DISSERTATIO

Estadística

COLEGIO DE GRADUADOS EN CIENCIAS ECONÓMICAS DE ROSARIO
CONSEJO PROFESIONAL DE CIENCIAS ECONÓMICAS
DE LA PROVINCIA DE SANTA FE CÁMARA II
FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA

TRABAJOS FINALES

RECENSIÓN DE TESINAS Y PRÁCTICAS
PROFESIONALES DE LA CARRERA
LICENCIATURA EN ESTADÍSTICA

CONSEJO PROFESIONAL DE CIENCIAS ECONOMICAS DE LA PROVINCIA DE SANTA FE CAMARA II

Colegio de Graduados en Ciencias Económicas de Rosario

Universidad Nacional de Rosario

AÑO V 2021, N.º 5, ISSN 2591-5215

ÍNDICE

AÑO V 2021, N.º 5, ISSN 2591-5215

Esta revista se pone a disposición de los profesionales matriculados al Consejo Profesional de Ciencias Económicas de la Provincia de Santa Fe Cámara II (CPCE), asociados del Colegio de Graduados en Ciencias Económicas de Rosario (CGCE), estudiantes y docentes de la Facultad de Ciencias Económicas y Estadística (FCEyE) de la Universidad Nacional de Rosario (UNR) y otras Instituciones vinculadas al quehacer profesional y académico.

El contenido de los trabajos finales no necesariamente refleja la opinión de los Comités responsables de esta publicación digital.

Las Instituciones no son responsables por el contenido de las informaciones y opiniones que viertan en esta revista quienes son identificados como autores de dichos trabajos finales, en todos los casos deberán ser cotejadas por los Profesionales y/o las fuentes.

DISEÑO: AXYOMA GLOBAL
Dir. Nahuel Gasparutti
DG. Victoria Salvarezza

Av. Luis Cándido Carballo 183
S2000DFB, Rosario, Argentina
+54 (341) 5899241 / +54 (341) 6155910

CONSTRUYENDO PUENTES

Transcurrido el segundo año de pandemia y adaptados en cierta medida a las nuevas maneras de trabajar y estudiar que posibilita la virtualidad, continuamos profundizando los lazos entre las instituciones que forman a los futuros graduados y las que luego los acompañarán durante su vida profesional.

En este sentido es que continuamos desarrollando el proyecto de las revistas digitales, elaboradas conjuntamente entre el Consejo Profesional en Ciencias Económicas de la Provincia de Santa Fe Cámara II, el Colegio de Graduados en Ciencias Económicas de Rosario y la Facultad de Ciencias Económicas y Estadística de la Universidad Nacional de Rosario.

Estas revistas, que compendian una selección de tesinas de grado y trabajos finales de las licenciaturas de las tres escuelas que integran la Facultad, tienen además como objetivo incentivar la investigación, fortalecer el progreso intelectual, y establecer un puente entre la vida académica y la profesional.

Nos enorgullece presentar la séptima edición de Dissertatio Economía, que dio el puntapié inicial de este camino en 2015, la quinta edición de Dissertatio Estadística y la cuarta de Dissertatio Administración, incorporadas en 2017 y 2018 respectivamente.

COMITÉ EDITORIAL

En la presente edición de Dissertatio Estadística (UNR-CGCE-CPCE) se publica la recensión de seis tesinas y de cuatro informes correspondientes a prácticas profesionales. En todos los casos, los trabajos fueron realizados por sus autores para cumplimentar con el último requisito obligatorio y poder alcanzar el título de Licenciado en Estadística, bajo la dirección y asesoramiento de docentes de la Escuela de Estadística, de la Facultad de Ciencias Económicas y Estadística (FCEyE) de la Universidad Nacional de Rosario (UNR) y en algunos casos, con la tutoría de profesionales estadísticos de instituciones públicas o privadas.

Esta quinta edición de la Revista Dissertatio se publica de manera ininterrumpida desde su primera aparición en el año 2017. Es de destacar la relevancia que tiene este tipo de revista digital al propiciar la difusión de una selección de trabajos finales de los egresados, teniendo además como objetivo incentivar la investigación y favorecer el tránsito de la vida académica a la profesional.

En los trabajos presentados en esta oportunidad, se puede encontrar una gran variedad de abordajes de problemáticas y metodologías. Esto es una muestra de la diversidad de campos en los que se aplica la Estadística. Se incluyen aplicaciones de métodos estadísticos en datos relacionados a la salud, como en los trabajos sobre sepsis materna (Lic. Gino Bartolelli) y fecundidad (Lic. Mariana Díaz). Aplicaciones en el área de agronomía (Lic. Lucas Peitton y Lic. Eugenia Settecase), en el área de seguros (Lic. Nadina Matteucci) y de la industria (Lic. Lucio Tinnirello). Así mismo, se muestra cómo puede colaborar la metodología estadística en problemas sociales y ambientales que atañen a la ciudad de Rosario: el manejo de los residuos sólidos (Lic. Cintia Cucco) y la vulnerabilidad ante la amenaza de inundaciones (Lic. Lautaro Ruiz). En esta misma línea, se incluye un trabajo sobre índices de inseguridad a nivel nacional, a partir de la Encuesta Nacional de Victimización 2017 (Lic. Danisa Rosatto). Por último, un trabajo se enfoca al estudio de técnicas de clasificación no supervisada en el área de grades bases de datos (Lic. Emanuel Ciardullo).

Es importante señalar que la versión completa de las tesinas y de los informes finales de las prácticas profesionales se puede consultar en la página web de la FCEyE: https://www.fcecon.unr.edu.ar/web-nueva/materias/tesina-o-practica-profesional-e-informe-final.

Desde el Comité Editorial aprovechamos la oportunidad para agradecer y felicitar a todas aquellas personas que hacen posible cada edición, en especial a quienes colaboraron en el presente año en un contexto tan particular como el dado por la pandemia COVID-19. Las instituciones participantes, Escuela de Estadística de la FCEyE, el Colegio de Graduados en Ciencias Económicas de Rosario y el Consejo Profesional en Ciencias Económicas de la Provincia de Santa Fe -Cámara II-, demuestran año a año un fuerte interés por seguir fortaleciendo este valioso espacio de difusión técnico estadístico, de acceso libre y gratuito al público en general. Una vez más, apostamos a seguir creciendo en próximas ediciones.

APLICACIÓN DE MODELOS JERÁRQUICOS MIXTOS EN UN ESTUDIO SOBRE SEPSIS MATERNA

Lic. Gino Bartolelli

Responsable de la Facultad de Ciencias Económicas y Estadística:
Mg. Cristina Cuesta
Responsable de la entidad: **Lic. Gabriela García Camacho**

Este trabajo es el producto de una Práctica Profesional llevada a cabo en el Centro Rosarino de Estudios Perinatales (CREP). Las técnicas estadísticas aplicadas y los resultados obtenidos responden a los objetivos planteados en un análisis secundario del Estudio Global de Sepsis Materna (GLOSS) dirigido por investigadores del CREP y del Departamento de Salud Sexual y Reproductiva de la Organización Mundial de la Salud (OMS).
El interés principal consistía en describir los centros de salud participantes del estudio GLOSS y evaluar su desempeño en administrar ciertas intervenciones, controlar los signos vitales de las pacientes y llevar a cabo exámenes clínicos y de laboratorio. Para esto, se calculó la distribución de las características de interés de los centros para cada nivel de ingreso de los países participantes y se evaluó en cada centro si las mujeres recibieron las prácticas mencionadas. Se ajustaron modelos jerárquicos mixtos con el fin de evaluar la asociación entre las características de los centros y la severidad de la infección que desarrollan las mujeres.
El análisis realizado es de gran importancia por ser el primero en estudiar la relación entre las características de los centros y el desenlace de interés.

INTRODUCCIÓN

La sepsis se define como una disfunción orgánica a causa de una respuesta desregulada del paciente ante una infección. Cuando esto ocurre en cualquier etapa del embarazo de una mujer recibe el nombre de sepsis materna. Actualmente, la sepsis materna es la tercera causa más importante de defunciones maternas en el mundo, representando alrededor del 11% de estas muertes (Say, *et al.*, 2014).

Con el objetivo de reducir los niveles de mortalidad y morbilidad a causa de sepsis en la población surge el Estudio Global de Sepsis Materna, el cual se llevó a cabo en centros de salud ubicados en áreas geográficas preespecificadas de los países participantes. Durante el período de reclutamiento desde el 28 de noviembre al 4 de diciembre de 2017, todas las mujeres admitidas en alguno de los centros participantes con sospecha de infección o infección confirmada en cualquier etapa de su embarazo fueron invitadas a formar parte del estudio.

En este análisis secundario titulado: "*Características de los centros de salud del estudio GLOSS y su relación con la identificación y el manejo de las infecciones maternas y su desenlace*", se intenta responder a tres objetivos:

- Caracterizar los centros de salud que formaron parte del estudio GLOSS.
- Evaluar el desempeño de los centros en administrar intervenciones específicas y en registrar los signos vitales de las pacientes y llevar a cabo exámenes clínicos y de laboratorio.
- Estudiar la asociación entre las características de los centros donde son admitidas las mujeres y la severidad de la infección que desarrollan.

A modo de caracterizar los centros de salud participantes del Estudio, se calcula la distribución de las características de los centros para cada nivel de ingreso de los países participantes, según la clasificación del Banco Mundial del 2017 (The World Bank, 2017).

En segundo lugar, para evaluar el desempeño de los centros en administrar las intervenciones, registrar los signos vitales de las pacientes y llevar a cabo los exámenes, se determina en cada centro su grado de cumplimiento a partir del porcentaje de mujeres que recibieron las prácticas mencionadas, clasificando los centros en cumplimiento "alto", "medio" o "bajo".

Por último, se ajustan modelos jerárquicos mixtos para estudiar la asociación entre las características de los centros de salud y la severidad de la infección que las mujeres desarrollan, modelando una función de la probabilidad de sufrir un evento materno severo.

DESCRIPCIÓN DE LOS CENTROS

La distribución de las características básicas de los centros para cada nivel de ingreso de los países participantes se muestra en Tabla 1. De forma análoga se describieron los centros de acuerdo a la disponibilidad de prácticas clínicas y recursos humanos, capacidad de cuidado obstétrico, adhesión a medidas de prevención y control de infecciones y disponibilidad de servicios de WASH.

Tabla 1. Características básicas de los centros por nivel de ingreso de los países participantes

Variable (n, %)	Total		Ingreso					
			Bajo		Medio-bajo		Medio-alto y Alto	
Nivel de la institución								
Información disponible	**445**	**(100,0)**	**108**	**(100,0)**	**193**	**(100,0)**	**144**	**(100,0)**
I	76	(17,1)	21	(19,4)	32	(16,6)	23	(16,0)
II	195	(43,8)	36	(33,3)	115	(59,6)	44	(30,6)
III	174	(39,1)	51	(47,2)	46	(23,8)	77	(53,5)
Tipo								
Información disponible	**446**	**(100,0)**	**108**	**(100,0)**	**193**	**(100,0)**	**145**	**(100,0)**
Público	342	(76,7)	81	(75,0)	171	(88,6)	90	(62,1)
Información disponible	**434**	**(100,0)**	**108**	**(100,0)**	**184**	**(100,0)**	**142**	**(100,0)**
Hospital Escuela	201	(46,3)	53	(49,1)	66	(35,9)	82	(57,7)
Localización								
Información disponible	**445**	**(100,0)**	**108**	**(100,0)**	**192**	**(100,0)**	**145**	**(100,0)**
Urbano	353	(79,3)	94	(87,0)	121	(63,0)	138	(95,2)
Tamaño (nacidos vivos/año)								
Información disponible	**436**	**(100,0)**	**108**	**(100,0)**	**188**	**(100,0)**	**140**	**(100,0)**
Pequeño (< 1000 NV)	115	(26,4)	11	(10,2)	60	(31,9)	44	(31,4)
Mediano (1000 ≤ NV < 2500)	123	(28,2)	33	(30,6)	46	(24,5)	44	(31,4)
Grande (2500 ≤ NV < 4500 NV)	90	(20,6)	20	(18,5)	34	(18,1)	36	(25,7)
Muy grande (≥ 4500 NV)	108	(24,8)	44	(40,7)	48	(25,5)	16	(11,4)

En países de ingreso medio-alto y alto, la fracción de centros de salud de nivel III (77/144; 53,5%), urbanos (138/145; 95,2%) y hospitales Escuela (82/142; 57,7%) es mayor en relación a países de menor ingreso. En países de ingreso medio-bajo, por ejemplo, sólo 46/193 (23,8%) son centros de nivel terciario, 121/192 (63,0%) son urbanos y 66/184 (35,9%) son hospitales Escuela (Tabla 1).

En los países de menor ingreso, hay una mayor proporción de centros de salud públicos (81/108; 75,0%) y centros muy grandes (44/108; 40,7%) comparado con los países de más alto ingreso, donde 90/145 (62,1%) son públicos y sólo 16/140 (11,4%) son centros muy grandes. Los centros de salud en países de altos recursos tienden a ser privados, pequeños y de alta

complejidad, mientras que los centros en países de menor ingreso suelen ser públicos, de mayor tamaño y menor complejidad (Tabla 1).

EVALUACIÓN DEL CUMPLIMIENTO DE LOS CENTROS

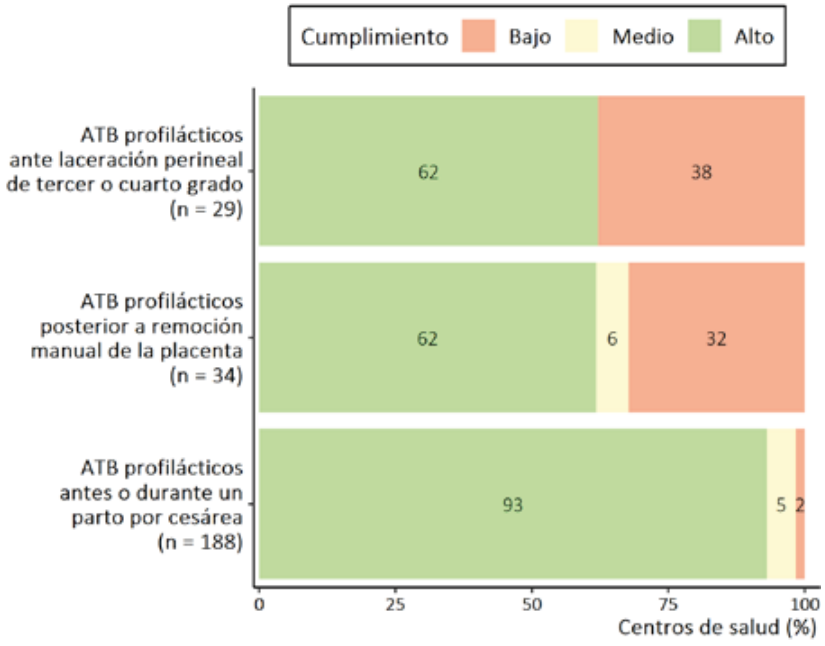
El desempeño de los centros se determina sobre dos prácticas diferentes:

- Uso profiláctico de antibióticos para prevenir infecciones en caso de parto por cesárea, desgarro perineal de tercer o cuarto grado y remoción manual de la placenta.
- Medición de los signos vitales de las pacientes, exámenes clínicos y de laboratorio.

Para medir el grado de cumplimiento de los centros en cada una de las prácticas se calcula el porcentaje de mujeres que la reciben, clasificando los centros en cumplimiento “alto” si al menos el 75% de las mujeres recibieron la intervención, cumplimiento “medio” si entre el 25 y el 75% de las pacientes recibieron la intervención y cumplimiento “bajo” si no más del 25% de las mujeres recibieron la intervención.

Con el propósito de evaluar el desempeño de los centros en cada práctica se calcula la distribución del grado de cumplimiento en el uso de antibióticos profilácticos (Figura 1) y la distribución del grado de cumplimiento en medir los signos vitales y llevar a cabo exámenes clínicos y de laboratorio (Figura 2).

Figura 1. Cumplimiento de los centros de salud en el uso de antibióticos profilácticos

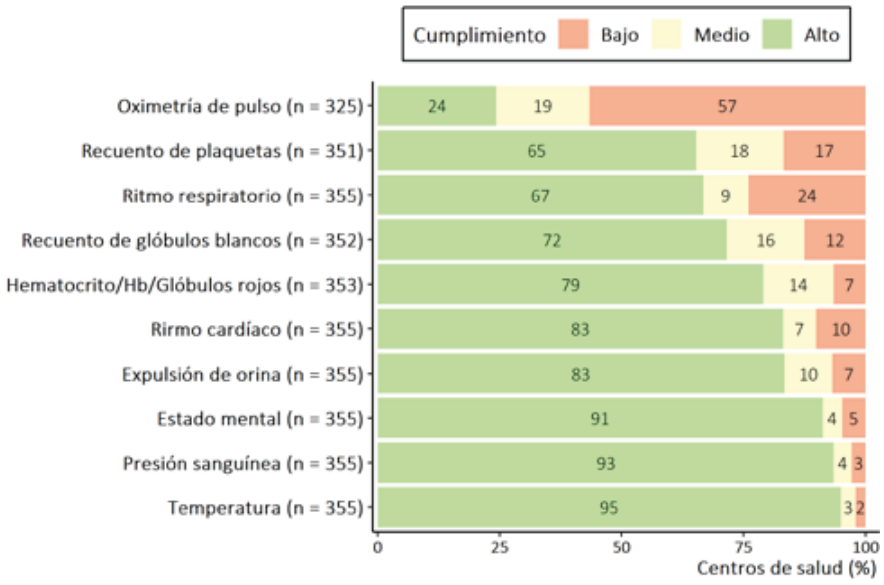


n = cantidad de establecimientos donde pudo realizarse la evaluación del cumplimiento

El grado de cumplimiento de los centros en cuanto al uso de antibióticos profilácticos antes o durante un parto por cesárea es muy alto. En efecto, 175/188 (93,1%) de los centros de salud tiene un cumplimiento alto en esta práctica. En cambio, el desempeño de los centros en cuanto al uso de antibióticos profilácticos posterior a la remoción manual de la placenta o ante una laceración perineal de tercer o cuarto grado es deficiente, ya que aproximadamente 1 de cada 3 centros muestra un bajo cumplimiento (Figura 1).

El grado de cumplimiento de los centros en la medición de los signos vitales y exámenes clínicos y de laboratorio varía entre los chequeos. Los centros muestran un buen desempeño en tomar la temperatura (337/355; 94,9% de alto cumplimiento), la presión sanguínea (332/355; 93,5% de alto cumplimiento) y el estado mental (324/355; 91,3% de cumplimiento alto). Su desempeño se deteriora en la medición de glóbulos rojos, expulsión de orina y ritmo cardíaco, ya que aproximadamente 1 de cada 5 centros tiene un cumplimiento medio o bajo en estos exámenes. El peor desempeño se observa en el recuento de glóbulos blancos, plaquetas, ritmo respiratorio y oximetría de pulso, donde el porcentaje de centros de cumplimiento intermedio o bajo supera el 25% en todos los exámenes. En el test de oximetría de pulso particularmente, más de la mitad de los centros (184/325; 56,6%) muestra un bajo cumplimiento (Figura 2).

Figura 2. Cumplimiento de los centros de salud en la medición de los signos vitales y en realizar exámenes clínicos y de laboratorio



n = cantidad de establecimientos donde pudo realizarse la evaluación del cumplimiento

No fue posible evaluar el cumplimiento en todos los centros debido a que no todos contaban con los protocolos necesarios para administrar los antibióticos y no todos disponían de la infraestructura para llevar a cabo algunos exámenes de laboratorio. Asimismo, no todos los centros admitieron mujeres que precisaran los antibióticos, por lo que tampoco fue posible evaluar el cumplimiento en estos centros.

ANÁLISIS DE ASOCIACIÓN

Se quiere estudiar la asociación entre las características de los centros de salud y la severidad de la infección que desarrollan las mujeres. La variable de interés principal mide si la mujer sufrió un evento materno severo (SMO, del inglés *severe maternal outcome*):

$$Y_i = \begin{cases} 1 \text{ en caso de SMO} \\ 0 \text{ en otro caso} \end{cases}$$

$$donde\, Y_i \sim Bernoulli\,(\pi_i)$$

Los eventos maternos severos incluyen las muertes maternas y los casos de *near-miss*, donde la mujer sobrevive a una condición potencialmente mortal (Pattinson, Say, Souza, Broek, & Rooney, 2009). Se modela una función de la probabilidad de SMO (π_i) conocida como función logit:

$$logit(\pi_i) = ln\left(\frac{\pi_i}{1-\pi_i}\right)$$

En el modelo se consideran predictores a nivel mujer, centro y país. Además, se incluye un efecto aleatorio correspondiente al país donde la mujer fue admitida y se estratifica el análisis en dos poblaciones de mujeres: las que ingresaron al estudio embarazadas o en trabajo de parto y las que fueron enroladas en período de posparto o posaborto. Los modelos se interpretan en términos de razones de *odds* acompañadas de su intervalo de confianza, calculados a partir de los coeficientes del modelo y su error estándar utilizando la aproximación Normal (Agresti, 2015).

El modelo elegido por los investigadores fue aquél que incluye todas las variables de interés en el análisis, denominado modelo completo:

$$logit\left(\pi_{ijk}\right) = \alpha + \beta_1 X_{1ijk} + \beta_2 X_{2jk} + \beta_3 X_{3k} + \delta_k\,, \qquad donde:$$

$$i = 1,\ldots, n_{jk} \quad j = 1,\ldots, n_k \quad k = 1,\ldots, K$$

π_{ijk} probabilidad de SMO de la i-ésima paciente, admitida en el j-ésimo centro del país k-ésimo

X_{1ijk}, X_{2jk} y X_{3k} predictores a nivel mujer, centro y país

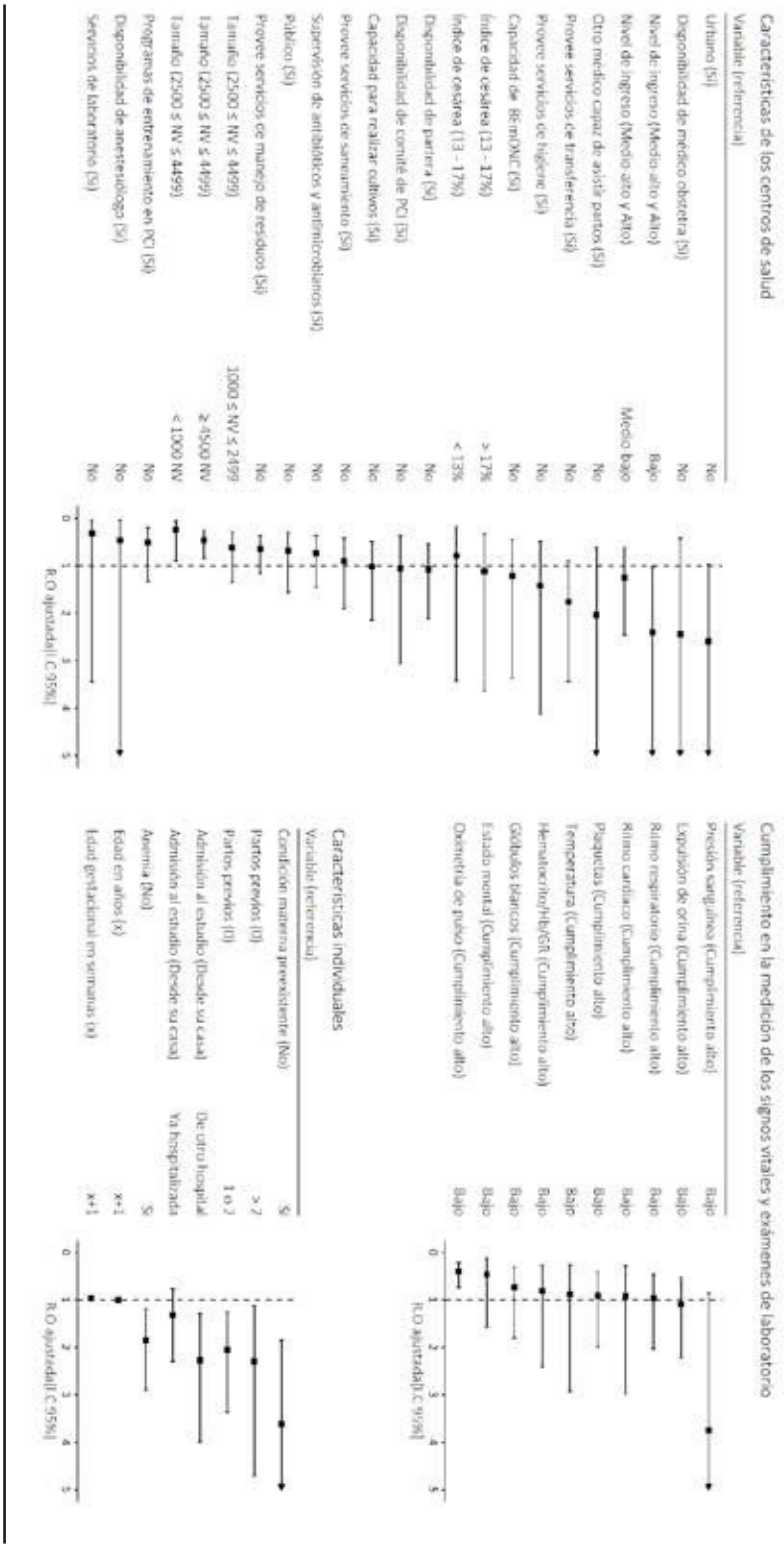
α, β_1, β_2 y β_3 coeficientes del modelo

δ_k efecto aleatorio del país k-ésimo / $\delta_k \sim N(0, \sigma_\alpha)$

El ajuste del modelo completo para la población de mujeres admitidas embarazadas o en trabajo de parto a la primer sospecha o diagnóstico de infección se presenta en Figura 3.

Figura 3. Ajuste del modelo completo - Mujeres embarazadas o en trabajo de parto a la primer sospecha o diagnóstico de infección

PCI = Prevención y control de infecciones
GR= Glóbulos rojos



Características de los centros de salud como el tamaño, el grado de cumplimiento en realizar el test de oximetría de pulso y el nivel de ingreso del país donde la mujer es admitida están

asociadas a la severidad de la infección que las pacientes desarrollan. El perfil individual de cada paciente está muy asociado a la chance de SMO (Figura 3).
Las mujeres admitidas en los centros más grandes y en los más chicos están sujetas a una menor chance de sufrir un SMO comparado con las admitidas en los centros de referencia. Esto podría deberse a que los centros más grandes están mejor preparados para recibir a las mujeres y los centros más pequeños no admiten los casos más graves, sino que los derivan a otros centros. Asimismo, la chance de SMO es mayor para las mujeres admitidas en centros de alto cumplimiento en la práctica de oximetría de pulso. Es probable que a las mujeres se les controle la saturación de oxígeno en sangre cuando presentan complicaciones, por esto los centros de alto cumplimiento en realizar la oximetría de pulso son los que reciben las pacientes más complicadas. Por último, la chance de SMO de las mujeres admitidas en países de bajo ingreso es más de 2 veces la chance para mujeres admitidas en países de ingreso medio-alto y alto.
Por otra parte, para cada población de mujeres se ajustó un modelo reducido siguiendo una estrategia de selección de modelo (Collet, 2003). Este proceso de selección consta de 5 pasos, donde los predictores entran o salen del modelo comparando las *Deviances* de modelos anidados:

1. Se ajustan modelos con un predictor por vez, comparando sus *Deviances* con la de un modelo sin predictores (modelo basal).
2. Todos los predictores significativos del paso anterior se utilizan para ajustar un modelo multivariado, reteniendo aquellos que en presencia de los demás sigan siendo significativos.
3. Sobre el modelo obtenido en el paso 2, se incluyen de a uno por vez los predictores excluidos en el primer paso (los que resultaron no significativos en los modelos univariados). Se retienen en el modelo aquellos predictores significativos en presencia del resto.
4. Se prueba incluyendo al modelo de a una por vez las interacciones de interés.
5. Como último paso se verifica que ningún predictor sale del modelo y ninguno ingresa en esta etapa.

El ajuste del modelo reducido a partir del proceso de selección para la población de mujeres embarazadas o en trabajo de parto se muestra en Tabla 2.

Tabla 2. Ajuste del modelo de regresión logística resultante del proceso de selección
Mujeres embarazadas o en trabajo de parto ($\hat{\sigma}_a = 0,25$)

Nivel	Variable (referencia)	Categoría		R.O [I.C 95%]
Institucional	Tamaño (2500 ≤ NV < 4500 NV)	≥ 4500 NV		0,56 [0,35 - 0,88]
		1000 ≤ NV < 2500 NV		0,56 [0,30 - 1,07]
		< 1000 NV		0,31 [0,12 - 0,81]
	Oximetría de pulso (Alto)	Bajo		0,47 [0,30 - 0,74]
Individual	Admisión (Desde su casa)	Desde otro hospital	0 partos previos	5,93 [2,73 - 12,89]
			1-2 partos previos	1,38 [0,35 - 5,34]
			>2 partos previos	1,63 [0,34 - 7,75]
		Ya hospitalizada	0 partos previos	1,79 [0,76 - 4,21]
			1-2 partos previos	1,76 [0,44 - 6,98]
			>2 partos previos	0,49 [0,03 - 7,86]
	Condición preexistente (No)	Si		4,62 [2,62 - 8,14]
	Anemia (No)	Sí	0 partos previos	3,26 [1,66 - 6,41]
			1-2 partos previos	1,72 [0,56 - 5,25]
			>2 partos previos	0,70 [0,18 - 2,82]
	Edad gestacional (x)	x+1		0,97 [0,95 - 0,99]

Aparte de los términos que resultaron significativos en el modelo completo, en el proceso de selección del modelo se detectaron interacciones asociadas a la chance de evento materno severo entre características individuales de la mujer (Tabla 2).
Entre mujeres que transcurren su primer embarazo, la chance de SMO es mayor para las que ingresan anémicas y para las transferidas desde otro centro de salud. En cambio, entre pacientes multíparas no se detectaron diferencias significativas entre ambos grupos.

COMENTARIOS FINALES

Los análisis estadísticos realizados en este trabajo responden a los objetivos planteados en un análisis secundario del Estudio Global de Sepsis Materna. El plan de análisis propuesto y llevado adelante fue discutido y consensuado en múltiples encuentros con profesionales de la Universidad Nacional de Rosario, del Centro Rosarino de Estudios Perinatales y de la Organización Mundial de la Salud. Los resultados fueron obtenidos con el software R, en el entorno de RStudio (R Core Team, 2020).
En un primer análisis descriptivo se caracterizaron los centros de salud participantes del estudio GLOSS, identificando perfiles institucionales diferentes en países de mayor y menor

ingreso. Los países de ingreso medio-alto y alto respecto a los de menor ingreso cuentan con centros más pequeños y especializados, tienen un mayor porcentaje de centros urbanos y privados, mayor disponibilidad de personal especializado y mayor adhesión a las medidas de prevención y control de infecciones. Por otra parte, los países de ingreso medio-bajo tienen el porcentaje más alto de centros rurales, públicos y de segundo nivel y los países de menor ingreso cuentan con el porcentaje más alto de centros de mayor tamaño.

