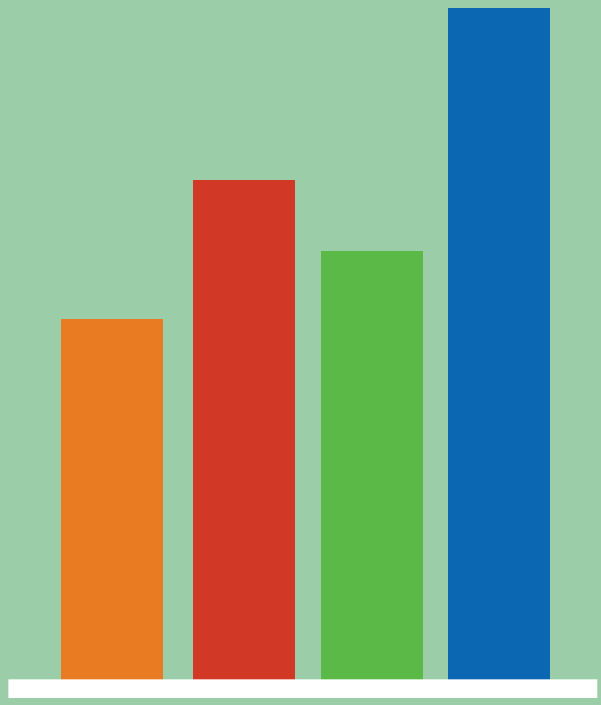
Instituto Nacional de Estadística y Geografía

Aportaciones a la estadística de los XXVII y XXVIII Foros Nacionales de Estadística



AME

INEGI
INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

Obras complementarias publicadas por el INEGI sobre el tema:

Memoria del Foro Nacional de Estadística 1993 a 2011 (Volumen del Núm. VIII al XXVI).

Catalogación en la fuente INEGI:

310.4 Foro Nacional de Estadística
Aportaciones a la estadística de los XXVII y XXVIII Foros Nacionales de Estadística / Instituto Nacional de Estadística y Geografía.-- México : INEGI, c2015.

155 p. : il.

ISBN: 978-607-739-623-9

"Universidad Autónoma del Estado de México. México, del 24 al 28 de septiembre de 2012"

"Instituto Nacional de Estadística y Geografía. Aguascalientes, Aguascalientes, del 23 al 27 de septiembre de 2013"

1. Estadística - Alocuciones, Ensayos, Conferencias. I. Instituto Nacional de Estadística y Geografía (México). II. Universidad Autónoma del Estado de México. III. Asociación Mexicana de Estadística

Conociendo México

01 800 111 46 34
www.inegi.org.mx
atencion.usuarios@inegi.org.mx

 **INEGI Informa**  **@INEGI_INFORMA**

Presentación

Este libro está constituido por 19 artículos que forman la Memoria de los Foros Nacionales de Estadística XXVII y XXVIII de manera conjunta. El primero se realizó los días del 24 al 28 de septiembre de 2012, en la Universidad Autónoma del Estado de México, con la temática “Estadística en la Industria y en los Sistemas de Información Geográfica”; y el segundo del 23 al 27 de septiembre del 2013 en el Instituto Nacional de Estadística y Geografía, con la temática "Estadística y Desarrollo”.

Como cada foro realizado por la Asociación Mexicana de Estadística (AME), la intención ha sido reunir a la comunidad de estadísticos mexicanos de todos los sectores de la sociedad, en particular, academia, gobierno e industria, así como a investigadores del resto del mundo, para facilitar intercambios profesionales, compartir y presentar los más recientes conocimientos e innovaciones en estadística.

Además en el 2013 se realizó la celebración del Año Internacional de la Estadística, por lo que el foro se realizó en el Instituto Nacional de Estadística y Geografía, creado por decreto presidencial el 25 de enero de 1983, institución en la cual se conjunto la responsabilidad de generar la información estadística y geográfica en México.

Los trabajos fueron sometidos a un proceso de arbitraje coordinado por la mesa directiva de la Asociación Mexicana de Estadística. En este proceso, todos los artículos fueron revisados en su forma y contenido; siguiendo, en todo momento, criterios mínimos para evaluar la calidad en sus propuestas, resultados y aplicaciones.

Agradecemos profundamente a todos los autores por su entusiasmo y por la calidad de los trabajos presentados. Agradecemos, además, a todos aquellos colegas que nos apoyaron participando como árbitros, ya que con su esfuerzo contribuyen a la calidad académica de estas Aportaciones. En nombre de la Asociación Mexicana de Estadística expresamos también nuestra gratitud a la Universidad Autónoma del Estado de México por el apoyo la realización del XXVII Foro y al Instituto Nacional

de Estadística y Geografía por el apoyo en la realización del XXVIII Foro y por patrocinar la edición de esta obra.

El Comité Editorial

Sergio Hernández González
Gabriel Núñez-Antonio
Ignacio Méndez Gómez Humarán

Índice general

Aportaciones del XXVII Foro Nacional de Estadística

Democracia y Estadística en México

EDMUNDO F. BERUMEN TORRES

Director General de Berumen y Asociados

dirección@berumen.com.mx, Tel. (52-55) 5093 8600

RESUMEN

El artículo hace una breve reseña de la creciente interrelación que en años y elecciones presidenciales recientes se ha establecido entre los procesos y procedimientos electorales establecidos por las Leyes, Reglamentos e Instituciones que rigen nuestra *democracia,* y el uso creciente de algunas técnicas y métodos de la *estadística,* que han apoyado la depuración de la infraestructura básica en que se sustenta cada elección (padrones electorales y listas nominales), así como las técnicas *estadísticas* para anticipar *estimaciones* de sus resultados mediante distintos tipos de encuestas y conteos rápidos.

Palabras clave

Padrón Electoral, Lista Nominal, Encuestas, Estimaciones, Resultados Oficiales

ABSTRACT

This paper gives a brief account of the increased interaction in recent years and presidential elections, between the established regulations and electoral processes dictated by relevant Laws and Regulations that administer our democracy, and the statistical tools that have played a key role in improving the quality of voter registry, election lists, as well as statistical tools to have early estimate of election results, such as exit polls, quick counts, and others.

Keywords

Voter Registry, Elegible Voters, Surveys, Poll Estimates, Official Results

Antecedentes

Todo país que se autodefine como *democrático* tiene una historia que relata porqué de origen así surgió o cómo evolucionó hasta ella alcanzar su democracia. Que yo sepa, se muy poco, ningún país hoy clasificado como democrático tiene historia idéntica a la de otro. Un denominador común que sí tienen es algún proceso que de manera periódica sirve para realizar elecciones de las autoridades del Poder Ejecutivo y Legislativo; el Poder Judicial se cuece aparte. Y son estos procesos los que hermanan *democracias* con *estadísticas*. En lo que sigue sólo me concentro en algunos incidentes recientes, últimas dos décadas, del largo camino que en México han recorrido de la mano *Democracia* y *Estadística* a la mexicana.

Aritmética. No se llega a la democracia ni a la estadística si antes no visitamos la aritmética, y aún antes a la simple actividad de *enumerar* cosas y casos. Son la esencia de instrumentos electorales para la cuenta de votos, de ciudadanos con derecho a votar, de opciones válidas que pueden ser votadas, de sumas que acumulan un resultado, y de la calificación oficial del resultado final una vez resueltas las impugnaciones y cuestionamientos sobre los mismos.

Padrones electorales y Listas Nominales. En México durante varios sexenios (periodo de 6 años que dura en el Poder Ejecutivo el candidato que gana la elección de Presidente) la discusión política resultaba monotemática: el mayor villano de toda elección según los partidos de oposición era la calidad del Padrón Electoral (PE) y posterior Lista Nominal (LN). En ellos residían muertos nunca dados de baja, migrantes fuera del país (sin derecho a votar en la época), y clones por doquier de aquellos que consistentemente votaban a favor del partido en el poder.

La diferencia entre el ciudadano empadronado y el que está incluido en las listas nominales de una elección particular, es que el primero llevó a cabo el trámite de registro, y el segundo se tomó la molestia de pasar por su credencial de elector una vez procesada. Por tanto las cifras del PE siempre son mayores a las de la LN.

Pues resulta que a mediados de los años 90's del milenio pasado se llevaron a cabo cerca de 50 "auditorías técnicas" al PE, algunas nacionales, otras regionales otras locales, cada una de las cuales daba pistas para depurar el PE y las LN. Resultado: las elecciones de 1994 (Presidente, senadores y diputados locales), se realizaron con la LN más depurada a la fecha (y quizá después de ella, a pesar de nuevas auditorías técnicas para actualizarlo). Más aún, dejó de ser el centro de discusión política las marrullerías electoreras atribuidas a la calidad del PE y LN. El cuestionamiento básico era que la LN adolecía de aberraciones como la ilustración siguiente:

$$LN = \{C_1, C_2, \ldots, C_i, \ldots, C_i, \ldots, C_i, \ldots, RIP_k, C_j, \ldots, C_j, \ldots, C_{j+1}, RIP_{j+2}, \ldots.\} \text{-------(1)}$$

Donde, C_i era un *i-ésimo ciudadano* que aparecía múltiples ocasiones en distintos listados de distintas casillas, y RIP_k era el *k-ésimo ciudadano fallecido* que nunca se dio de baja de la LN. Las auditorías técnicas depuraron el PE y LN al grado de borrarlos como tema central de la discusión política.

Sí, la estadística jugó papel de actor principal en estos ejercicios. Sobre-simplificando se tomaron muestras estrictamente probabilísticas del PE y LN y se fue a buscar en los domicilios registrados a los ciudadanos en muestra, y viceversa: se seleccionaron muestras estrictamente probabilísticas de personas de 18 años de edad cumplidos, empadronados o no según lo declaraban, y se les buscaba en el PE y LN.

Elecciones Presidenciales previas a la de *1994*. Previo a lo reseñado, las encuestas de opinión sobre intenciones de voto para la próxima elección presidencial se realizaban mucho antes del periodo que estamos cubriendo en esta nota, pero su consumo y conocimiento se restringía a una *élite* de funcionarios y políticos de primer nivel, así como algunos integrantes de cúpulas empresariales. Los ciudadanos no eran actores que merecieran conocer, mucho menos opinar al respecto.

En alguna reseña de cómo se forzó la salida del closet de los resultados de las encuestas sobre temas electorales, Miguel Basañez escribió "…el éxito de la encuesta de la elección de 1988 que por conducto de Federico Reyes Heroles me encargó La Jornada. Se abrió ahí la posibilidad que varios acariciábamos de contribuir a la democratización del país vía las encuestas. Se convertirían en martillazos numéricos para abrir la concha autoritaria. Dardos venenosos al viejo dinosaurio." Poco después nace la revista *Este País* justo con ese propósito, excelente publicación que sigue

auspiciando el tema pero enriquecido con muchos otros que abarcan ejercicios de prospectiva, ensayos temáticos de relevancia nacional, entrega de indicadores diversos y espacio para divulgar cultura.

Abierto el closet, con las acciones del IFE para depurar el PE y LN, y el buen resultado de los CR contratados por el IFE en la elección de 1994, ya nunca más retornaron las encuestas electorales a ser sólo para el consumo de élites. Los ciudadanos se empoderaron en el tema.

Elección Presidencial de 1994. Para este proceso electoral, por primera vez el IFE contrató a tres empresas para que el 21 de agosto de 1994 realizaran un ejercicio de conteo rápido (CR), todas con una muestra nacional estrictamente probabilística, del mismo tamaño, 500 secciones electorales (SE), 100 por Circunscripción, seleccionadas todas con el mismo diseño probabilístico (estratificado por Circunscripción y dentro de éstas según si la SE estaba clasificada como urbana, rural o mixta; con afijación proporcional y con igual probabilidad dentro de cada estrato), para tener una, en referencia contra la cual comparar las muchas otras que se realizarían por distintos actores interesados en el resultado del proceso.

El proceso de estimación del resultado de la elección de las tres empresas fue idéntico: recolección en campo de los resultados consignados en las actas de escrutinio y cómputo de cada casilla de cada SE en muestra, transcritas en formatos de campo para luego transmitirlas al centro de captura y validación de cada empresa, con protocolos de seguridad para minimizar riesgos de intrusos con intenciones de sembrar datos falsos. Merece la pena subrayar que este tipo de ejercicio *no entrevista a ningún tipo de informante,* se restringe a la transcripción de los datos consignados en la actas de escrutinio y cómputo de cada casilla en las SE en muestra, mismas que son exhibidas en "cartulinas" expresamente diseñadas para ello y que se hacen públicas al pegarlas al exterior de cada casilla una vez concluyen la elaboración de las actas de escrutinio y cómputo para cada elección.

La convergencia de las estimaciones de cada empresa contribuiría a la confianza en la estimación. Así sucedió, fue conocido por el Consejo del IFE, y al iniciar a salir de manera pública los resultados de terceros y ser congruentes con los contratados, el Consejo dejó que los "de fuera" tomaran el reflector para luego en hora prudente "cantar" lo estimado por, y responsabilidad de, las empresas contratadas por el propio IFE, para

luego ratificar que el miércoles siguiente iniciarían los cómputos distritales. Fue una elección con una noche y amanecer siguiente terso que dio confianza y tranquilidad al ciudadano que su voto fue contado y contó en el resultado de la elección.

Adicional a lo anterior se dieron a conocer algunas encuestas de salida (ES); instrumento cuyo diseño original fue para conocer el perfil de los votantes en una elección, pues es la única encuesta en cualquier proceso electoral que en efecto entrevista “votantes” (las previas entrevistan ciudadanos con credencial de elector vigente, que pueden o no convertirse en “votantes” el día de la jornada electoral), pero que igual puede ser un procedimiento para “de paso” estimar el resultado de la elección. Sus estimaciones están disponibles en cuanto cierran las últimas casillas y por ello son más oportunas que las estimaciones provenientes de los CR, más no más precisas, pues la calidad del dato cuyo origen es el acta de escrutinio y cómputo de cada casilla de las SE en muestra no están sujetas a la interacción entre encuestador y votante seleccionado y entrevistado, o no por rechazo. En 1994 las que se divulgaron no dieron resultados divergentes con las posteriores estimaciones provenientes de distintos CR.

Según consta en los anales históricos del IFE, fue justo en 1994 que la reforma electoral aprobada instituyó la figura de "Consejeros Ciudadanos", personalidades propuestas por las fracciones partidarias en la Cámara de Diputados y electos por el voto de las dos terceras partes de sus miembros sin considerar la profesión o título que poseyeran. Por su parte, los partidos políticos conservaron un representante con voz, pero sin voto en las decisiones del Consejo General. Ese año el Consejo General del IFE quedó organizado de la siguiente forma: Un Presidente del Consejo General (Secretario de Gobernación), seis consejeros ciudadanos, cuatro consejeros del poder legislativo, y representantes de los partidos políticos con registro.

También fue en 1994 la primera elección presidencial del IFE, donde se instauró por primera vez el Programa de Resultados Electorales Preliminares (PREP), implementado por la Dirección General del IFE, que tuvo la finalidad específica de captar los resultados del mayor número de casillas posible, de acuerdo al ritmo en que éstos llegaran a las sedes de los Consejos Distritales correspondientes. (En la elección del 6 de julio de 1988 tuvo su precedente, el Sistema de Información de Resultados Electorales, SIRE, de fatídica fama cuando el sistema se cayó y calló, dejando para siempre cuestionada la legitimidad del resultado oficial.) El PREP se basó en los resultados anotados en la primera copia de las actas de escrutinio y

cómputo de las casillas, elaborada por los funcionarios de casilla ante la presencia de los representantes de los partidos políticos. La copia del acta fue colocada por separado en un sobre llamado "sobre PREP", que el presidente de la mesa directiva de cada una de las casillas hizo llegar al Consejo Distrital. La coordinación general del PREP diseñó una red de transmisión con 300 Centros de Acopio y Transmisión de Datos (CEDAT), los cuales se instalaron en cada distrito electoral. En estos centros, la transmisión de los datos se hizo vía telefónica. Se instalaron dos Centros Nacionales de Recepción de Resultados Electorales Preliminares (CENARREP), uno principal y otro alterno. La difusión de la información al Consejo General del Instituto se realizó a través de diversos formatos como terminales computacionales, pantallas de televisión, medios magnéticos e impresos. El Programa cerró sus operaciones después de cuatro días (96 horas), y logró contabilizar aproximadamente el 92.27% de las casillas. NOTA: el PREP a diferencia de las ES y CR no es un ejercicio de estimación del resultado final, simplemente va dando cuenta de la *suma acumulada* de votos conforme se transmiten los datos de cada casilla que operó durante la elección.

