y el observado son Mar del Plata y San Luis (donde se subetimaron los valores reales en 9,4% y 9.5% respectivamente) y Gran Resistencia (donde se sobreestima el valor real en -11%)

Gráfico 4. Tasas de desocupación observadas en mayo de 2002 y octubre de 2002 y tasa de desocupación predicha para mayo 2003 con el modelo de efectos aleatorios con offset

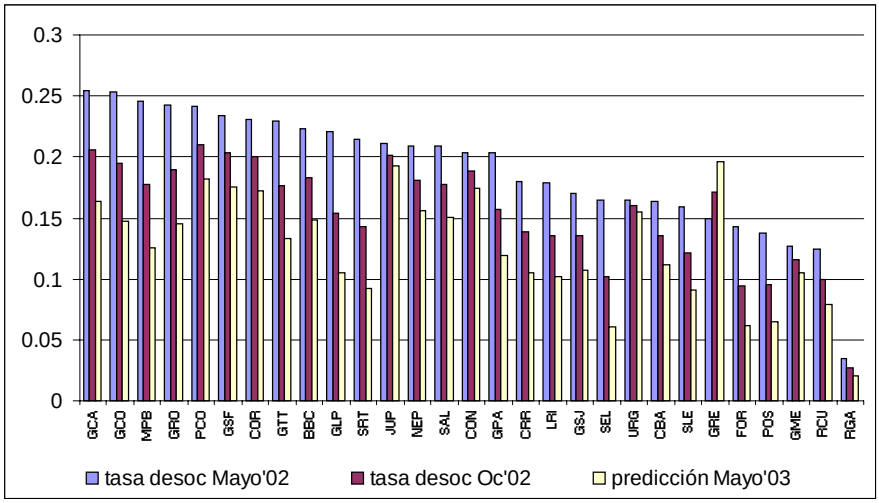
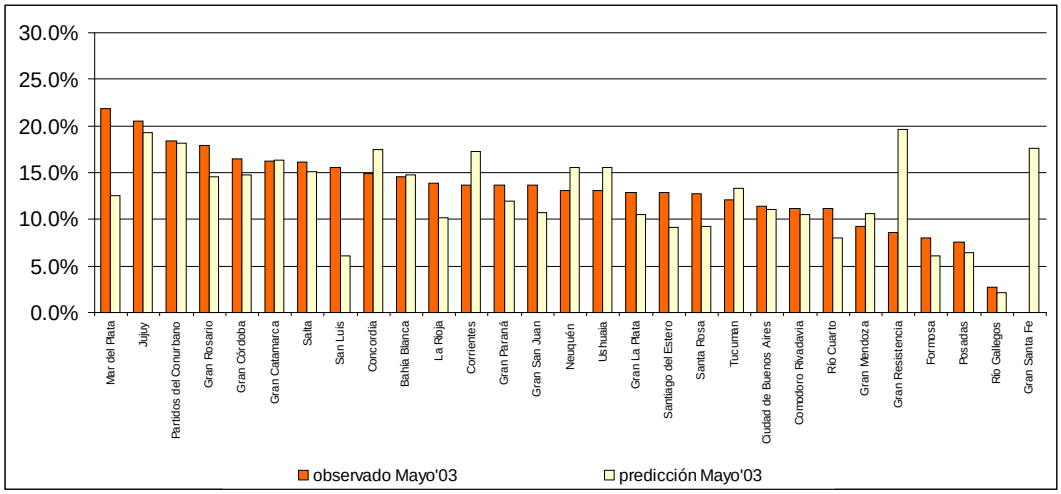


Gráfico 5. Comparación de la tasas de desocupación provista por INDEC a partir de la Onda mayo 2003 y las tasas estimadas con el modelo de efectos fijos con offset



4. DISCUSION

En este trabajo se muestra una forma de estimar proporciones a partir de un modelo de efectos aleatorios la cual resulta ser una adecuada metodología para suavizar estimaciones directas. En particular al utilizar información adicional se logra estimar mejor algunos valores alejados. La utilización de modelos de efectos aleatorios para realizar estimaciones es una herramienta muy útil especialmente cuando se trabaja con áreas o dominios pequeños o bien cuando la variabilidad entre ellas es baja.

Los modelos de efectos fijos se utilizan con otra finalidad: comparación. Se corroboraron diferencias entre las tasas de desocupación de los distintos aglomerados y se mostró la similitud en las chances de desocupación en octubre de 2002 en comparación con la onda previa.

En esta línea de investigación se está intentando determinar bajo qué situaciones particulares resulta preferible la elección de modelos de efectos aleatorios como metodología apropiada para la obtención de buenas estimaciones.

BIBLIOGRAFÍA

- Agresti, A.; Booth, J.; Hobert, J.; Caffo, B. (2000). Random –effects modeling of categorical response data. *Sociological Methodology*, 27-80.
- Agresti, A. (2002). “Categorical Data Análisis”. Second Edition. John Wiley and Sons
- Mc. Cullock, C.; Searle, S. (2001). “Generalized, Linear and Mixed Models”. John Wiley and Sons

ANEXO

Aglomerado		Tasa de desocupación	
Nombre	**Label**	**mayo 2002**	**octubre 2002**
Bahía Blanca	BBC	22.3	18.3
Gran La Plata	GLP	22.1	15.4
Mar del Plata	MPB	24.6	17.8
Gran Catamarca	GCA	25.5	20.5
Gran Córdoba	GCO	25.3	19.5
Río Cuarto	RCU	12.5	10.0
Corrientes	COR	23.1	20.0
Gran Resistencia	GRE	14.9	17.1
Comodoro Rivadavia	CRR	18.0	13.8
Concordia	CON	20.4	18.9
Gran Paraná	GPA	20.4	15.7
Formosa	FOR	14.3	9.5
Jujuy	JUP	21.1	20.2
Santa Rosa	SRT	21.4	14.3
La Rioja	LRI	17.9	13.6
Gran Mendoza	GME	12.7	11.6
Posadas	POS	13.8	9.5
Neuquén	NEP	20.9	18.1
Salta	SAL	20.9	17.8
Gran San Juan	GSJ	17.0	13.6
San Luis	SLE	15.9	12.1
Río Gallegos	RGA	3.5	2.7
Gran Rosario	GRO	24.3	18.9
Gran Santa Fe	GSF	23.4	20.3
Santiago del Estero	SEL	16.5	10.2
Ushuaia	URG	16.5	16.0
Tucuman	GTT	23.0	17.7
Ciudad de Buenos Aires	CBA	16.3	13.5
Partidos del Conurbano	PCO	24.2	21.0

*en miles

Fuente: EPH- INDEC