Evaluando el cumplimiento de los centros en llevar a cabo las prácticas de interés para los investigadores, se observa un buen desempeño en administrar antibióticos profilácticos antes o durante un parto por cesárea y en medir algunos signos vitales como la temperatura y la presión sanguínea. Sin embargo, en otras mediciones como el ritmo respiratorio y la saturación de oxígeno, los centros están preparados para llevar a cabo los exámenes, pero una porción importante de las mujeres no los recibe.

Para la población de mujeres que ingresaron al estudio embarazadas o en trabajo de parto, los resultados del análisis de asociación indican que algunas características de los centros como el tamaño y el nivel de ingresos del país donde las pacientes fueron admitidas están asociadas a la severidad de la infección que desarrollan. El perfil individual de cada paciente está muy asociado a la chance de SMO, siendo que las mujeres que ingresan con una condición materna preexistente, las multíparas, las transferidas desde otro centro, las anémicas y las de menor edad gestacional están asociadas a una mayor chance de evento materno severo.

Los resultados de este análisis hacen a un mejor entendimiento del contexto en el que son admitidas las pacientes que sufren infecciones y cómo esto afecta su desenlace.

El haber identificado perfiles distintos de centros de salud en países de mayor y menor ingreso podría explicar la diferencia en los niveles de SMO en unos y otros. Los resultados del análisis de cumplimiento de los centros podrían ser la base para impulsar políticas públicas dirigidas a mejorar la atención que reciben las mujeres. Por último, los resultados del análisis de asociación permiten identificar grupos de riesgo sobre los cuales actuar con anticipación y de este modo prevenir los eventos maternos severos.

REFERENCIAS

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models.* John Wiley & Sons.

Collet, D. (2003). *Modelling Survival Data in Medical Research.* Chapman & Hall/CRC.

Pattinson, R., Say, L., Souza, J., Broek, N., & Rooney, C. (2009). WHO maternal death and near-miss classifications. *Bulletin of the World Health Organization*.

R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Say, L., Chou, D., Gemmill, A., Tunçalp, Ö., Moller, A.-B., Daniels, J., . . . Alkema, L. (2014). Global causes of maternal death: a WHO systematic analysis. *Lancet Global Health*. doi:https://doi.org/10.1016/S2214-109X(14)70227-X

The World Bank. (2017). *The World by Income and Region.* Recuperado el 23 de Agosto de 2020, de https://datatopics.worldbank.org/world-development-indicators/the-world-by-income-and-region.html

ESTUDIO COMPARATIVO DE MÉTODOS DE CLASIFICACIÓN NO SUPERVISADA EN CONTEXTOS DE GRANDES BASES DE DATOS

Lic. Emanuel Ciardullo

Directora: **Dra. Marta Quaglino**
Co-Director: **Mg. Leandro Kovalevski**

En estadística se conoce como análisis *cluster* al estudio formal de los métodos para el agrupamiento de objetos según las características intrínsecas de los mismos. Estos métodos, tienen por objetivo obtener grupos dentro de los cuales los individuos, que a priori conforman un grupo heterogéneo, sean homogéneos entre sí y distintos de los pertenecientes a otro grupo.
En general cuando se pretende agrupar objetos no existe una forma única de lograrlo. Distintos métodos pueden derivar en diferentes configuraciones. El presente trabajo compara los métodos K-means, K-medoid, DBSCAN y algoritmo EM a fin de descubrir ventajas y desventajas en su aplicación.
Para evaluar la capacidad de las distintas estrategias de clasificación escogidas para identificar grupos que representen a ciertas estructuras que pueden aparecer en casos reales, se realiza un estudio por simulación. La bondad de la clasificación en cada caso, se estudia a través del estadístico de Hopkins, del criterio denominado Variación de la Información propuesto por Marina Meilă (2003), del Índice Silhouette y la proporción de datos correctamente clasificados.

INTRODUCCIÓN

Estadísticamente se pueden distinguir dos enfoques del problema de clasificación: clasificación supervisada o discriminante y clasificación no supervisada también conocida como análisis *cluster* o simplemente como clasificación.

En la clasificación no supervisada el objetivo es tratar de descubrir grupos de individuos que están naturalmente presente en los datos, pero no se dispone de información *a priori* que permita relacionar a cada individuo con un grupo.

Los grandes avances en la capacidad de almacenamiento y procesamiento de datos, así como el desarrollo de nuevos métodos y algoritmos de agrupamiento permitieron que muchos campos de la ciencia aprovecharan aún más el análisis *cluster*. Por ejemplo, en la biología al estudiar de qué manera se agrupan los genes y qué efectos tienen estas agrupaciones, en marketing cuando se busca agrupar a los clientes según sus hábitos de consumo y preferencias de compra o en seguros al clasificar la cartera de pólizas para identificar qué grupos de personas presentan más riesgo de tener siniestros.

Se pueden encontrar cientos de algoritmos de *clustering* propuestos a través de las distintas disciplinas científicas, además de las modificaciones y adaptaciones de estos a casos particulares. Esto hace extremadamente difícil hacer una revisión extensa y comparar entre sí todos los métodos publicados.

El presente trabajo compara los métodos *K-means*, *K-medoid*, DBSCAN y algoritmo EM a fin de descubrir ventajas y desventajas en su aplicación. Para evaluar la capacidad de las distintas estrategias de clasificación escogidas para identificar grupos que representen a ciertas estructuras que pueden aparecer en casos reales, se realiza un estudio por simulación.

OBJETIVO

En este documento se plantea como objetivo principal el estudio de un conjunto de técnicas de clasificación no supervisada, que permitan hacer frente a las dificultades que generalmente se presentan al tratar de identificar grupos de datos homogéneos en grandes bases de datos. Además de comparar la eficiencia de la clasificación a través de distintos criterios.

MÉTODOS DE CLASIFICACIÓN

Se proponen cuatro métodos de clasificación a comparar. Dos de los métodos propuestos, *K-means* y *K-medoids* se basan en optimizar un criterio definido *a priori* de forma iterativa hasta converger. Cuando se emplean estos algoritmos para buscar grupos en los datos se asume que los *clusters* son esféricos, es decir que las variables tienen la misma variancia y no están correlacionadas. Este supuesto resulta muy fuerte cuando se trabaja con datos multivariados, en los que existe correlación entre las variables dándole una forma elipsoidal a los *clusters*. Por este motivo, se incluye también el estudio de métodos de clasificación basados en modelos (algoritmo EM) y basados en densidades (DBSCAN) que pueden encontrar cualquier tipo de agrupaciones, no solo aquellas que corresponden a formas esféricas.

K-means es un algoritmo de clasificación no supervisada, que agrupa objetos inicialmente no clasificados, en K grupos (*clusters*), basándose en un vector multidimensional de características cuantitativas. El algoritmo requiere un parámetro de entrada K que representa el número de *clusters*, el cual debe ser conocido o fijado en un valor *a priori* antes de iniciar el análisis. El método es iterativo. Se conforman inicialmente y con un criterio arbitrario, *K* grupos, con el conjunto total de individuos. Luego se reasigna cada observación a uno de los K grupos, según un criterio de optimización predefinido. El criterio de optimización se basa en la similaridad entre las observaciones.

K-medoids es un algoritmo de *clustering* relacionado con *K-means*, la diferencia entre ellos es que *K-means* busca minimizar la suma de los errores al cuadrado, mientras que *K-medoids* minimiza la suma de disimilaridades entre cada uno de los puntos de un *cluster* y el centro del mismo, que también es un punto perteneciente al *cluster*. El algoritmo *K-means* utiliza al valor medio de los objetos del *cluster* como centroide, este punto generalmente no coincide con algún punto de la base original. En contraste a esto, el algoritmo *K-medoids* siempre utiliza como punto central (o *medoid*) a un punto de la base de datos.

El *clustering* basado en modelos, como su nombre indica, trata de identificar *clusters* entre los individuos asumiendo que los datos provienen de un modelo de probabilidad identificado, en general distribuciones mixtas. El algoritmo EM es un método de *clustering* basado en modelos, propuesto por Dempster *et al.* (1977). Este método, trata a los datos como si existiera información perdida; la información perdida es una variable indicadora que relaciona a cada observación particular con un *cluster*. El algoritmo que se emplea para realizar el agrupamiento, de forma iterativa, encuentra las estimaciones máximo verosímiles de un conjunto de parámetros de un modelo estadístico y calcula la probabilidad de que cada observación pertenezca al *cluster k*. Una vez finalizado el proceso iterativo cada observación se asigna al *cluster* con el que presenta la mayor probabilidad de pertenencia.

El método de *clustering* basado en densidades denominado DBSCAN (*Density-based spatial clustering of applications with noise*) forma los *clusters* con conjuntos de puntos que son vecinos cercanos e identifica como *outliers* a aquellos que se encuentran aislados en regiones de baja densidad. El algoritmo considera como criterio para la formación de los *clusters*, que una observación forma parte de un grupo si hay una cierta cantidad mínima de observaciones dentro de un radio de proximidad, y que los *clusters* deben estar separados por regiones de baja densidad de observaciones. Para su aplicación se requieren dos parámetros de entrada, el radio que define la región vecina a una observación y el número mínimo de observaciones dentro de dicha región denominada región épsilon.

EVALUACIÓN DE LA PERTINENCIA Y CALIDAD DE LA CLASIFICACIÓN

Los algoritmos de clasificación son utilizados para conformar grupos dentro de conjuntos de datos en situaciones donde el investigador ignora *a priori* si éstos realmente existen y cuántos son. Por lo tanto, es conveniente hacer evaluaciones *a priori*, para identificar la posible existencia de grupos y, una vez obtenido el agrupamiento, para analizar su estructura de modo de confirmar o modificar el número de grupos.

El estadístico de Hopkins es un criterio para evaluar si se identifican *a priori* agrupamientos en un conjunto de datos. Está basado en la estadística de prueba de un test que contrasta si los objetos del conjunto provienen de una distribución uniforme en el espacio multidimensional o si, por el contrario, responden a distribuciones diferentes, que justifiquen la búsqueda de grupos o *clusters*. El estadístico varía entre 0 y 1, y a mayor valor, mejor posibilidad de clasificación.

Como medida *a posteriori* se considera la denominada Silhouette, que pretende establecer una medida global de calidad del agrupamiento una vez aplicado un algoritmo de clasificación. Otra medida *a posteriori* es el denominado criterio de la Variación de la Información (VDI) que fue propuesto por Marina Meilă (2003) y mide la cantidad de información que se gana (o se pierde) al considerar agrupamientos diferentes.

SIMULACIÓN

A fin de evaluar la capacidad de distintas estrategias de clasificación para identificar grupos que representen a ciertas estructuras que pueden aparecer en casos reales, se realiza un estudio por simulación. Dentro de este trabajo se han considerado dos grupos de escenarios.

Escenarios con *p* = 2

En estos escenarios se ha restringido a dos el número de variables *(p)* medidas sobre cada individuo, con el único objeto de poder obtener una representación gráfica tanto de la estructura de los grupos originales a identificar por las técnicas de clasificación, como de los resultados logrados con la metodología de *clusters* utilizada. Los vectores fueron simulados a partir de densidades normales o uniformes con distintas estructuras de correlación. En total se simularon 8 casos distintos dentro de los cuales, los casos 1 a 4 estaban formados por grupos separables o cuasi separables linealmente y los casos 5 a 8 por grupos no separables linealmente.

Escenarios con *p* > 2

Se simularon cuatro escenarios considerando 10 variables cuantitativas continuas y cuatro poblaciones en cada uno. En todos ellos, los grupos de observaciones provienen de distintas poblaciones gaussianas mixtas con y sin contaminación, y distintos niveles de solapamiento entre las poblaciones. Todos los casos se simularon con una cantidad total de observaciones a clasificar de 1.000, 5.000, 10.000, 50.000, 100.000, 250.000, 500.000, 1.000.000 y 5.000.000 de observaciones. Los tamaños de grupo se mantuvieron iguales en todas las simulaciones.

RESULTADOS

Escenarios con *p* = 2

El estadístico de Hopkins para cada una de las poblaciones simuladas de datos bivariados se muestra en la Tabla 1.

Tabla 1. Valor del estadístico de Hopkins (H) en los escenarios bivariados

Casos	Caso 1	Caso 2	Caso 3	Caso 4	Caso 5	Caso 6	Caso 7	Caso 8
H	0,757	0,705	0,932	0,819	0,831	0,773	0,889	0,874

A fin de cuantificar la calidad de las clasificaciones, se muestra en la Tabla 2 las proporciones de observaciones correctamente clasificadas, el índice Silhouette y los VDI respectivamente, para los casos separables linealmente.

Tabla 2: Medidas de validación para escenarios con $p = 2$ separables linealmente

Medida de validación	Método	Caso 1	Caso 2	Caso 3	Caso 4
Proporción de aciertos	*K-means*	0,63	0,83	0,83	0,97
	K-medoid	0,98	1,00	1,00	0,97
	EM	0,99	1,00	1,00	0,96
	DBSCAN	0,71	0,86	0,77	0,34
Índice Silhouette	*K-means*	0,53	0,42	0,65	0,60
	K-medoid	0,63	0,57	0,72	0,60
	EM	0,63	0,57	0,72	0,60
	DBSCAN	0,55	0,5	0,58	0,42
Criterio de la Variación de la Información	*K-means*	0,43	0,14	0,31	0,25
	K-medoid	0,12	0,00	0,00	0,27
	EM	0,07	0,00	0,00	0,31
	DBSCAN	0,80	0,22	0,08	0,67

Las medidas de bondad de la clasificación para los casos no linealmente separables se muestran en la Tabla 3.

Tabla 3. Medidas de validación para escenarios con $p = 2$ no separables linealmente

Medida de validación	Método	Caso 5	Caso 6	Caso 7	Caso 8
Proporción de aciertos	*K-means*	0,75	0,90	0,68	0,45
	K-medoid	0,75	0,89	0,69	0,45
	EM	0,73	0,90	0,67	0,48
	DBSCAN	0,82	0,95	0,75	0,59
Índice Silhouette	*K-means*	0,51	0,61	0,52	0,59
	K-medoid	0,52	0,61	0,53	0,59
	EM	0,49	0,54	0,50	0,39
	DBSCAN	0,46	0,47	0,47	0,31
Criterio de la Variación de la Información	*K-means*	0,76	0,66	0,85	0,90
	K-medoid	0,84	0,67	0,88	0,90
	EM	0,69	0,61	0,82	0,94
	DBSCAN	0,36	0,06	0,73	0,26

Escenarios con *p* > *2*

En la Figura 1 se observan la proporción de aciertos en cada simulación para cada uno de los algoritmos de clasificación considerados en este trabajo. En la Figura 2 se muestran los valores del criterio de la variación en la información.

Figura 1. Proporción de aciertos por método de clasificación y escenario con $p = 10$

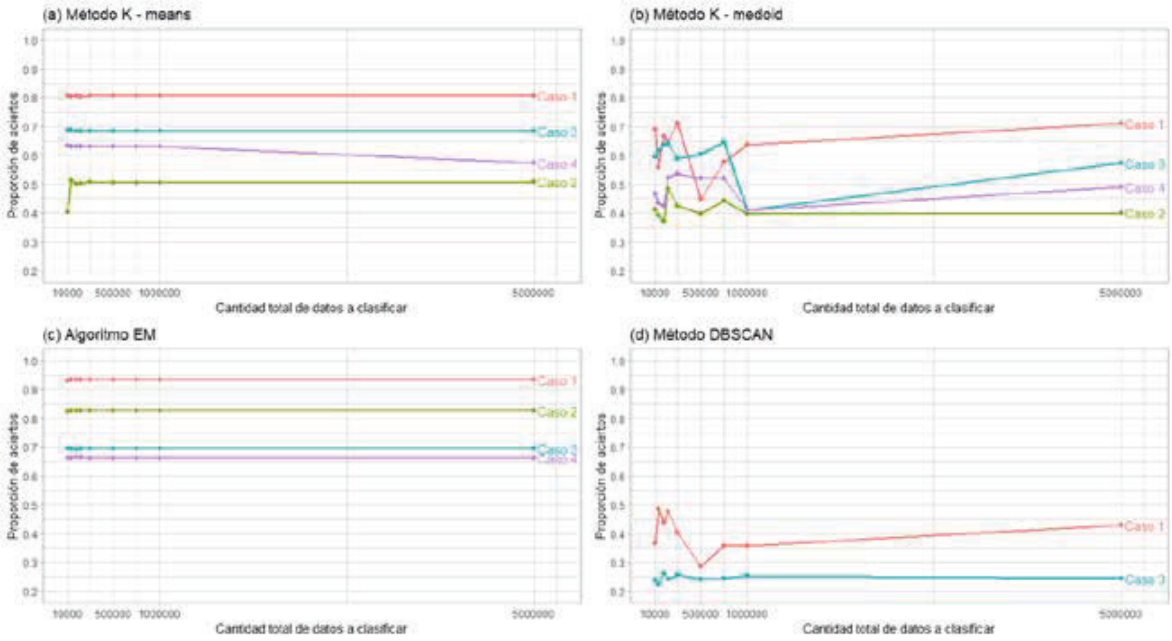
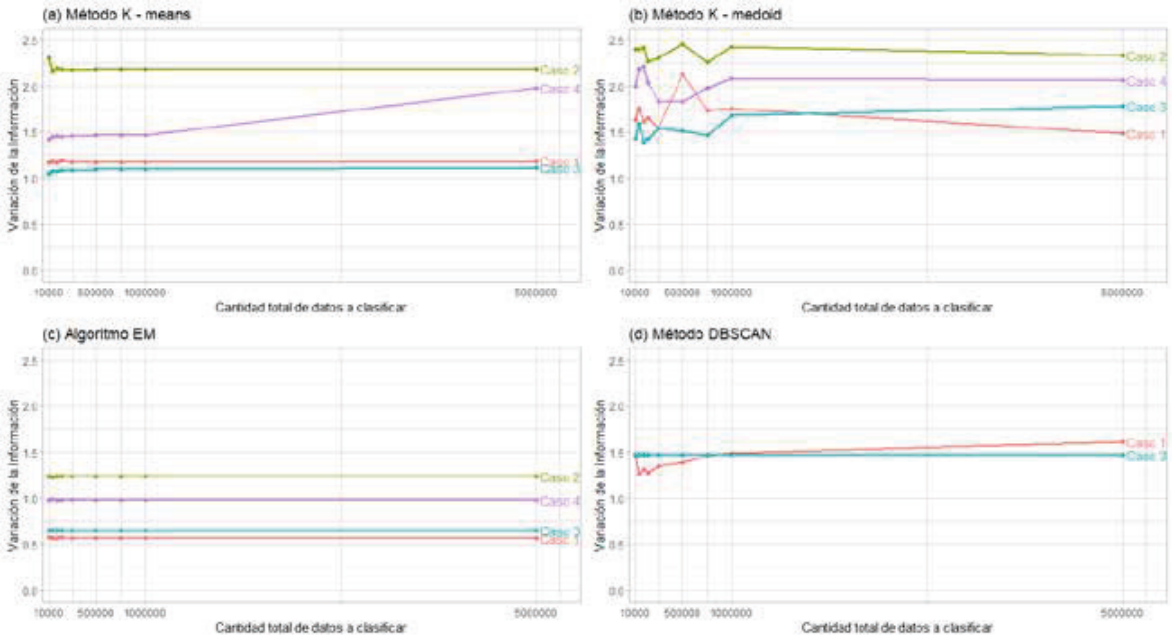


Figura 2. Criterio de la Variación de la información por método y simulación



CONCLUSIONES

De los resultados obtenidos a partir de las simulaciones en dos dimensiones es posible concluir que todos los métodos son capaces de identificar *clusters* cuando ellos son linealmente separables. En estos casos cuando los *clusters* están cercanos, la presencia de *outliers* puede producir distorsiones en el agrupamiento. En el caso de *K-means*, los *outliers* pueden producir un movimiento del centroide de los grupos identificados por el algoritmo, llevándolo a asignar

observaciones de distintos grupos a un mismo *cluster*. El método de *K-medoid*, en esta situación es más robusto y no distorsiona la clasificación lograda por el método, aun utilizando el algoritmo CLARA que trabaja sobre subconjuntos de los datos.

Con $p = 2$ y frente a *clusters* no separables linealmente, el único método capaz de identificar las agrupaciones originales es DBSCAN, todos los demás algoritmos generaron agrupaciones artificiales que se alejan de la representación de los verdaderos grupos existentes en los datos.

En forma general, considerando los ocho escenarios bivariados simulados, se deduce que no existe un único método que resulte consistentemente superior entre los cuatro considerados para la clasificación. Algunos de ellos presentan resultados inaceptables por ejemplo *K-means* frente a grupos parabólicos opuestos o DBSCAN frente a grupos solapados, los cuales configuran al final del proceso de clasificación, grupos completamente distintos a los originales. El método DBSCAN si bien funciona de forma aceptable en siete de los ocho escenarios, falla en los casos que presentan una pequeña superposición de los *clusters*.

Cuando se analizan datos provenientes de poblaciones de mayor dimensión y se aumenta el número de observaciones que conforman las bases de datos a clasificar, los mejores resultados se obtienen con el algoritmo EM. Este resultado puede deberse a que los datos son simulados a partir de distribuciones gaussianas mixtas, que es el modelo supuesto al aplicar la rutina para el algoritmo EM. En contraste con los resultados obtenidos de los datos bivariados, *K-medoid* resultó más desfavorable que *K-means* y EM cuando se analiza la proporción de aciertos entre la clasificación original y la encontrada por el método. Además, tanto *K-means* como EM demostraron ser estables con la calidad de los resultados obtenidos al aumentar el tamaño de los grupos a clasificar. Este resultado es consistente con lo evidenciado por los métodos *K-medoid* y DBSCAN.

El método DBSCAN, tanto en las simulaciones de datos bivariados como en las de mayor dimensionalidad, fue incapaz de identificar los agrupamientos cuando existía una cierta superposición de los grupos. Dada la posible sensibilidad del método frente a la definición de los parámetros de entrada, las simulaciones fueron replicadas considerando distintos valores y combinaciones de los parámetros de entrada, no logrando una mejora de los resultados.

BIBLIOGRAFÍA

Alpaydin, E. (2004*). Introduction to Machine Learning*. 3rd edition. Adaptive Computation and Machine Learning.

Campello, R.J.G.B., Moulavi, D. & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science, vol 7819. Springer, Berlin, Heidelberg.

Cebeci, Z. & Yildiz, F. (2015). Comparison of k-means and fuzzy c-means algorithms on different cluster structures. Agrárinformatika/journal of agricultural informatics, 6(3), 13-23.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological).

Everitt, B. S., Landau S., Leese M. & Stahl D. (2011). *Cluster Analysis*, 5th edition. John Wiley & Sons.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The elements of statistical learning, 2nd edition.* New York, NY: Springer.

Jain, A.K. (2010) Data Clustering: 50 Years Beyond K-means. Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science, vol 5211. Springer, Berlin, Heidelberg.

Jain, A. K. & Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice Hall Advance Reference Series.

Meilă, M. (2003). Comparing Clusterings by the Variation of Information. Learning Theory and Kernel Machines. Lecture Notes in Computer Science, vol 2777. Springer, Berlin, Heidelberg.

ANÁLISIS DE LA EVOLUCIÓN DE LOS RESIDUOS SÓLIDOS URBANOS DE LA CIUDAD DE ROSARIO

Lic. Cintia Belén Cucco

Responsable de la Facultad de Ciencias Económicas y Estadística:
Mg. Javier Bussi
Responsable de la entidad: **Lic. Nora Ventroni**

Este trabajo tiene como objetivo identificar el estado de situación de los Residuos Sólidos Urbanos (RSU) generados en la ciudad de Rosario. Se propone trabajar en la búsqueda de herramientas para predecir la cantidad de toneladas de residuos que la ciudad genera.
Se cuenta con una serie de datos mensuales que permite enfocar el estudio en la disposición final de RSU de la ciudad. En este marco, se propone desarrollar una metodología que permita construir modelos de series temporales para explicar la evolución histórica y el comportamiento de los RSU. El análisis de la serie se lleva a cabo utilizando el enfoque de los modelos *ARIMA* (Auto Regressive Integrated Moving Average). La finalidad de la construcción de estos modelos es la predicción a corto plazo. Luego se recurre a la búsqueda de información adicional que podría contribuir a mejorar estos pronósticos. A falta de datos medidos para la ciudad, se considera la variable Tasa de Desocupación para el Aglomerado Gran Rosario, para ajustar modelos *RegARIMA*.
Finalmente se llega a las conclusiones del estudio con algunas recomendaciones y sugerencias de la investigación y profundización en el tema vinculado a RSU de la ciudad de Rosario.

INTRODUCCIÓN

El presente informe es producto de la realización de una Práctica Profesional correspondiente a la carrera Licenciatura en Estadística, llevada a cabo entre febrero y mayo del año 2019 en las instalaciones de la Municipalidad de Rosario, alternando entre la Dirección General de Gestión Integral de Residuos y la Dirección General de Estadística.

Uno de los desafíos más importantes de las ciudades a nivel mundial está vinculado con el consumo y la excesiva generación de residuos. Se entiende por Residuo a "cualquier producto en estado sólido, líquido o gaseoso procedente de un proceso de extracción, transformación o utilización, al que su propietario decide abandonar o desprenderse, debido a que carece de valor para él o ya no puede ser utilizado para el uso que fue adquirido o creado". Qué hacer con los residuos y cómo aprovecharlos parece ser una de las metas cruciales para las nuevas generaciones.

El diseño de planes y programas vinculados a este objetivo demanda el estudio y desarrollo de herramientas que permitan estimar el estado de situación del cual parten los municipios y realizar un monitoreo riguroso que garantice un seguimiento vinculado a las toneladas de residuos que se generan de manera diaria y que varían de forma dinámica y constante, de acuerdo al consumo progresivo y los contextos socioeconómicos.

Las preocupaciones que genera el cambio climático global han inspirado a movimientos ambientalistas a desarrollar un concepto mundial de Basura Cero. En la ciudad de Rosario la ley vinculada con el tema es la N°8335 del año 2008, que obliga al Municipio a una reducción progresiva: "Se establece un cronograma de reducción progresiva de la cantidad de residuos depositados en rellenos sanitarios, sentando como base el total de los residuos dispuestos en el año 2006. La meta para el año 2010 será la reducción de un 15% del peso de los residuos sólidos urbanos de la ciudad de Rosario dispuesto en rellenos sanitarios, para el año 2012 de un 25% del total del peso y un 50% para el 2017. Se prohíbe para el año 2020 la disposición final en relleno sanitario de materiales tanto reciclables como aprovechables, incluyendo los residuos orgánicos".

OBJETIVO

Uno de los objetivos de la práctica profesional fue la búsqueda de herramientas para predecir la cantidad de toneladas de RSU que la ciudad genera. A partir de este dato, se podrá planificar y direccionar las acciones tendientes a la disminución progresiva de los residuos enterrados en relleno sanitario.

MATERIAL Y MÉTODO

Se trabaja con los datos disponibles de Toneladas de RSU de la ciudad de Rosario. Se utiliza la serie mensual "Cantidad de residuos sólidos urbanos enviados a relleno sanitario", para el período 2004-2018. También se evalúan datos provenientes de la Encuesta Permanente de Hogares (EPH), bajo la hipótesis de que un aumento en la Tasa de Desocupación podría estar relacionado con un menor consumo y una mayor presencia de recuperadores urbanos en la vía pública, redundando en una menor cantidad de RSU para su disposición final.

Modelos *ARIMA* y *RegARIMA*

Los modelos $ARIMA$ explican la progresión de una variable en función de la variación de ella misma. No toman en cuenta los factores exógenos. Así, en el caso de pronosticar una serie de tiempo, los modelos $ARIMA$ no se ven afectados explícitamente por la variación de otras variables. Si existe alguna variable relacionada con la variable en estudio se puede construir un

modelo que tenga en cuenta esta relación. Con este objetivo, se presentan los modelos $RegARIMA$ que consisten en modelar la variable dependiente con una regresión lineal y, los residuos obtenidos, con un modelo $ARIMA$.

Los datos

Para este estudio se cuenta con 186 datos mensuales. La serie a modelar corresponde a la cantidad (en Tn) de RSU de la ciudad de Rosario enviados a disposición final, en el período comprendido entre enero de 2004 y junio de 2019.

Se utilizan modelos $ARIMA$ para la serie univariada y luego se considera una variable independiente para ajustar modelos $RegARIMA$, con el fin de mejorar la predicción.

Entre los datos disponibles se consideró conveniente trabajar con los provistos por la EPH para el Aglomerado Gran Rosario referidos al tema Ocupación. Dichos datos son provistos en forma trimestral y la ciudad de Rosario representa aproximadamente un 78% del Aglomerado. La variable seleccionada para trabajar es la Tasa de Desocupación, calculada como porcentaje entre la población desocupada y la población económicamente activa.

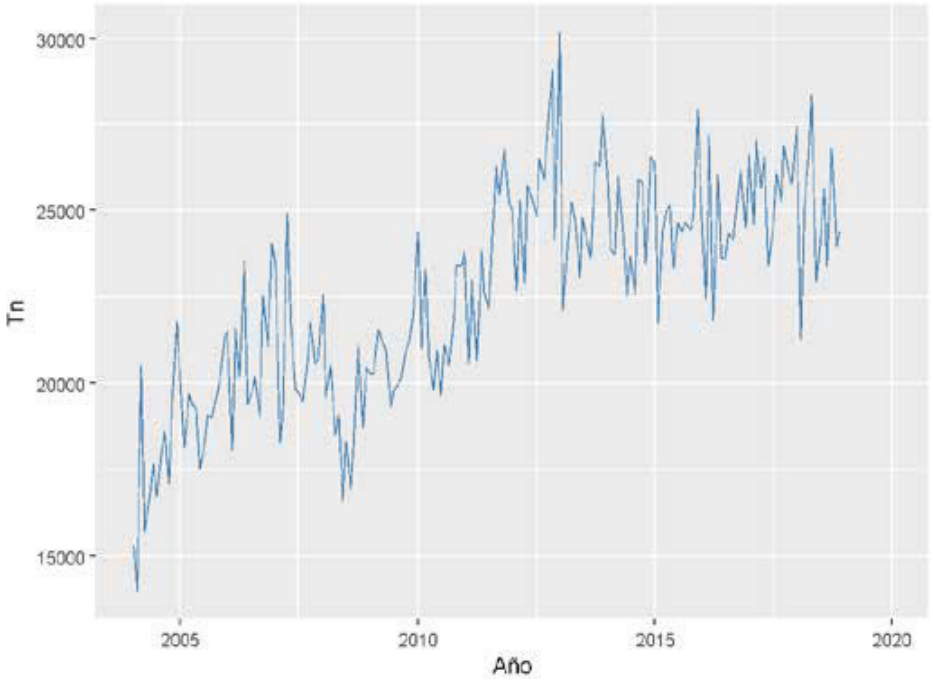
Para hacer previsiones, se utilizan las observaciones hasta diciembre 2018 y, se reservan los primeros 6 meses del año 2019, para evaluar la capacidad predictiva del modelo estimado. El análisis estadístico, se realiza utilizando el *software* R.

APLICACIÓN

Modelos *ARIMA*

Para realizar la identificación del modelo, se examina el comportamiento de la serie "cantidad de RSU de la ciudad de Rosario enviados a disposición final" por mes a través del tiempo mediante la representación gráfica (Gráfico 1).

Gráfico 1. Cantidad (en Tn) de RSU enviados a disposición final. Años 2004-2018



Mediante la observación de la Función de Autocorrelación Muestral ($FACM$) y la Función de Autocorrelación Parcial Muestral ($FACPM$) se determinan el número de diferencias y los órdenes de p y q del modelo. Una vez identificado el modelo se procede a su estimación y validación.