El auge

Una de las características de nuestra democracia, es que no nos gusta participar en las acciones y programas del gobierno en turno en la medida que podemos para contribuir al menos en aquellos programas de bienestar general que a nadie perjudican. Por mencionar uno trivial, más no por ello irrelevante, el no tirar basura y recoger la de otros para depositarla en su lugar, ya no digamos el clasificarla. No, nada de eso, salvo escasas y bienvenidas excepciones. Lo que nos fascina, quizá por largo ayuno de centurias, es el "juego del voto".

No bien toma posesión y se conoce a los integrantes del gabinete del Presidente recién electo, 87 millones de pares de ojos de mexicanos de 18+ escudriñan caras y nombres para iniciar el juego de quién de ellos será el próximo Presidente. ¿Acaso no está en el presídium y entrará en el primer ajuste; acaso será alguno de los Gobernadores invitados al acto?

Antes de finalizar el primer año de gobierno inician las encuestas que hurgan buscando ansiosos que en el arranque se queman, discretos que se les descubren virtudes ocultas, cordialidad especial del señor Presidente hacia fulano o zutano, etc.

Inician las series de a quiénes ve bien la ciudadanía, a quiénes mal, el esperado "ranking" del gabinete, y claro, no puede quedar atrás el de Gobernadores. Coloridas gráficas de distintos medios periódicamente dan cuenta de ello, plumas floridas de distintos analistas se alinean con unos u otros, y la pobreza y hambre de millones se olvida mientras las carreras de caballos, briosos o flacos, nos entretienen.

Encuestador que no tiene medio que lo auspicie, o Ministro de los que se sienten con posibilidades, o Comité Ejecutivo Nacional de algún Partido, o la Oficina de la Presidencia, o Gobernadores con recursos, o Empresario con intereses, o Cúpulas de poder, o la Secretaría de Gobernación, etc., casi es un paria en su gremio.

El cómo va la gestión del Presidente y su Gobierno, que igual se mide y exhibe por múltiples encuestas, son anécdotas colaterales que de inmediato se correlacionan con a quién favorecen, disminuyen o de plano destruyen en sus aspiraciones presidenciales.

Métodos

Como antaño en botica (hoy día farmacia), hay para escoger. Y bien que lo haya, pues casi desde su origen se reconoció que las encuestas más interesantes, las que miden hechos, opiniones y percepciones de cualquier sociedad, son tanto una ciencia como un arte. Así nos topamos con sondeos por cuotas levantados en centros de afluencia, otros en viviendas con distintos procedimientos de sustitución para cumplir cuotas, otros mediante entrevistas por teléfono a viviendas con "línea fija" residencial, otros mediante encuestas "en-línea", otros sustentados en muestras estrictamente probabilísticas sin sustitución en ninguna de sus etapas de selección.

Eso en cuanto a la recolección de datos, lo mismo sucede en cuanto a cómo procesarlos para arribar a resultados finales. Nuevamente hay de todo, desde los que ignoran el diseño del que provienen hasta los que usan cada incidencia del mismo (en selección de muestra, en campo, en datos de fuentes externas, etc.) para identificar ponderadores provenientes del diseño de muestra, ajustes de distinta naturaleza según las incidencias de campo, y fuentes externas a la encuesta cuando se justifique. Hay clientes para toda versión, desde las más económicas y prontas hasta las más caras y ortodoxas pero menos oportunas. Conforme se acercan las fechas de registro de candidatos para nueva elección, pre-campañas y campañas, y ya nombrados los candidatos de

cada partido, las más preferidas por quienes de esto saben – no muchos – y tienen los medios para financiarlas, son las que tienen carácter estrictamente probabilístico.

Dificultad intrínseca en encuestas previas a la jornada electoral. Las estimaciones de las encuestas pre-electorales, al igual que las levantadas el día de la elección, de origen están sujetas a incertidumbre vs las cuentas oficiales de votos, sujetas a reglas precisas. Son ciencia y arte vs hechos factuales futuros (lejanos o cercanos). Las poblaciones de una y otra son distintas.

Encuestas *previas* a la elección:

LN = Votantes + No Votantes - indistinguibles en las encuestas ----------(2)

Encuestas de Salida y Conteos Rápidos:

Votantes = Votos Válidos + Anulados, bien definidos en cómputos------(3)

Programa de Resultados Preliminares Oportunos, no es encuesta, (PREP):

LN = Votos Válidos + Anulados + No Votantes----------(4)

Cuando se seleccionan las muestras estrictamente probabilísticas de ciudadanos con credencial de elector vigente domiciliada al menos en el municipio donde está ubicada su vivienda en muestra, (de facto un marco muestral de áreas para la LN), el reto nada sencillo de superar es el de *distinguir* quiénes de los entrevistados el día de la elección se convertirán en “Votantes”, cuáles emitirán “Votos Válidos”, cuáles votos “Anulados”, y quiénes no acudirán a votar, los “No Votantes” a pesar de estar en la LN. Para completar el cuadro afloran los que en la pregunta de intención de voto deciden no responderla; ¿cómo tratarlos?

Y el reto no termina ahí, pues las mismas preguntas se debe hacer el encuestador respecto a los miembros de la población objetivo que resultaron seleccionados por el diseño estrictamente probabilístico y terminaron en alguna de las muchas variantes de “no-respuesta” total a la encuesta. Infelizmente el uso extendido de encuestas por cuotas, no probabilísticas, esconden e ignoran este creciente problema.

Luego viene el reto de comunicar al cliente las estimaciones resultantes (de manera particular insistir en el uso de intervalos de estimación, dando de manera explícita su nivel de precisión y confianza, para así subrayar el hecho de que efectivamente son estimaciones), sus limitaciones y virtudes. Probado está que buenos comunicadores no somos (los encuestadores); y aquellos que sí, sus clientes, dueños de la información, se encargan de *mal divulgar* los pocos resultados que seleccionan según estrategias o caprichos. En tanto no distorsionen o de plano mientan en lo difundido, muy su derecho, son los dueños; caso contrario tenemos (los encuestadores) el derecho y la obligación de salir, casi en tiempo real, a señalar la pifia o burdo engaño, ejemplos abundan.

Entre los errores comunes de comunicación está una de las etiquetas favoritas "los indecisos". ¿Quiénes son? Los que fueron entrevistados y rehusaron responder la pregunta sobre su intención de voto claman sesudos analistas; falso, quizá sean mayoría dentro de este grupo los que tiempo ha que ya decidieron por quién votar y optan por no compartir esta decisión, por la razón que sea. Sólo una encuesta tipo panel, que entrevista la misma muestra periódicamente, puede aproximar una respuesta al contrastar lo que el informante responde en una medición vs otra; quienes cambian con frecuencia quizá sean los *indecisos,* que nuevamente no sabemos si se convertirán luego en votantes y que emitan un voto válido.

Tantos escollos y problemas a salvar, imposible, ¡a la basura con las encuestas electorales! Pues no. Es una *virtud* y no una deficiencia el reconocer que nuestra actividad mide con incertidumbre, y que al hacerlo sustentado en muestras estrictamente probabilísticas tiene la virtud adicional de permitirnos *medirla* con los propios datos de la muestra a mano, para cada una de las estimaciones prioritarias, al nivel de confianza que se desee, y que esto es bueno.

Tarea, entre muchas, tenemos en aprender a comunicar mejor que las encuestas de pre-campaña y de campaña son ejercicios de estimación (que entrevistan electores, no votantes) totalmente diferentes al de las encuestas de salida (*únicas* que entrevistan votantes), al de los conteos rápidos (que no entrevistan a nadie), y al de los PREP que no entrevistan a nadie y *no son ejercicios de estimación,* que simplemente acumulan y suman datos hasta que se decide cerrarlos para esperar el resultado oficial proveniente de los cómputos distritales, cuyo resultado tiene obligación de "cantar" el consejero Presidente del IFE, *sin adjetivo ni juicio alguno,* y luego esperar a que el Tribunal

Federal Electoral (TRIFE) dictamine el resultado oficial de la elección una vez resueltas las impugnaciones que presenten, en su caso.

Por supuesto, como toda buena ensalada, surgen aderezos apetitosos provenientes de los profesionales en técnicas de investigación *cualitativa* "que le dan sabor al caldo". Pero eso es tema para otra charla.

Tiempos

Pasan con desenfado con sobresaltos esporádicos de entusiasmo espurio o decepción real los años del sexenio, hasta que los tiempos marcan fechas fatales próximas que revitalizan la proliferación de sondeos azarosos (que no probabilísticos) y encuestas de todo tipo: sea para "auscultar" y explorar posibilidades de suspirantes y posibles candidatos, presentando distintos escenarios del tipo "si XXX fuera el candidato del PPP para la próxima elección de … y NNN el de BBB y RRR el de ZZZ, por cuál de ellos votaría", con todas las variantes imaginables; para una vez nombrados los candidatos de cada partido, pasar a medir las preferencias de los ciudadanos con credencial vigente del Registro Federal Electoral (RFE), si las elecciones fueran el día en que son entrevistados: múltiples y variadas versiones de "carreras de caballos" se divulgan, algunas con el ánimo de influir en la intención de voto el día de la elección (a la fecha no hay evidencia que esto suceda); otros resultados que no se divulgan pues son para consumo interno de estrategas de los distintos candidatos-partidos para proponer ajustes a mensajes, discursos, imagen, publicidad, "slogans", etc. Las anécdotas abundan de resultados "sospechosos" por su gran similitud al grado de ser casi idénticos, rareza estadística, hasta divergentes en quién resultaría el ganador en la fecha de la encuesta, y todas las variantes intermedias (tendencias de series que al paso del tiempo se cruzan, a veces justo el día de la elección, una de las favoritas) que son insumos ansiosamente esperados por analistas y columnistas especializados de uno y otro bando, para especulaciones sin fin.

Claro, la diosa de la fecundidad es despertada para parir de inmediato todo tipo de acrónimos de supuestas encuestadoras que nunca antes se les conoció investigación alguna en éste o alguno otro tema, y oh maravilla, con recursos abundantes para pagar y publicar a página completa (a veces dos) sus resultados en diarios de circulación nacional. Terminado el proceso electoral en turno desconocido virus mortal sorpresivamente ataca a todas y en el acto mueren sin que nadie acuda a dar sus

condolencias a supervivientes (por cierto, muy difíciles de ubicar). Vuelve a dormir la diosa para salir de su letargo con precisión de reloj suizo cada nuevo periodo electoral y volver a parir engendros similares con igual destino, pero que mucho dañan al gremio de encuestadores con larga y conocida reputación de profesionalismo.

Entreverado con lo anterior el IFE realiza algunos estudios vía encuestas que permiten darle una manita de gato al PE y LN que depure lo más grave y notorio de las desactualizaciones que se dan de manera natural. Dicho sea de paso que tales desactualizaciones son por *irresponsabilidad* del *ciudadano* que no registra ante el IFE cambios en su situación (por ejemplo un cambio de domicilio) y datos de identificación. Hay errores que se detectan en estos ejercicios pero que no se clasifican como graves, en el sentido de que no impiden al ciudadano el ejercer su derecho a votar el día de la elección, ejemplos ilustrativos son: cambios de domicilio *dentro de la misma* SE, pues igual les toca votar en el mismo lugar; registro equivocado de edad, pues mismo error aparece en la LN contra la que se cotejan los datos de su credencial; incluso registro erróneo del sexo que igual se replica en su credencial y LN. Otros sin embargo si son graves y un obstáculo a su derecho a votar, por ejemplo: cambio de domicilio a otro fuera de la SE de origen, que si es lejano le imponen el recurrir a trasladarse a su "vieja" casilla donde está registrado su nombre en la LN para poder votar.

Jornada Electoral

El tiempo inexorable nos conduce inevitablemente a la fecha de la jornada electoral de cada elección y termina el periodo de gestación que da a luz *dos* instrumentos *estadísticos* y *uno aritmético* cuyo mejor destino es nacer y morir el mismo día después de cumplir con éxito su razón de ser: las encuestas de salida (ES), los conteos rápidos (CR) y el programa de resultados electorales preliminares (PREP), ya comentados con anterioridad.

Tropiezos

Elección Presidencial de 2000. Después de la historia de éxitos durante y después de las elecciones presidenciales de 1994, a pesar de un sismo político en las elecciones presidenciales del 2000 que puso a prueba nuestra democracia, algunas encuestas ya en campaña, las menos y de manera errática, daban estimaciones donde el partido en el poder durante más de siete décadas no resultaba el ganador. Aberraciones

estadísticas destilaban litros de tinta y mesas de discusión. Hasta que llegó la jornada electoral y las encuestas de salida anticipaban posibles cambios en el ánimo de los ciudadanos, que se reflejaron en un voto diferenciado que dio el triunfo en el Poder Ejecutivo a un partido de oposición, el Partido Acción Nacional (PAN), pero no el control en el Congreso, donde ningún partido obtuvo mayoría absoluta.

Consecuencia: sufrió la credibilidad en las encuestas pre-electorales. No así los ejercicios estadísticos de CR que nuevamente contrató el IFE con tres empresas, y que una vez que el Consejero Presidente (José Woldenberg) "cantó" públicamente los resultados que cada empresa reportó como su estimación, segundos después en cadena nacional el Presidente (del partido en el poder más de 7 décadas) en turno anunció que de acuerdo a esas estimaciones la oposición por primera vez salía triunfante en la elección de Presidente. Días después el TRIFE ratificó el resultado oficial a favor de la oposición, lo cual no levantó siquiera una ceja, y México todo en santa paz.

Elección Presidencial de 2006. Con el antecedente de la elección del 2000, la mal llamada "guerra de encuestas" se agudizó, el gremio de encuestadores sufrió mayor desgaste, incluyendo las ES. De paso el propio IFE se encargó de desacreditar *su* ejercicio de CR, pues en lugar de replicar el modelo probado en 1994 y 2000, contratando a terceros para que realizaran un ejercicio de estimación y luego cantar los resultados que le entregaran, el IFE decidió realizarlo con recursos propios, convirtiendo así al IFE en un actor más entre los muchos que *estiman* el resultado de la elección, en lugar de concentrase en lo suyo: la institución que organiza, cuenta y cuenta bien los votos emitidos por los ciudadanos para luego darlos a conocer. Y ello a sabiendas de que horas y días después tendría resultados necesariamente diferentes vía el PREP y los cómputos distritales. Empeoró la pifia al no divulgar el resultado de *su* CR por estar demasiado cerrado. Todo lo anterior sin menoscabo al profesionalismo, ética y experiencia de quienes realizaron el ejercicio para el IFE, contra ellos nada.

Las ES y CR ejercicios de terceros no corrieron mejor suerte, y hasta el mismo PREP resultó contaminado ante la falta de contraste con los resultados del CR ahora del propio IFE. Lo cerrado del resultado no permitió identificar al probable ganador el mismo día de la elección y ello provocó daños colaterales con acciones inmediatas de gran riesgo (multitudinarias y frecuentes manifestaciones, cierre de avenidas principales, impugnaciones, demandas de recuentos "voto por voto casilla por casilla",

etc.) y secuelas latentes pero sin mayor consecuencia, que brotan de vez en vez aún en el presente y algo más del futuro inmediato. Pero ésa es otra sabrosa historia para otra ocasión.