El *software* R dispone del procedimiento $ARIMA$ para realizar estimación y validación, así como para obtener predicciones de la serie.

El modelo $ARIMA\,(0,1,1)(1,0,0)_{12}$ es estadísticamente adecuado para explicar el comportamiento de la cantidad de RSU enterrados en el relleno sanitario de la ciudad de Rosario.

Modelo estimado: $(1-B)\left(1-0.4503B^{12}\right)\hat{z}_t = (1-(-0.7778\,B))\hat{a}_t$

Se calculan los errores porcentuales y el $PSMAPE$ (error medio absoluto porcentual post muestral). El ajuste se realiza sobre el período 2004-2018 y los pronósticos se extienden hasta el primer semestre del año 2019. Se encuentra que en todos los meses el intervalo de predicción contiene los valores originales de la serie. Los errores de los pronósticos resultan inferiores al 7,20%. El valor de la estadística $PSMAPE_6$ es de 2,92%, o sea que los pronósticos para los seis primeros meses del año 2019 difieren en promedio un 2,92% de los valores originales (Tabla 1 y Gráfico 2). El Gráfico 3 muestra la serie observada hasta diciembre 2018 y los pronósticos para los primeros seis meses del año 2019.

Tabla 1. Comparación de valores reales versus pronósticos obtenidos con el modelo $ARIMA\,(0,1,1)(1,0,0)_{12}$ para la serie cantidad de RSU enviados a disposición final

Mes	Valores originales	Pronósticos	Error Porcentual	PSMAPE	LI (95%)	LS (95%)
ene-19	26.843	25.591	4,67	4,67	22.463,68	28.717,72
feb-19	24.569	22.807	7,17	5,92	19.603,87	26.010,46
mar-19	24.923	24.504	1,68	4,51	21.226,14	27.781,74
abr-19	24.596	25.244	-2,64	4,04	21.893,31	28.594,60
may-19	25.864	25.996	-0,51	3,33	22.574,24	29.418,12
jun-19	23.751	23.555	0,83	2,92	20.063,02	27.046,58

Gráfico 2. Comparación de valores reales versus pronósticos con el modelo $ARIMA\ (0,1,1)(1,0,0)_{12}$ para la serie cantidad de RSU enviados a disposición final

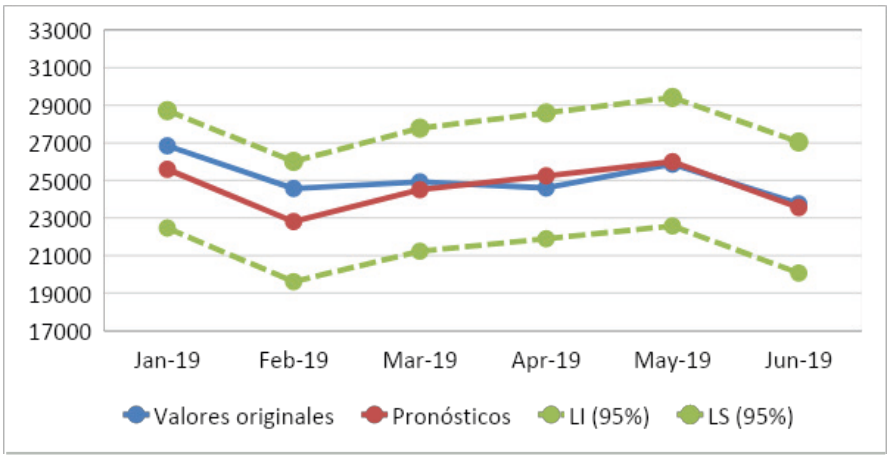
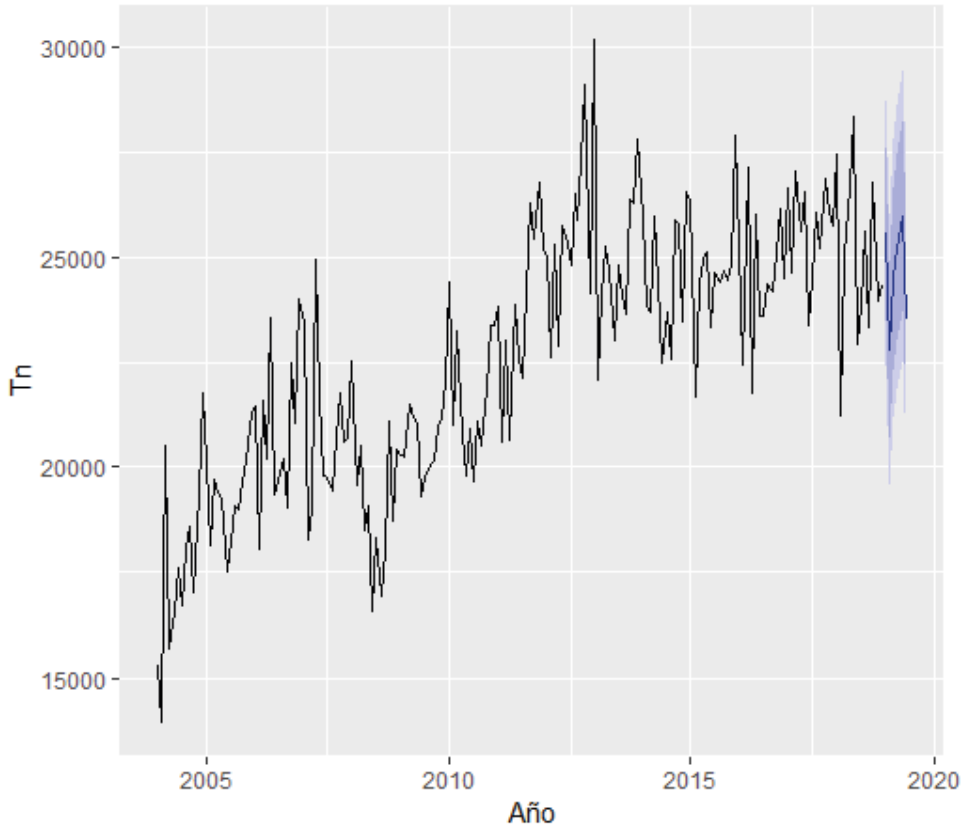


Gráfico 3. Pronósticos obtenidos con el modelo $ARIMA(0,1,1)(1,0,0)_{12}$ para la serie cantidad de RSU enviados a disposición final



Modelos RegARIMA

Para mejorar los pronósticos provistos por el modelo $ARIMA$ univariado, se planteó la posibilidad de construir modelos $RegARIMA$.

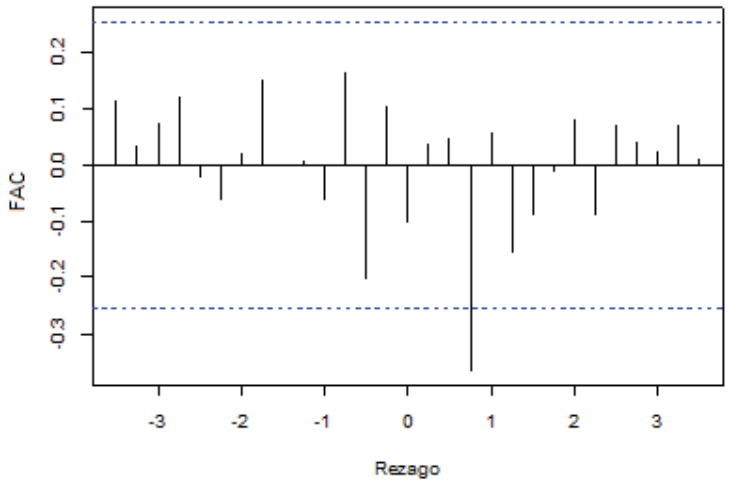
Entre un conjunto de variables se decidió trabajar con la Tasa de Desocupación (provista trimestralmente por la EPH para el Aglomerado Gran Rosario), como serie explicativa del comportamiento de la disposición final de RSU, con el objetivo de mejorar los resultados obtenidos con el ajuste $ARIMA$, bajo la premisa de que podía existir una correlación inversa, es decir, menor desempleo podría redundar en una mayor generación de residuos, y una mayor tasa, en una disminución de residuos.

Dado que la serie Tasa de Desocupación está disponible en forma trimestral, se construye la serie trimestral de la variable de interés, agregándose los valores mensuales por trimestre. A dicha serie trimestral se le ajusta un nuevo modelo $ARIMA$ univariado.

Teniendo en cuenta las correlaciones cruzadas significativas se ajustan modelos de regresión con errores $ARIMA$ para tratar de explicar la variable toneladas trimestrales de Disposición Final de RSU.

En el Gráfico 4 se presenta la correlación cruzada entre los residuos de ambas variables. En base a la misma, se deduce que la Tasa de Desocupación adelanta tres trimestres a la Disposición Final de RSU. Así, es posible incluir esta variable en un modelo $RegARIMA$, y comprobar si este nuevo modelo mejora la capacidad predictiva del modelo $ARIMA$ univariado aplicado a los datos trimestrales.

Gráfico 4. Correlación cruzada entre las series de residuos de cantidad de RSU enviados a disposición final y Tasa de Desocupación



Nota: El valor 1 implica un rezago de un año y está subdividido en 4 por los trimestres, por lo tanto el tercer valor corresponde al tercer trimestre.

El modelo $RegARIMA$ es:

$y_t = \beta_1 x_{t-3} + e_t$, donde e_t: $ARIMA\ (0,1,1)(1,0,0)_4$ y

x_t e y_t son las series trimestrales de Desocupación y de RSU.

Modelo estimado:

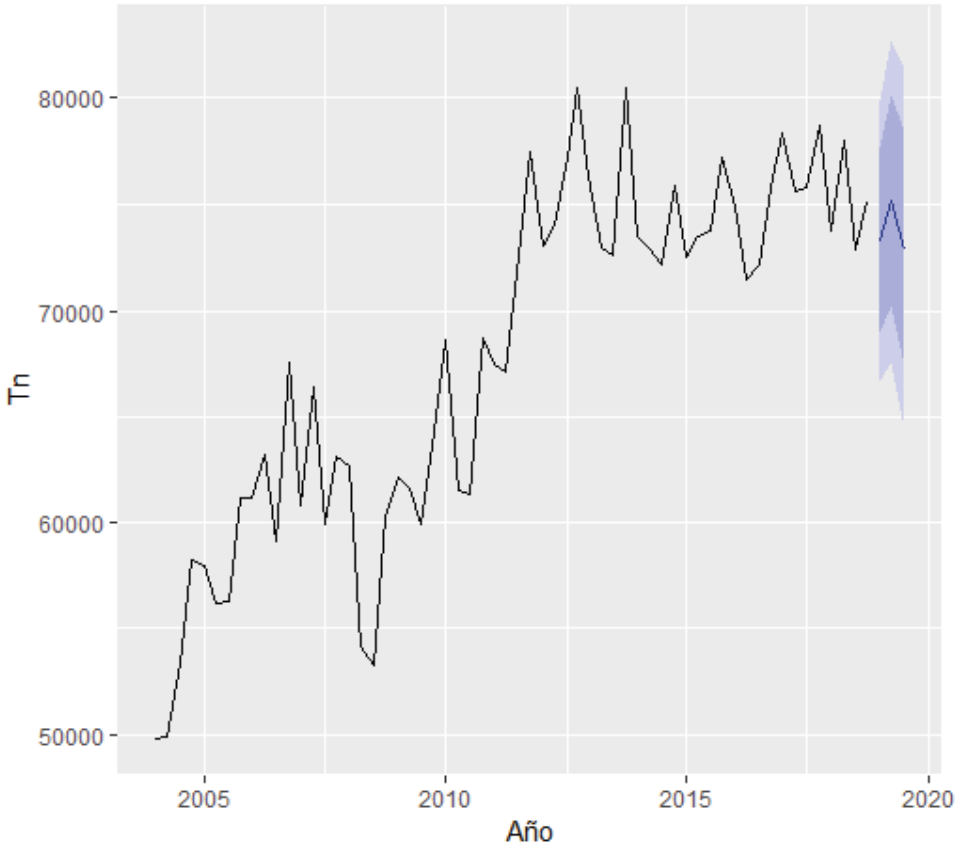
$\hat{y}_t = 112.2054\, x_{t-3} + \hat{e}_t$ donde $\hat{e}_t$: $(1 - B)\left(1 - 0.3962\, B^4\right)\hat{e}_t = (1 - (-0.4695\, B))\hat{a}_t$

La verificación de los pronósticos sólo es posible para dos períodos, ya que se trabaja con series trimestrales y se cuenta con 6 meses.
Se calculan los errores porcentuales y el error absoluto porcentual medio post muestral ($PSMAPE$) de los modelos $ARIMA\ (0,1,1)(1,0,0)_4$ y $RegARIMA$ para los primeros dos trimestres del año 2019. El intervalo de predicción contiene los valores originales de la serie (Tabla 2 y Tabla 3). El Gráfico 5 muestra la serie observada hasta diciembre 2018 y los pronósticos para los dos primeros trimestres del año 2019.

Tabla 2. Comparación de valores reales versus pronósticos obtenidos con el modelo $ARIMA\ (0,1,1)(1,0,0)_4$ para la serie cantidad de RSU enviados a disposición final

Trimestre 2019	Valores originales	Pronósticos	Error Porcentual	PSMAPE	LI (95%)	LS (95%)
Primer	76.335	73.298,41	3,98	3,98	66.698,97	79.897,95
Segundo	74.210	75.135,54	1,25	2,61	67.565,49	82.705,59

Gráfico 5.Pronósticos obtenidos con el modelo $ARIMA(0,1,1)(1,0,0)_4$ para la serie cantidad de RSU enviados a disposición final

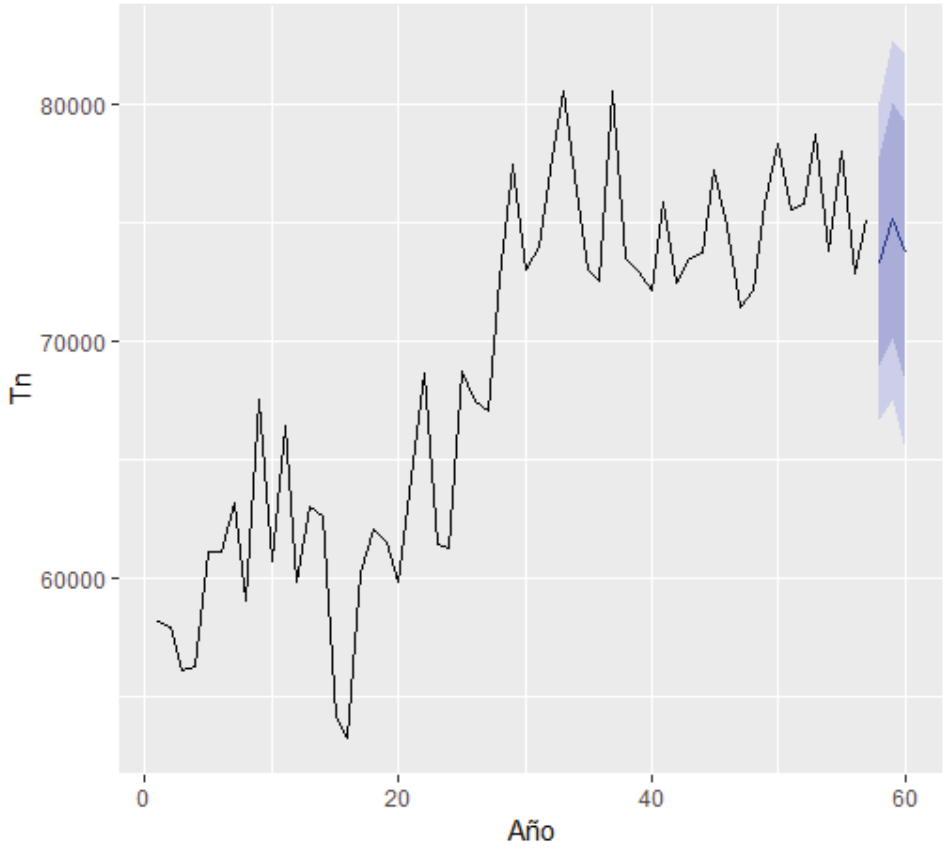


Para el modelo $RegARIMA$ el valor de la estadística $PSMAPE_2$ es de 2,56%, o sea que los pronósticos para los dos primeros trimestres del año 2019 difieren en promedio un 2,56% de los valores originales (Tabla 3). Se puede apreciar una mejora leve en el $PSMAPE_2$, posiblemente debido a la incorporación de la Tasa de Desocupación en el modelo. El Gráfico 6 muestra la serie observada hasta diciembre 2018 y los pronósticos para los primeros dos trimestres del año 2019.

Tabla 3. Comparación de valores reales versus pronósticos obtenidos con el modelo $RegARIMA$

Trimestre 2019	Valores originales	Pronósticos	Error Porcentual	PSMAPE	LI (95%)	LS (95%)
Primer	76.335	73.382,44	3,87	3,87	66.723,08	80.041,80
Segundo	74.210	75.143,87	1,26	2,56	67.605,58	82.682,15

Gráfico 6. Pronósticos obtenidos con el modelo $RegARIMA$



CONSIDERACIONES FINALES

El desarrollo de la Práctica Profesional consistió en la aplicación de conocimientos adquiridos durante el cursado de la carrera Licenciatura en Estadística y de nuevos conocimientos aprendidos durante la misma a un área municipal. A partir de los datos disponibles fue posible ajustar modelos que pueden resultar adecuados para generar pronósticos precisos que permitan prever la cantidad de residuos en el corto plazo y planificar estrategias apropiadas. A su vez, el análisis realizado puede ser utilizado por otros municipios interesados en estudiar la temática de los residuos sólidos urbanos.

Si bien el objetivo inicial era diseñar indicadores, debido a que los únicos datos que tenían continuidad en su medición eran los referidos a Disposición Final de Residuos Sólidos Urbanos, se centró el estudio en aplicar modelos a dicha serie de datos. Primeramente se aplicó un modelo $ARIMA$ univariado a esa serie. Como complemento, fue de interés estudiar la existencia de variables provenientes de otras fuentes que podrían ayudar en la descripción o a mejorar los pronósticos. De un conjunto de variables se decidió trabajar con la Tasa de Desocupación como variable auxiliar para plantear un modelo $RegARIMA$. Se observó que el agregado de la serie Desocupación al modelo univariado, no aporta significativamente desde el punto de vista del ajuste, pero mejora levemente los pronósticos. El modelo $RegARIMA$ muestra menor error, lo cual plantea la posibilidad de buscar otras variables cuya inclusión podría mejorar notoriamente los pronósticos.

Finalmente se puede mencionar que esta Práctica Profesional permitió evidenciar, además de los resultados obtenidos, la falta de sistematización de información y recolección de datos en la Dirección de Residuos, vinculados a otras fracciones de residuos. Se propone a la Dirección comenzar a sistematizar esta información para poder elaborar indicadores y realizar proyecciones futuras referidas a las distintas fracciones que componen los residuos sólidos urbanos.

BIBLIOGRAFÍA

Blaconá, María. T., García, María del. C., & Pellegrini, José. L. (1994). *La participación laboral de las cónyuges en el corto plazo: una explicación utilizando modelos REG-ARIMA*. Anales, XXIX Reunión Anual de la Asociación Argentina de Economía Política, Universidad Nacional de la Plata, Tomo 2, 301-318.

Box, G. y Jenkins, G. (1984). *Time Series Analysis: Forecasting and Control*. Third Edition. Prentice-Hall.

Dirección General de Estadística – Municipalidad de Rosario (2019). *Proyecciones de población*.

Dirección General de Gestión Integral de Residuos - Secretaría de Ambiente y Espacio Público - Municipalidad de Rosario (2019). *Disposición final de residuos.*

Hyndman, RJ. yAthanasopoulos, G. (2018). *Forecasting: Principles and Practice.*2da edición. Otexts: Melbourne, Australia. Otexts.com/fpp2.

INDEC (2014-2018). Mercado de trabajo. Tasas e indicadores socioeconómicos (EPH). *Tasa general de desocupación*. Aglomerado Gran Rosario.

Ley N°13055. (2008). *Basura Cero.* Provincia de Santa Fe.

Ordenanza N°8335. (2008). *Basura Cero*. Municipalidad de Rosario.

Peña, D. (2005). *Análisis de series temporales*. Alianza Editorial.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Secretaría de Ambiente y Espacio Público - Municipalidad de Rosario (2016). *Plan ambiental Rosario*. Sudamérica Impresos. Rosario. Disponible en: https://www.rosario.gob.ar/ArchivosWeb/libro%20PAR_10%2002.pdf

Uriel, E. (1985). *Análisis de series temporales, modelos ARIMA*. Editorial Paraninfo.

Wei, W. (1994).*Time Series Analysis: Univariate and Multivariate Methods*. Addison-Wesley.

APLICACIÓN DEL MODELO DE HADWIGER A LA FECUNDIDAD ARGENTINA

Lic. Mariana Díaz

Directora: **Dra. Lucía Andreozzi**

En la sociedad actual, quizás como nunca antes, los temas demográficos son motivo de debates, análisis e investigaciones. Dentro de la dinámica demográfica, las tendencias de la fecundidad y la mortalidad son componentes decisivos de una sociedad y los procesos de cambio en sus componentes, afectan de diferente manera a las poblaciones.
En el presente trabajo se hace hincapié en uno de los componentes principales del análisis demográfico, la fecundidad, a través de la aplicación del modelo simple de Hadwiger a las tasas específicas de fecundidad por edad simple de las madres (15 a 54 años), para el caso de Argentina, entre los años 1980 y 2011. Se analiza el ajuste del modelo de Hadwiger a los datos disponibles y se evalúan posibles interpretaciones demográficas a partir de las estimaciones de sus parámetros.

INTRODUCCIÓN

En décadas pasadas, la población vivió un alto crecimiento, que llevó a considerar que se experimentaba una explosión demográfica, cuyas consecuencias se conjeturaban atentatorias contra la sustentabilidad del desarrollo socioeconómico (Chackiel, 2004).
No mucho tiempo después, a fines del siglo XX e inicios del siglo actual, la población comenzó a experimentar una baja tasa de crecimiento y el emergente envejecimiento de la población, pensándolo tal vez, producto de la gran difusión de medios para controlar los nacimientos.
De acuerdo con Peristera & Kostaki (2007) y haciendo hincapié en uno de los componentes principales del análisis demográfico, el patrón de fecundidad específico por edad tiene una forma típica conocida en todas las poblaciones humanas, a través de los años. Para describir esta forma, los demógrafos e investigadores han utilizado una variedad de modelos.
Para algunas poblaciones se muestran distorsiones a su forma típica, en términos de un aumento en las tasas de fecundidad de las mujeres más jóvenes. De esta manera, se espera que los modelos existentes no puedan describir la nueva forma del patrón de fecundidad y, por lo tanto, se requiera el uso de representaciones más apropiadas (Peristeka & Kostaki, 2007).
En este contexto, los modelos de fecundidad encuentran uso en una amplia variedad de situaciones. En vista de ello, el interés se centra en el análisis de la fecundidad para Argentina en un período de tiempo determinado, a través de sus tasas específicas de fecundidad por edad simple. De esta manera, se propone un nuevo modelo flexible, el modelo de Hadwiger, para describir tanto los patrones antiguos de fecundidad como los nuevos.

OBJETIVO

Presentar y ajustar el modelo de Hadwiger para analizar las tasas específicas de fecundidad de la República Argentina, para los años 1980 a 2011 (con excepción del año 1990), correspondientes a edades simples de la madre que comprenden el rango de 15 a 54 años.

MARCO TEÓRICO

Conceptos relevantes

La utilización de tasas, en lugar de números absolutos, resulta muy importante para poder comparar una situación entre poblaciones en diferentes momentos. En demografía, según Preston *et al.* (2001), las tasas se definen comúnmente como tasas de ocurrencia/exposición. El numerador de este tipo de tasas contabiliza el número de ocurrencias de un evento de interés, mientras que el denominador combina dos factores: el número de personas en la población y la longitud del tiempo que enmarca el estudio.
Se hace hincapié en la tasa de fecundidad específica por edad, que se compone del total de mujeres $P(x)$ en edad x de procrear en un período de tiempo y el número de nacidos vivos $B(x)$ en el mismo período (generalmente un año calendario). Suponemos además que el número de nacidos vivos son generados por un proceso de Poisson. Los estimadores de máxima verosimilitud para cada tasa de fecundidad y su correspondiente variancia son:

$$F_x = \frac{B(x)}{P(x)} \quad , \quad Var(F_x) = \frac{B(x)}{P^2(x)} = \frac{F_x}{P(x)} \quad .$$

El estimador F_x es la tasa de fecundidad específica por edad, generalmente multiplicado por mil y su interpretación es la cantidad de nacimientos por cada mil mujeres de la población

pertenecientes a una misma edad. Su varianza $Var(F_x)$ es pequeña cuando la tasa en sí es pequeña, o cuando el número de mujeres en edad de procrear $P(x)$ es alto, o ambos.
La tasa de fecundidad específica por edad o grupo etario se define, como la razón entre el número anual de nacimientos en mujeres de determinada edad o grupo etario y la población de mujeres de la misma edad o grupo etario, en el mismo año, para un determinado país, territorio o área geográfica:

$$ {}_nF_x = \frac{{}_nB(x)}{{}_nP(x)} \times 1000 \ . $$

La notación incluye los subíndices n y x, donde n indica la cantidad de años a la que se refiere cada intervalo de edad y x señala el inicio del intervalo. Estas tasas indican el número medio de nacimientos por cada mil mujeres, en los diferentes grupos quinquenales de edad $(n = 5)$, comprendidos entre los 15 y 50 años (INDEC, 2015).
Se estima la Tasa Global de Fecundidad (TGF) definida como el número de hijos que en promedio tendría una mujer de una cohorte hipotética de mujeres que durante su vida fértil tuvieran sus hijos de acuerdo a las tasas de fecundidad por edad del período en estudio y no estuvieran expuestas al riesgo de mortalidad desde el nacimiento hasta el término de su período fértil (INDEC, 1999).

$$ TGF = \sum_{X=15}^{50} {}_nf_x \ . $$

Chandola *et al.* (1999) destacan que la fecundidad se puede comparar usando una amplia variedad de medidas existentes, sin embargo, pocas comparaciones se basan en la distribución detallada de la curva de fecundidad específica por edad (por edad individual,$n = 1$) y es aquí donde se va a centrar el interés de este trabajo.
El conjunto de las 40 tasas de fecundidad específicas de Argentina para cada año de la madre (15 a 54) se puede resumir por medio de una curva, que es una función de la edad. Por patrón de fecundidad, se entiende entonces a la distribución por edades de la fecundidad de las mujeres. Se presenta entonces la función de Hadwiger[1] como una alternativa interesante para modelar curvas de fecundidad específicas por año único de edad (15 a 54 años), de la población argentina en los años 1980 a 2011.
Se realiza un análisis de regresión para estimar los parámetros desconocidos del modelo de regresión. Se utiliza un modelo de regresión no lineal, siendo un caso particular de los modelos de regresión que son el resultado del proceso conceptualmente lógico de usar una ecuación para expresar la relación entre una variable de interés y un conjunto de variables predictoras/regresoras relacionadas.
Al igual que en los modelos de regresión lineal, la estimación de los parámetros en un modelo de regresión no lineal se realiza por el método de mínimos cuadrados o el método de máxima verosimilitud. Los cuales producen las mismas estimaciones de parámetros, cuando los errores en el modelo de regresión no lineal son normales, independientes y con varianza constante (Kutner *et al.*, 2005).

[1] Modelo de fecundidad, propuesto por Hadwiger (1940) y refinado por Gilje (1969) y Hoem *et al.,* (1981).

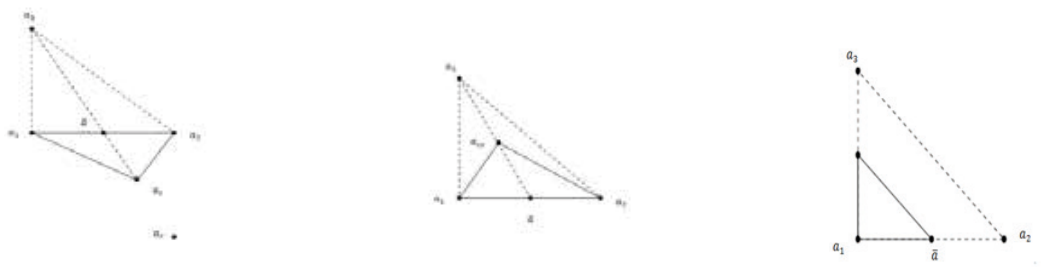
El análisis de estos modelos no lineales, usualmente, se lleva a cabo a través de programas de *software* estándar, ya que los procedimientos de búsqueda numérica que se deben usar con estos dos métodos de estimación requieren cálculos intensivos (Kutner *et al.*, 2005).
Hoem y Berge consideraron adecuada la implementación de O'Neill del algoritmo Simplex de Nelder y Mead para la minimización de funciones (Nelder & Mead, 1965; O'Neill, 1971 citado en Hoem & Berge, 1975) dado que proporciona un mejor ajuste a los propósitos de estimación en comparación con otros algoritmos.
El método de Nelder y Mead o Simplex es un algoritmo de optimización de búsqueda directa para la minimización sin restricciones de funciones multidimensionales. El objetivo del algoritmo es obtener un valor mínimo local, a partir de una estimación inicial de los parámetros. Dicho algoritmo busca este mínimo sin construir toda la superficie y utiliza el valor inicial y la información alrededor de este valor para moverse a un punto más cercano al valor mínimo y más bajo de la superficie.
En el análisis de este algoritmo se utiliza el concepto de símplex m-dimensional como una figura geométrica en dimensión m de volumen no nulo, que es la envolvente convexa de $m+1$ puntos.

El funcionamiento de este algoritmo se basa en la construcción de una sucesión de simplices para aproximarse al punto óptimo. Se muestran a continuación las figuras de las transformaciones de los simplices como ejemplo para el caso bidimensional (Figura 1 y 2).

Figura 1. Transformación de los simplices, (a) tras una reflexión y (b) tras una expansión.

(a) (b)

Figura 2. Transformación de los simplices, (c.1) tras una contracción hacia afuera, (c.2) tras una contracción hacia adentro y (d) después de una reducción.



(c.1) (c.2) (d)

El modelo de Hadwiger

Se presenta la función de Hadwiger para explorar e intentar modelar las curvas de fecundidad específicas por edad individual, como una alternativa interesante en el abanico de posibilidades existentes. El modelo simple propuesto por Hadwiger (1940) en Chandola *et al.* (1999), que fue refinado por Gilje (1969) y Hoem *et al.* (1981), se basa en la siguiente función:

$$f(x) = \frac{ab}{c}\left(\frac{c}{x}\right)^{3/2} exp\left[-\ b^2\left(\frac{c}{x} + \frac{x}{c} - 2\right)\right] \quad (1)$$

Una aclaración importante es que frecuentemente en la literatura demográfica se plantean los modelos a partir de la función que subyace en los datos y luego se menciona la existencia de un error que no se hace explícito junto a la función, mediante, por ejemplo, una componente ξ, como se emplea habitualmente en estadística. La forma correcta luego de plantear el modelo es:

$$Y_i\ = f\left(x_i,\ a,\ b,\ c\right) + \xi_i\ para\ i\ =\ 1,\ 2,\ \ ...,\ k. \quad (2)$$

Es un modelo de tres parámetros, que se denotan $a,\ b$ *y* $c,$ siendo x la edad de la madre.

En la descripción de los parámetros de la función, el parámetro "a" se relaciona con la TGF, en otras palabras, se puede decir que se asocia con el nivel de fecundidad; "b" se relaciona con el punto máximo de la distribución de esta función y "c" está asociado con la edad media de la maternidad.

Estos modelos de fecundidad encuentran su uso en una amplia variedad de situaciones, como, por ejemplo, para suavizar los datos observados y también como insumos en proyecciones de población, u otros ejercicios analíticos (UNFPA s/f).

Se pueden encontrar curvas de fecundidad que presenten excesos de fecundidad en edades tempranas, representando una población no homogénea. Si eso es así, una combinación de un par de curvas, en lugar de sólo una, puede ser más apropiada como modelo. En consecuencia, se introduce un término “mixto” en la función de Hadwiger para separar dos distribuciones. La función mixta o de mezcla se expresa como:

$$f(x) = m\left(\frac{b_1}{c_1}\right)\left(\frac{c_1}{x}\right)^{3/2} exp\left[-\ b_1{}^2\left(\frac{c_1}{x} + \frac{x}{c_1} - 2\right)\right] + (1\ -\ m)\left(\frac{b_2}{c_2}\right)\left(\frac{c_2}{x}\right)^{3/2} exp\left[-\ b_2{}^2\left(\frac{c_2}{x} + \frac{x}{c_2} - 2\right)\right] \quad (3)$$

Donde “m” es el parámetro mixto que determina el tamaño relativo de las distribuciones de los dos componentes y b_1, c_1, y b_2, c_2 son los otros parámetros. Nuevamente la forma correcta de explicitar el modelo es:

$$Y_i\ = f\left(x_i, m,\ b_1, b_2, c_1, c_2\right) + \xi_i\ \ para\ \ i\ =\ 1,\ 2,\ \ ...,\ k. \quad (4)$$

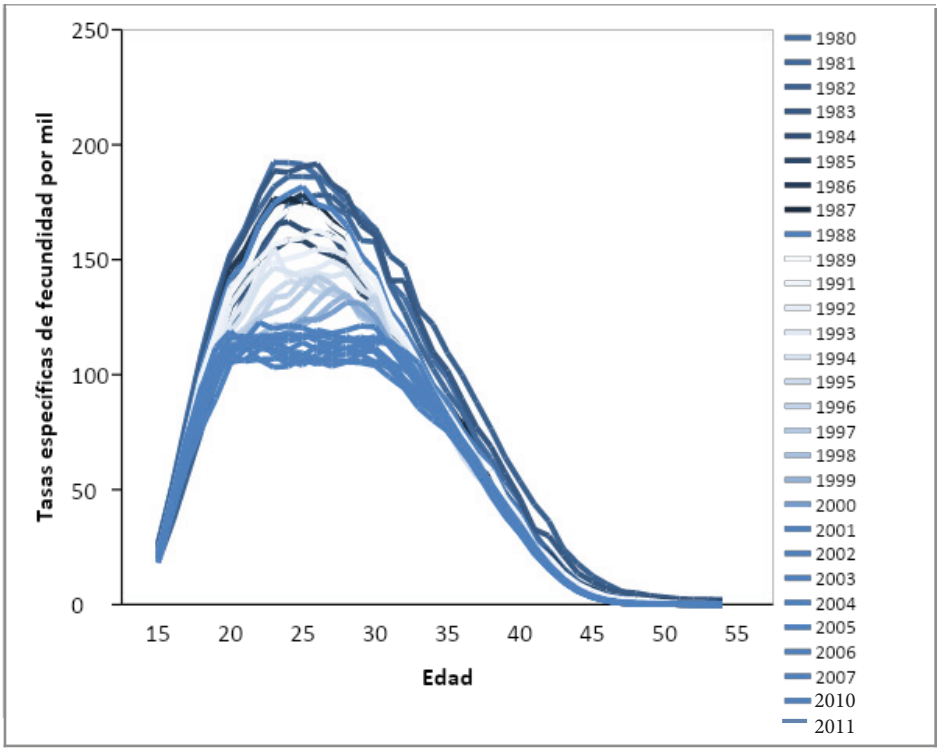
Por último, se tiene la medida del error, que es un indicador de bondad de ajuste del modelo, es decir, cuán cerca se encuentran los datos observados de los valores predichos del modelo. Es una parte inevitable del proceso y por eso se desea minimizar la raíz del cuadrado medio del error, que mide la cantidad de error que hay entre dos conjuntos de datos.

ANÁLISIS EMPÍRICO

En este trabajo, el modelo simple propuesto de Hadwiger (Gilje 1969, Hoem *et al.*, 1981 en Chandola *et al.*, 1999) se aplica a tasas específicas de fecundidad por edades simples de la madre, elaboradas a partir del total de nacidos vivos por edad de las madres (15 a 54 años) durante un año calendario y la población de mujeres para las mismas edades, al 30 de junio del mismo año. Se utiliza una base de datos proveniente de un estudio previo[2], que fue proporcionada por la Dirección de Estadística e Información de Salud (DEIS), y los datos de la población que fueron suministrados por el Centro Latinoamericano y Caribeño de Demografía (CELADE). El período analizado incluye desde el año 1980 hasta el año 2011, a excepción del año 1990 para el cual no se cuenta con los registros de nacidos vivos.

Con la información disponible, se construyeron las tasas específicas de fecundidad que se muestran en la Figura 3, en ellas es posible observar una asimetría a la derecha y una clara disminución a través del tiempo. Para los cuatro primeros años analizados (1980-1984), correspondiendo a las curvas de color azul oscuro en la escala elegida, se presentan los picos más altos entre las edades de 22 y 28 años. La forma característica de campana se hace visible hasta el año 2000 aproximadamente, y a partir de allí se nota un achatamiento de las curvas a comienzos del siglo XXI, asemejándolas a una forma de mesetas.

Figura 3. Tasas específicas de fecundidad (por mil) por edad simple en Argentina, 1980-2011



A partir de la construcción de las tasas específicas de fecundidad, se puede obtener la TGF que presenta un decrecimiento en el transcurso de los años, encontrando el mayor descenso entre 1983 y 1985 y manteniéndose estable en el último quinquenio con un valor promedio de 2, 3 hijos.

[2] "Proyecciones probabilísticas en Demografía". Maestría en Estadística Aplicada. Facultad de Ciencias Económicas y Estadística. U.N.R, realizado por Andreozzi, Lucía. (2016).

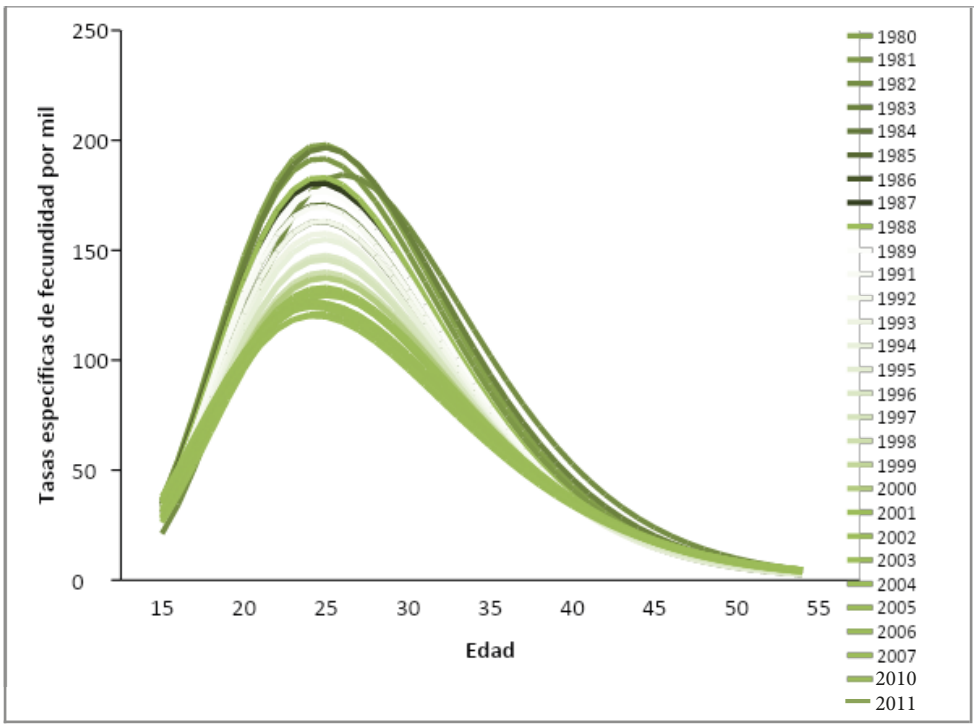
Para procesar la información antes descripta se aborda la técnica de Hadwiger utilizando el lenguaje de programación R[3], que tiene un amplio enfoque al análisis estadístico, haciendo uso de varios paquetes específicos de fecundidad con funciones necesarias para encarar el trabajo de la programación.
De esta manera, se obtienen los parámetros estimados del modelo y la raíz del cuadrado medio del error ($\sqrt{CME}$).

RESULTADOS

Ajuste del modelo

Para las curvas individuales correspondientes a las tasas específicas de fecundidad del modelo de Hadwiger, se observa (Figura 4) un comportamiento similar en las tasas específicas de fecundidad observadas. Es notorio el suavizado a partir del modelo Hadwiger, para los últimos años calendario. Bajo este modelo no se ajusta la forma de meseta que identificaba a las curvas de los últimos casi seis años, sino que estas nuevas curvas presentan un alisado mucho más moderado para el mismo rango de edad. A partir de los 30 años de la mujer, las tasas específicas de fecundidad comienzan a disminuir acercándose gradualmente a cero a medida que las mujeres se acercan a la menopausia.
Se observa una particularidad en el año 1982, ya que la curva presenta un desplazamiento hacia la derecha, o sea hacia edades mayores. El detalle de las tasas específicas de fecundidad observadas da cuenta de tasas más grandes a partir de los 30 años de edad de la madre. Este comportamiento en comparación con los demás años analizados puede deberse a diversos factores, cuyo análisis excede a este trabajo.

Figura 4. Tasas específicas de fecundidad (por mil) por edad simple en Argentina según el modelo de Hadwiger, 1980-2011



[3]R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Disponible en https://www.R-project.org/

Luego, la tasa global de fecundidad estimada por el modelo, no presenta grandes diferencias en lo que respecta a su trayectoria, con aquellas tasas observadas, aunque al final de la serie, la separación de las curvas es mayor en comparación con los primeros años.

Interpretación de los parámetros estimados

A partir de los parámetros estimados, es posible explorar algunas interpretaciones de acuerdo a la relación que evidencian en su forma con ciertos términos demográficos tradicionales o familiares (Chandola *et al.*, 1999).

El parámetro estimado "a" se comporta de manera similar a la TGF de la base de datos original. Se vislumbra la misma trayectoria a través de los años, con un descenso muy marcado en el año 1984 y manteniéndose estable al final de la serie.

En consonancia con la comparación propuesta por Chandola *et al.*, (1999), el parámetro estimado "b", contribuye a determinar la altura de la curva en la fórmula: $\frac{a*b}{c}$ explícita en el modelo simple de Hadwiger, que está asociada con la tasa específica de fecundidad máxima por edad. De acuerdo con sus valores, es posible deducir que hay una asociación en el comportamiento que presentan los parámetros estimados de manera conjunta en la fórmula "a*b/c" con la trayectoria de las tasas específicas de fecundidad máxima.

Finalmente, el comportamiento de los valores del parámetro estimado "c", a través de los años, teniendo en cuenta la relación plasmada por algunos autores (Chandola *et al.*, 1999) con la edad media de la maternidad para los datos de Argentina, se vislumbra una correspondencia en la forma de la trayectoria de ambas medidas en lo que respecta al principio de la serie analizada hasta el año 1990 y luego desde los años 2002 a 2008. No así en los años medios, donde el parámetro estimado "c" presenta una tendencia en aumento y la edad media de la maternidad muestra un pequeño descenso. Esto puede deberse a otros factores o aspectos de la fecundidad que pueden influir en este parámetro y no están siendo considerados.

Al examinar en mayor profundidad las tasas estimadas para el año 1982, se registran valores que son muy diferentes, las tasas específicas de fecundidad son bajas de 15 a 26 años y son altas de 30 a 46 años aproximadamente. También se detecta que las diferencias entre las tasas específicas de fecundidad máximas para el año 1982, con sus años adyacentes, presenta valores más altos en comparación con las diferencias en los demás años.

Análisis de residuos

Los valores de la raíz del CME no superan un valor de 0,012 en ningún año, para tasas específicas de fecundidad que tienen un rango de valores entre 0,0025 y 0,0198 nacimientos. Se destaca un aumento a partir del año 1991.

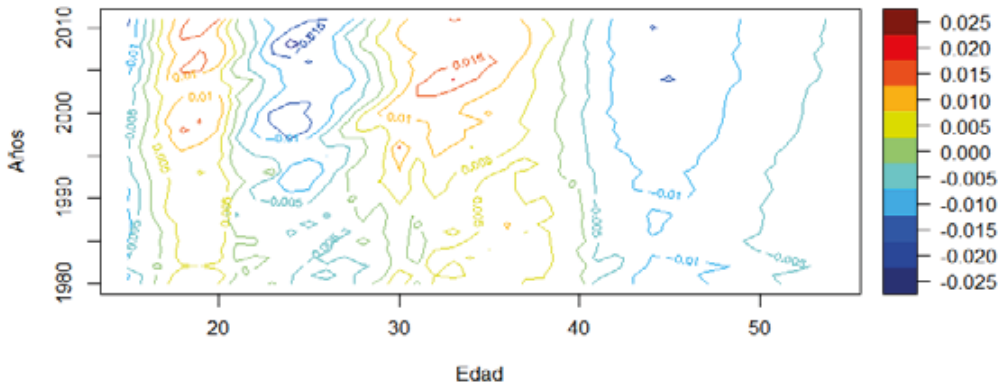
Finalmente, se construye un gráfico de contorno para los residuos del modelo (siendo las diferencias entre los valores de los datos observados y el modelo ajustado) y de esta manera verificar su independencia y comprobar si existe algún patrón sin explicar.

La Figura 5 presenta en términos generales colores alternados para los residuos, aunque se observan franjas que indican residuos negativos (gama de azules) y por lo tanto una sobreestimación de estos. También se pueden ver residuos positivos (gama de amarillos) indicando subestimación, lo cual se traduce en que las tasas estimadas resultaron menores a las observadas.

Los errores más grandes se presentan para los últimos años que se ubican en la parte superior de la figura y se indican en color rojo fuerte (residuos positivos) y en azul fuerte (residuos

negativos). En base a esto, se podría pensar que los mismos se deben a la forma de meseta que presentaban las distribuciones de las tasas de fecundidad observadas correspondiente a los últimos años y por lo tanto concluir que no sería un buen ajuste.
Por el contrario, al observar los primeros años de la serie, se encuentran colores más claros que referencian valores más pequeños. Con estas diferencias visibles en el gráfico de contorno de residuos, se puede decir que el ajuste no sería bueno para los últimos 10 años aproximadamente.

Figura 5. Residuos para el modelo de Hadwiger aplicado a datos de Argentina, 1980-2011



REFLEXIONES FINALES

A partir de la aplicación del modelo de Hadwiger a los datos de Argentina se obtuvieron distintos resultados que se sintetizan a continuación.
A través de los parámetros estimados, fue posible explorar algunas interpretaciones de acuerdo a la relación que evidenciaron en su forma con algunos términos demográficos tradicionales como, la tasa global de fecundidad, la tasa específica de fecundidad máxima por edad y la edad media de la maternidad.
Por otra parte, el análisis de los residuos presentó un buen ajuste para los primeros años de la serie analizada, pero no así para los últimos 10 años aproximadamente, donde los errores presentaron valores más grandes. Esto puede referirse al comportamiento observado de las tasas específicas de fecundidad que presentan una forma de meseta, indicando de esa manera que el modelo no presenta un buen ajuste de los datos.
Se observó que las poblaciones actuales muestran otro comportamiento y por lo tanto el modelo simple de Hadwiger no sería el indicado para el análisis. En consecuencia, se podría profundizar el análisis del modelo mixto de Hadwiger, considerando información más actualizada.
Por lo tanto, como futuras investigaciones, se propone ahondar en el uso de los modelos mixtos de Hadwiger para comparar y explicar los patrones de fecundidad en los diferentes países a lo largo del tiempo. Sin dejar de lado, lo que aún queda por describir respecto a la varianza, el sesgo, la curtosis y la simetría de las curvas de fecundidad. Resultaría muy interesante disponer de una perspectiva de género, incorporando a los hombres como población objetivo, que acompañe el análisis de la fecundidad teniendo en cuenta que actualmente es un factor poco considerado en los estudios de nuestro país.

Por último, en relación a la fecundidad, es menester la continua implementación de políticas públicas que garanticen que todos los embarazos sean deseados y todos los nacimientos sean seguros.

BIBLIOGRAFÍA

Andreozzi, L. (2016). "Proyecciones probabilísticas en Demografía". Maestría en Estadística Aplicada. Facultad de Ciencias Económicas y Estadística, Universidad Nacional de Rosario.

Andreozzi, L (2018). "Métodos probabilísticos de pronóstico de la mortalidad y su aplicación a tres Departamentos de la Provincia de Córdoba". Tesis de Doctorado en Demografía. Facultad de Ciencias Económicas, Universidad Nacional de Córdoba.

Centro Latinoamericano y Caribeño de Demografía (CELADE) (2000). Boletín demográfico n66. América Latina. "Población por año calendario y edad simple". 1995-2005 (la base de datos fue suministrada para todo el período 1950-2050).

Chackiel, J. (2004). "La dinámica demográfica en América Latina". Centro Latinoamericano y Caribeño de Demografía (CELADE) – División de Población. Naciones Unidas. Santiago de Chile.

Chandola, T., Coleman, D., & Hiorns, R. (1999). "Recent European Fertility Patterns: Fitting Curves to 'Distorted' Distributions". Population Studies, 53(3), 317-329. Disponible en: http://www.jstor.org/stable/2584702

Dirección de Estadísticas e Información en Salud (DEIS). "Estadísticas vitales" – información básica años 1980-2011. Ministerio de Salud y Desarrollo Social de la Nación.

Fondo de Población de las Naciones Unidas (UNFPA). "Modelos demográficos: fecundidad". Consultado el 29/5/2019. Disponible en: http://papp.iussp.org/sessions/papp103_s03/PAPP103_s03_010_010.html

Gómez, S., Patiño, D. & Vélez, C. (2012). "Variante del método de Nelder & Mead para optimización de funciones multivariadas". Cuaderno ACTIVA. No. 4, Julio-diciembre 2012, pp. 73-62. Tecnológico de Antioquia, Medellín (Colombia).

Hoem, J. & Berge E. (1975). "Some problems in Hadwiger fertility graduation". Scandinavian Actuarial Journal. 3, 129-144. Disponible en: 10.1080 / 03461238.1975.10405091

Hoem, J. M., Madien, D., Nielsen, J. L. et al. (1981). "Experiments in modeling recent Danish fertility curves". Demography, 18: 231. Disponible en: https://doi.org/10.2307/2061095

Instituto Nacional de Estadística y Censos (INDEC) (1999). "Situación y evolución social (Síntesis N°4)". Buenos Aires, Argentina.

Instituto Nacional de Estadística y Censos (INDEC) (2015). "Población e inclusión social en la Argentina del Bicentenario: Indicadores demográficos y sociales" – 1a ed. - Ciudad Autónoma de Buenos Aires.

Kutner, M. H., Nachtsheim, C., Neter, J. & Li, W. (2005). "Applied linear statistical models". McGraw-Hill/Irwin series operations and decision sciences.

Martínez Varela, A. (s/f). "Algoritmos de optimización para funciones con ruido". Departamento de Matemática aplicada II. Universidad de Vigo. Disponible en: http://www.dma.uvigo.es/~aurea/transparencias3.pdf

Montgomery, D., Peck, E. & Vining, G. (2006). "Introducción al Análisis de Regresión Lineal". Compañía editorial Continental. Tercera reimpresión, México.

Peristera, P. & Kostaki, A. (2007). "Modeling fertility in modern populations". Demographic Research, 16, 141-194. Disponible en: http://www.jstor.org/stable/26347932

Preston, S., Heuveline, P. & Guillot, M. (2001). "Demography, Measuring and Modelling Population Processes". Blackwell, Oxford.

USO DE MODELOS LINEALES GENERALIZADOS PARA DATOS BINOMIALES SOBREDISPERSOS EN EL ESTUDIO DEL FRAUDE EN SINIESTROS DEL RAMO AUTOMOTOR EN UNA COMPAÑÍA ASEGURADORA

Lic. Nadina Matteucci

Directora: **Dra. Gabriela Boggio**

Uno de los principales desafíos de la industria aseguradora es combatir el fraude. En el presente trabajo se analizan los datos de una compañía que ha desarrollado dentro de su organización un área específica en pos de combatir este delito.
El objetivo es estudiar la probabilidad de que un siniestro que ingresa al área para su análisis, sea finalmente detectado como fraude. Para esto, se recurre al análisis estadístico mediante el ajuste de un modelo lineal generalizado, considerando la respuesta binomial y el enlace logit. A fines de corregir la aparente sobredispersión de los datos, se abordó el estudio del modelo binomial corregido y del modelo beta-binomial.

INTRODUCCIÓN

El fraude en la industria del seguro es un delito que para las grandes empresas ocasiona anualmente pérdidas millonarias. Consiste en cualquier acción que tenga como finalidad engañar a la compañía aseguradora para obtener un beneficio indebido.

La compañía aseguradora se obliga a responder por el suceso que haya afectado al asegurado. Sin embargo, si logra evidenciar que el siniestro es un fraude, podrá rechazar el reclamo, discontinuar la póliza o realizar acciones legales según lo amerite el caso. En pos de combatir el fraude, una compañía aseguradora procedió a la construcción de un área específica dentro de la organización, formada por un equipo multidisciplinario de investigadores internos (pertenecientes a la compañía) encargados de evaluar cada siniestro derivado al sector con motivo de ser un caso sospechoso de fraude. Pese a la capacidad operativa de cada uno de ellos, en algunos casos, debido a la complejidad de los mismos, a la dificultad de acceder a determinada documentación o a la necesidad de realizar pericias en el lugar del siniestro, es necesario derivar la investigación a estudios de investigadores externos (contratados por la compañía).

El objetivo es analizar el desempeño del área de fraude de la compañía aseguradora estudiando la probabilidad de que un siniestro que ingresa al sector finalmente sea detectado como fraude, en función de quién fue el investigador a cargo del caso y del requerimiento de que éste sea derivado, o no, a un estudio de investigación externa.

El análisis estadístico para abordar este objetivo comienza mediante el ajuste de un modelo lineal generalizado, considerando la respuesta binomial y el enlace *logit*. Surge como disparador el estudio del modelo binomial corregido y del modelo beta-binomial a fines de considerar la sobredispersión intrínseca de los datos, incapaz de ser capturada mediante un modelo binomial clásico.

MATERIAL Y MÉTODOS

Los siniestros objeto de investigación pertenecen al rubro automotor y son aquellos que, por sus características, ameritan ser investigados. Es así que, de los 1769 siniestros que ingresaron al sector en el período febrero-agosto de 2018, 342 conforman el grupo bajo estudio. Cabe aclarar que de los siniestros que ingresan al área, no todos merecen la pena ser investigados, ya sea porque no tienen indicadores fuertes de indicios de fraude, o bien, porque no es económicamente redituable realizar la investigación. Por este motivo en el área existe una instancia de examinación (a cargo de examinadores capacitados para la tarea) previa a la instancia de investigación, en la cual se determina si el caso en cuestión amerita ser investigado.

Dado que se desea estudiar la probabilidad de que un siniestro sea detectado como fraude en función de quién fue el investigador interno que analizó el caso y si fue necesario que el siniestro sea derivado a investigación externa, se definen las variables involucradas en el análisis. Como dato adicional, se cuenta con información de quién fue el examinador que dio curso a la investigación del siniestro, aunque se presupone, según la experiencia del negocio, que esta variable no influye en la decisión de determinar si un siniestro es fraude o no.

La variable respuesta es *Estado final del siniestro* (con categorías: fraude, no fraude) y las variables explicativas son *Investigador interno* (A, B, C, D, E), *Derivación a investigación externa* (no, si), *Examinador,* la persona que da curso a la investigación del siniestro, (a, b).

La evaluación de la importancia de estas variables explicativas sobre la respuesta de interés se realiza mediante el ajuste de apropiados modelos estadísticos, tal como se describe a continuación.

Modelos lineales generalizados. Caso respuesta binaria

Un modelo lineal generalizado se define en término de tres componentes: variables respuestas $Y_1, \ldots Y_N$ que se asume tienen la misma función de distribución perteneciente a la familia exponencial, un conjunto de parámetros y variables explicativas y una función monótona y diferenciable g tal que $g(\mu_i) = x_i^T \beta$ donde $\mu_i = E(Y_i)$ para $i = 1, \ldots, N$. En este caso $Y_1, \ldots Y_N$ corresponden al número de éxitos en *N* diferentes grupos, con $Y_{i\sim} binomial(n_i, p_i)$, $i = 1, \ldots N$. Si se quiere describir la proporción de éxitos $\pi_i = Y_i/n_i$ en cada grupo, como $E(Y_i) = n_i p_i$, y por lo tanto $E(\pi_i) = p_i$, se pueden modelar las probabilidades p_i a través de un modelo lineal generalizado $g(p_i) = x_i^T \beta$**,** pudiendo elegir entre diferentes funciones de enlace g. La más habitual es la que conduce al modelo *logit:*

$$logit\, p_i = log\left(\frac{p_i}{1-p_i}\right) = x_i^T \beta.$$

La elección de este enlace, permite realizar la interpretación de los coeficientes del modelo en términos de razones de *odds*.

Los estimadores máximo verosímiles de los parámetros β, y consecuentemente de las probabilidades $p_i = g^{-1}(x_i^T \beta)$ se pueden obtener maximizando la función de log-verosimilitud mediante métodos iterativos y la bondad de ajuste del modelo puede ser evaluada a través de las siguientes estadísticas: *Deviance* o Chi-Cuadrado.

Estudio de la sobredispersión

Si un modelo *logit* ajustado a *N* proporciones binomiales es adecuado, la *Deviance* tendrá distribución aproximada Chi-cuadrado con *(N-q)* grados de libertad (gl), siendo *q* el número de parámetros estimados. Debido a que el valor esperado de una Chi-cuadrado con *(N-q)* gl es justamente *(N-q)*, es de esperar que la *Deviance* correspondiente a un modelo con buen ajuste sea aproximadamente igual a sus gl, o equivalentemente, que la *Deviance* sobre los gl sea un valor cercano a 1. Si la *Deviance* sobre los gl excede la unidad, se dice que los datos presentan sobredispersión. También se suele utilizar el cociente de la estadística Chi-cuadrado sobre sus gl para detectar la presencia de sobredispersión.

La sobredispersión puede deberse a que la componente sistemática del modelo es inadecuada ya sea por omisión de covariables, interacciones o problemas de escala en alguna variable o bien porque la función de enlace no es la adecuada. Una vez eliminados los posibles motivos causantes de provocar que los datos presenten un valor grande de *Deviance* sobre gl ó Chi-Cuadrado sobre gl, la sobredispersión que persista puede deberse a la ocurrencia de variación entre las probabilidades de respuesta o bien, a la correlación entre las respuestas binarias, no cumpliéndose los supuestos del modelo binomial. Collett (2002) describe de manera detallada estas causas:

1. Sobredispersión debida a la variación entre las probabilidades de respuesta

En esta situación, se asume que la probabilidad de respuesta para la *i-ésima* observación, π_i, varía alrededor de una media p_i. Entonces π_i es una variable aleatoria no observable con $E(\pi_i) = p_i$ y variancia de la forma $Var(\pi_i) = Q p_i(1 - p_i)$ donde $Q \geq 0$ es un parámetro desconocido.

Se puede obtener fácilmente la expresión de la esperanza y variancia de Y_i en términos de valores conocidos resultando $E(Y_i) = n_i p_i$ y $Var(Y_i) = n_i p_i(1 - p_i)\{1 + (n_i - 1)Q\}$.

2. Sobredispersión debida a la correlación entre respuestas binarias

En este caso, la consideración de la correlación entre las respuestas da lugar a las siguientes expresiones de la esperanza y variancia de la binomial: $E(Y_i) = n_i p_i$ y $Var(Y_i) = n_i p_i(1 - p_i)\{1 + (n_i - 1)\delta\}$ donde δ es el coeficiente de correlación entre pares de respuestas binarias.

Ambas causas conducen a la misma expresión para la variancia de Y_i, por lo tanto los efectos de "correlación entre respuestas binarias" y "variación entre probabilidades de respuesta" no pueden ser distinguidos teóricamente; se dice entonces que una causa conduce a otra y viceversa (Collett, 2002). Es decir, se puede concluir que si existe variación entre las probabilidades de respuesta o si hay falta de independencia entre pares de observaciones binarias, la variancia de Y_i excede la variancia bajo distribución binomial por un factor de $\{1 + (n_i - 1)Q\}$.

Una forma de corregir el problema de la sobredispersión es simplemente multiplicar la variancia de las estimaciones de los coeficientes del modelo por un parámetro $\emptyset$. Modificada así la variancia, la componente aleatoria del modelo no se corresponde con ninguna distribución de probabilidad, por lo que Wedderburn (1974) propuso un enfoque cuasi verosímil para la estimación de $\emptyset$ que viene dada por $\sqrt{X^2/(N - q)}$, siendo N la cantidad de proporciones binomiales y q el número de parámetros estimados en el modelo.

Otra forma de captar la sobredispersión de los datos es considerar la variancia de Y_i como $n_i p_i(1 - p_i)\sigma_i^2$ donde $\sigma_i^2 = 1 + (n_i - 1)Q$. El parámetro Q es desconocido y hay que estimarlo. Williams (1982) muestra que su estimación se puede encontrar igualando la estadística Chi-cuadrado de Pearson del modelo maximal, con su valor esperado. Como el valor de X^2 depende de $\hat{Q}$, se trata de un procedimiento iterativo de dos pasos: (1) se resuelven las ecuaciones cuasi-verosímiles para β dado $\hat{Q}$ y (2) se utiliza el valor obtenido de $\hat{\beta}$para obtener el valor de $\hat{Q}$ en la ecuación siguiente:

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - n_i \hat{p}_i)^2}{n_i \hat{p}_i(1-\hat{p}_i)\{1+(n_i-1)\hat{Q}\}} = N - q \ .$$

Las soluciones presentadas anteriormente provocan que los errores estándares de los valores estimados de los parámetros β se aumenten, lo que puede traer como consecuencia que efectos que antes aparecían como significativos ya no lo sean.