Elección Presidencial de 2012. Se agudiza *guerra de encuestas;* periodista-periódico nacional como gallito de pelea kikirikea medición diaria de encuestadora contratada para ello y reta que *ya se verán las caras el día de la elección.* Carrera de caballos muestra a salidor a la cabeza desde el arranque ganando puntos hasta ser de más de un cuerpo, llega a ser de dos dígitos cercanos a 20 puntos porcentuales. Un par de mediciones se atreven a publicar que no, que la ventaja es de sólo un dígito y con tiempo para cerrarse aún más. Surge de nuevo el calificativo burlón de "aberrantes", quizá ni quien va en segundo lugar lo cree pues no se nota que haga algún ajuste para achicar distancia … y llega el día de la elección. Los aberrantes eran el resto, gana el puntero con cómoda distancia, pero de un solo dígito. El acabose para las casas encuestadoras, todas cuestionadas, descrédito que afecta a todo el gremio, aún a quienes no se dedican a este tipo de mediciones. TRIFE confirma resultado oficial con ventaja del orden de magnitud entre los dos aberrantes. Periodista-periódico kikirikero no deja de escribir su columna diaria entre semana, eso sí, corre a encuestador y abandona las encuestas.

El IFE repite el numerito del 2006 y en 2012 vuelve a entrarle al juego de *estimación,* que no es lo suyo, y le pone sal al no apegarse a simplemente *cantar el resultado,* única atribución que el COFIPE (Código Federal de Instituciones y Procedimientos Electorales) le da, sino que agrega de su ronco pecho …y por tanto quien ganó la elección fue …. Desastre anunciado que de rebote desacredita aún más a las encuestas y casas encuestadoras.

Culpables. ¡Encuestadoras! Claman los actores y partidos políticos participantes. ¡Encuestadoras! Clama el círculo rojo. ¡Encuestadoras! Claman los medios. ¡Encuestadoras! Clama el resto.

Falso digo yo. La realidad reside en la dificultad de distinguir lo expuesto en las tres sencillas expresiones (2), (3) y (4) expuestas al tratar los *métodos*. Intentos vía la ruta de "votantes probables" hay muchos (nadie revela el suyo, pues es "know how" propio, you know), pero todos se quedan cortos al no preocuparse por aplicar esfuerzos en la creciente *no-respuesta total* a las encuestas. Fácil exponerlo, difícil resolverlo,

sobre todo porque las aproximaciones con más expectativas de aproximación a algo mejor conducen inevitablemente a *encuestas panel,* estrictamente probabilísticas, con varias revisitas para ubicar y convencer al seleccionado a responder la encuesta; posibles pero resultan costosas y hasta ahora no hay quien esté dispuesto a pagarlas.

Voces calificadas y sensatas como la de José Woldenberg en su nota editorial semanal en el periódico Reforma de fecha 18 de julio de 2013 (sin desperdicio leerla completa), termina afirmando en último párrafo: "A pesar de ello, las encuestas se siguieron realizando en serio y en serie. Pero, dado el escándalo que se produjo en 2012, cuando un puñado de importantes encuestadoras estuvo dando a lo largo del proceso un posible escenario que resultó mucho más estrecho el día de la elección, ahora también han menguado de manera considerable las encuestas que se hacen públicas sobre las intenciones del voto. Total: que el mecanismo que tan buenos resultados dio a lo largo de un periodo, parece que –por miedo– se empieza a desmantelar." Terrible advertencia de que puede darse un regreso al origen, para nuevamente encerrarlas en el closet al que sólo tienen acceso élites de siempre.

Epílogo

Elecciones Locales del 7 de julio de 2013. Tan reciente la experiencia y continuada permanencia en medios que seguro los detalles siguen en la mente de muchos. Resumo uno que se dio en la elección que más atención concentró, la de Gobernador en Baja California: ansias de novillero en reconocidos toreros de larga y exitosa trayectoria política los llevó, contra consejas del más alto nivel, a salir a declararse ganadores, según encuestas de salida y conteos rápidos por ellos conocidos (nunca nombraron las casas encuestadoras ni los resultados estimados para cada partido-coalición-candidato), para pocas horas después, el mismo día de la jornada electoral recular en público y hacer un llamado a la prudencia para esperar el resultado oficial.

La cereza en el pastel fue que el PREP, que como ya dijimos se restringe a recibir, acumular y sumar los resultados conforme los van reportando de los distintos Distritos Electorales, sumaba mal y dividía mal al transformar votos en porcentajes, al menos en lo exhibido en sus pantallas (luego se arguyó que en las bases todo cuadraba en las sumas, más no así en el tiempo en que se actualizaba cada renglón; y que el tema de porcentajes era una simple variante entre dos maneras de redondear). Vuelta a la aritmética y el simple ejercicio de enumeración. Los errores reconocidos, eran pequeños

y no afectaban el resultado que con ellos se especulaba (que no es una estimación estadística), pero bastó para que el Instituto Electoral y de Participación Ciudadana de Baja California (IEPC-BC) descalificara el PREP y a la empresa contratada por el propio IEPC-BC. Los novilleros por su lado vociferaron la frase menos afortunada al reclamar un recuento: "voto por voto, casilla por casilla". Reclamo en total contradicción a la reseña victoriosa que momentos antes dieron citando frases imputadas a Luis Donaldo Colosio cuando reconoció el triunfo del PAN en1989.

Pendientes. Es momento de convencer a los clientes (y algunos colegas) a abandonar el "muestreo de cuotas" dentro de las manzanas en muestra y continuar con esquemas estrictamente probabilísticos, aunque esto incremente los costos del trabajo de campo de manera significativa, pues *no* permite ningún esquema de "sustitución", por sofisticado que sea, ante cualquier tipo de no-respuesta; que implica varias visitas en diferentes horas y días a los hogares seleccionados para intentar encontrar y lograr entrevistar al miembro específico que resulte seleccionado mediante un esquema estrictamente probabilístico; y donde la experiencia de cada encuestador le dirá qué tanta sobre-muestra requiere de origen para que al final del trabajo de campo se cuente con un número "cercano" al deseado de entrevistas completas.

Urge un compromiso de transparencia entre medios de comunicación y agencias encuestadoras. Es fundamental que la encuesta que sea pagada y publicada por un medio de comunicación tenga el entero reconocimiento del grupo editorial: la casa encuestadora y el medio deben asumir la responsabilidad de los datos que arrojen sus mediciones. Incluye el concertar *a priori* el formato y contenido de la difusión y/o publicación de algunos de los resultados y así evitar sorpresas *a posteriori*.

Incluye el examinar si debemos arribar a un convenio-contrato básico que usemos toda la industria (o al menos los agremiados en AMAI) donde se estipulen cláusulas preventivas de excesos, abusos dolosos o incluso groseras manipulaciones en la difusión (difusión a la que por cierto tienen derecho al ser los dueños de los resultados).

Necesitamos acercarnos aún más a los medios, sus conductores y plumas especializadas en el tema, para de manera conjunta aprender unos de otros a comunicar mejor todo lo anterior así como los resultados sin demérito de hacerlo en un contexto noticioso. Debemos diversificar los temas que medimos. Durante la elección presidencial no medimos temas específicos relacionados con las propuestas

de los candidatos. Simplemente nos enfocamos a medir la carrera de caballos para saber quien encabezaba las preferencias electorales, pero dejamos a un lado lo que los mexicanos pensaban sobre temas fundamentales coyunturales o estructurales, ejemplos abundan.

Aportaciones del XXVIII Foro Nacional de Estadística

Un modelo semiparamétrico bayesiano para datos circulares*

Gabriel Núñez Antonio[a], Gabriel Escarela
Universidad Autónoma Metropolitana, unidad Iztapalapa, México

Mike Wiper, Concepción Ausín
Universidad Carlos III de Madrid, España

Clasificación: Trabajo de Investigación.
Área: Estadística Bayesiana.
Subárea: Datos circulares.
Palabras clave: Distribución Normal envuelta; Métodos MCMC.

1. Introducción

Los datos circulares, pueden ser interpretados como direcciones angulares, tales como la dirección del viento, se producen en muchas áreas en las ciencias ecológicas y ambientales. Una revisión de las principales características y modelos para tales datos se pueden encontrar, por ejemplo, en Fisher (1993), Mardia Jupp (2000), y Pewsey et al. (2013).

Aunque la mayoría de los modelos para datos circulares son paramétricos, en muchas aplicaciones prácticas, como por ejemplo, la descripción de direcciones de migración de aves o direcciones de viento, los datos observados presentan características tales como asimetría y multimodalidad, características que son difíciles de describir usando modelos paramétricos estándar. En tales casos, puede ser preferible considerar modelos no paramétricos o semiparamétricos como una buena alternativa.

En el contexto de inferencia Bayesiana, el enfoque usual es considerar mezclas de procesos Dirichlet (DP). En el caso de datos circulares se han analizado mezclas DP con distribuciones base distribuciones von Mises, ver por ejemplo, Ghosh et al. (2003). Sin embargo, la

*Este trabajo fue apoyado parcialmente por el SNI, México. El apoyo del Departamento de Matemáticas de la UAM-I también es reconocido ampliamente.

[a]gabnunez@xanum.uam.mx

constante de normalización de la densidad von Mises en cierto es sentido compleja, lo cual puede producir dificultades cuando se realizan inferencias. En este artículo se propone una alternativa de mezclas DP basadas en distribuciones normales *wrapped* (envueltas).

2. La distribución normal wrapped

Una manera de generar distribuciones para datos circulares es el enfoque *wrapping*. Dada una distribución conocida sobre la recta real, esta se puede *envolver* (wrapped) alrededor de la circunferencia del circulo de radio uno, $\mathbb{S}$. Así, si X es una variable aleatoria con función de distribución $F_X(x)$, entonces la correspondiente variable aleatoria Θ sobre el círculo unitario esta definida por $\Theta = \mod(X, 2\pi)$, y la función de densidad de Θ es $f_\Theta(\theta) = \sum_{k=-\infty}^{\infty} f_X(\theta + 2\pi k)$, para$\theta \in \mathbb{S},$, donde $f_X(\cdot)$ es la función de densidad de X. En el caso particular de que X tenga una distribución normal $N(\mu, \sigma^2)$, entonces se dice que Θ tiene una *distribución normal envuelta*, $\Theta \sim WN(\mu, \sigma^2)$. Así, la correspondiente función de densidad circular esta dada por

$$\phi^{WN}\left(\theta \mid \mu, \sigma^2\right) = \sum_{k=-\infty}^{\infty} \frac{1}{\sigma}\phi\left(\frac{\theta + 2\pi k - \mu}{\sigma}\right), \qquad 0 \leq \theta < 2\pi, \tag{1}$$

donde $\phi(\cdot)$ es la función de densidad de probabilidad de la distribución normal estándar.

2.1. Inferencia Bayesiana

Para una revisión, desde una perspectiva frecuentista, de los procedimientos de inferencia para la distribución normal envuelta el lector se puede referir a Mardia (1972) y Mardia y Jupp (2000). Inferencia bayesiana para la distribución normal envuelta se puede revisar por ejemplo en Coles (1998), Ferrari (2009) y Ravindran y Ghosh (2011).

En primer lugar, se debe notar que una variable $\Theta \sim WN(\mu, \sigma^2)$ puede ser *desenvuelta* definiendo $X = \Theta + 2\pi K$ donde $X|\mu, \sigma^2 \sim N(\mu, \sigma^2)$ y

$$P\left(K = k|\mu, \sigma^2\right) = P\left(2\pi k \leq X \leq 2\pi(k+1) \mid \mu, \sigma^2\right). \tag{2}$$

En segundo lugar, hay que notar que una distribución $WN(\mu, \sigma^2)$ es la misma distribución que

$WN(\mu + 2\pi k, \sigma^2)$ para cualquier $k \in \mathbb{Z}$, lo cual implica que para propósitos de identificabilidad se debe definir una distribución inicial para μ con soporte sobre $\mathbb{S}$. Por lo tanto, se define una inicial uniforme circular para μ y una inicial gamma inversa para σ^2, o equivalentemente, si $\tau = 1/\sigma^2$, entonces $\tau \sim Ga\left(\frac{a}{2}, \frac{b}{2}\right)$ para $a, b > 0$.

Dado una muestra $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)^t$ de la distribución normal envuelta, $WN(\mu, \sigma^2)$, se pueden llevar a cabo inferencias generando el número de *envolturas* no observadas, k_i, para $i = 1, ..., n$. En particular, la distribución final condicional para estas variables latentes es

$$P(k_i \mid \theta_i, \mu, \tau) \propto \phi\left(\sqrt{\tau}(\theta_i + 2k_i\pi - \mu)\right), \qquad \text{para } k_i \in \mathbb{Z} \text{ y } i = 1, ..., n. \tag{3}$$

Condicional sobre el número de envolturas, k_i, se pueden definir los datos desenvueltos

$$x_i = \theta_i + 2\pi k_i, \qquad \text{para } i = 1, ..., n.$$

Finalmente, condicional sobre los datos desenvueltos $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, la distribución final de los parámetros del modelo se puede factorizar como,

$$f(\mu, \tau \mid \boldsymbol{x}) \propto f(\tau \mid \mu, \boldsymbol{x}) f(\mu \mid \boldsymbol{x}),$$

donde la distribución final condicional del parámetro de concentración τ, está dada por

$$\tau \mid \boldsymbol{x}, \mu \sim Ga\left(\frac{a+n}{2}, \frac{b + (n-1)s^2 + n(\mu - \bar{x})^2}{2}\right), \tag{4}$$

donde $\bar{x}$ y s^2 son la media muestral y la varianza de los datos $\boldsymbol{x}$, y la distribución final del parámetro de localización μ, es una distribución t no-estándar tal que si se define

$$\vartheta = \frac{\mu - \bar{x}}{\sqrt{\frac{b+(n-1)s^2}{n(a+n-1)}}},$$

entonces la distribución final de ϑ es una distribución t estándar con $(a + n - 1)$ grados de libertad, truncada sobre la región

$$-\frac{\bar{x}}{\sqrt{\frac{b+(n-1)s^2}{n(a+n-1)}}} \leq \vartheta \leq \frac{2\pi - \bar{x}}{\sqrt{\frac{b+(n-1)s^2}{n(a+n-1)}}},$$

de la cual es fácil de muestrear. Así, se pueden obtener inferencias usando un muestreador de Gibbs para muestrear sucesivamente de las envolturas (3), calcular los valores desenvueltos $\boldsymbol{x}$ y entonces muestrear de la distribución final conjunta $\mu, \tau | \boldsymbol{x}$. El único paso marginal complicado está en (3), pero en este caso, el muestreo se puede llevar a cabo, por ejemplo, ya sea por truncar la distribución en algún valor grande de k o usando un paso de Metropolis.

3. Modelo de mezcla DP circular

Supóngase que se desea definir un modelo Bayesiano no-paramétrico para la variable $\boldsymbol{X}$ con soporte en la algún espacio $\mathbb{C}$. Se puede suponer que $\boldsymbol{X}|H \sim H$, donde H es una función de distribución sobre variables con soporte $\mathbb{C}$, y entonces asignar una distribución inicial para H. Una forma de hacer lo anterior es usar un proceso Dirichlet (Ferguson, 1973) como inicial. Es decir, $H \sim DP(\alpha, H_0)$. Sin embargo, esta especificación inicial produce distribuciones discretas, las cuales no son apropiadas para distribuciones con soporte continuo. Para resolver este problema Antoniak (1974) introdujo mezclas de DP como iniciales. Estos modelo se definen de manera jerárquica y son equivalentes a mezclas infinitas contables de densidades paramétricas (Sethuraman, 1994).

3.1. El modelo normal envuelto

A continuación se define un modelo de mezcla DP de normales envueltas. Se asume que Θ es la envoltura de X sobre el circulo, donde

$$\begin{aligned} X|\mu,\tau &\sim N\left(\mu, \frac{1}{\tau}\right) \\ \mu,\tau \mid H &\sim H \\ H \mid \alpha, H_0 &\sim DP\left(\alpha, H_0\right). \end{aligned}$$

Aquí, H_0 es la distribución base del modelo de mezcla DP. En este caso, ésta se define como el producto de una distribución uniforme sobre $[0, 2\pi)$ y una $Ga\left(\frac{a}{2}, \frac{b}{2}\right)$. Para el parámetro de concentración del proceso Dirichlet, α, se considera una distribución gamma, $Gamma(a_0, b_0)$. Bajo el modelo anterior, la variable Θ sigue una mezcla infinita contable de distribuciones normales envueltas, es decir,

$$f\left(\theta \mid \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\tau}\right) = \sum_{s=1}^{\infty} \rho_s \phi^{WN}\left(\theta \mid (\mu, \tau)_s\right).$$

Usando los resultados de la sección 2.1 y la técnica de *slide sampling* propuesta por Walker (2007), se pueden llevar a cabo inferencias y predicciones vía un muestreador de Gibbs, Así, para $i = 1, ..., n$, condicional sobre el número de *wrapping*, k_i, los datos se pueden reducir a datos escalares univariados x_i generados de una mezcla DP de normales univariadas.