Por último, se plantea la consideración de la distribución beta-binomial, (Zangiacomi Martínez *et al.*, 2015; Agresti, 2015) para poder modelar la sobredispersión de los datos. Sea *Y* el número de ocurrencias de una variable aleatoria *Y* en *n* ensayos Bernoulli con probabilidad de éxito *p,* donde *p* es una variable aleatoria que sigue una distribución Beta con parámetros *a* y *b*. Entonces la función de probabilidad de una distribución beta-binomial viene dada por:

$$P(Y = y) = (n\ y)\frac{\beta(y+a, n-y+b)}{\beta(a,b)}.$$

Una parametrización muy usada de este modelo, considera $a = \theta\tau^{-1}$ y $b = (1 - \theta)\tau^{-1}$, donde $\tau > 0$ y $0 < \theta < 1$. Así, la media y variancia de Y vienen dadas por $E(Y) = n\theta$ y

$Var(Y) = n\theta(1-\theta)\left[1 + (n-1)\frac{\tau}{1+\tau}\right]$. El parámetro τ es interpretado como un parámetro de sobredispersión, Cuando $\tau = 0$, la variancia es equivalente a la variancia de una variable aleatoria que sigue una distribución binomial.
La estimaciones máximo verosímiles de θ y de τ no tienen una solución explícita, por lo que las mismas surgen de métodos iterativos.
Por último, en base a esta parametrización de la distribución beta-binomial, se puede postular un modelo de regresión tal como el modelo lineal generalizado binomial (Liang y McCullagh, 1993). Es decir: $logit\,\theta_i = log\left(\frac{\theta_i}{1-\theta_i}\right) = x_i^T\beta$.

RESULTADOS

Para el estudio del desempeño del área de fraude de la compañía aseguradora, se presenta en primera instancia una descripción de la información disponible en términos de las proporciones de fraude según el investigador interno a cargo, el examinador y el requerimiento o no de investigación externa.

Análisis descriptivo

De la Tabla 1 se puede observar que diferentes números de siniestros fueron investigados para cada combinación de investigador interno y requerimiento de investigador externo. Por otro lado, la cantidad de siniestros investigados, así como las proporciones de fraude para cada combinación posible de variables explicativas varía considerablemente, teniendo mayor precisión aquellas proporciones observadas que se refieren a las combinaciones con mayor número de siniestros investigados.

Tabla 1. Cantidad de siniestros detectados como fraude (y), cantidad total de siniestros (n) y probabilidad observada (y/n) según el examinador, el investigador interno y el requerimiento de mandar el caso a investigación externa

		Investigación Externa: NO			Investigación Externa: SI		
Investigador Interno	Examinador	y	n	prop obs	y	n	prop obs
A	a	6	63	0,10	0	7	0,00
	b	14	116	0,12	4	14	0,29
B	a	31	182	0,17	2	16	0,13
	b	26	168	0,15	9	27	0,33
C	a	19	143	0,13	3	23	0,13
	b	45	223	0,20	19	51	0,37
D	a	11	73	0,15	7	12	0,58
	b	18	130	0,14	6	15	0,40
E	a	25	102	0,25	13	28	0,46
	b	56	289	0,19	28	87	0,32

Ajuste de modelos

Luego de haber ajustado diferentes modelos *logit* se llega a la conclusión de que la probabilidad de que un siniestro, que pasó la instancia de examinación, sea finalmente detectado como fraude, podría modelarse según quién haya sido su investigador interno y el requerimiento de ser derivado o no a investigación externa, a través del siguiente modelo:
$logit(p_{jk}) = \beta_0 + \beta_1 InvInt_A + \beta_2 InvInt_B + \beta_3 InvInt_C + \beta_4 InvInt_D + \beta_5 InvExt_{NO}$

donde $InvInt_A$, $InvInt_B$, $InvInt_C$, $InvInt_D$ son las variables indicadoras asociadas a la variable *Investigador interno* (tomando como categoría de referencia la E) e $InvExt_{NO}$, la variable indicadora asociada a la variable *Derivación a investigación externa* (tomando como categoría de referencia, "sí"). Cabe aclarar que se decide tomar el nivel E de la variable *Investigador interno* como categoría de referencia dado que corresponde al investigador con más antigüedad en la empresa y más experiencia en la detección de fraude.
Para este modelo se evalúa la función de enlace y la misma resulta adecuada. La *Deviance* es igual a 23,47 por lo que, en base a una distribución Chi-cuadrado con 14 gl, no se rechazaría la hipótesis de que el modelo ajusta adecuadamente los datos, con una probabilidad asociada igual a 0,053, aunque es un valor demasiado cercano al nivel de significación del 5% establecido. Es así que el valor de *Deviance*/gl=1,68 y de X^2/gl=1,53 muestran indicios de sobredispersión. Es por ello que se decide continuar el análisis en búsqueda de alternativas que brinden un mejor ajuste.

Comparación de los resultados obtenidos bajo las distintas alternativas
La Tabla 2 resume los resultados de los 4 análisis realizados. Los modelos ajustados con la corrección mediante la inflación de la variancia y utilizando el método de Williams tienen similar errores estándares y más grandes que los correspondientes a los de los modelos binomial y beta-binomial. Sin embargo, los resultados en términos de significación de los efectos son prácticamente iguales. La única diferencia es la relacionada al efecto del investigador B con respecto al E (tomado como referencia) que resulta significativa solo para los modelos binomial y beta-binomial.

Tabla 2. Estimación (errores estándares) y significación de los diferentes parámetros para los cuatro ajustes presentados

Parámetro		Tipo de modelo ajustado							
		Binomial		Binomial con corrección por inflación de la variancia		Binomial con corrección por método de Williams		Beta-Binomial	
InvInt	A	-0,74 (0,24)	✔	-0,74 (0,30)	✔	-0,83 (0,36)	✔	-0,74 (0,24)	✔
InvInt	B	-0,32 (0,17)	✔	-0,32 (0,21)		-0,41 (0,29)		-0,32 (0,17)	✔
InvInt	C	-0,22 (0,16)		-0,22 (0,20)		-0,33 (0,27)		-0,22 (0,16)	
InvInt	D	-0,26 (0,20)		-0,26 (0,25)		-0,17 (0,30)		-0,26 (0,20)	
InvExt	NO	-0,80 (0,15)	✔	-0,80 (0,18)	✔	-0,83 (0,20)	✔	-0,81 (0,15)	✔

✔ Significativo al 10%

En cuanto a la interpretación de los modelos respecto al problema planteado, teniendo en cuenta el signo de los coeficientes y su significación, se puede concluir que la probabilidad de detectar fraude en un siniestro es especialmente mayor cuando el caso se deriva a un investigador externo, en concordancia con la lógica del negocio. En cuanto a la comparación de los investigadores respecto al de mayor experiencia y antigüedad, el investigador E, se encuentra que la mayor diferencia se da con el investigador A.

Así es que la chance de detectar un fraude es el doble cuando interviene un investigador externo en el estudio cualquiera haya sido el investigador interno y, por otro lado, también es aproximadamente el doble la chance de detectar un fraude por parte del investigador E en comparación con el A.

CONSIDERACIONES FINALES

El análisis del desempeño del área de fraude de la compañía aseguradora se basó en la indagación de los factores que pueden influir en la detección de fraude en un siniestro sobre el rubro automotor. A tal fin se estudió la probabilidad de que un siniestro sea finalmente detectado como fraude en función de quién fue el investigador a cargo del caso y del requerimiento de derivación a un estudio de investigación externa en base a modelos estadísticos. Para ello, en principio se recurrió al clásico modelo lineal generalizado para respuesta binomial, el modelo *logit*, pero la supuesta sobredispersión de los datos motiva el estudio de diferentes alternativas disponibles para su tratamiento ya sea a partir de correcciones de la variancia o bien postulando una distribución de probabilidad con variancia mayor, la distribución beta-binomial.

El análisis realizado desestima la existencia de sobredispersión en los datos por lo que los ajustes a partir de las diferentes alternativas consideradas condujeron a resultados muy similares. Sin embargo, desde el punto de vista metodológico, resultó enriquecedor ahondar en el tratamiento de datos binomiales sobredispersos.

En relación al problema planteado por la compañía aseguradora, resultó clara la influencia del efecto de requerir o no derivación a un estudio de investigación externa sobre la probabilidad de detección de fraude en un siniestro. También se detectaron diferencias según cuál haya sido el investigador interno que analizó cada caso, aunque se aprecian con menor contundencia.

BIBLIOGRAFÍA

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models,* First Edition.John Wiley & Sons, Inc.

Collett, D. (2002). *Modelling Binary Data,* Second Edition.Chapman & Hall/CRC Texts in Statistical Science.

Dobson, A.J. (2001). *An introduction to generalized linear models,* Second Edition.Chapman & Hall/CRC Texts in Statistical Science.

Liang, K., McCullagh, P. (1993).Case Studies in Binary Dispersion. *Biometrics, 49(2), 623-630.*

McCullagh, P., Nelder, J.A. (1989). *Generalized Linear Models*, Second Edition.Chapman & Hall/CRC Texts in Statistical Science.

Nelder, J.A., Wedderburn, R.W.M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, 135(3), 370-384.*

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika. 61(3), 439-447.*

Williams, D.A. (1982). Extra-Binomial Variation in Logistic Linear Models. *Journal of the Royal Statistical Society, 31(2), 144-148.*

Zangiacomi Martinez, E., Achcar, J.A., Casale Aragon, D. (2015). Parameter estimation of the beta-binomial distribution: an application using the SAS software. *Ciência e Natura, 37(4), 12-19.*

COMPARACIÓN DE MÉTODOS DE ANÁLISIS ESPACIAL DE EXPERIMENTOS DE CAMPO EN PROGRAMAS DE MEJORAMIENTO DE CULTIVOS

Lic. Lucas Peitton

Directora: **Mg. María Gabriela Borgognone**
Co-Directora: **Dra. Luciana Magnano**

La variabilidad en experimentos en el campo debe ser modelada para asegurar que la comparación entre las líneas de cultivo evaluadas sea precisa. Tradicionalmente se utiliza un modelo lineal mixto propuesto por Gilmour et al. (1997) de modelación multi-etápica basada en herramientas gráficas y tests formales. Este análisis puede realizarse habitualmente contando con una licencia de software. Rodríguez-Álvarez et al. (2016a) propusieron un modelo lineal mixto que se obtiene en un solo paso. Ambos modelos fueron comparados en ensayos de sorgo siguiendo diseños parcialmente replicados (Velazco et al. 2017), obteniendo resultados consistentes. En esta tesina, el interés es evaluar si tales resultados pueden extenderse a otras circunstancias, dado que el nuevo modelo puede ajustarse con un programa de acceso gratuito. En el marco de programas de mejoramiento australianos de cultivos de poroto mung y garbanzo, se analizan 32 ensayos realizados en el campo. Se presentan los dos modelos espaciales y se los compara entre sí con respecto a la precisión y eficiencia en la selección de los mejores genotipos. Además, se demuestra la importancia de modelar las tendencias espaciales al comparar los modelos espaciales con un modelo no-espacial.

INTRODUCCIÓN

Un programa de mejoramiento de cultivos tiene como objetivo identificar nuevas variedades que se adapten a distintas regiones de producción, que tengan mayor rendimiento, resistencia a enfermedades y tolerancia a herbicidas, como también que alcancen parámetros de alta calidad para su utilización de consumidores finales. El rendimiento de las nuevas líneas es comparado con las variedades comerciales. Con el fin de encontrar genotipos de mejores cualidades, se conducen experimentos de campo donde se evalúan diversas líneas de cultivos. En este caso, el mejoramiento se enfoca al rendimiento medio y la estabilidad de los genotipos.

En general, los programas de mejoramiento de cultivos en Australia, entre ellos, los programas de mejoramiento de porotos mung (NMIP, *National Mungbean Improvement Program*) y de garbanzo de PBA (*Pulse Breeding Australia*), siguen una estructura que consta de la selección progresiva del material genético inicial en varias etapas pre-comerciales. La etapa final de pruebas se desarrolla en el campo y consiste en series de ensayos donde las líneas se van evaluando a través de varios años. Estos ensayos son llevados a cabo en ambientes naturales, es decir, en campos donde diversas fuentes de variabilidad no pueden ser controladas por el investigador. Cada año de prueba, se seleccionan los mejores genotipos para seguir avanzando en el programa. En el primer año, se evalúa un extenso número de genotipos en un reducido número de localidades y, según el rasgo evaluado, las mejores líneas son seleccionadas para la evaluación del año siguiente. Este grupo de genotipos es evaluado en el segundo año en una cantidad más grande de localidades, para luego ejecutar nuevamente la selección de los mejores. Una vez alcanzados el tercer y cuarto año, sólo un número reducido respecto de la cantidad inicial de genotipos es evaluado en una mayor cantidad de localidades. El conjunto de genotipos finalmente retenidos, es analizado por equipos interdisciplinarios y entidades comerciales para asegurar que, a la hora de introducir las nuevas variedades al mercado, tengan características comerciales relevantes.

El análisis estadístico de los ensayos de mejoramiento resulta sumamente útil para describir la respuesta observada de los genotipos evaluados y realizar así la selección más eficiente de las mejores variedades. A la hora de analizar la información obtenida, es importante distinguir cómo se componen los términos del modelo de acuerdo a cómo se llevó a cabo el experimento y a las posibles fuentes de variabilidad presentes, asegurando así los mejores resultados posibles.

Una de las principales fuentes de variabilidad que puede afectar las decisiones de la selección de genotipos, es la variabilidad espacial en el campo. La variabilidad espacial es la dependencia de la respuesta observada a través del campo experimental. Esto puede deberse a los cambios en la fertilidad del suelo, la humedad de la tierra, el relieve del campo, posibles depresiones y pendientes presentes en el mismo, etc.; lo cual introduce variabilidad y correlación entre parcelas cercanas.

Gilmour et al. (1997) identificaron tres fuentes principales de variabilidad espacial, identificadas como tendencias global y local, y variabilidad externa. La tendencia global consiste en la variabilidad a lo largo del campo en dirección de las filas y/o columnas debida a características del suelo, como pendientes en el campo y gradientes de fertilidad. La tendencia local refiere a la correlación entre parcelas cercanas; parcelas vecinas tendrán características más similares mientras que, parcelas distantes tenderán a ser diferentes. La variabilidad externa se debe principalmente a técnicas y procedimientos operativos propios de la realización del ensayo, como siembra, cosecha, riego, entre otros. Generalmente, la tendencia global se modela con polinomios y/o funciones *spline* en dirección de las filas y/o columnas, la tendencia local se

modela con un proceso separable AR1 para los residuales en dirección de filas y columnas y, la variabilidad externa, a través de efectos de diseño.

El modelo lineal mixto de Gilmour et al. (1997) se ha convertido en el método estándar de análisis espacial de ensayos de variedades y consta en identificar los términos a incluir en el modelo utilizando herramientas gráficas y pruebas de hipótesis formales en un procedimiento multi-etápico, asegurando que la estructura de covariancia de los residuos sea la más adecuada al ensayo. Su aplicación se realiza mediante el software ASReml (VSN *International*), aunque también puede realizarse a través del paquete de R, llamado ASReml-R (Butler et al. 2009), el cual requiere de una licencia paga, que puede ser costosa para los investigadores.

Rodríguez-Álvarez et al. (2016) presentaron un nuevo método llamado Análisis Espacial de Ensayos de campo con *Splines* (en inglés, *Spatial Analysis of field Trials with Splines*, SpATS). En el marco de los modelos lineales mixtos, este método utiliza la teoría de *splines* penalizados en dos dimensiones (2D P-Splines; Eilers y Marx 1996) y la representación *P-spline Smooth-ANOVA* (PS-ANOVA; Lee et al. 2013). Las fuentes de variabilidad global y local son modeladas simultáneamente en un único paso y, penalizando las componentes a través de parámetros de suavizado, se evita el sobre-ajuste. El análisis mediante un modelo SpATS puede realizarse utilizando el paquete de R, llamado SpATS (Rodríguez-Álvarez et al. 2016b) que es de acceso gratuito.

Velazco et al. (2017) aplicaron el método SpATS y el método estándar de Gilmour et al. (1997) a grandes ensayos de líneas de sorgo utilizando diseños parcialmente replicados (p-rep; Cullis et al. 2006). En esa investigación, los autores presentaron resultados eficientes y confiables del modelo SpATS al compararlo con el modelo estándar (referido como *Best Spatial Standard model*, BSS).

En la presente investigación, se analizan ensayos de campo en el marco de los programas de mejoramiento de poroto mung y garbanzo en Australia utilizando los modelos BSS y SpATS. Estos modelos son expuestos y evaluados de manera individual y comparados entre sí. Esta comparación tiene como finalidad evaluar si la eficiencia y confiabilidad del modelo SpATS mencionadas por Velazco et al. (2017) para los ensayos de sorgo se extiende a otros contextos, tales como diferentes cultivos, diferentes tamaños de ensayos y otros diseños de experimentos. La comparación de los modelos BSS y SpATS se hace en función de la concordancia en los rankings de predicciones de los efectos genotípicos y la precisión y variabilidad genotípica de los ensayos. Además, ambos modelos espaciales se comparan contra un modelo no-espacial (es decir, un modelo que no incluye términos adicionales para ajustar la variabilidad espacial), con el objetivo de destacar la mejora en la precisión y reducción en la variancia residual obtenidas al modelar la variabilidad espacial en experimentos de mejoramiento.

OBJETIVOS

1. Describir, aplicar y comparar los modelos de análisis espacial estándar (BSS) y SpATS bajo diferentes escenarios.
2. Describir las ventajas de modelar la variabilidad espacial comparando ambos modelos con uno que no incluye términos para modelar la variabilidad espacial.

MÉTODOS

El modelo estándar

El método tradicionalmente utilizado en el análisis espacial en ensayos de campo, fue propuesto por Gilmour et al. (1997). Estos autores consideraron la descomposición de la

variabilidad espacial total en tres fuentes principales de variación: la variabilidad no estacionaria a través del campo (tendencia global), la variabilidad estacionaria dentro del ensayo (tendencia local) y la variabilidad externa.
La tendencia global generalmente refiere a tendencias (lineales o no) a lo largo del campo. Esta tendencia es modelada por polinomios unidimensionales y, de ser necesario, términos *spline* cúbicas, que siguen la dirección de las filas y/o columnas. Los términos lineales son incluidos en el modelo como efectos fijos mientras que, los términos *spline*, como efectos aleatorios.
La tendencia local, debida a la similitud de parcelas cercanas, generalmente es modelada por un proceso autorregresivo de primer orden ($AR1$) separable en dos dimensiones, considerando la autocorrelación entre las mismas a través de filas y columnas. Las parcelas vecinas tenderán a ser más similares que parcelas alejadas, debido a las características del suelo. El modelo $AR1 \times AR1$ (o su variante con una estructura de independencia en una de las direcciones) generalmente provee de una estructura de variancia adecuada para el ajuste de la tendencia local (Smith et al. 2001).
La variabilidad externa es introducida por causas no naturales, como resultado de las labores que se realizan en el campo, como ser sistemas de riego, orientación de cosecha, circuitos de sembrado, entre otros. Esta fuente se modela a través de efectos de diseño, como los efectos fila y/o columna, teniendo en cuenta cómo se opera en el campo a través de las filas y columnas, técnicas que suelen tener un patrón visible y recurrente. Dependiendo del patrón observado, se incluyen como efectos fijos y/o aleatorios.
El método estándar, se desarrolla en un procedimiento multi-etápico en el cual, a través de herramientas gráficas sobre los residuos y la estructura espacial de covariancia y tests formales sobre los términos del modelo, se selecciona el modelo de mejor ajuste. Para identificar las fuentes de variación, no existe un único camino. La selección del modelo, los factores a identificar y su forma varían de acuerdo al contexto. Es importante la distinción entre la tendencia global y la local. El proceso de identificación de las fuentes de variabilidad es extenso y secuencial, y se realiza utilizando el semivariograma muestral y los residuos del modelo ajustado.

El modelo SpATS

El modelo SpATS (Rodríguez-Álvarez et al. 2016a), modela las tendencias y variaciones del campo mediante una función de suavizado bivariada de las coordenadas espaciales.
Cuando las estructuras de variación son complejas, los métodos basados en regresiones multidimensionales utilizando líneas de suavizado polinómicas (*splines*) constituyen una buena alternativa de modelación. Esto se debe a la flexibilidad con la cual permiten modelar superficies de interacción. Estas regresiones basadas en *splines* son funciones eficientes para ajustar curvas, y están compuestas por porciones de polinomios (en general, cuadráticos o cúbicos) que se unen en puntos llamados nodos (*knots*).
Uno de los métodos que utiliza las *splines* está basado en líneas de suavizado penalizadas en dos dimensiones (Eilers y Marx 1996). El suavizado a través de *spline* penalizadas (*P-spline*) puede realizarse en dos o más dimensiones utilizando uno o más parámetros de suavizado que controlan el grado de suavizado de la superficie espacial ajustada para prevenir el sobreajuste.
El modelo P-*splines* en dos dimensiones (2D *P-spline*), tiene una formulación en el marco de los modelos lineales mixtos (Eilers 1999; Currie y Durbán 2002). En este contexto, pueden estimarse los parámetros de suavizado como cocientes de las componentes de variancia obtenidas por el método REML (Patterson y Thompson 1971). Estos parámetros controlan el

grado de suavizado de la superficie espacial; valores altos implican gradientes espaciales más suaves, mientras que valores bajos implican tendencias ajustadas más rugosas.
La superficie espacial de suavizado puede descomponerse en una suma de cinco componentes mutuamente independientes presentada por Lee et al. (2013) bajo el nombre PS-ANOVA. En consecuencia, el grado de suavizado de la superficie espacial es regulado por cinco parámetros diferentes.

Implementación de los modelos

El análisis de los ensayos comienza ajustando un modelo no-espacial utilizando ASReml-R (Butler et al. 2009), el cual se compone de efectos aleatorios de genotipo y bloque resoluble y errores independientes. Este ajuste no necesariamente debe hacerse utilizando el paquete mencionado.
Luego es ajustado un modelo espacial base utilizando ASReml-R, el cual considera como efectos aleatorios de genotipo y bloque resoluble y errores autocorrelacionados con una estructura separable $AR1$. En un proceso de modelación secuencial, se identifican tendencias y patrones en los gráficos de residuos y el semivariograma muestral asociados al modelo y se añaden términos de acuerdo a las interpretaciones gráficas. Por esta razón, la forma final del modelo BSS no incluye los mismos términos de ensayo a ensayo.
En paralelo, se ajusta un modelo SpATS siguiendo la descomposición PS-ANOVA (Lee et al. 2013), basado en configuraciones por defecto utilizadas en el marco de *P-splines* y estableciendo el número de nodos, utilizando el paquete de R, SpATS (Rodríguez-Álvarez et al. 2016b) de acceso gratuito.
En el marco del suavizado penalizado, el número exacto de nodos no resulta crucial una vez que cierta cantidad es excedida (Ruppert et al. 2003; Eilers et al. 2015). La única limitación en este contexto, es el tiempo computacional; a mayor cantidad de nodos, deben estimarse más parámetros y, por ende, la carga computacional resulta mayor. Es importante aclarar que un gran número de nodos proveerá de mayor flexibilidad, pero en la práctica, la penalización evitará el sobre-ajuste. Si el número de nodos es menor de lo necesario, el modelo no tendría suficiente flexibilidad para modelar la superficie espacial. En este estudio, para cada dimensión (filas y columnas) se utilizaron aproximadamente un nodo por cada dos filas y un nodo por cada dos columnas.
La bondad del modelo SpATS es evaluada por la similitud de sus resultados respecto al modelo BSS. Además, los dos modelos espaciales son comparados con el modelo no-espacial en cuanto a la variancia residual independiente y la precisión del ensayo.

Herramientas de comparación de los modelos

Los modelos BSS y SpATS son comparados utilizando las siguientes herramientas:

1. Correlación de Spearman entre los rankings de predicciones de los efectos genotípicos obtenidos por ambos modelos. Con el objetivo de medir el grado de acuerdo entre los rankings de predicciones de ambos modelos, se calcula la correlación de Spearman (Hollander y Wolfe 1999) entre los mismos para cada ensayo, la cual mide el grado de acuerdo y dirección de la misma.
2. Heredabilidad en sentido amplio. La heredabilidad proporciona una medida del grado de comprensión de los efectos genotípicos. Como la heredabilidad es una proporción, toma valores entre 0 y 1. Para que la selección de variedades en un ensayo sea efectiva, es deseable que el valor de heredabilidad del ensayo sea lo más alto posible.

3. Reducción de la variancia de los residuos independientes respecto al modelo no-espacial (NS). Mayores reducciones de la variancia de los residuos independientes reflejarán grandes mejoras en el modelo y, por ende, una mayor necesidad de modelar la variabilidad espacial.

MATERIALES

Los datos utilizados en la presente investigación provienen de 14 ensayos de porotos mung llevados a cabo en cinco localidades desde 2015 hasta 2019 en el contexto del NMIP y datos de 18 ensayos de garbanzos realizados en nueve localidades entre 2008 y 2015 en el contexto del programa de mejoramiento de garbanzos del PBA.

Los ensayos de porotos mung siguieron diseños completamente replicados y variaron de 90 a 900 parcelas y de 30 a 443 genotipos. Los ensayos realizados en 2015 y 2016 siguieron diseños fila-columna latinizados y de 2017 en adelante siguieron diseños óptimos basados en modelos que tienen en cuenta el pedigrí de los genotipos (Butler 2013). En general, los ensayos tuvieron un rendimiento medio estable alrededor de 1,045 t/ha con una desviación estándar de 0,508 t/ha.

Los ensayos de garbanzo variaron de 372 a 1128 parcelas y de 213 a 768 genotipos, donde 10 ensayos siguieron diseños completamente replicados con diseños bloques completamente aleatorizados (2008 – 2011) y de 2012 en adelante látices generalizados; otros 8 ensayos siguieron diseños parcialmente replicados. El rendimiento medio en estos ensayos fue 1,987 t/ha, con una desviación estándar de 0,908 t/ha.

RESULTADOS

En general, los rankings de las predicciones son muy consistentes para los modelos SpATS y BSS, con la media de los coeficientes de correlación de Spearman estimados de 0,98, con algunas leves desviaciones.

La heredabilidad calculada utilizando el modelo SpATS fue consistente con aquella obtenida con el modelo BSS. Al comparar las heredabilidades obtenidas con un modelo NS y con ambos modelos espaciales, hubo una gran mejora en la precisión de los ensayos al modelar la variabilidad espacial

Ambos modelos espaciales redujeron la variancia residual independiente, notando que, en general, las menores reducciones se dieron para ensayos de poroto mung para los cuales la variabilidad espacial fue de baja magnitud. En esos ensayos, la mayoría de la variabilidad fue explicada por los efectos genotípicos y la variancia residual era muy baja. A su vez, importantes reducciones se dieron para ensayos de garbanzo, donde la variabilidad espacial era mucho más intensa.

COMENTARIOS FINALES

En todos los casos, el modelo SpATS de Rodríguez-Álvarez et al. (2016a) demostró un desempeño similar al propuesto por Gilmour et al. (1997). En la presente investigación, al igual que en programas de sorgo, demostró gran capacidad para la selección eficiente de nuevas variedades y presentó gran flexibilidad ante los diferentes escenarios. Ambos modelos demostraron un fuerte acuerdo en la identificación de los mejores genotipos evaluados en un ensayo, con algunas discrepancias leves. Las heredabilidades calculadas bajo ambos modelos mostraron un gran acuerdo y las correlaciones de los rankings de las predicciones de los efectos genotípicos fueron muy altas.

Al comparar los modelos espaciales (BSS y SpATS) y el modelo NS, en la mayoría de los ensayos, los modelos espaciales redujeron la variancia residual independiente y mejoraron la precisión. Utilizando una metodología NS, los resultados sobre el rendimiento de las variedades evaluadas estarán influenciados por variabilidad espacial no modelada. Cuando la magnitud de la variabilidad espacial es grande, la aplicación de un método NS, no sería la decisión más apropiada.

Al modelar la variabilidad espacial con la función 2D P-splines y efectos aleatorios en filas y columnas, el modelo SpATS demostró gran capacidad para diferenciar los efectos genotípicos sin estas influencias, equiparando la capacidad del método estándar para el análisis espacial de los ensayos evaluados. Una cuestión pendiente es evaluar si el nuevo modelo es inclusive mejor en términos de precisión en las predicciones.

Generalmente, en programas de mejoramiento, la inclusión de información genética en el análisis es muy importante. Con el método de Gilmour et al. (1997) puede utilizarse información genética (pedigrí y/o marcadores moleculares) para explicar la relación entre los genotipos. Además, a través de este método pueden realizarse análisis combinado de ensayos realizados en distintas localidades o ambientes (MET, *Multi-Environment Trials*) evaluando así la interacción Genotipo-Ambiente, es decir, el comportamiento de los genotipos en distintos ambientes.

El modelo SpATS es una alternativa atractiva para el análisis de experimentos de genotipos realizados en el campo debido a la sencillez y rapidez en su ajuste, accesibilidad gratuita, flexibilidad ante distintas circunstancias y gran fiabilidad en sus resultados. La metodología estándar, por el contrario, requiere de un análisis más sofisticado y demanda mayor conocimiento por parte del analista ya que, además de ser necesario contar con las coordenadas de fila y columna en el campo de cada observación, para llevar a cabo el análisis, también se debe tener experiencia en el área para poder reconocer los patrones de variabilidad espacial y saber cómo modelarlos.

BIBLIOGRAFÍA

Butler, D.G., Cullis, B.R., Gilmour, A.R. & Gogel, B.J. (2009). *ASReml-R Reference Manual. Technical Report 3.* Queensland Department of Agriculture, Fisheries & Forestry.

Butler, D.G. (2013). *On the Optimal Design of Experiments Under the Linear Mixed Model.* PhD thesis, School of Mathematics and Physics, The University of Queensland.

Cullis, B.R. & Gleeson, A.C. (1991). Spatial Analysis of Field Experiments - An Extension to Two Dimensions. *Biometrics*, 47, 1449-1460.

Cullis, B.R., Smith, A.B. & Coombes, N.E. (2006). On the Design of Early Generation Variety Trials with Correlated Data. *Journal of Agricultural, Biological and Environmental Statistics*, 11, 381-393.

Currie, I.D. & Durbán, M. (2002). Flexible Smoothing with P-Splines: a Unified Approach. *Statistical Modelling Society*, 2, 333-349.

Eilers, P.H.C. (1999). Discussion on: the Analysis of Designed Experiments and Longitudinal Data by Using Smoothing Splines (by Verbyla et al.). *Journal of the Royal Statistical Society*, Ser C, 48, 307-308.

Eilers, P.H.C. & Marx, B.D. (1996). Flexible Smoothing with B-Splines and Penalties, *Statistical Science*, 11, 89-102.

Eilers, P.H.C., Marx, B.D. & Durbán, M. (2015). Twenty Years of P-Splines. *Statistics and Operations Research Transactions*, 39(2), 149-186.

Gilmour, A.R., Cullis, B.R. & Verbyla, A.P. (1997). Accounting for Natural and Extraneous Variation in the Analysis of Field Experiments. *Journal of Agricultural, Biological, and Environmental Statistics*, 2, 269-293.

Hollander, M. & Wolfe, D. A. (1999). *Nonparametric Statistical Methods*, Wiley, Chichester.

Lee, D.J., Durbán, M. & Eilers, P.H.C. (2013). Efficient Two-Dimensional Smoothing with P-Spline ANOVA Mixed Models and Nested Bases. *Computational Statistics and Data Analysis*, 61, 22-37.

Patterson, H.D. & Thompson, R. (1971). Recovery of Inter-Block Information when Block Sizes are Unequal. *Biometrika*, 58, 545-554.