Por otro lado, para estimar la densidad predictiva, una posibilidad es usar el algoritmo de Walker (2007). Sin embargo, una mejor alternativa se puede obtener empleando el siguiente estimador:

$$f\left(\theta_{n+1} \mid \theta_1, \ldots, \theta_n\right) \approx \frac{1}{M} \sum_{m=1}^{M} \phi^{PN}\left(\theta \mid (\mu, \tau)^{(m)}\right) \tag{5}$$

donde M es el tamaño del algoritmo MCMC y $(\mu, \tau)^{(m)}$ son los parámetros de la normal envuelta de la componente de la mezcla de la cual son muestreados en cada paso del algoritmo.

4. Ejemplo

En este ejemplo, se simuló una muestra de tamaño 1000 de la distribución con densidad:

$$f(\theta) = 0.5\ \phi^{WN}(\theta \mid \pi/2, \frac{1}{0.3^2}) + 0.2\ \phi^{WN}(\theta \mid \pi, \frac{1}{0.2^2}) + 0.3\ \phi^{WN}(\theta \mid 5\pi/4, \frac{1}{0.2^2}).$$

El histograma lineal de este conjunto de datos se muestra en la Figura 1. Se puede notar que esta especificación produce un conjunto de datos con tres modas. En la Figura 1 se muestra la densidad predictiva obtenida con el procedimiento descrito en este trabajo, sobrepuesta al histograma de datos reales. Se puede observar que con la metodología presentada se describe adecuadamente la forma de la verdadera densidad.

Bibliografía

Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, **6**, 1152–1174.

Coles, S. (1998). Inference for circular distributions and processes. *Statistics and Computing*, **8**, 105–113.

Ferguson, T. (1973). Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209–230.

Ferrari, C. (2009). T*he Wrapping Approach for Circular Data Bayesian Modeling*. Doctoral Thesis, Bologna University, Bologna, Italy.

Fisher, N.I. (1993). *Statistical Analysis of Circular Data*. Cambridge: University Press.

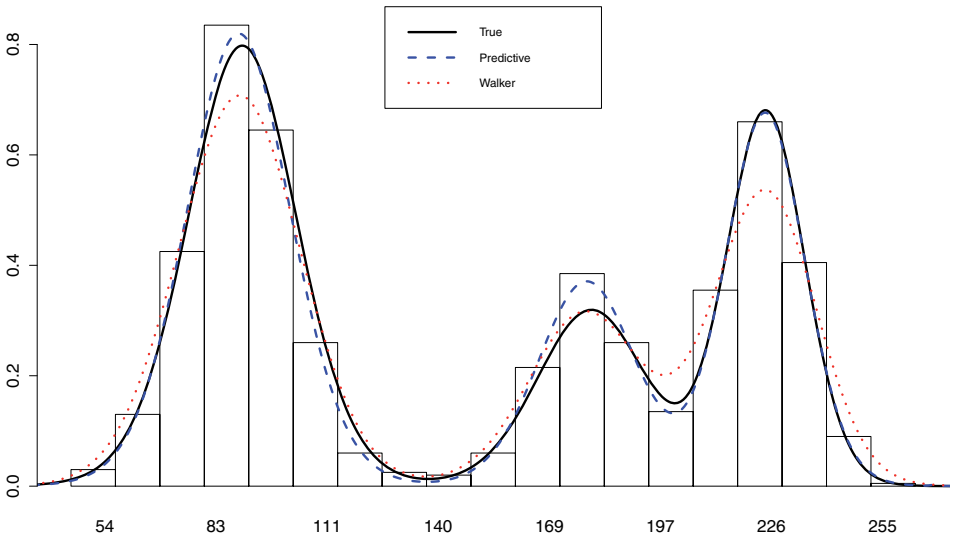


Figura 1: *Histograma lineal de los datos simulados, verdadera densidad (línea sólida), densidad predictiva obtenida con el modelo de mezcla DP de normal envuelta (guión) y, densidad predictiva estimada como en Walker (2007) (punteada).*

Ghosh, K., Jammalamadaka, S.R. y Tiwari, R.C. (2003). Semiparametric Bayesian techniques for problems in circular data. *Journal of Applied Statistics*, **30**, 145–161.

Mardia, K.V. y Jupp, P.E. (2000). *Directional Statistics*. Chichester: Wiley.

Pewsey, A., Neuhäuser, M. y Ruxton, G.D. (2013). *Circular Statistics in R*. Oxford: University Press.

Ravindran, P. y Ghosh, S.K. (2011). Bayesian Analysis of Circular Data Using Wrapped Distributions. *Journal of Statistical Theory and Practice*, **5**, 547–561.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–650.

Walker, S.G. (2007). Sampling the Dirichlet Mixture Model with Slices. *Communications in Statistics. Simulation and Computation*, **36**, 45–54.

Diseño Robusto en Teoría de Control*

Armando Mares Castro[a], Jorge Domínguez Domínguez[b]
Ciatec, A.C., Cimat, A.C.

Clasificación: Trabajo de Investigación,Tesis de Doctorado.
Área: Diseño de Experimentos.
Subárea: Diseño Experimental Robusto de Parámetros.
Trabajo presentado en: XXVIII Foro Nacional de Estadística.

1. Introducción

Un concepto de importancia en el enfoque moderno de la mejora de la calidad es la utilización del diseño de experimentos para hacer productos robustos, es decir, elaborar productos que sean poco sensibles a imperfecciones de manufactura, deterioro por aspectos ambientales, entre otros factores, los diseños que permiten obtener éste tipo de productos se llaman robustos. La ingenieria de control estudia el diseño de sistemas que juegan un papel importante en muchas actividades que realizamos cada día. Este tipo de sistemas consta de 3 características claves, las entradas, las salidas y los parámetros de control. Éstos últimos se emplean para perturbar el sistema para ponerlo en un estado estable. La conjunción de éstas temáticas genera un tema de investigación para establecer un proceso de optimización que permita encontrar situaciones robustas de operación. El diseño de parámetros por sí solo nos lleva a una calidad lo suficientemente alta. Una mejora más a fondo puede ser lograda mediante el control de las causas de variación donde sea económicamente justificable (Phadke (1989)). El objetivo de éste trabajo es aplicar la metodología del diseño robusto en un sistema dinámico, para ello será necesario desarrollar y analizar las técnicas que permitan generar un producto robusto en teoría de control. Así como aplicar éstas ideas a procesos reales.

* Este trabajo fue realizado con el auspicio del Consejo Nacional de Ciencia y Tecnología CONACYT
[a] _amares.picyt@ciatec.mx
[b] _jorge@cimat.mx

2. Marco Teórico

El estudio del diseño robusto de parámetros ha pasado por etapas de desarrollo en base a las necesidades y cuestionamientos que se fueron generando particularmente en su introducción a occidente. La historia de aplicación del diseño experimental en la industria con fines de optimización de procesos ha pasado por varias etapas tales como la generación de diseños para modelos de segundo orden, por ejemplo el diseño central compuesto Box Behnken y el Box-Draper con los cuales se pueden buscar condiciones óptimas de operación para un determinado proceso, de aquí pasamos al enfoque multirespuesta en la cual ya se considera la optimización de dos respuestas de manera simultánea en la cual se han manejado diversos esquemas tales como la función de deseabilidad y la deseabilidad doble exponencial (Wu (2005)), todas éstas técnicas fueron retomadas por diversos autores estadísticos como alternativas a la metodología original del Dr. Taguchi.

2.1 Diseño Robusto de Parámetros

La metodología del diseño robusto de parámetros fué desarrollada por el Dr. Genichi Taguchi a mediados del siglo XX en Japón, no fué hasta mediados de la década de los 90´s que la metodología se introdujo en occidente siendo el autor que logró la aceptación del diseño de experimentos en las empresas de Estados Unidos. La metodología original planteada por Taguchi considera un proceso experimental para estudios de robustez en un proceso, se caracteriza por el uso de arreglos cruzados separados para los factores de control y de ruido. Los factores de control son fáciles de controlar, mientras que los factores de ruido son difíciles o costosos de mantener bajo control. Se busca una solución en la cual se mantenga el valor promedio cercano a un valor Target pero con la menor variación posible alrededor de él, ésto indica que los llamados factores de ruido en ésta solución tienen el menor efecto sobre la variación de la característica de calidad analizada, lo cual equivale a mencionar que el proceso es robusto, la medida de desempeño (PerMIA) por sus siglas en inglés) (León et al.(1987)), utilizada para la optimización es la llamada señal a ruido, la cual se maneja para diferentes casos como lo son: el target es lo mejor, entre más pequeño es mejor, entre más grande es mejor y de proporción de defectuosos. Autores como Phadke (1989) fueron discípulos directos del Dr. Taguchi y analizaron a detalle la metodología del Diseño Robusto

de Parámetros, en su libro Phadke (1989) proporciona directrices sobre el uso de los arreglos tipo Taguchi y las medidas de desempeño. Otros autores como Wu y Hamada (2000) analizan a detalle la metodología y participan en las discusiones sobre el tema.

2.2 Críticas a la metodología de Taguchi

Easterling (1985), Pignatiello y Ramberg (1985) y Nair (1992) analizaron y criticaron la metodología de Taguchi. Entre las críticas realizadas se menciona que los diseños de arreglo cruzado tienen al menos dos debilidades. Primero, a menudo requieren un número largo de corridas. Segundo, generalmente el arreglo interno es usualmente un fraccionado a 3 niveles y el externo a dos niveles con lo que los experimentadores deben estimar el efecto lineal y cuadrático de los factores de control pero no de las interacciones. Box (1988) observó que en muchos casos, Var(y) es una función de μ y el procedimiento a dos pasos de Taguchi no funciona. Nair y Pregibon (1988) criticaron a Taguchi por el uso de la señal razón a ruido y junto con Box sugirieron un enfoque diferente al problema del diseño de parámetros. El error cuadrado medio $ECM = \left((x_2 - T)^2 + x_1\right)$ mediante la selección de un valor que minimice la varianza x_1 y un valor de la variable x_2 para alcanzar la mínima desviación.

2.3 Modelos de Optimización Basados en el doble Arreglo Ortogonal

El esquema del doble arreglo ortogonal propuesto por Taguchi fue seguido por varios autores los cuales siguieron utilizando el formato del arreglo interno para los factores de control y el arreglo externo se utiliza para los factores de ruido, bajo éste esquema se generan modelos para la $\mu, \sigma^2, \log \sigma^2$ e incluso para la razón S/R, los modelos pueden ser orden 1 o superior y se muestra en ec (1) y ec (2):

$$Y_1 = \beta_0 + \mathbf{x}'\beta + \mathbf{x}'\mathbf{B}\mathbf{x} + \varepsilon_1 \tag{1}$$

$$Y_2 = \gamma_0 + \mathbf{x}'\gamma + \mathbf{x}'\mathbf{D}\mathbf{x} + \varepsilon_2 \tag{2}$$

Donde $\mathbf{x}' = (x_1, ..., x_k)$ k factores,β_0 es la constante, $\beta' = (\beta_1, ..., \beta_k)$ un vector de parámetros, $\mathbf{B} = (\beta_{11}, ..., \beta_{1k}, \beta_{k1}, ..., \beta_{kk})$, matríz de parámetros de segundo orden, y $\varepsilon_1 \approx N(0, \sigma_1^2)$, γ_0 la

constante, $\gamma' = (\gamma_1, ..., \gamma_k)$, un vector de parámetros, $\mathbf{D} = (\gamma_{11}, ..., \gamma_{1k}, \gamma_{k1}, ..., \gamma_{kk})$, matríz de parámetros de segundo orden, y $\varepsilon_2 \approx N(0, \sigma_2^2)$. Bajo éste esquema Vining y Myers (1990) recurrieron al problema de optimización de respuesta Dual de Myers y Carter. Tanto Copeland y Nelson (1996) como Del Castillo y Montgomery (1993) utilizaron técnicas de programación no lineal para determinar las condiciones de operación óptimas mediante la minimización directa de la función $\sigma_y + \varepsilon$., Lin y Tu (1995) propusieron modelos para la media y la varianza, su criterio consiste en dos términos sesgo de la distribución en la variable de respuesta y varianza del error.

2.4 Modelos de Optimización Basados en el arreglo Combinado

Welch et al (1990) propusieron una alternativa al uso del doble arreglo ortogonal en el cual se manejan en el mismo arreglo los factores de control y los factores de ruido, éste esquema tiene ventajas sobre el doble arreglo ortogonal ya que se reduce el número de corridas además de que se pueden estimar las interacciones de control por ruido lo cual es básico para la optimización, el modelo se muestra en ec (3)

$$Y(\mathbf{x}, \mathbf{z}) = \beta_0 + \mathbf{x}'\beta + \mathbf{x}'\mathbf{B}\mathbf{x} + \mathbf{z}'\delta + \mathbf{x}'\mathbf{C}\mathbf{z} + \varepsilon \tag{3}$$

Donde $\mathbf{x}' = (x_1, ..., x_k)$ k factores de control, $\mathbf{z} = (z_1, ..., z_q)$ q factores de ruido, β_0 la constante, los vectores de los parámetros $\beta' = (\beta_1, ..., \beta_k)$, $\delta' = (\alpha_1, ..., \alpha_q)$, $\mathbf{B} = (\beta_{11}, ..., \beta_{1k}, \beta_{k1}, ..., \beta_{kk})$ y $\mathbf{C} = (\alpha_{11}, ..., \alpha_{1q}, ..., \alpha_{k1}, ..., \alpha_{kq})$ son las matrices de parámetros de segundo orden, y $\varepsilon \approx N(0, \sigma_e^2)$. Además se plantean los siguientes supuestos para los factores de ruido $E(z) = 0$, y $Var(z) = V$; $diag(V) = \sigma_z^2$. Bajo éste esquema también se propusieron alterativas para la optimización, algunas contemplan el escribir en forma general en forma reparametrizada para los factores de ruido, se utiliza la simulación para generarlos; otras técnicas de optimización son basadas en análisis numérico, su método parte del concepto de función de pérdida, Vining y Myers (1990) utilizan superficies de respuesta para los modelos de la media y la varianza, y para resolver el problema de diseño robusto de parámetros utilizan la pérdida cuadrática estimada.