Rodríguez-Álvarez, M.X., Boer, M.P., van Eeuwijk, F.A. & Eilers, P.H.C. (2016a). *Spatial Models for Field Trials*.

Rodríguez-Álvarez, M.X., Boer, M.P., Eilers, P.H.C. & van Eeuwijk, F.A. (2016b). SpATS: spatial analysis of field trials with splines. *R-package version* https://cran.r-project.org/web/packages/SpATS/index.html

Ruppert, D., Wand, M. & Carroll, R. (2003). *Semiparametric Regression*, Cambridge University Press, Cambridge.

Smith, A.B., Cullis, B.R. & Thompson, R. (2001). Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics*, 57, 1138-1147.

Velazco, J.G., Rodríguez-Álvarez, M.X., Boer, M.P., Jordan, D.R., Eilers, P.H.C., Malosetti, M. & van Eeuwijk, F.A. (2017). Modelling Spatial Trends in Sorghum Breeding Field Trials Using a Two-Dimensional P-Spline Mixed Model. *Theoretical and Applied Genetics*, 130, 1375-1392.

CONSTRUCCIÓN DE INDICADORES A PARTIR DE VARIABLES PERTENECIENTES A LA ENCUESTA NACIONAL DE VICTIMIZACIÓN 2017 Y ANÁLISIS DE LOS MISMOS MEDIANTE MAPAS DE DATOS GEORREFERENCIADOS POR JURISDICCIÓN DEL PAÍS

Lic. Danisa María Rosatto

Responsable de la Facultad de Ciencias Económicas y Estadística:
Mg. Guillermina B. Harvey
Responsables de la entidad: **Lic. Sabrina Balbi; Est. Nora Daruich**

La Encuesta Nacional de Victimización 2017 (ENV 2017) relevó delitos cometidos contra el hogar y contra la persona durante el año 2016. En el presente trabajo se elaboran mapas con datos georreferenciados, a fin de representar la distribución de indicadores construidos a partir de variables pertenecientes a dicha encuesta, en el territorio nacional. Además, se compara el comportamiento de los indicadores en las distintas jurisdicciones del país.
Entre los resultados obtenidos se identifica, en líneas generales, que en el año 2016 los delitos contra la persona fueron más habituales que los delitos contra el hogar. En la mayoría de las jurisdicciones el más frecuente fue hurtos personales, para delitos contra la persona, y robo o hurto en vivienda, para delitos contra el hogar. En general, las jurisdicciones menos victimizadas fueron: San Luis, La Pampa, Río Negro, Santa Cruz y Tierra del Fuego, y las más victimizadas: la Ciudad de Buenos Aires, Jujuy, Salta y Tucumán.

INTRODUCCIÓN

El presente trabajo es el resultado de una Práctica Profesional llevada a cabo en el Instituto Provincial de Estadística y Censos (IPEC), Delegación Rosario.

El objetivo de interés es la elaboración de mapas con datos georreferenciados, a fin de representar la distribución de indicadores construidos a partir de variables pertenecientes a la Encuesta Nacional de Victimización 2017 (ENV 2017), en el territorio nacional. Además, interesa comparar el comportamiento de los indicadores en las distintas jurisdicciones del país.

Es importante mencionar que, si bien la principal fuente de información oficial sobre los hechos delictivos son los registros administrativos policiales y judiciales, es claro que dichos registros no contabilizan aquellos casos en los cuales los delitos no son reportados a las autoridades competentes. Para contrarrestar esta limitación, las encuestas de victimización relevan los delitos que no fueron denunciados y los motivos por los cuales no se realizaron las correspondientes denuncias, complementando la información generada por los registros administrativos y proporcionando una estimación de los delitos no denunciados. Esto aporta una herramienta de diagnóstico para evaluar el vínculo entre la ciudadanía y el sistema de seguridad pública.

En particular, la ENV 2017 es la primera encuesta diseñada en conjunto entre el INDEC y el Ministerio de Seguridad de la Nación (MSN). Ambos organismos convinieron, durante el año 2016, mancomunar esfuerzos para promover la generación y análisis de datos sobre aspectos de seguridad ciudadana, a fin de ampliar y fortalecer las capacidades de implementación y monitoreo de políticas públicas en la materia.

Para cumplir con el objetivo propuesto en la Práctica Profesional, se analiza la base de datos de la ENV 2017 a través del programa estadístico SPSS. Previamente se seleccionan las variables para la construcción de los indicadores a georreferenciar. Luego, se construyen los mapas de interés utilizando el programa de georreferenciación Quantum GIS (QGIS). Finalmente, con las regiones mapeadas se realiza un análisis descriptivo y comparativo de los indicadores construidos, según las jurisdicciones del país.

MATERIALES Y MÉTODOS

Generalidades de la Encuesta Nacional de Victimización 2017

La ENV 2017 se llevó a cabo durante el primer semestre del año 2017. La población objetivo abarcó a las personas de 18 años o más, residentes en viviendas particulares de las localidades de 5.000 y más habitantes de la República Argentina.

Entre otros temas de interés, se indagó al encuestado sobre la ocurrencia de dos tipos de delitos:

Delitos contra el hogar: Robo o hurto de automóvil/camioneta/camión, Hurto de autopartes de automóvil/camioneta/camión, Robo o hurto de motocicleta o ciclomotor, Robo o hurto en vivienda, Secuestros.

Delitos contra las personas: Robo con violencia, Hurtos personales, Fraude bancario, Estafa/fraude, Secuestro virtual, Agresión física, Amenazas, Corrupción (soborno pasivo), Ofensas sexuales.

En cuanto a Secuestro virtual, el INDEC lo excluyó de los indicadores ya que ha demostrado ciertas características que requieren de un análisis más particular.

Vale aclarar que, para los delitos contra el hogar, se consideró como víctima al hogar afectado; mientras que, para los delitos contra las personas, se consideró como víctima al encuestado.

Base de datos

Para generalizar los resultados de la ENV 2017 a partir de los datos por muestra, y así poder componer las estimaciones definitivas, es necesario ponderar empleando factores de expansión.

En virtud de que se requieren estimaciones para distintas unidades (hogares y personas) la ENV 2017 dispone de dos factores de expansión, precisamente, uno asociado al hogar y otro a las personas.

Consideraciones para la construcción de indicadores

En primer lugar, es pertinente detallar la definición de prevalencia, la cual se define como el porcentaje de hogares o personas que fueron víctimas de al menos un delito durante el período de referencia. Este indicador contabiliza una única vez a cada víctima, así haya sufrido uno o más delitos, o más de un episodio de un mismo tipo de delito durante el período. Las prevalencias pueden calcularse para cada delito de manera separada o agrupada. Concretamente, las prevalencias agrupadas indican qué porcentaje de la población fue víctima de al menos uno de los delitos considerados en el agrupamiento durante el período de referencia. La prevalencia general se calcula considerando la totalidad de los delitos contra el hogar o la persona.

Para obtener las prevalencias separadas respecto a los delitos contra el hogar, en el caso de Robo o hurto de automóvil/camioneta/camión, Hurto de autopartes de automóvil/camioneta/camión y Robo o hurto de motocicleta o ciclomotor, se seleccionan los hogares que sufrieron ese delito y, además, se debe verificar si el vehículo en cuestión pertenecía a alguna persona del hogar. Para que el hogar sea considerado víctima, deben cumplirse ambas condiciones. Para el resto de los delitos contra el hogar y los delitos contra la persona no se impone condición alguna.

La prevalencia general de delitos contra la persona o contra el hogar total del país (P) se calcula como:

$$P = \frac{\hat{t}}{\hat{T}} \times 100,$$

donde, en el caso de delitos contra la persona, $\hat{t}$ es el total estimado de personas que presentaron al menos un delito contra la persona y $\hat{T}$ es el total estimado de personas a nivel país. En el caso de delitos contra el hogar, $\hat{t}$ es el total estimado de hogares que presentaron al menos un delito contra el hogar y $\hat{T}$ es el total estimado de hogares a nivel país.

Es claro que si lo que se pretende es calcular las prevalencias según las jurisdicciones del país, tanto en el numerador como en el denominador se consideran los valores correspondientes a cada una de dichas jurisdicciones.

Implementación computacional

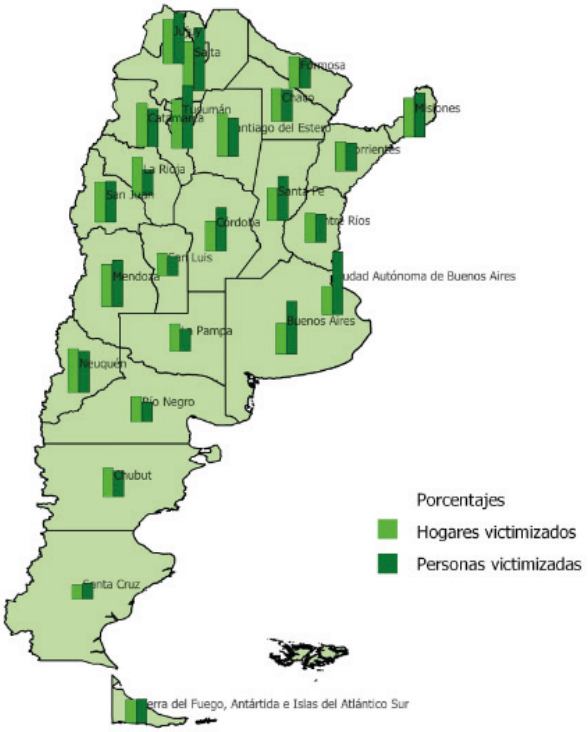
Se trabaja con el programa estadístico SPSS (versión 22), Excel y los mapas se obtienen con el Sistema de Información Geográfica QGIS (versión 3.4 Madeira).

RESULTADOS

Entre los resultados obtenidos se observó que, en el año 2016, a nivel país, alrededor del 14% de los hogares sufrió al menos un delito contra el hogar y alrededor del 20% de las personas fue víctima de al menos un delito contra la persona.

Si bien son muchos los gráficos realizados y presentados en el informe de la Práctica Profesional, se seleccionaron a modo de ejemplo sólo tres, con distintas formas empleadas para representar a los indicadores por jurisdicción del país. En el Gráfico 1 el indicador de prevalencia general de delitos contra el hogar y contra la persona se representa en forma de barras, en el Gráfico 2 el indicador de prevalencia de robo o hurto en vivienda se representa en degrade y en el Gráfico 3 el indicador de delitos más frecuentes contra la persona se representa de forma categorizada.

Gráfico 1. Prevalencia general de delitos contra el hogar y contra la persona según "Jurisdicción del país". Año 2016



Las jurisdicciones con mayor porcentaje de hogares afectados por delitos contra el hogar son Salta (20,5%) y Tucumán (20,3%), en cambio, con los porcentajes más bajos se encuentran San Luis y Tierra del Fuego (9% aproximadamente), seguida de Santa Cruz (6,3%). El resto de las jurisdicciones presentan porcentajes entre 10,0% y 18,5% (Gráfico 1).

En cuanto a los delitos contra la persona, las jurisdicciones con mayor porcentaje de víctimas son: Salta (26,4%), Tucumán (26,2%) y la Ciudad de Buenos Aires (26,1%), a las que le suceden las provincias de Buenos Aires (22,1%) y Jujuy (21,3%). El resto de las jurisdicciones tienen una prevalencia menor que la correspondiente al total país (19,9%) (Gráfico 1).

Al comparar el porcentaje de hogares afectados con el porcentaje de personas victimizadas, según cada jurisdicción, se puede decir que:

✔ en algunas jurisdicciones son más prevalentes los delitos contra la persona que contra el hogar, como ocurre en: la Ciudad de Buenos Aires, Buenos Aires, Salta, Tucumán, Córdoba, Santa Fe, Jujuy, Misiones y Mendoza;

✔ en otras jurisdicciones son más prevalentes los delitos contra el hogar, como ser en: La Rioja, La Pampa, Santiago del Estero, Catamarca y Río Negro;

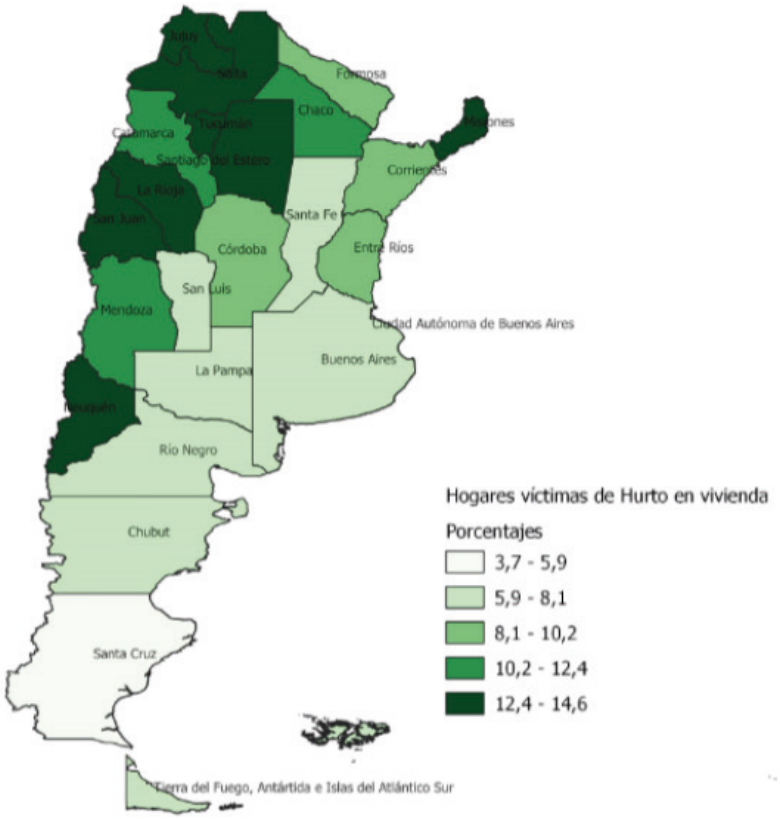
✔ San Juan, Tierra del Fuego, Chaco, Entre Ríos, Santa Cruz, Corrientes, Formosa Neuquén, San Luis y Chubut, presentan prevalencias similares en ambos tipos de delitos (Gráfico 1).

En general, La Pampa, Tierra del Fuego, Río Negro, San Luis y Santa Cruz presentaron menor porcentaje de delitos que el resto de las jurisdicciones, mientras que Salta, Tucumán, Jujuy y la Ciudad de Buenos Aires fueron las más victimizadas (Gráfico 1).

A nivel país se registró, en el año 2016, un mayor porcentaje de delitos contra la persona comparado con el porcentaje de delitos contra el hogar (Gráfico 1).

Los hogares de la mayoría de las jurisdicciones han sufrido en mayor medida Robo o hurto en vivienda, mientras que en la provincia de Buenos Aires el mayor porcentaje de hogares victimizados sufrió Hurto de autopartes de automóvil/camioneta/camión. Para visualizar cómo se distribuye el porcentaje de este delito más frecuente en el país, por jurisdicción, se realiza el mapa que se presenta en el Gráfico 2.

Gráfico 2. Prevalencia de Robo o hurto en vivienda según “Jurisdicción del país”. Año 2016

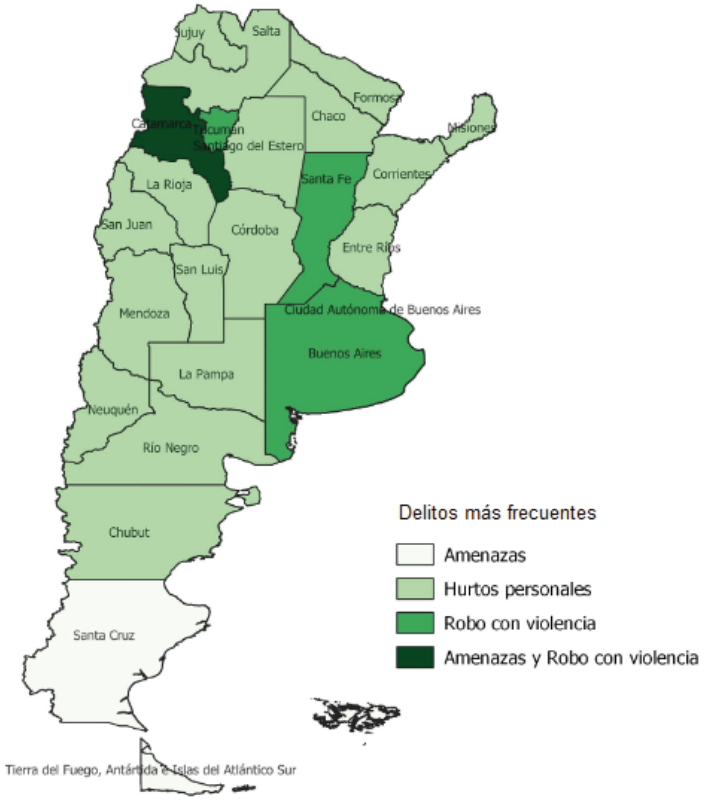


Las provincias del Noroeste argentino (Catamarca, La Rioja, Jujuy, Santiago del Estero, Tucumán y Salta) presentan los porcentajes más elevados de hogares que han sufrido Robo o hurto en vivienda (entre 12,3% y 14,0%), también las provincias del Nordeste (Corrientes, Formosa, Chaco y Misiones) junto con Entre Ríos, Córdoba, San Juan, Mendoza y Neuquén, muestran altos porcentajes (entre 8,9% y 14,6%) (Gráfico 2).

Excepto Neuquén, las provincias de la Patagonia (Santa Cruz, Tierra del Fuego, Chubut y Río Negro) al igual que la Ciudad de Buenos Aires, San Luis, Santa Fe, Buenos Aires y La Pampa,

poseen los porcentajes más bajos de hogares víctimas de Robo o hurto en vivienda (entre 3,7% y 7,6%) (Gráfico 2).

Gráfico 3. Delitos más frecuentes contra la persona según “Jurisdicción del país”. Año 2016



Al observar el Gráfico 3 se puede decir que, en la mayoría de las jurisdicciones, el delito que presentó el porcentaje más elevado de víctimas es Hurtos personales, mientras que, en Santa Fe, Buenos Aires y Tucumán el mayor porcentaje de personas victimizadas fue a causa de Robo con violencia, y de Amenazas en Tierra Del Fuego y Santa Cruz.
En cuanto a Catamarca el porcentaje de víctimas de amenazas es el mismo que el de robo con violencia.

CONSIDERACIONES FINALES

Entre los resultados obtenidos se consiguió identificar, en líneas generales, que en el año 2016 los delitos contra la persona fueron más frecuentes que los delitos contra el hogar; el más reiterado en la mayoría de las jurisdicciones del país fue Hurtos personales y con respecto a los delitos contra el hogar el más usual, excepto en la provincia de Buenos Aires, fue el Robo o hurto en vivienda. En general, las jurisdicciones menos victimizadas resultaron ser: San Luis, La Pampa, Río Negro, Santa Cruz y Tierra del Fuego, y las más victimizadas: La Ciudad de Buenos Aires, Jujuy, Salta y Tucumán.
Para finalizar, vale una aclaración acerca del caso de Santa Cruz, ya que el INDEC manifiesta que el relevamiento en esta jurisdicción estuvo afectado por dificultades, tanto climáticas como operativas, dando una muestra efectiva menor a la deseada. Al observar los diversos mapas construidos en la Práctica Profesional, Santa Cruz se destaca por presentar porcentajes en general con un patrón diferente al de la mayoría de las jurisdicciones del país. No obstante,

siguiendo la línea de análisis del INDEC en la Práctica Profesional, se decidió incluir sus resultados en el estudio, haciendo la salvedad de que deben ser considerados con cautela.

BIBLIOGRAFÍA

Bezos, I. & Retamosa, E. (2008*). "Introducción a los SIG. Producción y gestión de la información geográfica".* Recuperado de: https://www.santafe.gob.ar/idesf/portal/index.php/capacitacion-interna

Comari, C. & Hoszowski, A. "*Ponderación de la muestra y tratamiento de valores faltantes en las variables de ingreso en la Encuesta Permanente de Hogares".* Recuperado de: file:///C:/Users/DANISA/Downloads/EPH-Encuesta-Permanente-de-Hogaresmetodologia-N%C2%BA-15%20(1).pdf

Francois, A. (2017). "*QGis: junturas en Excel".* Recuperado de: https://www.sigterritoires.fr/index.php/es/qgis-junturas-con-tablas-excel/

Herrador, M. & Saralegui, J. (marzo de 2005). "La estimación en áreas pequeñas para la estadística oficial". *BEIO, 21*(1), 47-53. Recuperado de: http://www.seio.es/BEIO/files/BEIO_Vol21Num1_EO_MHerrador+JSaralegui.pdf

Instituto Geográfico Nacional. *Capas SIG. Provincia.* Recuperado de: https://www.ign.gob.ar/NuestrasActividades/InformacionGeoespacial/CapasSIG

Instituto Nacional de Estadística y Censos - I.N.D.E.C. (2017). *"Encuesta Nacional de Victimización 2017". Diseño general*. Recuperado de: https://www.indec.gob.ar/ftp/cuadros/sociedad/env_aspectos_metodologicos.pdf

Instituto Nacional de Estadística y Censos - I.N.D.E.C. (2017). *"Encuesta Nacional de Victimización 2017* [Base de datos]". Recuperado de: https://www.indec.gob.ar/indec/web/Institucional-Indec-BasesDeDatos-5

Instituto Nacional de Estadística y Censos - I.N.D.E.C. (2017). *"Informes Técnicos. Condiciones de vida. Encuesta Nacional de Victimización 2017. Resultados preliminares y provisorios* (2545-6660)". Recuperado de: https://www.indec.gob.ar/ftp/cuadros/publicaciones/env_2017_07_17.pdf

Instituto Nacional de Estadística y Censos - I.N.D.E.C. (2018). "*Encuesta Nacional de Victimización 2017".* Recuperado de: https://www.indec.gob.ar/uploads/informesdeprensa/env_2017_02_18.pdf

Instituto Nacional de Estadística y Censos - I.N.D.E.C. "*Documento para la utilización de la base de datos usuario".* Recuperado de: https://www.indec.gob.ar/ftp/cuadros/menusuperior/env/env2017_diccionario_registros.pdf

Instituto Nacional de Estadística y Censos- I.N.D.E.C. "*Encuesta Nacional de Victimización 2017, Manual del Encuestador".*

Instituto Nacional de Estadística y Censos- I.N.D.E.C. "*Encuesta Nacional de Victimización 2017, Cuestionario".* Recuperado de: https://www.indec.gob.ar/ftp/cuadros/sociedad/env_cuestionario_06_17.pdf

Instituto Nacional de Estadística y Censos- I.N.D.E.C. "*Encuesta Nacional de Victimización 2017: Cuadros estadísticos del informe, coeficientes de variación e intervalos de confianza".* Recuperado de: https://www.indec.gob.ar/indec/web/Nivel4-Tema-4-34-155

Instituto Provincial de Estadísticas y Censos, delegación Santa Fe. "*Sobre IPEC".* Recuperado de http://www.estadisticasantafe.gob.ar/sobre-ipec/

Montoya, S. (2017, enero 5). "Tutorial para la representación mejorada de Campo de Valores en QGIS". Recuperado de: https://www.youtube.com/watch?v=4WYSu4xFOp8

QGIS (2017). "*QGIS Training Manual. QGIS Project".* Recuperado de: file:///C:/Users/DANISA/Downloads/QGIS-2.14-QGISTrainingManual-es%20(1).pdf

QGIS (2019). "*QGIS User Guide: QGIS Project".* Recuperado de: https://docs.qgis.org/2.18/pdf/es/QGIS-2.18-UserGuide-es.pdf

QGIS. Recuperado de: https://qgis.org/es/site/

ELABORACIÓN DE ÍNDICES DE VULNERABILIDAD ANTE AMENAZA DE INUNDACIONES PARA LA CIUDAD DE ROSARIO, PROVINCIA DE SANTA FE, ARGENTINA, APLICANDO DISTINTAS METODOLOGÍAS MULTIVARIADAS

Lic. Lautaro Ruiz

Responsable de la Facultad de Ciencias Económicas y Estadística:
Mg. Virginia Laura Borra
Responsable de la entidad: **Lic. Laura Rita Balparda**

En el presente estudio se plantea la obtención de un índice de vulnerabilidad ante inundaciones para la ciudad de Rosario, Santa Fe, Argentina; utilizando datos censales y de los portales de Datos Abiertos. El conjunto inicial de indicadores a usar se identificó en el marco de un trabajo interdisciplinario entre diferentes áreas de la Municipalidad de Rosario y el Área de Sensores Remotos (Universidad Nacional de Rosario), considerando la definición de Cardona (2001) sobre vulnerabilidad. Un primer índice se calcula siguiendo la propuesta metodológica de la Secretaría de Protección Civil y Abordaje Integral de Emergencias y Catástrofes del Ministerio de Seguridad de la Nación Argentina. En una segunda propuesta, se aplica un análisis de componentes principales (ACP) y por selección o combinación de las componentes se construyen cuatro índices de vulnerabilidad, clasificados considerando el agrupamiento resultante del análisis de cluster (clasificación jerárquica de Ward y k-medias). Finalmente, en una tercera propuesta se utiliza un análisis factorial múltiple (AFM), seleccionando al primer factor obtenido como el índice de vulnerabilidad. Se obtuvo que el índice que considera la primera componente principal (explica 58% de la variación total) es el que mejor cuantifica la situación de vulnerabilidad percibida en la ciudad de Rosario.

INTRODUCCIÓN

La vulnerabilidad, según Cardona (2001), es “la predisposición o susceptibilidad física, económica, política o social que tiene una comunidad de ser afectada o de sufrir daños en caso de que un fenómeno desestabilizador de origen natural o antrópico se manifieste”. Se considera que los distintos aspectos o dimensiones de la vulnerabilidad se pueden subdividir en tres categorías: susceptibilidad física, fragilidades socioeconómicas y falta de resiliencia. La primera se corresponde a un riesgo “duro” relacionado con el daño potencial a la infraestructura física y en el ambiente. La segunda y la tercera contribuyen a un riesgo “blando” que se relaciona con el impacto potencial sobre el contexto social y las comunidades (Cardona, 2006). Este autor, al igual que otros, tales como Wisner *et al.* (2004), consideran al riesgo como una función que depende de la amenaza principal (por ejemplo: inundaciones, terremotos, entre otras) y de las condiciones de vulnerabilidad asociadas a ésta.

Es necesario que las sociedades evalúen su situación de vulnerabilidad previamente al impacto de una amenaza natural o antrópica, de forma tal que se puedan llevar a cabo medidas preventivas y evitar serios daños. El monitoreo y cuantificación de la vulnerabilidad se pueden utilizar para identificar comunidades y poblaciones relativamente más frágiles. Esta información se podría considerar en primer término en la planificación o diseño de protocolos, ya sea preventivos o de acción frente a la ocurrencia de una amenaza natural.

Diversas metodologías han sido usadas para evaluar la vulnerabilidad, tanto a nivel nacional como internacional. En este sentido, se destacan las propuestas de Cutter *et al.* (2003), Fernández *et al.* (2016) y Apostos (2019) en Estados Unidos, Portugal y Sudáfrica, respectivamente. Localmente, se podrían nombrar los trabajos del Sistema Municipal de Epidemiología dependiente de la Secretaría de Salud Pública de la Municipalidad de Rosario (2005) para la construcción del índice de condiciones saludables para la ciudad de Rosario, Merello (2010) para la elaboración del mapa de vulnerabilidad social para la provincia de Santa Fe y Cardoso (2017) para generar un índice de vulnerabilidad socio-ambiental para los distritos Santa Fe, Recreo y Monte Vera.

En Argentina, ante la sanción de la ley 27.287/2016 de creación del Sistema Nacional para la Gestión Integral del Riesgo y la Protección Civil, surgió la necesidad de plantear una metodología de planeamiento a fin de obtener eficacia en la prevención y mitigación de efectos adversos en todo el territorio nacional. En este sentido, la Secretaría de Protección Civil y Abordaje Integral de Emergencias y Catástrofes del Ministerio de Seguridad de la Nación Argentina desarrolló el “Manual para la elaboración de mapas de riesgo” (2017), con el fin de contar con una herramienta de análisis para la determinación de escenarios de riesgos.

En consideración a la definición de vulnerabilidad mencionada, los estudios previos, el estado de situación local y las prioridades establecidas por Defensa Civil de la Municipalidad de Rosario, el presente trabajo tiene el propósito de realizar un aporte al desarrollo del mapa de riesgo de la ciudad. Así, se plantean distintas metodologías para la elaboración del índice de vulnerabilidad ante inundaciones.

OBJETIVOS

Cuantificar y caracterizar la vulnerabilidad frente a la amenaza de inundaciones en la ciudad de Rosario, provincia de Santa Fe, Argentina; utilizando datos del Censo Nacional de Población, Hogares y Viviendas, año 2010, y de los portales de Datos Abiertos de Argentina y Rosario.

MATERIALES Y MÉTODOS

Área de estudio y unidad de análisis

La ciudad de Rosario está ubicada en la zona sur de la provincia de Santa Fe, República Argentina. El municipio de Rosario ocupa una superficie de 178,69 km^2, y se divide en 1070 radios censales. Estas unidades, son definidas por el Instituto Nacional de Estadísticas y Censos de la República Argentina (INDEC) a los fines operativos del censo y cuentan en promedio con 300 viviendas, en zonas urbanas. Se selecciona el radio censal como unidad de análisis geográfica dado que son las áreas más pequeñas y homogéneas para las cuales se dispone de datos.

Datos

En primer lugar, los datos del Censo 2010 se procesan con REDATAM+SP, para la obtención de los indicadores. En segundo lugar, desde los portales de Datos Abiertos de la Secretaría de Modernización Administrativa de la Presidencia de la Nación y de la Municipalidad de Rosario se descargan los datos geoespaciales, en formato abierto referidos a: barrios populares, centros de salud, centros de convivencia barrial, puntos de acceso a *wifi* público y contenedores de recolección pública de residuos. Además, se utiliza la base cartográfica por radio censal de la ciudad de Rosario, en formato ESRI *Shapefile*, que almacena la gráfica y la ubicación geográfica de los mismos, acompañados de su código y estadísticas básicas; digitalizada por el INDEC para el Censo 2010.

A partir de los datos en formato abierto, el Área de Sensores Remotos y la Dirección de Cartografía junto con la Dirección General de Estadística y la Dirección General de Gestión Integral de Residuos de la Municipalidad de Rosario, realizaron geoprocesos para el cálculo del porcentaje de área cubierta en el caso de barrios populares y de distancias promedio para los datos restantes por radio censal, en metros, utilizando el software QGIS, versión 2.16.3 Nødebo y 3.6.3 Noosa.

Indicadores

En consideración con la definición de vulnerabilidad propuesta por Cardona (2001), la revisión bibliográfica y en un marco de trabajo interdisciplinario se seleccionaron inicialmente 39 indicadores. Los mismos corresponden a las dimensiones de infraestructura, servicios básicos y localización en la categoría de susceptibilidad física; demografía, educación, económica y régimen de tenencia de la tierra en la de fragilidades socioeconómicas y por último comunicación, institucional y transporte, para falta de resiliencia.

Esquema Metodológico

Teniendo en cuenta los antecedentes, para el presente trabajo se proponen tres metodologías para la elaboración del índice de vulnerabilidad ante inundaciones y su posterior representación geográfica en un mapa, clasificando las áreas en baja, media y alta vulnerabilidad.