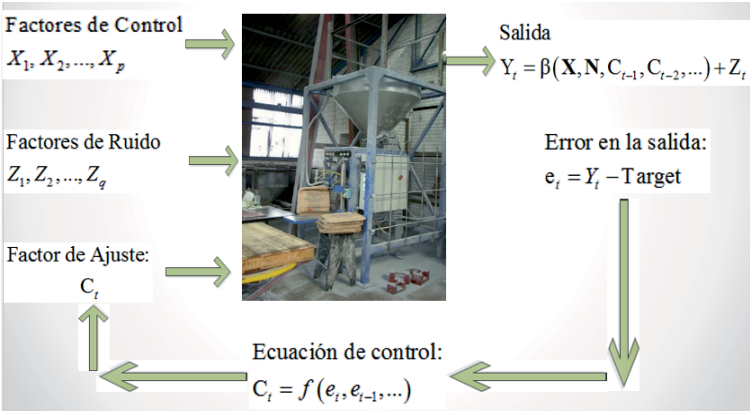


Figura 1: Esquema de un proceso con un elemento de control Feedback

2.5 Análisis para Sistemas Señal-Respuesta

Existen sistemas en los cuales la característica de interés está en función de otro factor, el ejemplo clásico es el acelerador el cual nos da una respuesta de aceleración en función de la pisada que se dé al pedal, éste tipo de sistemas son llamados sistemas dinámicos o sisteas señal-respuesta, Taguchi también abordó éste tipo de situaciones en las cuales se busca hacer robusta la respuesta a la variación que generan los factores de ruido. Los autores que más han trabajado en ésta metodología sentaron las bases para los sistemas con control. Miller y Wu (1996) aplicaron la metodología para el análisis de sistemas señal respuesta y de medición, así como el uso adecuado de los PerMIA para su análisis mediante el uso de arreglos simples y combinados y selección de factores por gráficos Half Normal. Joseph y Wu (2002) probaron la validez del modelo $Y = \beta(\mathbf{X}, \mathbf{Z})M$ para sistemas de múltiple objetivo, β(X,Z) es el modelo de la ec (3) pero se agrega la característica dinámica M, la cual es un factor que la persona varía y al hacerlo obtiene una respuesta Y diferente. Joseph (2003) Presentó una investigación para sistemas dinámicos con control realimentado (feed forward) en el cual analiza sistemas de medición, sistemas estáticos y sistemas dinámicos con control. Los dos autores mencionados y Dasgupta et al (2010) presentaron un artículo sobre sistemas de control adelantado (feedback) sobre sistemas de medición. Los sistemas de control se definen como "Sistemas de doble señal"dado el efecto que tiene el componente dinámico y el componente de control sobre la respuesta, el esquema de un sistema con control feed forward se muestra en la figura 1. La cual se tomó de referencia de un proyecto

de C.F.J. Wu. El elemento de control debe estar conectado en línea para poder realizar las correcciones necesarias en función de un tiempo t, el elemento de control realiza una corrección de acuerdo a la comparación de la respuesta contra el valor deseado, ésto garantiza que aquella variable que está en línea estará siendo corregida de manera constante y reducirá o eliminará su variabilidad, lo cual en conjunto con la solución que proporciona el diseño robusto de parámetros ofrece una solución mejorada.

3. Conclusiones

Existen muchos sistemas que no pueden trabajar sin un esquema de control y se hace necesario el definir y aplicar metodologías aplicables a éste tipo de situaciones, en el presente trabajo se hace una investigación sobre las aportaciones en éste tipo de sistemas, en el campo de la investigación sobre el diseño robusto y sistemas de control actualmente no existen ejemplos de aplicaciones reales en comparación con otras metodologías como la superficie de respuesta, lo cual genera un área de oportunidad en el campo de aplicación.

Referencias

1. Box, G.E.P. "Signal to noise ratios, performance, criteria and transformations". Technometrics, 1988: 30:1-17.

2. Copeland, K.A.F. and Nelson, P.R. "Dual response optimization via direct function minimization". J Quality Techno, 1996 : 28:331–336.

3. Dasgupta, T., Miller, A. and Wu, C.J.F. Robust Design of Measurement Systems". Technometrics, 2010: 52:80-93.

4. Del Castillo, E. and Montgomery, D.C. " A nonlinear programming solution to the dual response problem". J Quality Techno, 1993: 25:199–204.

5. Easterling, R.B. "Discussion of off-line Quality Control, Parameter Design and the Taguchi Methods". J Quality Techno, 1985: 17:198-206.

6. Joseph, V.R. Robust Parameter Design With Feed-Forward Control". Technometrics, 2003: 45:284–291.

7. Joseph, V.R. and Wu, C.F.J. "Performance Measures in Dynamic Parameter Design". J Japanese Quality Eng Soc, 2002: 10:82-86.

8. León, R.V.,Shoemaker, A.C., and Kacker, R.N. "Performance Measures Independent of Adjustment: An Explanation and Extension of Taguchi's Signal-to-Noise Ratios". Technometrics,1987:29:253-265.

9. Lin, D.K.J., Tu, W. "Dual response surface optimization". J Quality Techno, 1995: 27:34–39.

10. Miller, A.E. and Wu, C.F.J. "Parameter Design for Signal-Response Systems: A Different Look at Taguchi's Dynamic Parameter Design".Statistical Science, 1996 : 11:122–136.

11. Nair, V.N. and Pregibon, D. " Analyzing dispersion effects from replicated factorial experiments".Technometrics 1988; 30:, 1988: 30:247-257.

12. Nair, V.N. "Taguchi's Parameter Design: A panel Discussion. Technometrics". 1992:34:127-161.

13. Phadke, M.S. "Quality Engineering Using Robust Design". Englewood Cliffs, New Jersey:Prentice Hall, 1989.

14. Pignatiello, J.S. and Ramberg, J.S. "Discussion of Off-line Quality Control, Parameter Design, and the Taguchi Method by R.N. Kacker". J Quality Techno, 1985:17:198-206.

15. Vining, G.G. and Myers, R.H. " Combining Taguchi and response surface philosophies: A dual response approach". J Quality Techno, 1990: 22:38–45.

16. Welch, W.J., Yu, T.K., Kang and S.M. Sacks J. " Computer experiments for quality control by parameter design". J Quality Techno, 1990 : 22:15–22.

17. Wu, C.F.J. and Hamada, M. " Experiments: Planning, Analysis and Parameter Design Optimization". New York: Wiley Interscience, 2000.

18. Wu, C.F.J. " Optimization of Correlated multiple Quality Characterístics using Desirability Function". Quality Engineering, 2005.17:119-126.

Análisis de Correlación Canónica Regularizada Generalizada: Una aplicación en bosques de mangle

Brenda Catalina Matías Castillo[a], Hortensia J. Reyes Cervantes
Facultad de Ciencias Físico Matemáticas, Benemérita Universidad Autónoma de Puebla

Gladys Linares Fleites
Departamento de Investigación en Ciencias Agrícolas, Benemérita Universidad Autónoma de Puebla

Clasificación: Tesis de Doctorado.
Área: Análisis Multivariado.
Subárea: Análisis de Correlación Canónica.
Trabajo presentado en: XXVIII Foro Nacional de Estadística.

1. Introducción

El Análisis de Correlación Canónica (ACC) es un método exploratorio de datos multivariados y, de manera semejante al Análisis de Componentes Principales (ACP), se basa en resultados del Álgebra Matricial. El ACC fue inicialmente estudiado por Hotelling (1936). El propósito del ACC es la exploración de las correlaciones muestrales entre dos conjuntos de variables cuantitativas observadas sobre el mismo conjunto de individuos o unidades experimentales, a través de combinaciones lineales de las variables iniciales, las que permiten reducir la dimensionalidad. Esta técnica ha sido ampliamente estudiada por Hair et al. (1998), Johnson & Wichern (2007) y Mardia et al. (1979).

Se han desarrollado algunas extensiones del ACC clásico, esto es debido a la presencia de diferentes problemáticas. Como un primer problema, suele ocurrir en algunos casos que el número de unidades experimentales o individuos es menor al número de variables; una

[a] caty_b26@hotmail.com

manera para tratar con este problema es incluir un paso de regularización en el cálculo del ACC, obteniendo un método llamado Análisis de Correlación Canónica Regularizada (ACCR), aplicado por González et al. (2008) y Samarov (2009). Como un segundo problema se puede observar que tanto el ACC y el ACCR son utilizados cuando se tienen dos grupos de variables, sin embargo, el Análisis de Correlación Canónica Generalizada (ACCG) es aplicado a tres o más conjuntos de variables, observados en el mismo conjunto de individuos. Y como una combinación de los dos métodos antes mencionados se encuentra el Análisis de Correlación Canónica Regularizada Generalizada (ACCRG).

En este trabajo se presenta una aplicación del ACCRG del algoritmo realizado por Tenenhaus (2011) a un ecosistema de manglares. El objeto de estudio son los ecosistemas de manglar, que corresponden a la vegetación arbórea que se localiza en la zona de mareas en las regiones tropicales y subtropicales, siendo uno de estos ecosistemas el sistema lagunar de Chacahua - Pastorías, en el estado de Oaxaca. En estos ecosistemas, el interés principal es conocer el grado de relación entre las variables físicas y químicas del agua intersticial y la estructura forestal del bosque de manglar.

2. Análisis de Correlación Canónica Regularizada Generalizada

En el Análisis de Correlación Canónica Regularizada Generalizada se quieren encontrar las relaciones lineales entre varios bloques o grupos de variables, que se supone están conectadas (todas las variables pertenecen a un mismo conjunto de individuos), es decir, se quiere encontrar combinaciones lineales de las variables del grupo, tales que estas nuevas variables están altamente correlacionadas. Siguiendo el trabajo de Tenenhaus (2011), se obtiene el desarrollo siguiente.

Considere J grupos de variables, $\mathbf{X}_1, ..., \mathbf{X}_J$, en un conjunto de n individuos. Una fila de $\mathbf{X}_j$ representa una realización del vector fila aleatorio $\mathbf{x}_j^i$, una columna $\mathbf{x}_{jh}$ de $\mathbf{X}_j$ es considerado como una variable observada en n individuos, x_{jhi} es el valor de la variable de $\mathbf{x}_{jh}$ para el individuo i.

Sea $\mathbf{C} = \{c_{jk}\}$ donde $c_{jk} = 1$ si dos grupos están conectados y 0 en el otro caso, $\mathbf{C}$ es llamada matriz de diseño, está matriz describe las relaciones existentes entre los grupos.

Para encontrar una mejor estimación de la matriz de covarianza $\mathbf{\Sigma}_{jj}$ para cada grupo, es necesario considerar la clase de combinaciones lineales $\left\{\hat{\mathbf{S}}_{jj} = \tau_j \mathbf{I} + (1-\tau_j)\mathbf{S}_{jj}, 0 \leq \tau_j \leq 1\right\}$ de la matriz identidad $\mathbf{I}$ y la matriz de covarianza muestral $\mathbf{S}_{jj}$. $\hat{\mathbf{S}}_{jj}$ es denominada estimación de la contracción de $\mathbf{\Sigma}_{jj}$ y τ_j es la constante de contracción, donde

$$\hat{\tau}_j = \frac{\sum_{k\neq l=1}^{p_j} Var(s'_{j,kl}) + \sum_{k=1}^{p_j} Var(s'_{j,kk})}{\sum_{k\neq l=1}^{p_j} (s'_{j,kl})^2 + \sum_{k=1}^{p_j} (s'_{j,kl})^2},$$

aquí $s'_{j,kl}$ es una entrada de $\mathbf{S}'_{jj} = \frac{n}{n-1}\mathbf{S}_{jj}$.

Para cada grupo se presenta un vector de peso exterior $\mathbf{a}_j$, una componente exterior $\mathbf{y}_j = \mathbf{X}_j \mathbf{a}_j$ (que representa las combinaciones lineales mediante las cuales se determinaran las relaciones entre los grupos) y una componente interior definida como sigue

$$\mathbf{z}_j = \sum_{k=1, k\neq j}^{J} c_{jk} w(Cov(\mathbf{y}_j, \mathbf{y}_k))\mathbf{y}_k,$$

donde la función $w(x)$ será igual a 1, x o $sign(x)$, denominados esquema de Horst, factorial o centroide, respectivamente.

Además sea g la función identidad, cuadrada o valor absoluto (para el esquema de Horst, factorial o centroide, respectivamente). Por último sean $\tau_1, ..., \tau_J$ constantes de contracción.

Entonces, el problema de querer encontrar las relaciones existentes entre cada grupo de variables, se reduce a querer maximizar:

$$\sum_{j,k=1, j\neq k}^{J} c_{jk} g(Cov(X_j \mathbf{a_j}, X_k \mathbf{a}_k)), \tag{1}$$

sujeto a las restricciones

$$\tau_j \left\|\mathbf{a}_j\right\|^2 + (1-\tau_j) Var(X_j \mathbf{a}_j) = 1,$$

para $j = 1, ..., J$. Mediante esta maximización obtendremos las ecuaciones $\mathbf{a}_1, ..., \mathbf{a}_J$ donde

$$\mathbf{a}_j = \left[\mathbf{z}'_j \mathbf{X}_j \left[\tau_j \mathbf{I} + (1-\tau_j)\frac{1}{n}\mathbf{X}'_j \mathbf{X}_j\right]^{-1} \mathbf{X}'_j \mathbf{z}_j\right]^{-1/2} \left[\tau_j \mathbf{I} + (1-\tau_j)\frac{1}{n}\mathbf{X}'_j \mathbf{X}_j\right]^{-1} \mathbf{X}'_j \mathbf{z}_j \tag{2}$$

para $j = 1, ..., J$.

3. Aplicación

Se cuenta con una base de datos con información obtenida de un ecosistema de manglar, del sistema lagunar de Chacahua-Pastorías en Oaxaca, Chan et al. (2012), específicamente se trabajó con los datos de la laguna de Pastorías. Se aplicó el ACCRG a tres grupos de variables, cada variable con 10 individuos: el primer grupo con las variables físicas (ramas, misceláneos y hojarasca) de la estructura forestal del bosque de manglar, el segundo grupo con las variables físico-químicas (salinidad, nitrato, sulfato, amonio, fosfato, PH, REDOX y temperatura) del agua intersticial y el tercer grupo con las variables de productividad primaria de los bosques de mangle (hojas, flores, ramas y estípulas). El objetivo es determinar el grado de relación entre los grupos de variables.
Los datos fueron procesados con el paquete RGCCA en el lenguaje R, desarrollado por Tenenhaus (2011).

Se usó como matriz de diseño la siguiente

$$\mathbf{C} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

Se obtuvieron los siguientes valores óptimos de las constantes de contracción

$$\tau_1 = 0.9124262 \quad \tau_2 = 0.2359993 \quad \tau_3 = 0.3842536.$$

Así también se obtuvieron los siguientes primeros vectores canónicos para cada grupo denotados por $\mathbf{a}_j$ para $j = 1, ..., J$, dados en (2). Estas nuevas variables son presentadas en la Tabla 1.

Las $\mathbf{a}_i$ para $i = 1, 2, 3$, son los coeficientes estimados tal que maximizan (1). Se puede observar en la tabla que en $\mathbf{a}_1$ la variable PO4 con el valor -0.691 es la más representativa del grupo así como salinidad y PH. En $\mathbf{a}_2$ se encuentran los frutos con el valor del coeficiente de 0.839 así como las hojas con 0.651. Finalmente en $\mathbf{a}_3$ las ramas con el valor del coeficiente de 0.373 y los miscelaneos con 0.304.

En la Figura 1 se presentan las relaciones existentes entre las $\mathbf{Y}_i$, que son las combinaciones lineales que son obtenidas para cada grupo, una vez obtenidas las $\mathbf{a}_i$. Las combinaciones

Vectores Canónicos					
$\mathbf{X}_1$	$\mathbf{a}_1$	$\mathbf{X}_2$	$\mathbf{a}_2$	$\mathbf{X}_3$	$\mathbf{a}_3$
NO3	0.233	Estípulas	0.228	Rama	0.373
PO4	-0.691	Hoja	0.651	Misceláneos	0.304
SO4	-0.242	Flor	-0.070	Hoj/día	0.257
NH4	-0.282	Fruto	0.839	Hoj/mes	0.142
Salinidad	-0.383			Hoj/año	0.257
Redox	0.060				
PH	-0.340				
Temperatura	0.008				

Tabla 1: Primer vector canónico para cada grupo de variables, dados en $\mathbf{a}_i$.

$\mathbf{Y}_2$ y $\mathbf{Y}_3$ presentan una relación lineal mas notable con respecto a las otras parejas. Esto indicaría que el grupo de las variables físico-químicas y el grupo de las variables de producción primaria presentan un grado de relación alto.

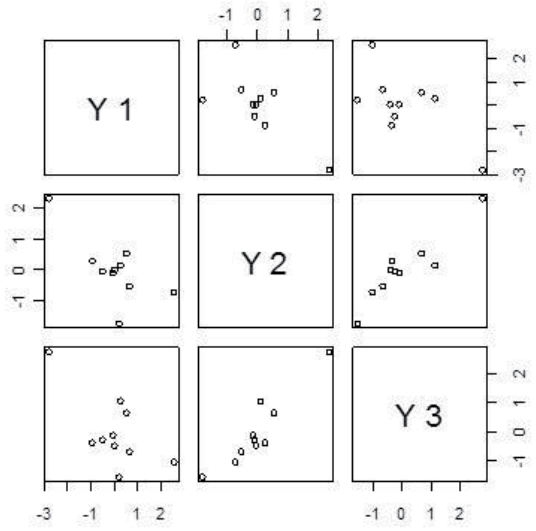


Figura 1: Gráfica de las $\mathbf{Y}_i$, que representan la combinación lineal para cada grupo.

4. Conclusiones

Se presentó el Análisis de Correlación Canónica Regularizada Generalizada para dar solución al ACC cuando de tienen tres o mas grupos de variables. En la laguna Pastorías se pueden observar las relaciones existentes entre algunas variables químicas del agua intersticial con la producción de frutos, que indica que estos nutrientes se ven reflejados en la producción de frutos. Además se muestran altas relaciones entre los parámetros de producción y la estructura forestal del bosque de manglar, principalmente en la producción de hojas. Se observó que el grupo de las variables físico-químicas está altamente relacionado con el grupo de las variables de producción primaria.

Referencias

1. Chan, C.A., Linares, G. y Agraz, C. (2012). "Los manglares del Complejo Lagunar Chacahua Pastorías, Oaxaca". Memorias del XI Congreso Internacional y XVII Congreso Nacional de Ciencias Ambientales. ISBN: 923-546-687-4.

2. González, I., Déjean, S., Martín, P. y Baccini, A. (2008). "CCA: An R package to extend canonical correlation analysis". *Journal of Statistical Software*, 23(12): pp. 1-14.

3. Hair, J. F., Anderson, R. E., Tatham, R. L. y Black, W. C. (1998). "Multivariate Data Analysis". Quinta edición. Editorial Prentice Hall.

4. Hotelling, H. (1936)."Relations between two sets of variates". *Biometrika* 28: pp. 321-377.

5. Johnson, R. A. y Wichern, D. W. (2007). "Applied Multivariate Statistical Analysis". Sexta edición. Editorial Prentice Hall. New Jersey.

6. Mardia, K., Kent, J. y Bibby, J. (1979). "Multivariate Analysis". Academic Press.

7. Samarov, D. V. (2009). "The Analysis and Advanced Extensions of Canonical Correlation Analysis". Tesis de Doctorado en Matemáticas. Departamento de Estadística e Investigación de Operaciones. Universidad de Carolina del Norte.

8. Tenenhaus, Z. (2011). "Regularized generalized canonical correlation analysis". *Psychometrika*, 76(2): pp. 257-284.

Diagnóstico de la Educación Estadística en la Universidad Veracruzana

Cecilia Cruz López[a], Mario Miguel Ojeda Ramírez
Universidad Veracruzana

Clasificación: Tesis de Doctorado
Área: Educación Estadística
Subárea: Análisis de Correspondencia Múltiple
Trabajo presentado en: XXVIII Foro Nacional de Estadística

1. Introducción

La Universidad Veracruzana (UV) tiene 69 años de antigüedad y desde su creación ha sido la Institución de Educación Superior con la mayor cobertura en el Estado de Veracruz, ya que se encuentra distribuida en cinco campus universitarios y con presencia en 28 municipios. Dentro de su oferta académica cuenta con 172 programas de licenciatura en los que atienden una matrícula de 59 476 estudiantes. La UV desde 1999 utiliza un Modelo Educativo Integral Flexible (MEIF) y los programas de asignatura dentro de la institución son considerados como Experiencias Educativas (EE), en ellos se promueven diversos aprendizajes que permiten al estudiante trascender más allá del aula y llevar el conocimiento adquirido a la práctica profesional. El área Económico-Administrativa dentro de su oferta académica tiene la carrera de Ciencias y Técnicas Estadísticas que cuenta con un grupo de profesores preocupados por la Educación Estadística, y han creado una red en colaboración con otras instituciones de Educación Superior, para realizar el Diagnóstico de la educación estadística en el país, comenzando por la UV. Esto nos llevará a tomar medidas de mejora en el proceso de enseñanza-aprendizaje dentro de la institución. Para comenzar el diagnóstico, Batanero (2000) recomienda que se lleve a cabo el análisis del currículo para tomar acciones que lleven a la mejora del proceso de Enseñanza-Aprendizaje, por lo que, se realizó el análisis de los

[a] autor_responsable ceccruz@uv.mx

programas de las EE de Estadística. Posteriormente se evaluó a los profesores y estudiantes a través de las Metas de Aprendizaje de la Estadística establecidas por Gal y Garfield (1997), éstas son reconocidas como líneas de innovación internacional, donde se indica que un estudiante que haya llevado un curso introductorio de estadística debe dominar ocho metas. El estudio en la UV se llevó a cabo durante el periodo enero-junio 2013 y el objetivo es determinar, respecto a las líneas internacionales de innovación de la estadística, el estado general de la educación en esta disciplina en la Universidad Veracruzana.

2. Materiales y Métodos

La recolección de los programas de las EE se llevó a cabo descargando de internet los que estaban disponibles a través del Sistema Integral de Información Universitaria de la UV (SIIU), Finalmente se recolectaron 116 programas que están clasificados por área académica. Para el análisis de los programas se diseño una lista de cotejo que consta de dos secciones, la primera hace la evaluación de las competencias en cada programa definidas segun Tobón, 2006. La segunda sección en la lista evalúa en qué medida son consideradas las metas de aprendizaje.

Para la evaluación de las competencias y sus componentes se crearon seis preguntas con cuatro respuestas de clasificación. Y para evaluar en qué medida son consideradas las metas se utilizaron como dimensiones las ocho metas y cómo variables los diversos ámbitos que en ellas se manejan. De tal forma que se construyeron indicadores que nos permitieron medir el grado de inclusión de las metas. Las categorías para cada índice son nivel básico, medio y avanzado. Para la aplicación de la encuesta a los profesores y estudiantes se acudió con los directores y jefes de carrera en las 5 regiones de la UV y ellos proporcionaron los nombres de los profesores que imparten las EE de estadística, en total se tuvieron 84 profesores y 258 estudiantes. Finalmente, para todos los casos se construyó un indicador global llamado Dedicación Total de las Metas de Aprendizaje de la Estadística (DTMAE).

3. Resultados

3.1 Evaluación de Competencias y Metas en los programas de las EE

Al analizar los programas observamos que más de la mitad (56 %) tienen suficientemente bien definida la Unidad de Competencia, sin embargo, un porcentaje amplio muestra programas que no tienen bien definido este concepto o simplemente no lo contienen (18 %). El mismo comportamiento se presenta en la Descripción de los Contenidos, debido a que el 72 % de los programas los describen suficientemente bien para cumplir la competencia. Asimismo, las Evidencias de Desempeño (68 %) también son lo suficientemente acordes para cumplir la competencia. Un aspecto importante en donde se debe considerar especial atención, es en la definición de los Criterios de Desempeño, ya que 40 % de los programas cayó en la categoría de: solo levemente son acordes a las Evidencias. En cuanto a las Fuentes de Información Básica el 89 % de los programas especifíca que los libros de texto son de acuerdo al contexto del curso. El 85 % de los programas especifican el uso de un Software Estadístico, aunque la mayor parte de ellos (60 %) no menciona el nombre del software que utilizan. En cuanto a las Metas de Aprendizaje de la Estadística la meta 3 es la más considerada en todos los programas de las EE con un 81 % en el nivel avanzado y la meta 7 la menos considerada con un 12 % en el mismo nivel. El Índice de Inclusión Total de las Metas indica que el 56 % de los programas incluyen en un nivel avanzado las metas de aprendizaje, este comportamiento se da con más frecuencia en los programas del área técnica (74 %) y el 22 % de los programas las incluyen en un nivel básico, los programas que observan este patrón son del área de Ciencias de la Salud (54 %). En el análisis de correspondencia múltiple se observa que el área de Ciencias de la Salud solamente incluye las metas en su nivel básico especialmente en los programas de Epidemiología y Bioestadística, Estadística en ciencias de la salud, Estadística básica y Taller de estadística. En cuanto al área Económico-Administrativa únicamente se están incluyendo las metas de aprendizaje en el nivel medio, específicamente en los programas de Estadística y Estadística inferencial. (ver Figura 1).

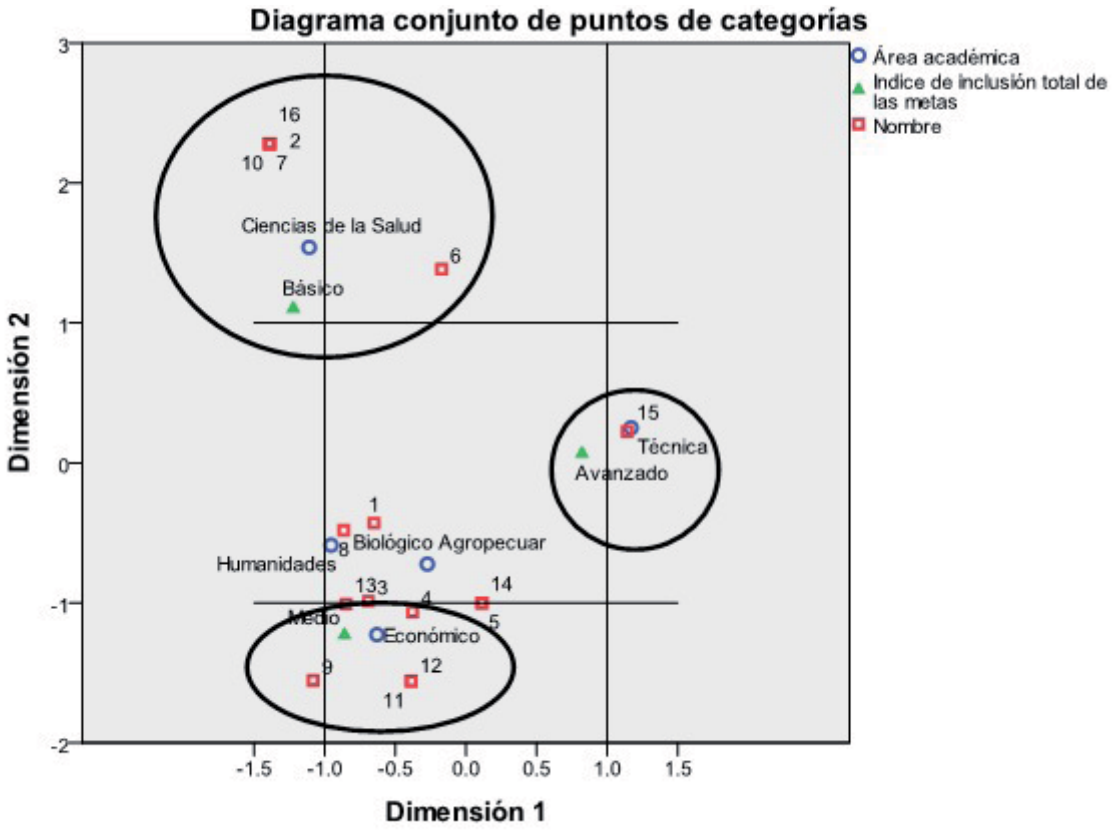


Figura 1. Análisis de correspondencia múltiple entre área académica, índice ITMAE y nombre del programa de la EE.

3.2 Análisis de los profesores

La mayoría de los profesores participantes son de sexo masculino (61 %). Del total de profesores participantes el 90 % tiene estudios de posgrado y el 43 % es de tiempo completo. La edad promedio de los profesores participantes es de 47 años con una desviación estándar de 12. Al analizar los indicadores creados para cada meta se observa que la meta 2 y 3 son las que presentan en su categoría avanzada el mayor porcentaje de dedicación ambas con el 83 %, y la meta que menos dedicación le dan es la meta 4 con un 42 % en su categoría avanzada. El DTMAE indica que el 44 % de los profesores mencionan dedicar en un nivel avanzado al uso de las metas, este comportamiento se da en su mayoría en los profesores del área Económico-Administrativa (14.3 %) y el 23 % de los profesores indican el uso de las metas en un nivel básico, esto se da en profesores del área Técnica (9.5 %).

En el análisis de correspondencia para los profesores se observa (Figura 2), que en las carreras de Física (13), Instrumentación electrónica (18) y Pedagogía (24) se aplican todas las metas en un nivel básico. Otra relación importante en el gráfico es que en las carreras de Ambiental (3), Educación física (11), Ingeniería industrial (16), Ingeniería mecánica- eléctrica (19) y Producción agropecuaria (25) las metas son usadas en su nivel medio.

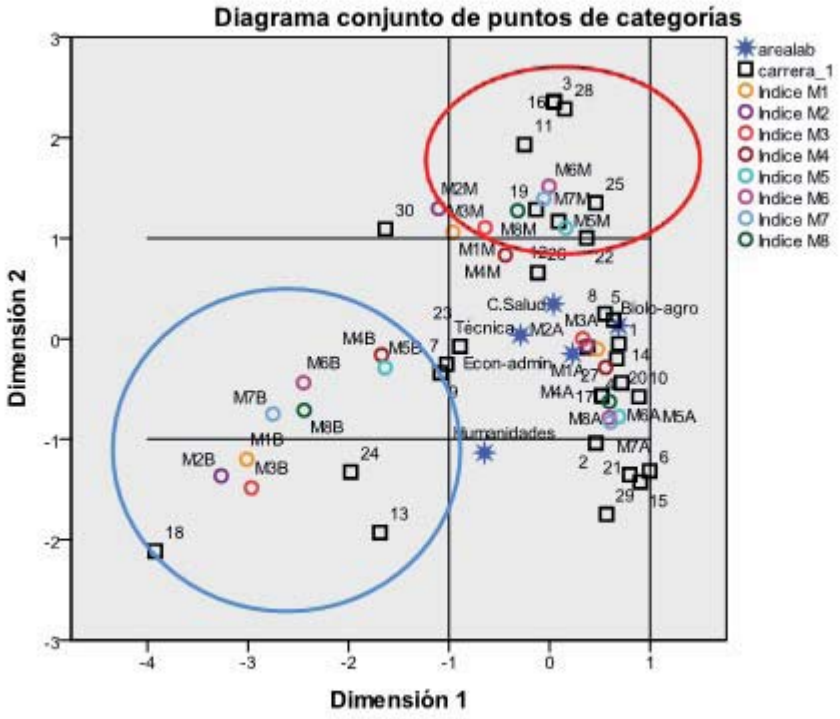


Figura 2. Diagrama de Correspondencia múltiple entre área laboral, carrera e índices.

3.3 Análisis de los estudiantes

Los estudiantes mencionaron que las metas 2, 3, 4, 7 y 8 son cumplidas en mayor porcentaje en sus niveles medios. El indicador global de dedicación total de las metas DTMAE indica que el 44 % de los estudiantes mencionan en un nivel avanzado al uso de las metas de aprendizaje en los cursos, este comportamiento se da en el área técnica (34.5 %) y el 8 % de

los estudiantes indican el uso de las metas en un nivel básico, esto en su mayoría se da en el área de Ciencias de la salud (40.9 %).

En el análisis de correspondencia múltiple se observa que en los cursos Estadística descriptiva (7) del área Económico y Taller de estadística (15) del área de Ciencias de la Salud se encuentra una relación con las categorías bajas de todas las metas. Otra relación importante es que en los cursos de Bioestadísica (1) (área Biológico-Agropecuaria) y Estadística inferencial (11) (área Humanidades) las metas son usadas en su nivel alto, en particular la 5, 6 y 7.