En la primera propuesta para la obtención del índice de vulnerabilidad se siguen las recomendaciones de la Secretaría de Protección Civil de la Nación en el “Manual para la elaboración de mapas de riesgos” (2017), que consisten en:

1. Seleccionar variables o indicadores de mayor representatividad en cada dimensión teniendo en cuenta su variabilidad y correlación.
2. Calcular un indicador porcentual para cada variable o indicador seleccionado (cantidad de hogares o personas por radio censal sobre la suma total de hogares o personas, multiplicado por 100).

3. Clasificar cada indicador por radio censal, asignando el score:
    - 1: valores del indicador mayores o iguales al 0% y menores al 30%.
    - 5: valores del indicador mayores o iguales al 30% y menores al 70%.
    - 10: valores del indicador mayores o iguales al 70% y menores o iguales al 100%.
4. Sumar los scores de los indicadores por radio censal. Rango de variación del índice de vulnerabilidad:
    - Valor mínimo, igual al número de indicadores.
    - Valor máximo, igual al número de indicadores, multiplicado por 10.
5. Normalizar el índice de vulnerabilidad dividiendo por el valor máximo del índice y multiplicado por 100. El índice toma valores entre 10 y 100.
6. Clasificar el índice de vulnerabilidad normalizado en:
    - Baja: valores del índice normalizado mayores o iguales a 0 y menores a 30.
    - Media: valores del índice normalizado mayores o iguales a 30 y menores a 70.
    - Alta: valores del índice normalizado mayores o iguales a 70 y menores o iguales a 100.

En la segunda propuesta, inicialmente se realiza un análisis de correlación en cada dimensión para eliminar indicadores redundantes, como así también *boxplots* y mapas temáticos para conocer las distribuciones espaciales de los mismos.

Posteriormente se aplica un análisis de componentes principales (ACP), y a través de la combinación o selección de las componentes principales retenidas (Fernandez *et al.*, 2016) se elaboran cuatro índices de vulnerabilidad. Uno de ellos, deriva en un criterio de clasificación cualitativo, mientras que los restantes son expresados cada uno como una variable cuantitativa.

El índice de vulnerabilidad cualitativo se obtiene siguiendo la estrategia desarrollada por Lebart *et al.* (1995) para formar agrupamientos de radios censales con características similares en términos de vulnerabilidad. Esta estrategia combina el ACP con un análisis de cluster, seleccionando los ejes más relevantes obtenidos en un ACP y utilizando un algoritmo de clasificación mixto (métodos de Ward y K-medias). Luego, los agrupamientos se caracterizan en función de los promedios de cada uno de los indicadores que difieren de manera significativa del valor promedio global (promedio en función de los indicadores para la totalidad de radios censales), acompañado de estadísticas de comparación; esta caracterización permite etiquetar a los agrupamientos según su nivel de vulnerabilidad (baja, media y alta).

Por su parte, los tres índices de vulnerabilidad numéricos se clasifican en baja, media y alta vulnerabilidad a partir de sus valores en cada uno de los agrupamientos obtenidos en el análisis cluster (según la estrategia propuesta en la investigación de César, 2014). Es decir, observado el rango de variación de cada índice cuantitativo (3 en total) en cada categoría de vulnerabilidad del índice cualitativo se establecen los límites de los intervalos de clasificación.

En la tercera propuesta se aplica un análisis factorial múltiple (AFM), técnica desarrollada por Escofier y Pagès (1990), con el objetivo de obtener un índice de vulnerabilidad donde el cálculo considere la estratificación de los indicadores estandarizados en grupos. Ésta consiste en un ACP global que busca equilibrar la influencia de los grupos, ponderando a cada uno de ellos por la raíz cuadrada del autovalor obtenido de realizar un ACP de manera independiente en cada grupo. Para este trabajo se considera el agrupamiento de los indicadores según las categorías de vulnerabilidad definidas por Cardona (2001).

Finalmente, observando la distribución del índice de vulnerabilidad resultante del AFM en las categorías de agrupamiento del análisis de cluster de la propuesta metodológica anterior, quedan definidos grupos no mutuamente excluyentes que conducen a la determinación de nuevos puntos de corte a usar en la clasificación de baja, media y alta vulnerabilidad.

En cada una de las propuestas, el resultado se representa cartográficamente según los niveles de vulnerabilidad.
Para la realización del análisis descriptivo, así como también, en el ACP, AFM y análisis cluster se utilizan los programas estadísticos R y RStudio, versión 3.5.3 y 1.1.442, respectivamente. Además, en el ACP y AFM se usan los paquetes FactoMineR y factoextra, mientras que en el análisis de clusters, el paquete FactoClass.

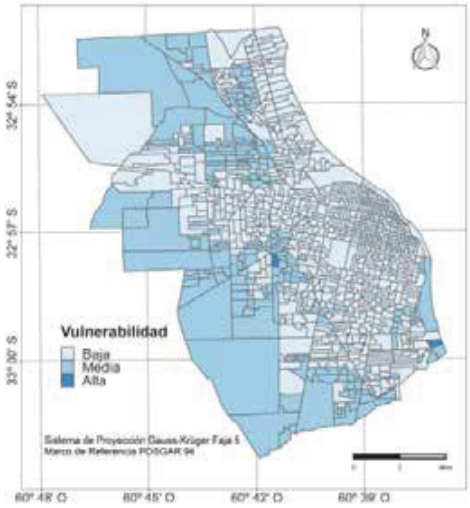
RESULTADOS

Siguiendo a Apotsos (2019), quien recomienda separar los radios censales poco poblados, se excluyen de los análisis aquellos con menos de 20 personas o menos de 10 hogares; puesto que los errores en la recolección de los datos en estas unidades producen un mayor sesgo, en comparación a las unidades de mayor tamaño. Los radios 4, 5 y 2 de las fracciones 22, 41 y 56, respectivamente, cumplen con al menos uno de estos criterios y se eliminan del proceso de construcción del índice de vulnerabilidad. Posteriormente, estos son clasificados en la clase de baja vulnerabilidad del mapa temático debido a que se encuentran ubicados en zonas de la ciudad consideradas con buenas condiciones.
Con respecto a la primera propuesta, a partir de encuentros interdisciplinarios realizados se seleccionó un subconjunto de indicadores, a saber:

- % de hogares que se alojan en viviendas cuya calidad de los materiales responde a la categoría 3 y 4 (materiales poco resistentes).
- % de hogares que se alojan en viviendas con calidad de conexión a servicios básicos correspondiente a las categorías básica o insuficiente (sin conexión a agua de red y cloaca).
- % de área del radio censal con barrios populares.
- % de población menor a 14 años.
- % de jefes de hogar con nivel primario completo como máximo nivel de instrucción alcanzado.
- % de hogares con al menos un indicador de NBI.
- % de hogares sin tenencia segura de la tierra.
- % de hogares sin computadora.

Cada indicador se clasificó considerando los scores 1, 5 y 10. Luego, se calculó el índice normalizado y se lo clasificó en las clases baja, media y alta (Figura 1).
Los radios censales clasificados con baja vulnerabilidad se localizaron hacia la zona norte (barrio Alberdi), centro, algunas zonas del distrito sur y noroeste (barrio Fisherton) de la ciudad de Rosario. Por su parte, los radios censales con vulnerabilidad media se encontraron hacia la periferia de la ciudad, sumando áreas al interior donde, en algunas de ellas, se sitúan asentamientos irregulares. Cabe resaltar que sólo 2 radios censales se clasificaron con alta vulnerabilidad.

Figura 1. Mapa de vulnerabilidad siguiendo la primera propuesta metodológica



Un total de 16 indicadores (Tabla 1) se incluyeron en el ACP y el AFM, los cuales fueron seleccionados a partir de estudiar su comportamiento y distribución espacial, utilizando análisis de correlación (identificando correlaciones mayores a 0,80 en valor absoluto), *boxplot* y mapas temáticos, así como también, de un trabajo interdisciplinario con diferentes actores de la Municipalidad de Rosario y de la Universidad.

Tabla 1. Indicadores incluidos en los análisis multivariados

Categoría	Dimensión	Indicador
Susceptibilidad Física	Infraestructura	% hogares que se alojan en viviendas referidas a casas tipo B, ranchos y casillas
		% hogares que se alojan en viviendas cuya calidad de los materiales responde a la categoría 4
		% hogares que se alojan en viviendas cuya calidad constructiva responde a la categoría básica o insuficiente
	Servicios Básicos	% hogares sin procedencia del agua de red pública para beber y cocinar
		% hogares sin desagüe de inodoro a red pública
		% hogares sin gas de red
		Distancia promedio al contenedor de residuos más cercano en el área urbana (metros)
Fragilidades Socioeconómicas	Demográfica	% población menor a 14 años y mayor a 70 años
	Educación	% jefes de hogar con nivel primario incompleto como máximo nivel de instrucción alcanzado
		% jefes de hogar con nivel primario completo como máximo nivel de instrucción alcanzado
	Económico	% hogares con al menos un indicador NBI
		Tasa de desocupación
	Régimen de Tenencia	% hogares sin tenencia segura de la tierra
Falta de Resiliencia	Comunicación	% hogares sin computadora
		Distancia promedio a una zona wifi pública en el área urbana (metros)
	Institucional	Distancia promedio a un centro de salud o centro de convivencia barrial más cercano en el área urbana (metros)

Siguiendo con la segunda propuesta metodológica, se aplicó un ACP y se retuvieron las primeras cuatro componentes principales (CP), dado que explican al menos un 80% de la variabilidad total. Cabe resaltar que la primera componente (CP1) explicó aproximadamente el 58% de la variación total de las observaciones. Además, esta componente presentó una fuerte correlación en sentido positivo con 8 indicadores y moderadamente con 7 de ellos (6 en sentido

positivo y 1 en sentido negativo). Dada la correlación de los indicadores con la CP1 y su variancia explicada, la misma podría constituirse por sí sola en un índice de vulnerabilidad. No obstante, se observaron 3 indicadores correlacionados moderadamente con la segunda componente y 2 con la tercera componente. Por último, la cuarta componente principal evidenció una correlación moderada en sentido negativo con un único indicador, por lo cual participa con signo negativo.

Las cuatro primeras CPs para cada radio censal se estandarizaron y se calcularon los 3 índices de vulnerabilidad para Rosario, a saber:

- Índice de vulnerabilidad 1: suma de las CPs estandarizadas retenidas.
- Índice de vulnerabilidad 2: CP1 estandarizada.
- Índice de vulnerabilidad 3: suma ponderada de las CPs estandarizadas retenidas, con ponderación igual a la proporción de variancia explicada por la CP.

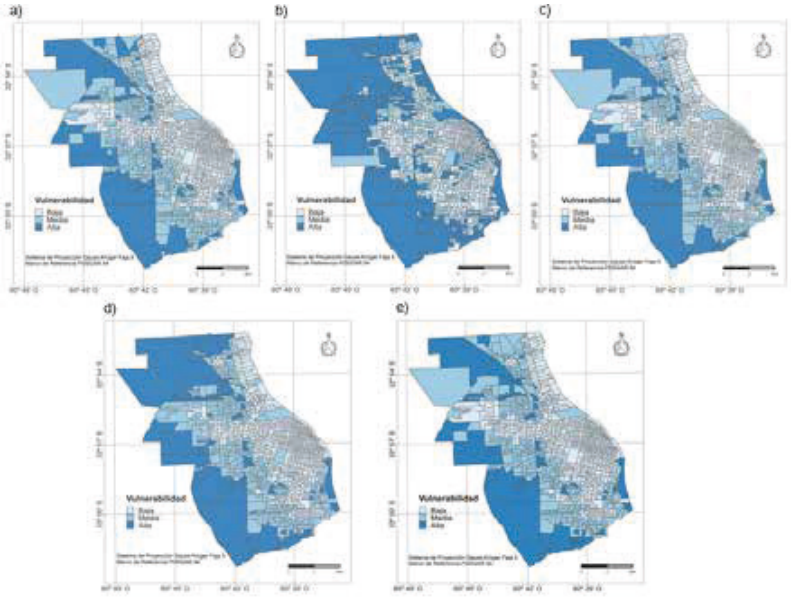
Con el fin de contar con más información para la clasificación de los índices de vulnerabilidad, se realizó el análisis de cluster considerando las primeras 6 CPs, que explicaron el 88,5% de la variación total de los datos originales. Se aplicó entonces el algoritmo de clasificación mixto y se obtuvieron tres agrupamientos de radios censales. Cada grupo se caracterizó en función de los valores promedios de los indicadores que difieren de manera significativa del valor promedio global, etiquetando estos agrupamientos de radios censales en baja, media y alta vulnerabilidad, y obteniendo de este modo un índice de vulnerabilidad cualitativo (Figura 2a).

Llegado a este punto del análisis, el desafío consistió en clasificar los 3 índices de vulnerabilidad obtenidos con el ACP (expresados como variables continuas), considerando el agrupamiento de los radios censales según el análisis cluster. De este modo estos pueden ser comparados en el mapa temático con el indicador obtenido siguiendo las recomendaciones de la Secretaría de Protección Civil, según los niveles de vulnerabilidad baja, media y alta. Sin embargo, los rangos de variación de cada índice de vulnerabilidad cuantitativo en cada uno de los tres agrupamientos del análisis cluster no eran mutuamente excluyentes. Este punto se resolvió utilizando los promedios de cuantiles, obteniendo intervalos mutuamente excluyentes para la clasificación de los índices de vulnerabilidad 1, 2 y 3 en baja, media y alta vulnerabilidad. A partir de esta clasificación se obtuvieron los mapas que se presentan en la Figura 2b, c y d.

Por otro lado, siguiendo con la tercera propuesta, en el análisis global del AFM se obtuvo que el porcentaje de variabilidad total explicada por el primer factor fue del 58% aproximadamente. Asimismo, se pudo observar una fuerte relación de este factor con cada una de las categorías de vulnerabilidad (susceptibilidad física, fragilidades socioeconómicas y falta de resiliencia), en el sentido de constituir una dirección de inercia (variabilidad) importante para cada una de ellas. De esta manera, el primer factor obtenido del AFM puede ser considerado un factor común para los tres grupos de indicadores.

Este índice de vulnerabilidad (expresado como variable continua o compleja), también se clasificó en los niveles de baja, media y alta vulnerabilidad de acuerdo con el agrupamiento de los radios censales del análisis cluster, de la misma forma que los índices cuantitativos de la propuesta anterior, ya que se consideró que logra una buena clasificación. Finalmente se obtuvo su representación cartográfica (Figura 2e).

Figura 2. Mapas de Vulnerabilidad: a) Análisis Cluster; b) Índice de vulnerabilidad 1; c) Índice de vulnerabilidad 2; d) Índice de vulnerabilidad 3; e) Índice de vulnerabilidad del AFM



A partir del trabajo interdisciplinario con el área de Defensa Civil de la Municipalidad de Rosario se logró identificar que el agrupamiento de los radios censales obtenido en el análisis de cluster, fue el que mejor se aproximó a la situación de vulnerabilidad percibida en la ciudad de Rosario (Figura 2a).

Por su parte, el índice de vulnerabilidad 2 (sólo incluye la CP1) fue el que mejor cuantificó la situación de vulnerabilidad percibida en la ciudad, donde se observó un alto porcentaje de coincidencia (98,50 %) en la clasificación de los radios censales en las tres clases en comparación con el agrupamiento por cluster (Figuras 2c y 2a, respectivamente).

En el mapa de vulnerabilidad del índice obtenido en el AFM (Figura 2e) también se obtuvo un alto porcentaje de coincidencia (97,30 %) en la clasificación de los radios censales en las tres clases de vulnerabilidad en comparación con el agrupamiento obtenido en el análisis cluster.

Al igual que en la Figura 1, en los mapas de las Figuras 2a, 2c y 2e, los radios censales clasificados en la categoría de baja vulnerabilidad se localizaron principalmente en las zonas centro, noreste (barrio Alberdi), noroeste (Fisherton) y parte de la zona sur de Rosario. En tanto que los radios censales con vulnerabilidad media se ubicaron bordeando las zonas de baja vulnerabilidad. Por último, hacia la periferia de la ciudad se encontraron los radios censales con alta vulnerabilidad, y se sumaron algunas excepciones al interior del municipio que se condicen, en general, con la localización de asentamientos irregulares.

CONCLUSIONES

Se evidenció que a partir de las tres propuestas metodológicas seleccionadas se logró cuantificar y caracterizar la vulnerabilidad, a través de la obtención de índices (cuantitativos continuos) y su posterior clasificación en las clases baja, media y alta.

La realización de reuniones interdisciplinarias propició un espacio de intercambio, que permitió la identificación de los indicadores iniciales a considerar, el posterior seguimiento en los avances del estudio y el intercambio de saberes entre lo académico y lo profesional-operativo, que resultaron en aportes significativos para el desarrollo del presente trabajo.

Los hallazgos en este estudio inducen a continuar en un futuro con el análisis de la vulnerabilidad ante inundaciones en la ciudad de Rosario, de modo tal de incluir indicadores actualizados, como así también, buscar alternativas para la obtención de indicadores no disponibles al momento de ejecución del presente estudio.

BIBLIOGRAFÍA

Apotsos, A. (2019). Mapping relative social vulnerability in six mostly urban municipalities in South Africa. *Applied Geography*, 105, 86-101.

Cardona, O. D. (2001). *Estimación holística del riesgo sísmico utilizando sistemas dinámicos complejos*. Tesis Doctoral, Universidad Politécnica de Catalunya.

Cardona, O. D. (2006). Midiendo lo inmedible. Indicadores de vulnerabilidad y riesgo. *Boletín Ambiental. Instituto de estudios ambientales IDE*. Recuperado: www.bdigital.unal.edu.co/48628/1/boletin53.pdf.

Cardoso, M. M. (2017). Estudio de la vulnerabilidad socio-ambiental a través de un índice sintético/Caso de distritos bajo riesgo de inundación: Santa Fe, Recreo y Monte Vera, Provincia de Santa Fe, Argentina. *Caderno de Geografia*, 27(48), 156-183.

Césari, R. y Césari, M. (2014). La Estrategia de Lebart. Disponible en web: https://www.researchgate.net/publication/264995434_La_Estrategia_de_Lebart.

Cutter, S. L., Boruff, B. J., y Shirley, W. L. (2003). Social vulnerability to environmental hazards. *Social science quarterly*, 84(2), 242-261.

Escofier, B. y Pagès, J. (1992). Análisis factoriales simples y múltiples: objetivos, métodos e interpretación (Abascal Fernandez, E. *et al.*, trad.). *Servicio Editorial Universidad del País Vasco*. (Obra original publicada en 1990).

Fernandez, P., Mourato, S., Moreira, M., y Pereira, L. (2016). A new approach for computing a flood vulnerability index using cluster analysis. *Physics and Chemistry of the Earth*, Parts A/B/C, 94, 47-55.

Lebart, L., Morineau, A. y Piron, M. (1995). *Statistique exploratoire multidimensionnelle* (Vol. 3). Paris: Dunod.

Merello, J. (2010*). Aproximaciones al diagnóstico de la situación social de la provincia de Santa Fe mediante un análisis de clusters*. Tesina de Grado, Facultad de Ciencias Económicas y Estadística, Universidad Nacional de Rosario.

Pardo, C. E. y Del Campo, P. C. (2007). Combinación de métodos factoriales y de análisis de conglomerados en R: el paquete FactoClass. *Revista colombiana de estadística*, 30(2), 231-245.

Peña, D. (2002). Análisis de datos multivariantes. Mc Graw Hill Interamericana de España, SAV.

Renda, E., Garay, M. R., Moscardini, O. y Torchia, N. P. (2017). *Manual para la elaboración de mapas de riesgo*. Buenos Aires: Programa Naciones Unidas para el Desarrollo. Recuperado: www.mininterior.gov.ar/planificacion/pdf/AS_13662310131.pdf.

Sistema Municipal de Epidemiología. Municipalidad de Rosario (2005). Índice de Condiciones Saludables usando Sistemas de Información Geográfica en salud. *Boletín de Epidemiología*.

Wisner, B., Blaikie, P., Cannon, T. Davis, I. (2004). At risk: natural hazards, people's vulnerability and disasters. Routledge.

EXPLORACIÓN DE LAS RELACIONES ENTRE CARACTERÍSTICAS DEL CULTIVO MUNGBEAN UTILIZANDO UN MODELO LINEAL MIXTO MULTIVARIADO

Lic. Eugenia Settecase

Directora: **Mg. María Valeria Paccapelo**
Codirectora: **Mg. Cristina Cuesta**

Los programas de mejoramiento de cultivos abarcan la evaluación de posibles nuevos genotipos en experimentos llevados a cabo en múltiples locaciones y posiblemente a lo largo de varios años. En un ensayo de variedades estándar, el objetivo es identificar los genotipos con mejor desempeño para su lanzamiento comercial en términos de características de interés como lo son el rendimiento o la resistencia a enfermedades. Al ser común evaluar en paralelo más de una característica, es deseable que el análisis de las variables se realice en forma conjunta.

Con el fin de estudiar dos características en simultáneo, Ganesalingam et al. (2013) propusieron un modelo mixto lineal bivariado que incorpora diferentes estructuras de covariancia para el efecto genotipo y el error aleatorio, así como correlaciones genotípicas y residuales entre las características.

En este trabajo se desarrolla una extensión del modelo propuesto por Ganesalingam et al. (2013) al caso multivariado. El modelo permite comprender la relación entre las variables y seleccionar los genotipos con mejor desempeño a partir de la obtención de predicciones más precisas. La propuesta resulta una herramienta de análisis apropiada para otros conjuntos de datos con un número pequeño de variables y en otras áreas de estudio, además de la agronomía.

INTRODUCCIÓN

Los modelos lineales mixtos son altamente utilizados en el análisis de datos provenientes de diseños de experimentos en la agricultura debido a que permiten analizar datos con estructuras de dependencia, falta de homogeneidad de variancias y desbalances en el número de observaciones en cada unidad experimental. Estas consideraciones resultan en un modelo muy flexible respecto a la estructura de covariancias considerada.

Cuando más de una característica es evaluada simultáneamente, puede resultar de interés estudiar si estas características están correlacionadas. El análisis de más de una variable puede llevarse a cabo utilizando distintos métodos que se basan en análisis individuales de cada variable. Sin embargo, dichos enfoques no consideran la correlación entre las medidas tomadas sobre la misma unidad experimental, por lo que se necesita la aplicación de modelos multivariados que posibiliten su inclusión. En caso de que haya correlación significativa, estos modelos brindan la ventaja de obtener mejores predicciones respecto de las obtenidas bajo modelos univariados.

Ganesalingam *et al*. (2013) desarrollaron un modelo lineal mixto bivariado aplicado en el análisis de un ensayo de campo para evaluar una enfermedad en múltiples genotipos del cultivo canola. Las variables respuesta fueron el número de plantas que emergieron y el número de plantas que llegaron a la madurez. El método contempla el ajuste de un modelo bivariado mediante el análisis de ambas variables en conjunto, utilizando los datos de cada parcela del experimento. El modelo desarrollado por Ganesalingam *et al*., permite especificar diferentes variancias residuales y genotípicas para cada variable. Además, tiene la ventaja de que modela la correlación entre las variables a nivel genotípico y residual, dando como resultado un modelo bivariado que incrementa la exactitud de las predicciones de los genotipos en comparación con los modelos univariados o de variables no correlacionadas.

Los programas de mejoramiento de cultivos tienen como objetivo principal identificar potenciales nuevas variedades con desempeño superior para su lanzamiento al mercado. Los genotipos con mejor desempeño para múltiples variables de interés se identifican mediante la evaluación de varios genotipos en un rango de ambientes distintos. La selección de los genotipos con mejor desempeño usualmente se basa en la medición de múltiples variables o características sobre la misma unidad experimental.

El presente trabajo busca extender el modelo lineal mixto bivariado al caso multivariado, y comparar las predicciones con los resultados provenientes de modelos univariados. La aplicación de este modelo se realizó sobre datos provenientes del Programa Nacional de Mejoramiento de *Mungbean* (NMIP del nombre en inglés *National Mungbean Improvement Program*) de Australia, con el fin de entender las relaciones genotípicas entre las variables estudiadas y seleccionar los genotipos con mejor desempeño.

OBJETIVOS

- Extender el modelo lineal mixto bivariado propuesto por Ganesalingam *et al*. (2013) a un modelo multivariado.
- Ajustar un modelo lineal mixto multivariado con variables respuestas correlacionadas, a datos provenientes del NMIP (Australia) con el fin de entender las relaciones genotípicas entre las variables estudiadas.
- Comparar las predicciones que se obtienen bajo los modelos multivariados con respuestas correlacionadas y no correlacionadas.

MATERIALES

Los datos utilizados en este estudio pertenecen al ensayo de una población diversa de genotipos de *mungbeans* (porotos *mung*), pertenecientes al NMIP, donde se seleccionó un conjunto de 25 genotipos. El experimento consistió en un arreglo rectangular de parcelas en el campo siguiendo un diseño en bloques completos aleatorizados con dos repeticiones. El experimento se llevó a cabo durante el periodo de verano del año 2018 en la estación experimental Hermitage, situada en la localidad de Warwick, en el estado de Queensland.

Para cada parcela se seleccionan múltiples chauchas y se ajusta un modelo sobre tres características que, en base a la bibliografía, se cree que se encuentran relacionadas al rendimiento del cultivo. Dichas variables son: el largo de la chaucha, el número de semillas dentro de cada chaucha y el peso de estas semillas.

En el análisis se utilizó el *software* R (versión 3.6), el paquete para el ajuste de los modelos es ASReml-R (versión 3). La licencia del paquete fue provista por el Departamento de Agricultura y Pesca (*DAF*), de Queensland, Australia.

METODOLOGÍA

Extensión del modelo mixto al caso multivariado

El modelo mixto bivariado presentado por Ganesalingam en 2013 propone el análisis de un vector de respuestas bivariadas modeladas a través de un modelo mixto. En situaciones empíricas, puede ser necesario considerar tres o más variables respuestas. Se propone la extensión del modelo lineal mixto bivariado a uno con respuesta multivariada, de la forma:

$$y = X\tau + Z_g u_g + Z_b u_b + Z_p u_p + e,$$

donde y es el vector de observaciones de las variables respuestas.

La matriz de incidencia X corresponde al vector τ de efectos fijos, que para este estudio sólo incluye una media general para cada característica, ya que no hay otros efectos fijos considerados en el modelo.

Los efectos genotípicos se asumen aleatorios y están contenidos en el vector u_g cuya matriz de incidencia es Z_g. El vector u_b está asociado a la parte aleatoria del modelo referido a los efectos de bloque con matriz de incidencia Z_b. Mientras el vector u_p está asociado a los efectos aleatorios de la parcela con matriz de incidencia Z_p.

Se asume que los efectos aleatorios son independientes entre sí y que cada efecto sigue una distribución normal con su respectiva matriz de variancias y covariancias, es decir, $u_g \sim N(0, G_g)$, $u_b \sim N(0, G_b)$, $u_p \sim N(0, G_p)$.

El supuesto que se establece sobre los errores es que estos siguen una distribución normal multivariada con vector de medias 0 y matriz de variancias y covariancias R.

Particularmente es de interés que el modelo incorpore las correlaciones existentes entre las variables respuesta, tanto a nivel genotípico como residual. Por este motivo, bajo el modelo multivariado se evalúan distintas estructuras de covariancia para las matrices del efecto genotípico y los errores. Mientras que las matrices de variancias y covariancias asociadas a los efectos aleatorios de parcela y bloque conservan una estructura diagonal.

Estadística r

Una de las cualidades más destacada del modelo que se presenta, especialmente en el área de mejoramiento, es el incremento en la medida "*accuracy*" de las predicciones de los efectos aleatorios con respecto a las predicciones obtenidas bajo los modelos univariados o con variables no correlacionadas. Para poder medir estos incrementos es que se presenta la medida definida por Mrode y Thompson (2005), que será denominada en esta tesina como estadística r,

$$r_{i_w} = \sqrt{1 - \frac{sep^2_{i_w}}{\sigma^2_{g_w}}},$$

siendo $sep^2_{i_w}$ la variancia asociada a la predicción del i-ésimo genotipo en la w-ésimavariable y $\sigma^2_{g_w}$ la componente de variancia genotípica asociada a la *w-ésima* variable.

A partir de los valores de r_{i_w} se puede calcular el porcentaje de ganancia en r que se obtiene para el *i-ésimo* genotipo en la *w-ésima* variable, bajo un modelo multivariado con estructura diagonal respecto de un modelo multivariado con matriz de covariancias no estructurada (para R y G_g). Esta medida es de la forma:

$$\left(\frac{r_{i_{w_{modelo\ multivariado}}} - r_{i_{w_{modelo\ univariado}}}}{r_{i_{w_{modelo\ univariado}}}} \right). 100.$$

RESULTADOS

Para la selección del mejor modelo, se comparan las medidas de la prueba de verosimilitud restringida y el criterio de Akaike (Tabla 1). Los modelos comparados difieren en la estructura de la matriz de variancias y covariancias, para la cual se consideraron estructuras diagonales (DIAG) y/o sin estructura (US) para la modelización de cada efecto aleatorio.

El modelo elegido para representar a los datos es el número 4, que considera matrices de covariancias no estructuradas para G_g y R, mientras asume matrices diagonales para G_b y G_p.

Las pruebas de verosimilitud restringida resultan ser significativas al comparar este modelo con los restantes, además presenta el menor valor de AIC.

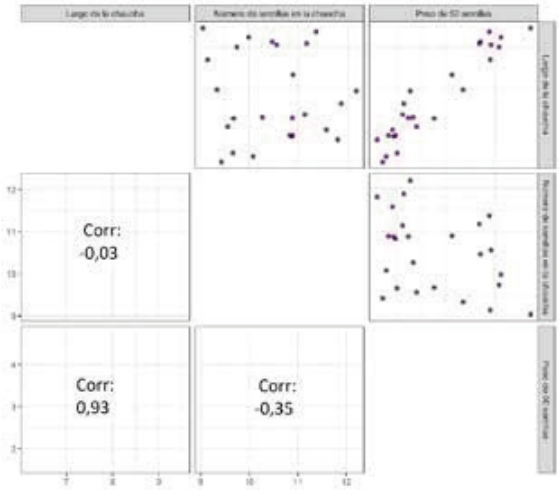
A partir del ajuste del modelo 4 se calculan las predicciones para cada variable bajo estudio, y se obtienen las correlaciones genotípicas entre las variables (Figura 1). A partir de estos resultados se observan las relaciones existentes entre las variables a nivel genotípico.

En el caso de las variables peso de 50 semillas y largo de la chaucha, su relación es positiva y alta (0,93). Luego existe una relación moderada negativa entre el peso de 50 semillas y el número de semillas por chaucha (-0,35). Mientras que parece no existir relación lineal entre largo de la chaucha y el número de semillas por chaucha (-0,03).