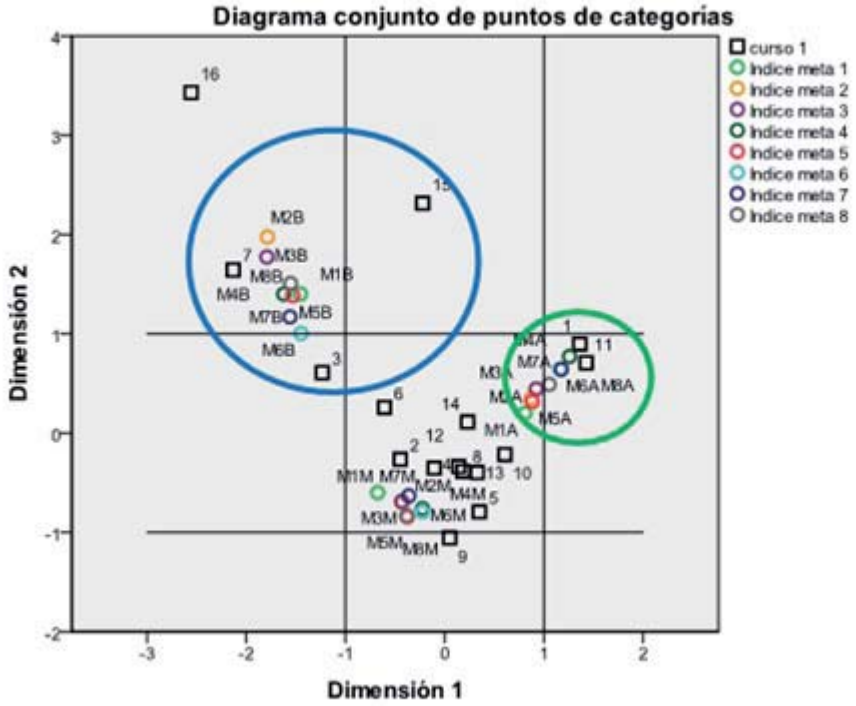


Figura 3. Diagrama de correspondencia múltiple de estudiantes entre cursos e índices de las metas.

4. Conclusiones

Se identificaron las áreas académicas en donde se requiere especial atención en la inclusión de las metas, así como los programas de asignatura que deben ser reestructurados tomandolas en

cuenta. Asimismo, se detectaron las carreras en las cuales los profesores necesitan intensificar el uso de las metas de aprendizaje, por lo que se harán las recomendaciones necesarias a las academias encargadas de los cursos de estadística. Finalmente, se destaca que los resultados reflejan que se están haciendo esfuerzos significativos para mejorar la impartición de los cursos y cada vez son más profesores que incluyen de manera global las metas da aprendizaje, logrando con ello el mejor aprovechamiento por parte del estudiante y quizá con el tiempo poder asegurar que la enseñanza-aprendizaje de la estadística está logrando salir del rezago en el que se encuentra.

Referencias

1. Batanero, C. (2000). ¿Hacia dónde va la educación estadística? Blaix 15, 2-13.

2. Gal, I. and Garfield J. (1997). Curricular Goals And Assessment Challenges. In Gal I. Garfield J. (1997). (Edts) The Assessment Challenge in Statistics Education. IOS Perss, ISI, Voorburg. The Netherlands.

3. Tobón, S. (2006). Aspectos básicos de la formación basada en competencias. Documento de trabajo, 2006, 1-8.

Identificación de Áreas de Riesgo para Citologías Positivas del Programa IMSS Solidaridad del Estado de Durango, Mediante el Programa SIGEPI*

Edgar Felipe Lares Bayona [a], Luis Francisco Sánchez Anguiano[b], Francisco Sandoval Herrera[c],

1. Introducción

El Papanicolaou es barato, indoloro y preciso para el diagnóstico de infecciones, lesiones premalignas y malignas del cérvix, por lo que ha formado parte de la rutina en la exploración anual de las mujeres en la etapa reproductiva de la vida. La citología exfoliativa se ocupa del estudio de las células descamadas de los tejidos, tanto en condiciones normales como patológicas. En la mayor parte de los casos estos tejidos son epiteliales. La facilidad del acceso al cuello uterino, para el estudio de las células, los tejidos y para el examen físico, ha permitido el estudio microscópico de las células exfoliadas del cérvix y la vagina con tinción de Papanicolaou, que constituye el método de elección para identificar oportunamente lesiones inflamatorias, precancerosas, cancerosas e infecciosas; (Sánchez, Reyes, and Lares, B.E.F. 2010). SIGEpi es un Sistema de Información Geográfica (SIG) diseñado para aplicaciones en Epidemiología y Salud Pública. Ofrece una compilación de técnicas, procedimientos y métodos para el análisis de datos epidemiológicos. Los mismos se presentan de manera simplificada, en un ambiente amigable y en múltiples idiomas; (SIGEpi 2005). Dentro de un marco de análisis de la vulnerabilidad, herramientas como SIGEpi permiten la integración de las medidas y de los indicadores de diferentes fuentes, y los colocan en un espacio común

*
[a] elbfelipe@hotmail.com
[b] I. de Investigación Científica de la Universidad Juárez del Estado de Durango
[c] I. Mexicano del Seguro Social de Durango

para la estadística y análisis geográfico, (Nájera 2001). Utilizando un Sistema de Información Geográfica en los problemas de Salud Pública y Salud Reproductiva, se tendría a la mano información valiosa con determinación clara para realizar estrategias adecuadas en la solución de dicha problemática.

2. Marco Teórico

Citología Cervicovaginal de Papanicolaou. Esta prueba fue descrita por George Papanicolalou del cual lleva su nombre, es una prueba de tamizaje no de diagnóstico, es relativamente sencilla de tomar y de procesar por personal adecuadamente entrenado, su interpretación microscópica requiere citotecnólogos y patólogos muy bien capacitados, para que sea un estudio de calidad. Se recomienda que la toma se haga del exo (es la parte del cérvix que se continúa con la vagina) y el endocervix utilizando espátula de Ayre y Citobrush, (A.L.F., R.M., and B.E.F. 2012). La prueba de mayor utilidad para diagnosticar displasias y cáncer cervicouterino fue la citología cervical, ya que detectó todos los casos verdaderos positivos confirmados por biopsia, comparada con la inspección visual con ácido acético, ésta última técnica tuvo mayor porcentaje de casos falsos positivos que la citología cervical, (H.N., A.L.F., and Lares B.E.F. 2010). SIGEpi es un producto desarrollado por el Área de Análisis de Salud y Sistemas de Información (AIS) de la Organización Panamericana de la Salud (OPS) como parte del Proyecto de Cooperación Técnica, Aplicación de los Sistemas de Información Geográfica en Epidemiología y Salud Pública, (SIGEpi 2005). El poder de los SIG radica en su capacidad de mostrar la distribución espacial de una predicción o un resultado relacionado a la salud. Estos mapas se pueden utilizar ya sea para generar o probar hipótesis que no se podría de otra manera producir por el investigador sin visualizar las relaciones espaciales. El tipo de aplicación SIG utilizado depende del tipo de datos que el investigador tiene y la pregunta de investigación. Los datos Geoestadísticos proveen el conteo o numeros en un lugar determinado, (Parchman 2002).

Objetivos. Describir la distribución de las citologías cervicales positivas en el Estado de Durango del programa IMSS Solidaridad, e identificar áreas de riesgo adherentes a la positividad citológica y sus rangos en promedios de edad mediante el software de sistemas de información geográfica SIGEPI.

Método. Es un estudio descriptivo y transversal. Se capturó mediante una base de datos

(Visual Fox Pro 5.0) un total de 29601 hojas de solicitud y reporte de resultados de citología cervical durante 2005 y 2011 para el Estado de Durango. Para su localización geográfica se utilizó el programa informático SIGEPI versión 1.4. Se generaron mapas específicos que permitieron identificar, cuantificar y visualizar las áreas con mayor representatividad para la citología positiva.

3. Resultados

Se encontraron 100 citologías positivas (0.4 %) en el Estado de Durango del 2005 al 2011, Tabla 1. De las citologías cervicales positivas se obtuvo lo siguiente: frecuencias más altas en los municipios de Vicente Guerrero, Mezquital y Guadalupe Victoria, con 13, 12 y 9 casos respectivamente, Figura 1; sobre los porcentajes más altos fueron en los municipios de Nuevo Ideal, Canelas y San Pedro del Gallo, con 1.26 %, 1.21 % y 1.1 % respectivamente, Figura 2. Un 37 % tuvo inicio de vida sexual antes de los 18 años, Tabla 2; en cinco municipios todas iniciaron vida sexual antes de los 18 años, Figura 3. Un 17 % tenían múltiples parejas sexuales, Tabla 3; en dos municipios todas tenían múltiples parejas sexuales, Figura 4. La edad promedio y desviación estándar (DE) fue de 49.4 ± 16.6 años respectivamente, mínimo de 22 años y máximo de 85 años, Tabla 4; ocho municipios tuvieron promedios por arriba de los 50 años y tres por debajo de los 35 años. En nueve municipios no se obtuvieron resultados citológicos positivos, Figura 5.

Tabla 1. Prevalencia de citologías cervicales positivas en el Estado de Durango.

	Frecuencia	Porcentaje	Porcentaje Válido
Positivas	100	0.3	0.4
Negativas	28200	95.3	99.6
Total	28300	95.6	100
Perdidos del sistema	1301	4.4	
Total	29601	100	

Figura 1.Población del programa IMSS Solidaridad durante 2005 a 2011 y frecuencias citologías positivas cervicales por municipio de Durango.

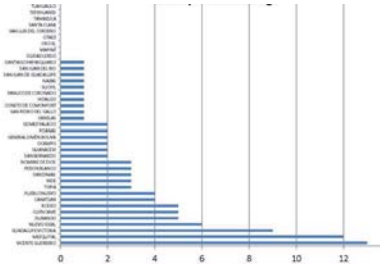


Figura 2. Mapa con rangos porcentuales de las citologías positivas por municipio del estado de Durango. Población del programa IMSS Solidaridad.

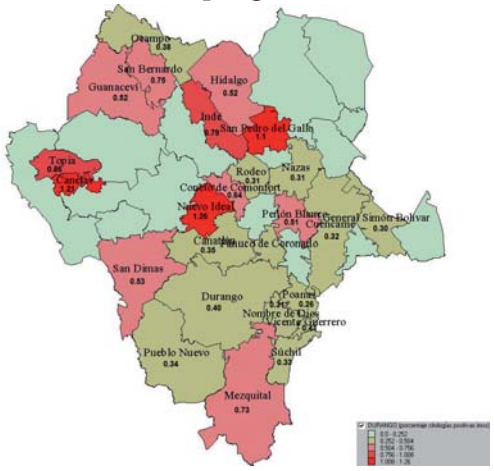


Tabla 2. Inicio de vida sexual activa antes de los 18 años de las citologías cervicales positivas de Durango.

	Frecuencia	Porcentaje
Si	37	37.0
No	63	63.0
Total	100	100

Población del programa IMSS Solidaridad durante 2005 a 2011.

Figura 3. Mapa de rangos porcentuales de las citologías cervicales positivas con inicio de vida sexual activa antes de los 18 años en Durango. Población cautiva del programa IMSS Solidaridad.

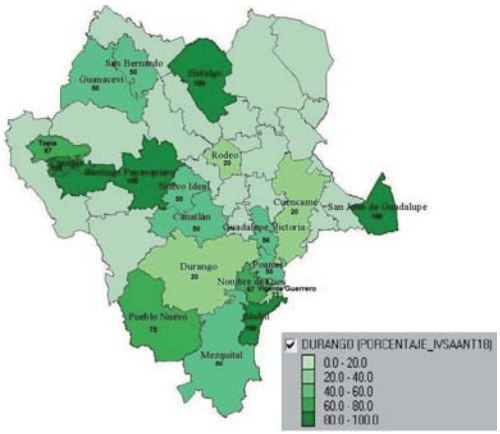


Tabla 3. Múltiples parejas sexuales de las citologías cervicales positivas de Durango.

	Frecuencia	Porcentaje
Si	17	17.0
No	83	83.0
Total	100	100

Población del programa IMSS Solidaridad durante 2005 a 2011.

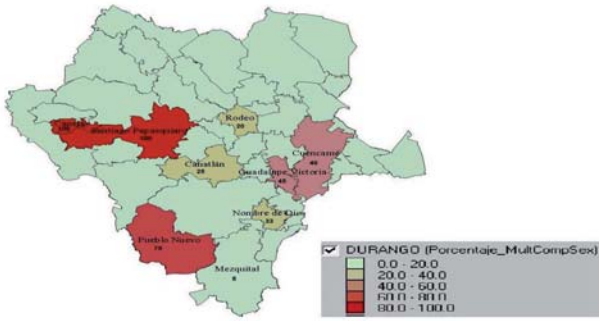


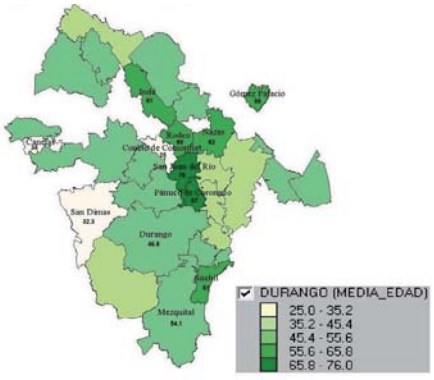
Figura 4. Mapa de rangos porcentual de las citologías positivas con múltiples parejas sexuales en el estado de Durango. Población cautiva del programa IMSS Solidaridad.

Tabla 4. Descripción de la población de citologías cervicales positivas por edad del estado de Durango.

	N	Mínimo	Máximo	Media	Desv. Est.
Edad	94	22	85	49.4	16.6
Perdidos en el sistema	6				
Total	10				

Población del programa IMSS Solidaridad durante 2005 a 2011.

Figura 5. Mapa de Intervalos de Medias de Edad de las positivas del Estado de Durango. Población cautiva del programa IMSS Solidaridad.



NOTA: Municipios que no tienen datos de positividad son eliminados del mapa de medias de edad.

Conclusiones

La citología es un procedimiento importante para determinar las complicaciones de aspecto ginecológico que pueden conducir a una mortalidad cuando no se tiene el seguimiento adecuado. El municipio de Canelas fue identificado con los factores de riesgo mas notables de acuerdo a los mapas. Los porcentajes de mayor riesgo de acuerdo a los mapas de citologías cervicales positivas, múltiples parejas sexuales, inicio de vida sexual antes de los 18 fueron y edad de las participantes, fueron en aquellos municipios menos poblados. El programa SIGEPI constituye una opción adecuada para estudiar los problemas de salud publica y resulto eficaz para identificar las áreas con mayor representatividad porcentual de las citologías positivas.

Referencia

A.L.F., Sánchez, Reyes R.M., and Lares B.E.F. , 2012. *Genotificación del Virus del Papiloma Humano en Durango México.* México: Editorial Publicia, Impreso en Alemania por AV KADEMISKERVELAG GMBH & CO.KG ALLE RECHTE VORBEHALTEN. Todos los derechos reservados. SAARBRUCKEN.

H.N., Velázquez, Sánchez A.L.F., and Lares B.E.F. 2010. "Comparación de la utilidad diagnóstica entre la inspección visual con ácido acético y la citología cervical." *Revista Ginecología y Obstetricia de México* 78:261–267.

Nájera, P., and Martínez R. and Vidaurre M. and Loyola E. and Castillo Salgado C. and Eisner, C. 2001. "Use of SIGEpi for the Indentification of Localities Vulnerable to Environmental Risks in México." *PAHO/WHO,Epidemiological Bulletin* 22:3.

Parchman, M.L., and Ferrer RL. and Blanchard, K.S. 2002. "Geography and geographic information systems in family medicine research." *PubMed indexed for MEDLIN, http://www.ncbi.nlm.nih.gov/pubmed/11874023.*

Sánchez, A.L.F., R. M. Reyes, and Lares, B.E.F. 2010. "El Virus del Papiloma Humano y el Carcinoma Cervicouterino." *Investigación y Educación en Salud Pública. La casa editorial de Durango, 1ra. edición, México, http://www.ncbi.nlm.nih.gov/pubmed/1187402*, pp. 149–164.

SIGEpi. 2005. *Organización Panamerica de la Salud.Sistemas de Información Geográfica en Epidemiología y Salud Pública.* http://ais.paho.org/sigepi/index.asp?xml=sigepi/index.htm.

Una carta EWMA con intervalo de muestreo variable para monitorear procesos de calibración*

María Guadalupe Russell Noriega[a]
Universidad Autónoma de Sinaloa

Enrique Villa Diharce[b]
Centro de Investigación en Matemáticas, A.C.