Tabla 1. Comparación del ajuste de los modelos multivariados con estructuras de covariancia diagonal (DIAG) y no estructurada (US) para R y G_g

	Estructura de Covariancia				Número de parámetros	$-\ log(REML)$	AIC
Modelo	G_g	G_b	G_p	R			
1	DIAG	DIAG	DIAG	DIAG	12	-203,65	429,30
2	DIAG	DIAG	DIAG	US	15	-115,73	259,46
3	US	DIAG	DIAG	DIAG	15	-179,47	386,94
4	US	DIAG	DIAG	US	18	-91,58	217,17

Figura 1. Predicciones genotípicas bajo el modelo multivariado 4 para cada variable de interés (diagonal superior). Correlaciones genotípicas entre las variables bajo estudio (diagonal inferior)



El modelo elegido nos permite además analizar las correlaciones residuales entre las variables por separado. Cabe destacar que las correlaciones residuales entre el peso de 50 semillas y las dos variables restantes, el largo de la chaucha y el número de semillas por chaucha, son débiles (0,19 y 0,20 respectivamente). Mientras que la correlación entre estas dos últimas variables es moderada (0,59).

Como fue mencionado, una de las ventajas del uso de un modelo multivariado sobre los modelos univariados es la mejora en precisión de las predicciones. Por lo que es de interés evaluar cuál es la ganancia promedio, para cada variable respuesta, en base a la estadística *r*, cuando se ajusta un modelo multivariado. Por lo tanto, se presentan las ganancias obtenidas al comparar el modelo 4 con el modelo 1, siendo este último equivalente a ajustar modelos univariados para cada variable (Tabla 2).

La ganancia promedio llega a ser de hasta el 6,5%, la cual se obtiene para la variable número de semillas por chaucha, característica que presentó mayor variabilidad residual. A contrapunto, la variable peso de 50 semillas presenta la menor ganancia según r de 0,1%, esta variable fue la que presentó menor variabilidad residual en el modelo.

Tabla 2.Porcentaje de la variancia explicada por los efectos genotípicos, estructurales y residuales; ganancia en exactitud promedio para cada una de las variables relacionadas bajo estudio; y errores estándar de predicción en promedio (*sep*) bajo los modelos 1 y 4

		Largo de la chaucha	Número de semillas por chaucha	Peso de 50 semillas
Variabilidad atribuible al/a los efecto/s (%)	Genotípico	65,80	28,60	89,40
	Residual	17,90	57,50	7,70
	Estructurales	16,30	13,90	2,90
Ganancia en exactitud (%)		3,30	6,50	0,10
sep	Modelo 1	0,43	0,61	0,14
	Modelo 4	0,34	0,57	0,13

Al comparar los errores estándar de predicción en promedio bajo cada modelo (Tabla 2), se nota que estos disminuyen moderadamente para todas las variables bajo el ajuste del modelo 4. La variable que presenta mayor incremento en precisión es el largo de la chaucha (de 0,43 a 0,34), seguida del número de semillas por chaucha (de 0,61 a 0,57). Mientras que, para el peso de 50 semillas, la precisión no presenta una mejora notable (de 0,14 a 0,13).

CONSIDERACIONES FINALES

El modelo lineal mixto bivariado de Ganesalingam *et al* (2013) fue extendido al caso multivariado. Este modelo proporciona con éxito la predicción de los efectos genotípicos de forma individual para cada variable, teniendo en cuenta las correlaciones entre las características genotípicas y residuales mediante el uso de estructuras de covariancias apropiadas.

La ganancia respecto a la estadística *r* para el modelo multivariado frente al análisis univariado, presenta resultados contrastantes para el conjunto de datos de *mungbeans*. Esto es una ventaja para el programa, ya que le permite al mejorador entender los genotipos bajo estudio, dándole una base para que, en futuros ensayos, pueda discernir cuáles son los genotipos que se comportan de diferente manera para utilizarlos en una nueva cruza.

Aunque es necesario seguir investigando, este estudio refleja que las principales ganancias por parte del análisis multivariado se producen para las características que presentan menor variabilidad genotípica cuando se modela adecuadamente la variabilidad residual. Al mismo tiempo, el modelo propuesto posibilita un mejor entendimiento de las relaciones entre las variables a nivel genotípico y residual.

Si bien la aplicación realizada en el presente trabajo considera tres variables respuesta, este modelo puede ser aplicado sobre un número mayor de variables. Sin embargo, si se imponen matrices de covariancia no estructuradas para un modelo con un gran número de variables respuesta, pueden presentarse problemas debido a la demanda computacional.

BIBLIOGRAFÍA

Ganesalingam, A., Smith, A. B., Thompson, R., Cullis, B. R. (2013). A bivariate mixed model approach for the analysis of plant survival data. 371-383.

Mrode R., Thompson R. (2005). *Linear models for the prediction of animal breeding values, 2nd edn.* CABI Publishing, Wallingford.

Tao, Y., Mace, E., George-Jaeggli, B., Hunt, C., Cruickshank, A., Henzell, R., Jordan, D. (2018). Novel Grain Weight Loci Revealed in a Cross between Cultivated and Wild Sorghum. *The Plant Genome*.

Welham, S., Cullis, B., Gogel, B., Gilmour, A., Thompson, R. (2004). Prediction in linear mixed models . *Australian & New Zealand Journal of Statistics*, 46(3), 325 - 347.

ÍNDICES DE CAPACIDAD DE PROCESOS BAJO DISTRIBUCIONES NO NORMALES

Lic. Lucio Tinnirello

Directora: Dra. Daniela Dianda

Los índices de capacidad de procesos (PCI) son métricas utilizadas mundialmente con el propósito de cuantificar el comportamiento del proceso en relación a sus especificaciones. Si bien hoy en día estos PCI son muy utilizados, un requisito esencial para los índices de capacidad tradicionales o mayormente utilizados en la industria, es que la característica de calidad analizada pueda ser adecuadamente modelada con una distribución normal. En este trabajo, se profundiza en el estudio y aplicación de metodologías para llevar a cabo el análisis de capacidad cuando este supuesto no se cumple.
Se exponen metodológicamente cinco de los métodos mayormente tratados en la literatura para este propósito, los cuales son además evaluados mediante un estudio por simulación, considerando diversos escenarios no normales, con el objetivo de identificar sus ventajas y desventajas según la distribución con la que se trabaje. Se destacan además los errores que pueden cometerse al ignorar la falta de normalidad en los datos y utilizar índices de capacidad tradicionales para evaluar la capacidad del proceso, práctica usual en la industria de nuestro medio, que puede llevar a conclusiones equivocadas con la consecuente toma de decisiones incorrectas sobre los procesos.

INTRODUCCIÓN

Hoy en día existe una intensa competencia en el plano nacional e internacional entre las organizaciones empresariales, la cual obliga a las mismas a fabricar productos sin defectos. Para lograr este objetivo las grandes empresas comenzaron a adoptar diferentes estrategias de mejora continua de la calidad, como la implantación de Sistemas de Gestión de la Calidad Total y la adaptación a los lineamientos sugeridos por el reconocido procedimiento Seis Sigma. Una parte esencial de esta filosofía de mejora continua, requiere el monitoreo del desempeño de los procesos individuales en función de los requerimientos que el mercado impone para ellos. Las métricas utilizadas mundialmente para este propósito son los conocidos Índices de Capacidad del Proceso (PCI), los que, esencialmente, cuantifican el comportamiento del proceso en relación a sus especificaciones. En el campo aplicado, es muy usual que sólo se consideren las alternativas de índices de capacidad desarrollados bajo modelos gaussianos, es decir, aquellos que asumen que la variable de calidad puede modelarse adecuadamente por una distribución normal. Como ya se mencionó, los PCIs intentan resumir el rendimiento del proceso, por lo tanto, son función de la distribución del mismo y de las especificaciones que para éste se hayan definido. Por su construcción, dichos índices pueden verse seriamente afectados si el supuesto de normalidad no es válido, pudiendo llevar a conclusiones completamente erróneas respecto del estado de capacidad real del proceso. La presencia de valores atípicos en los datos, de asimetría en la distribución de los registros, de correlación entre las observaciones, entre muchos otros, son factores que usualmente suceden en las situaciones prácticas reales, e ignorar dichos problemas puede llevar a que el proceso parezca mejor o peor de lo que realmente es. En tales situaciones, es extremadamente importante ser capaces de reconocer estas violaciones y actuar en consecuencia, considerando métodos alternativos que no hagan uso de tal supuesto. Los denominados índices de capacidad robustos, buscan suavizar o bien eliminar las distorsiones que ocasione el trabajar con distribuciones no normales, asimétricas o con una correlación en el proceso. En este trabajo se profundiza en el estudio y aplicación de las diferentes propuestas que han ido surgiendo para llevar a cabo un análisis de capacidad de procesos bajo distribuciones no normales.

METODOLOGÍA

La importancia de los estudios de capacidad de procesos deriva de la filosofía de adecuación a los requerimientos del cliente presentes en todo proyecto de calidad. La calibración de un proceso será satisfactoria, siempre que los productos respondan a los estándares fijados, ya sea por consideraciones técnicas, como por exigencias para satisfacer al mercado.

Entendiendo como proceso al conjunto de actividades, tareas, decisiones, etc. que se combinan para obtener un resultado determinado, es claro y natural que las características de dicho resultado estarán sujetas a una cierta variabilidad aleatoria, de modo que no siempre se logrará el nivel ideal deseado para cada una de ellas. De allí, la necesidad de controlar, o al menos cuantificar, la proporción de veces que el proceso dará resultados indeseados. El análisis de capacidad permite estudiar el comportamiento del proceso en relación a los requerimientos establecidos para las características de calidad de interés en el resultado y resumir dicho comportamiento en medidas conocidas como índices de capacidad.

Los primeros intentos de cuantificar la capacidad de un proceso estuvieron inmersos en un contexto bastante simplificado, en el que se consideraba que el proceso bajo estudio poseía una única característica de calidad de interés, y más aún, que el comportamiento de dicha característica podía ser modelado adecuadamente por una distribución normal. Esto dio origen al desarrollo de los índices de capacidad univariados bajo distribución normal.

Estos índices fueron en su mayoría definidos para proveer una medida relativa de la magnitud de la variación total del proceso respecto de la variación permitida por especificación, algunos de ellos teniendo en cuenta además el centrado del proceso. Dos de los primeros índices univariados propuestos, y de los más habitualmente utilizados, son el índice C_p y el índice C_{pk} (Kane, 1986).

Si se asume que la variable de calidad de interés X, puede ser adecuadamente modelada por una distribución normal de media μ y desviación estándar σ, y que para ella se tienen requerimientos dados por un valor objetivo (T) y límites inferior y superior de especificación $(L,\ U)$, estos índices se definen como:

$$C_p = \frac{U-L}{6\sigma}$$

$$C_{pk} = min\left\{\frac{U-\mu}{3\sigma}, \frac{\mu-L}{3\sigma}\right\} = min\{C_{pu},\ C_{pl}\}$$

Claramente, valores superiores a la unidad en ambos índices señalan procesos capaces, siendo el primero un índice potencial, ya que solo tiene en cuenta la dispersión del proceso, mientras que el segundo incorpora información del centrado, conformando una medida de capacidad real.

Uno de los pilares fundamentales para la construcción de los índices de capacidad tradicionales, es el supuesto de normalidad de la variable de calidad analizada, ya que, en ese caso, la cantidad 6σ es representativa de la tolerancia natural del proceso, es decir, de la amplitud del rango de operación habitual del mismo. Sin embargo, es usual en muchas situaciones prácticas, encontrar variables cuyas distribuciones distan del modelo normal. Bajo estas situaciones, las propiedades de las medidas de capacidad tradicionales podrían verse afectadas. Por ello, es necesario contar con métodos alternativos que permitan realizar una evaluación correcta, aun cuando las variables bajo estudio no respondan al supuesto de normalidad tradicional. A continuación, se presentan algunas de las alternativas de medidas robustas frente a la falla de este supuesto distribucional, propuestas en la literatura.

Método de Clements

Se desarrolló bajo el supuesto de que la distribución de la variable de interés puede representarse adecuadamente por una distribución Pearsoniana (Clements, 1989), es decir, una distribución de la familia de distribuciones cuyas densidades satisfacen la ecuación diferencial dada por:

$$\frac{d\,f(x)}{dx} = \frac{(x-a)f(x)}{dx^2+cx+b}$$

Los parámetros a, b, c y d determinan el tipo de distribución de Pearson al que corresponde $f(x)$ y sus valores dependen de la media, desviación estándar, asimetría y kurtosis de los datos (Lahcene, 2013). La idea básica de este método es reemplazar en la ecuación del índice tradicional, el valor 6σ por la longitud del intervalo entre los puntos percentiles 0.135 (L_p) y 99.865 (U_p) de la distribución de Pearson que se haya identificado para X, esto es, se reemplaza el valor 6σ por la cantidad $U_p - L_p$. Así, el índice de Clements alternativo al C_p tradicional se define de la siguiente manera:

$$CC_p = \frac{U-L}{U_p-L_p}$$

De la misma manera, para obtener una alternativa robusta del índice tradicional C_{pk}, Clements propone estimar la media del proceso μ por la mediana *Mna* de la distribución de Pearson identificada y reemplazar los valores de 3σ por $Mna - L_p$ y $U_p - Mna$, en C_{pl} y C_{pu}, respectivamente. Los percentiles para una variedad de distribuciones Pearsonianas estandarizadas, se encuentran tabulados y pueden ser identificados en cada caso particular a partir de la asimetría y la kurtosis de los datos muestrales (Gruska *et al.*, 1989).

Método de Burr

Este método es similar al de Clements, pero se propone la utilización de la familia de distribuciones de Burr XII, en lugar de considerar el sistema de distribuciones de Pearson que utiliza Clements. El sistema de distribuciones de Burr (Burr, 1942) incluye a las distribuciones continuas, cuya función de distribución responda a la ecuación diferencial $F'(x) = F(x)(1 - F(x))g(x)$, donde $g(x)$ es alguna función no negativa.

Para este método, el mismo autor tabuló los valores necesarios para identificar la distribución de Burr adecuada y hallar los percentiles requeridos para el cálculo de los índices de capacidad, nuevamente como función de la asimetría y kurtosis de los datos muestrales (Burr, 1973). Finalmente, para la estimación de los índices de capacidad, la propuesta de estos autores es similar a la de Clements, pero utilizando los percentiles obtenidos a través de la distribución de Burr XII que corresponda en cada caso.

Método de Wright

El índice tradicional C_{pmk} (Pearn *et al.*, 1992) está diseñado para advertir cuando el proceso presenta cambios tanto en la media como en la variancia. Sin embargo, no tiene en cuenta alteraciones en la distribución de los datos, por lo que podría originar inconvenientes cuando se lo aplica en procesos propensos a la asimetría. Peter Wright (1995) presentó un índice, basado en C_{pmk}, que incorpora una corrección por la asimetría, utilizando el momento centrado de orden 3, $\mu_3 = E(X - \mu)^3$, como medida de asimetría. El índice propuesto se define como:

$$C_s = \frac{D - |\mu - T|}{3\left[\sigma^2 + (\mu - T)^2 + \left|\frac{\mu_3}{\sigma}\right|\right]^{1/2}}$$

De la definición del índice C_s, se deduce que éste siempre arrojará un valor menor al del C_{pmk}, o a lo sumo igual en el caso de distribuciones simétricas. El índice C_s conserva varias de las cualidades deseables del índice C_{pmk}, como, por ejemplo, tiene en cuenta que la media del proceso podría no coincidir con el punto medio del intervalo de especificación, a la vez que penaliza las desviaciones de la media respecto del valor objetivo T. Pero, además, posee la ventaja adicional de incorporar información sobre la forma de la distribución, a través de la penalización por asimetría adicionada en el denominador.

Método de la variancia ponderada

Chang, Choi y Bai (2002) fundamentaron la idea de que el desvío estándar de la población (σ) se puede dividir en dos desviaciones σ_U^W y σ_L^W que representan los grados de dispersión de los datos por encima y por debajo de la media μ respectivamente, y cuya suma es igual a la desviación original. Así, una función de densidad de probabilidad asimétrica se puede aproximar mediante dos funciones de densidad de probabilidad normales, ambas con la misma

media μ pero diferentes desviaciones estándares, $2\sigma_U^W$ y $2\sigma_L^W$. Los autores demuestran que estas variancias ponderadas σ_U^W y σ_L^W necesarias para construir las densidades superior e inferior se obtienen como $\sigma_U^W = p_x\sigma$ y $\sigma_L^W = \left(1 - p_x\right)\sigma$, donde $p_x = Pr[X \leq \mu]$.

Los autores retoman esta idea, inicialmente diseñada para proponer gráficos de control para distribuciones asimétricas, para definir también medidas de capacidad. Los índices C_p y C_{pk} basados en el método de la variancia ponderada se definen como:

$$C_p^{WSD} = min\left\{\frac{U-L}{6*2\sigma_U^W}, \frac{U-L}{6*2\sigma_L^W}\right\} = \frac{U-L}{6\sigma} min\left\{\frac{1}{2p_x}, \frac{1}{2\left(1-p_x\right)}\right\}$$

En la fórmula de C_p^{WSD}, se utiliza $2\sigma_U^W$ y $2\sigma_L^W$ en lugar de σ para tener en cuenta el grado de asimetría. Si la distribución subyacente es simétrica ($p_x = 0.5$), $C_p^{WSD} = C_p$, pero si existe asimetría, entonces $C_p^{WSD} < C_p$.

Por otro lado, el índice C_{pk} bajo este método se define como: $C_{pk}^{WSD} = \left\{C_{pku}^{WSD}, C_{pkl}^{WSD}\right\}$.

Método de las transformaciones

La última alternativa que se presenta está relacionada con uno de los métodos más usuales en diversas técnicas de análisis estadístico, frente al incumplimiento del supuesto de normalidad: la aplicación de transformaciones normalizadoras. La idea principal es la transformación de la variable dependiente mediante alguna función paramétrica que logre normalizar el comportamiento de la misma, es decir, que la distribución empírica de los datos observados transformados puede adaptarse correctamente a un modelo gaussiano.

Existen dos métodos ampliamente reconocidos para este propósito. El método de Box y Cox (Box y Cox, 1964), en el que se propone una familia de transformaciones de tipo potencia sobre la variable de interés Y, especificada por un parámetro λ que define la transformación. Específicamente, la propuesta de los autores consiste en transformar los valores observados y de la variable de interés a valores nuevos $y^{(\lambda)}$ mediante la función:

$$y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda} (\lambda \neq 0) \\ log\, y \ \ (\lambda = 0) \end{cases}$$

donde el valor del parámetro λ se determina de modo que los datos transformados $y^{(\lambda)}$ se distribuyan aproximadamente normales.

El otro método de transformación es el propuesto por Johnson (Johnson, 1949), quien desarrolló un sistema de distribuciones basado en el método de los momentos, similar al introducido por Pearson. La forma general de la transformación que propone este autor viene dada por:

$$z = \gamma + \eta\tau(x; \varepsilon, \lambda);\ \eta > 0;\ \ -\infty < \gamma < \infty;\ \lambda > 0;\ \ -\infty < \varepsilon < \infty$$

donde z es una variable aleatoria con distribución normal estándar y x es la variable de interés que se pretende transformar. Los cuatro parámetros γ, η, ε y λ deben ser estimados y τ es una función arbitraria, también a determinar, que puede tomar una de la siguientes tres formas: sistema lognormal (S_L): $\tau_1(x, \varepsilon, \lambda) = log(x - \varepsilon),\ x \geq \varepsilon$; sistema no acotado ($S_U$):

$\tau_2(x;\varepsilon,\lambda) = sinh^{-1}\left(\frac{x-\varepsilon}{\lambda}\right)$, $-\infty < x < \infty$; sistema acotado (S_B):

$\tau_3(x;\varepsilon,\lambda) = log\left(\frac{x-\varepsilon}{\lambda+\varepsilon-x}\right)$, $\varepsilon \leq x \leq \varepsilon + \lambda$.

En cualquiera de los dos métodos, una vez que los datos han sido transformados, se realiza el análisis de capacidad utilizando los índices tradicionales, definidos bajo normalidad, pero aplicados sobre los datos transformados.

APLICACIÓN

A pesar de existir una amplia variedad de propuestas para la construcción de índices de capacidad en situaciones de no normalidad, son escasos los antecedentes de estudios comparativos entre las mismas, y, por consiguiente, no existen lineamientos o criterios unificados que guíen en la selección de una u otra alternativa frente a un caso práctico particular. Por tal motivo, se realizó un estudio comparativo que permita evaluar el comportamiento de cada método frente a diferentes situaciones de no normalidad.

Diseño del estudio comparativo

Es importante mencionar que un índice de capacidad representativo para datos no normales debería ser compatible con los índices calculados bajo normalidad, en el sentido de que, a valores similares de los índices, las proporciones de productos fuera de especificación en los respectivos procesos, sean también similares. Por lo tanto, resulta conveniente utilizar, el índice de capacidad unilateral C_{pu}, para llevar a cabo las comparaciones de los métodos en el estudio de simulación, ya que para este índice es relativamente sencillo hacer la correspondencia entre su valor y la proporción de productos fuera de especificación, cualquiera sea la distribución de los datos.

Teniendo en cuenta esto, el esquema general del estudio comparativo consiste en definir valores de referencia para el índice C_{pu} y estimar dicho índice a partir de muestras aleatorias de una distribución dada, empleando los diferentes métodos considerados en este trabajo, incluyendo el método tradicional definido bajo normalidad. El objetivo es identificar el o los métodos que arrojan estimaciones del índice lo más cercanas posible al valor de referencia y con la menor dispersión.

Para realizar las comparaciones, se diseñó un procedimiento de simulación en el que se tuvieron en cuenta los siguientes factores: el tamaño de muestra n (n =50, 100 y 200), los valores de referencia de C_{pu} ($C_{pu} = 1,\ 1.5\ y\ 2$), la distribución de la variable de calidad, considerando los modelos Beta, Gamma, Weibull y Lognormal. Estas distribuciones tienen parámetros cuyos valores pueden generar densidades con desviaciones de la normalidad de diferente magnitud, y los comportamientos en sus colas pueden influir en la capacidad del proceso. A su vez, son modelos que suelen representar adecuadamente variables reales en el contexto de procesos industriales.

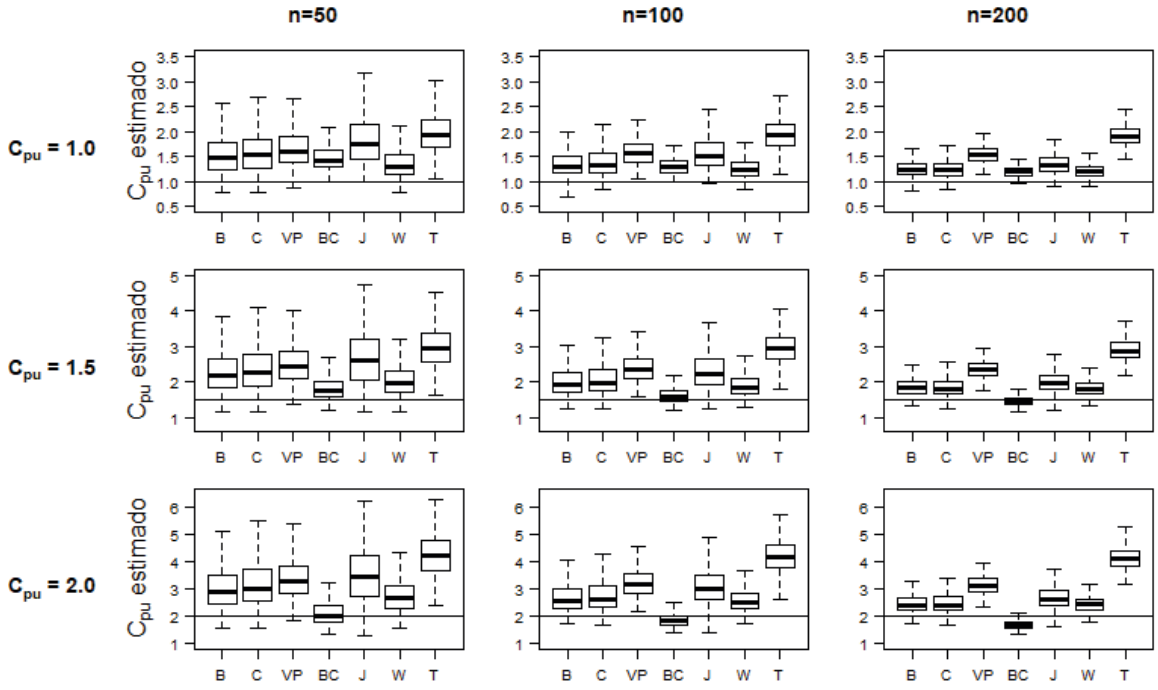
El procedimiento de simulación consistió en generar, para cada escenario, una muestra aleatoria de valores de la distribución especificada, estimar el índice de capacidad C_{pu} con cada uno de los métodos estudiados, repetir el proceso 500 veces y obtener el promedio, mediana y desvío estándar de los índices de capacidad, en cada combinación de tamaño de muestra y valor de referencia de C_{pu}.

RESULTADOS

El objetivo del estudio comparativo es determinar el o los métodos que muestren un mejor comportamiento, en el sentido de brindar estimaciones que en promedio coincidan con el valor de referencia y que, a su vez, tengan la menor dispersión. Por cuestiones de espacio, se presentan los resultados detallados para sólo una de las distribuciones consideradas en el estudio. Para el resto de los casos se hace mención a los hallazgos más importantes.

La Figura 1 muestra los boxplots para las estimaciones del índice con cada uno de los métodos, como así también para el índice tradicional, en cada combinación de tamaño de muestra y valor objetivo del C_{pu}, considerando datos provenientes de una distribución Gamma, con parámetros de forma y escala ambos iguales a 1.

Figura 1. Diagramas de caja para las estimaciones de C_{pu} bajo cada método, para datos provenientes de una distribución $\gamma(r = 1, \mu = 1)$



Referencias: B: Burr, C: Clements, VP: Variancia ponderada, BC: Box-Cox, J: Johnson, W: Wright, T: tradicional.

Para esta situación, los métodos de Wright y Box y Cox son los que ofrecen estimadores menos dispersos y también los que más se acercan a los valores de referencia, a través de los diferentes escenarios. Las alternativas de Burr y Clements también se acercan en promedio a estos valores objetivo, pero tienen desvíos estándares mayores para tamaños de muestra pequeños. Entre los dos métodos destacados, Wright y Box y Cox, la elección de uno de ellos no es directa. Para tamaños de muestra grandes, el método de Wright genera una pequeña sobreestimación, mientras que el de Box y Cox subestima la capacidad real, sobre todo para procesos muy capaces ($C_{pu} = 2$). Para finalizar, se observa que el método tradicional no proporciona buenos resultados, arrojando valores del índice por encima del valor de referencia, sobreestimando la verdadera capacidad del proceso.

El mismo análisis replicado a través de las distribuciones consideradas, permitió obtener un panorama general del comportamiento de las diferentes propuestas y delinear algunas consideraciones para su utilización en situaciones prácticas.

Los resultados mostraron que, dependiendo del tipo de distribución que se tenga, los métodos que mejor se comportan pueden variar. Así, por ejemplo, en la distribución Beta, si bien el índice tradicional es el que mejor se ajusta a los valores esperados, otros métodos como el de Burr y Clements, o así mismo el de Variancia Ponderada, arrojaron buenos resultados. Para el caso de las distribuciones gamma, se observó una marcada diferencia entre el método tradicional y los alternativos, siendo el método de Wright uno de los que mejor se comportó para la variante más asimétrica $(r = 1, \mu = 1)$, y para la variante con asimetría más leve, se suman los métodos de Box y Cox e incluso Burr y Clements, con un comportamiento aceptable. Siguiendo con la distribución Lognormal, los métodos de Burr y Clements presentaron resultados similares, ambos se acercan a los valores esperados, como así también tienen los menores desvíos estándares. Estos mismos métodos mostraron buen comportamiento también para la distribución Weibull, sumándose en este caso el método de Wright, probablemente por la marcada asimetría de la distribución.

CONCLUSIÓN

Ante la gran variedad de alternativas disponibles para el estudio de capacidad de procesos no normales, surge el interrogante de cuál de ellos utilizar en una situación práctica real. La literatura ofrece algunos estudios comparativos limitados (Liu y Chen, 2006; Tang, 1999; Swamy *et al.*, 2016), por lo que en este trabajo se propuso también como objetivo realizar un estudio comparativo, considerando algunas distribuciones con diferentes grados de alejamiento de la normalidad.

Los resultados dejaron en evidencia que, si bien no hay un método superador a través de todos los casos, es posible derivar algunas reglas de comportamiento generales. Los métodos basados en identificación de la distribución subyacente de los datos funcionan de manera aceptable a través de todos los escenarios, a menos que la asimetría de la distribución sea muy marcada, en cuyo caso, el método de Wright ofrece buenos resultados. Los métodos de transformaciones resultaron adecuados sólo en algunas situaciones particulares, con resultados medios similares, pero diferentes en dispersión, siendo el estimador basado en la transformación de Johnson más variable. El método de la variancia ponderada sólo mostró buen comportamiento en el caso de la distribución beta.

Es de destacar también, que el índice tradicional, basado en el supuesto de normalidad, mostró un muy buen comportamiento en el caso de la distribución Beta, que, para los parámetros establecidos, corresponde a un modelo aproximadamente simétrico, pero empezó a fallar a medida que se consideraron modelos más asimétricos.

Queda a la vista que la elección del método dependerá en parte del tipo de distribución que muestre la variable de calidad analizada. Pero más importante aún es resaltar la necesidad de chequear adecuadamente los supuestos necesarios antes de emplear la metodología tradicional, ya que los resultados podrían verse seriamente afectados.

REFERENCIAS

Box, G. & Cox, D. (1964). An analysis of transformations, *J. Roy. Stat. Soc. B,* 26: 221-252.

Burr, I. (1942). Cumulative frequency distribution*, Ann. Math. Stat.*, 13: 215-232.

Burr, I. (1973). Parameters for a general system of distributions tomatch a grid. *Commun. Stat.*, 2:1–21.

Chang, Y., Choi, I. & Bai, D. (2002). Process capability indices for skewed populations. *Qual. Reliab. Eng. Int.*, 18: 383-393.
Clements, J. (1989). Process Capability Calculations for Non-Normal Distributions. *Qual. Prog.*, 22: 95-100.
Gruska, G., Mirkhani, K. & Lamberson, L. (1989) Non-normal data analysis. Applied Computer Solutions, Inc, St.Clair Shores, Michigan.
Johnson, N. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*. 36: 149-176.
Kane, V. (1986). Process capability indices. *J. Qual. Technol.*, 18, pp.41-52.
Lahcene, B. (2013). On Pearson families of distributions and its applications. *Afr. J. Math. Comp. Sc. Res.* 6(5): 108-117.
Liu, P. & Chen, F. (2004). Process capability analysis of non-normal process data using the Burr XII distribution. *Int. J. Adv. Manuf. Technol*., 27: 975-984.
Pearn, W., Kotz, S. & Johnson, N. (1992). Distributional and inferential properties of process control indices. *J. Qual. Technol.,* 24(4): 216-231.
Swamy, D. (2016). Process Capability Indices for Non-Normal Distribution- A Review. *International Conference on Operations Management and research.* Mysuru, India.
Tang, L. & Than, S. (1999) Computing process capability indices for nonnormal data: A review and comparative study. *Qual. Reliab. Eng. Int.,* 15: 339–353.
Wright, P. (1995). A process capability index sensitive to skewness. *J.Stat. Comput. Sim.,* 52: 195-203.

CONSEJO PROFESIONAL
DE CIENCIAS ECONOMICAS
DE LA PROVINCIA DE SANTA FE
CAMARA II

CPCE

Colegio de Graduados
en Ciencias Económicas
de Rosario

UNR
Universidad
Nacional
de Rosario