Clasificación: Trabajo de Investigación.
Área: Control Estadístico de Procesos
Subárea: Cartas de control y Procesos de calibración de instrumentos de medición
Trabajo presentado en: XXVIII Foro Nacional de Estadística

1. Introducción

El monitoreo de los procesos de calibración de instrumentos de medición se realiza como parte de cualquier programa de aseguramiento de calidad de mediciones, para garantizar la calidad de los resultados de las calibraciones, Croarkin y Varner (1982). Las cartas de control propuestas inicialmente por Shewhart (1931), para el monitoreo de procesos, son cartas univariadas, que se utilizan para monitorear una característica de calidad de un producto, que se mide periódicamente y cuya variación sigue una distribución normal. Estas cartas son gráficas que tienen tres líneas horizontales: una línea central que se encuentra a una altura igual a la media de la distribución normal y dos más ubicadas a una distancia de tres desviaciones estándar de la media. En esta carta se grafican los valores de la característica de interés, que se miden periódicamente, en cada una de las unidades que se monitorean. Si el

* Este trabajo fue realizado con el auspicio de la Universidad Autónoma de Sinaloa, a través del proyecto PROFAPI2013/181, "Métodos Estadísticos con Aplicaciones en Metrología"
[a] mgrussell@uas.edu.mx
[b] villadi@cimat.mx

patrón gráfico de dichos valores no corresponde a una distribución normal, lo interpretamos como una señal que nos indica que el proceso está fuera de control. Cuando esto ocurre, revisamos el proceso para encontrar y eliminar las fuentes específicas de variación que han causado la salida de control del proceso. Con esta operación, mantenemos el proceso en control estadístico.

En 1982 C. Croarkin propuso una carta de control para el monitoreo de procesos de calibración, basada en las desviaciones de los valores medidos y los patrones, dadas en la Sección 2. A esta carta, la denominamos carta NIST, debido a que ha sido publicada en el manual de estadística del NIST (National Institute of Standards and Technology) NIST/SEMATECH (2011). Posteriormente en la literatura de control estadístico de procesos (CEP) se publicaron varias cartas de control para perfiles lineales que se aplicaron al monitoreo de procesos de calibración lineal en la fase II. Al comparar la carta de Croarkin, con las cartas anteriores, la carta EWMA3 de Kim et al. (2003), supera claramente a la primera. Esta ventaja se debe a que incorpora el procedimiento de medias móviles con ponderación exponencial (EWMA, por sus siglas en inglés), que acelera la detección de cambios pequeños en el proceso monitoreado.

Para cada uno de los métodos utilizados en la construcción de cartas, representamos por x_i el valor del $i-$esimo nivel del patrón de referencia y por y_{ij} la $j-$esima lectura del $i-$esimo nivel del patrón de referencia utilizado en el proceso de calibración. Cuando el proceso de calibración esta en control estadístico, el modelo fundamental es,

$$y_{ij} = \beta_0 + \beta_1 x_i + \epsilon_{ij},$$

donde $i = 1, 2, .., n, \quad j = 1, 2, ..., m$. Suponemos que los términos ϵ_{ij} son variables aleatorias i.i.d. con distribución normal, con media cero y varianza σ^2. Dado que estamos considerando la Fase II del monitoreo, suponemos que los valores de los parámetros de control β_0, β_1 y σ^2 son conocidos. En este trabajo proponemos una modificación a la carta NIST, incorporando el procedimiento EWMA a la estadística de control de Croarkin y Varner, resultando la carta que denotamos por EWMA/NIST. Así mismo incorporamos el esquema de intervalos de muestreo de longitud variable en la carta EWMA/NIST obteniendo así, la carta VSI-EWMA/NIST.

2. Métodos de monitoreo de perfiles

En la construcción de la carta de control NIST de Croarkin y Varner, suponemos que los datos siguen el modelo de regresión lineal dado en la Introducción. El método esta basado en el principio de calibración inversa, de esta manera obtenemos la estadística Z_{ij} que graficamos en la carta de control en el tiempo de la jésima muestra,

$$Z_{ij} = \frac{y_{ij} - \beta_0}{\beta_1} - x_i, \quad i = 1, 2, ..., n.$$

Los autores recomiendan que se hagan tres mediciones ($n = 3$) en el proceso de calibración, dos en los extremos y una en el centro. Los límites de control inferior y superior para la carta son,

$$LCL = -s_c z_{\alpha^*}, \quad UCL = s_c z_{\alpha^*},$$

donde $s_c = \sigma/\beta_1$, siendo σ la desviación estándar conocida y β_1 el valor conocido en control de la pendiente. El valor de z_{α^*} corresponde al cuantil superior de la distribución normal estándar. Obtenemos la probabilidad α^* con la corrección de Bonferroni, $\alpha^* = \{1 - \exp[\log(1 - \alpha)/n]\}/2$, donde α se elige de tal manera que tengamos una Longitud Promedio de Corrida (ARL) en control, usando la relación $ARL_0 = 1/\alpha$. Los límites de control, se construyen utilizando el cuantil de la distribución normal estándar, en vez de la distribución t de Student, como lo hacen en Croarkin y Varner (1982), ya que los valores de los parámetros de control son conocidos. En las siguientes subsecciones mostramos las dos modificaciones que proponemos para la carta NIST.

2.1 Método EWMA/NIST

Nuestra propuesta consiste en una versión EWMA de la carta NIST, para acelerar la detección de cambios en el modelo de calibración lineal. En esta carta, que denominamos EWMA/NIST, monitoreamos los valores medidos corregidos Z_{ij}, definidos anteriormente, utilizando la estadística EWMA, denotada por $W_i(j), i = 1, 2, ..., n,\ j = 1, 2, ..., m$ y dada por

$$W_i(j) = \theta Z_{ij} + (1 - \theta) W_i(j - 1),$$

con $W_i(0) = 0$. Como en todas las cartas, θ es una constante de suavizamiento cuyo valor determina la magnitud de los cambios que detecta. Es decir, se detecta una señal de fuera

de control, tan pronto como $W_i(j)$ sea menor que el límite de control inferior LCL, o mayor que el límite de control superior UCL, donde

$$LCL = -L\frac{\alpha}{\beta_1}\sqrt{\frac{\theta}{2-\theta}}, \quad UCL = L\frac{\alpha}{\beta_1}\sqrt{\frac{\theta}{2-\theta}}.$$

La constante L se elige de tal forma que el ARL en control, tenga un valor específico.

2.2 Método VSI-EWMA/NIST

Con el fin de aumentar la capacidad de detección de la carta EWMA/NIST, hacemos una modificación adicional que consiste en considerar variable el intervalo de muestreo, obteniendo así la carta NIST tipo EWMA con intervalo de muestreo variable, esto es, la carta VSI-EWMA/NIST (esto es similar a la propuesta de Li y Wang para la carta EWMA3). En esta carta, el intervalo de muestreo se toma de forma variable, con una longitud mayor si el punto de la carta tipo EWMA cae cerca del valor objetivo (target) y con una longitud menor si el punto cae cerca de los límites de control, ya que en este caso asumimos que esto significa que el proceso ha cambiado.

Sean l_1 y l_2 las longitudes de los intervalos de muestreo, donde $0 < l_1 < l_2$. Cuando tomamos una muestra y calculamos la estadística EWMA correspondiente *Wi(j),* la muestra siguiente se tomará después de un intervalo determinado por la siguiente función del intervalo de muestreo,

$$\left\{ \begin{array}{ll} l_1 & \text{si } W_i(j) \in R_1 \\ l_2 & \text{si } W_i(j) \in R_2 \end{array} \right\}.$$

La región que corresponde a un estado en control, dada por (LCL, UCL), se subdivide en las dos regiones R_1 y R_2, donde $R_1 = (-LCL, -w) \cup (w, UCL)$ y $R_2 = (-w, w)$. Cuando el proceso de calibración inicia, es recomendable que la primera muestra se tome con el intervalo de muestreo pequeño l_1.

3. Comparaciones

Comparamos el desempeño de las cartas mencionadas, para la Fase II, mediante un estudio de simulación en el cual se generan 10 000 muestras bajo el modelo de calibración lineal

$y_{ij} = 3 + 2x_i + \epsilon_{ij}$, en control, donde los ϵ_{ij} son v.a.i.i.d. como normales con media cero y varianza 1. Consideramos los siguientes desplazamientos de los parámetros del modelo de regresión lineal, que se muestran en la Tabla 1.

Tabla 1: Cambios considerados en la comparación de los métodos.

Cambio en	Notación	Valores
Intercepto	de β_0 a $\beta_0 + \lambda\sigma$	$\lambda = 0.2, 0.4, ..., 2.0$
Pendiente	de β_1 a $\beta_1 + \gamma\sigma$	$\gamma = 0.025, 0.050, .., 0.250$
Desv. estándar	de σ a $\delta\sigma$	$\delta = 1.2, 1.4, .., 3.0$

En las Figuras 1 (a y b) observamos que el método de monitoreo EWMA/NIST propuesto supera notablemente el desempeño de la carta NIST y resulta equiparable con el método EWMA3 que ha sido el de mejor desempeño en las comparaciones que encontramos en la literatura. En la Figura 1 (c) observamos que los tres métodos resultan comparables en la detección del cambio en la desviación estándar. Este es un comportamiento ya conocido.

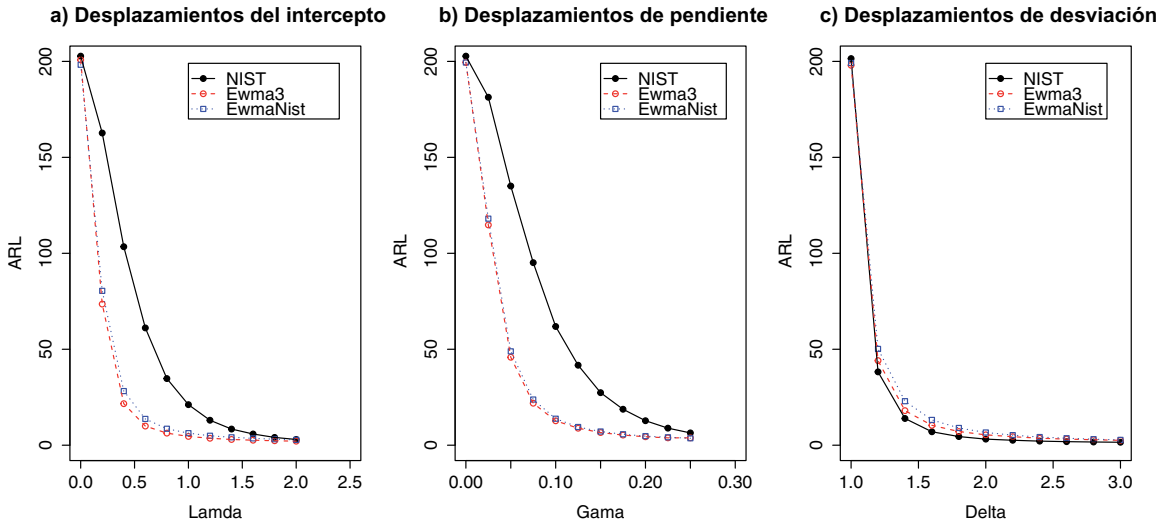


Figura 1: Valores de ARL para diferentes desplazamientos del intercepto (a), la pendiente (b) y la desviación estándar (c), para los procedimientos NIST, EWMA/NIST y EWMA3.

4. Conclusiones

El monitoreo de los procesos de calibración de instrumentos de medición es de gran importancia porque nos permite verificar que la exactitud y la precisión se mantienen estables y en control estadístico. Proponemos y evaluamos dos modificaciones a la carta NIST que pueden utilizarse para el monitoreo de procesos de calibración. La primera modificación corresponde a incorporar el esquema EWMA a la estadística de monitoreo de la carta de Croarkin y Varner. Esta modificación dio por resultado la carta EWMA/NIST cuyo desempeño es comparable con la carta EWMA3, identificada en la literatura como la de mejor desempeño.

La segunda modificación se basa en la consideración de intervalos de muestreo variable en la carta EWMA/NIST y dió por resultado ligeras ventajas en relación al desempeño de esta carta bajo cambios en los parámetros. Desde el punto de vista práctico, resulta de gran utilidad el considerar intervalos muestreo de longitud variable, ya que con esto podemos reducir los costos en tiempo y dinero.

Referencias

1. Croarkin, C. y Varner, R. (1982). "Measurement assurance for dimensional measurements on integrated-circuit photomasks", NBS Technical Note 1164, US Department of Commerce, Washington DC, USA.

2. Kim, K., Mahmoud, M.A. y Woodall, W.H. (2003). "On the monitoring of linear profiles", Journal of Quality Technology, 5, pp. 317-328.

3. Li, Z. y Wang, Z. (2010). "An exponentially weighted moving average scheme with variable sampling intervals for monitoring linear profiles". Computers and Industrial Engineering. 59, pp. 630-637.

4. NIST/SEMATECH (2011). http://www.itl.nist.gov.div898/handbook/mpc/section3/mpc37.htm.

5. Shewhart, W. A. (1931). "Control of Quality of Manufactured Product", New York Van Nostrand.

Evaluador de Eficiencias de Técnicas de Clasificación en R

Francisco Javier Landa Torres[a]

Sergio Hernández González[b], Genaro Rebolledo Méndez[c], Héctor Francisco Coronel Brizio[d], Nery Sofía Huerta Pacheco[e]

Universidad Veracruzana

Clasificación: Trabajo de Investigación

Área: Cómputo Estadístico

Subárea: Lenguaje R

Trabajo presentado en: XXVIII Foro Nacional de Estadística

1. Introducción

La clasificación es una tarea utilizada en el proceso de descubrimiento de conocimiento durante la etapa de Minería de Datos, en estudios cuyo principal objetivo es la de encontrar patrones que suceden en un fenómeno de estudio dado la existencia de una variable discreta que representa la clase. El interés de esto radica en poder realizar discriminación de observaciones basados en características semejantes, o bien tener la capacidad de desarrollar un modelo para predicción.

Este trabajo consiste en la implementación de una herramienta de evaluación de 5 clasificadores en el lenguaje R (R Core Team 2003) con diferentes enfoques, con el objetivo de verificar el rendimiento de cada uno en diferentes conjuntos de datos. Posteriormente, éstos fueron sometidos a un proceso de evaluación que permitió observar el rendimiento de

[a] fco.j.landa@gmail.com
[b] sehernandez@uv.mx
[c] grebolledo@uv.mx
[d] hcoronel@uv.mx
[e] nehuerta@uv.mx

cada uno con cada conjunto de datos con clases binarias, empleando la técnica de Validación Cruzada (dejando uno fuera)(Kohavi et al. 1995). Las métricas utilizadas para validar el rendimiento fueron: precisión, exactitud, recuerdo, valor-F y el Error Cuadrático Medio (ECM) (Tan, Steinbach, and Kumar 2005).

2. Marco Teórico

Un clasificador es una función $f(x)$ cuya función principal es mapear las variables de un conjunto de datos de acuerdo al valor de la clase $c_1, ...c_n \in C$, por lo que también es conocido como modelo de clasificación. (Tan, Steinbach, and Kumar 2005)

Los K vecinos más cercanos (KNN por sus siglas en inglés) es un modelo de clasificación perezoso, ya que recuerda la distancia euclideana de cada instancia con respecto a un punto, para posteriormente obtener los K puntos más cercanos. Cuando el valor de $K = 1$, la técnica asignará la clase perteneciente al punto más cercano, mientras que si el valor de $K > 1$ entonces la clase electa será la que predomine entre el vecindario conformado por los K puntos más cercanos (Tan, Steinbach, and Kumar 2005).

El Discriminante Lineal de Fisher (DLF) es una técnica de clasificación que consiste en proyectar un conjunto de datos de alta dimensionalidad en una, para llevar a cabo en ese espacio la clasificación, maximizando la distancia entre las medias de dos clases (interclase) y a la vez minimizando la varianza intraclase, cuando la cantidad de clases es igual a 2 (Murty and Devi 2011).

Ingenuo Bayesiano (IB) es un método de clasificación que consiste en la determinación de la clase tomando como base a la probabilidad condicional, asumiendo que los atributos son condicionalmente independientes entre si según el valor de la variable clase. Este es considerado un clasificador probabilísta, ya que proporciona el porcentaje de ocurrencia (Tan, Steinbach, and Kumar 2005)

C4.5 es un árbol de decisión utilizado para la tarea de clasificación, cuyo proceso de contrucción se basa en la maximización del Porcentaje de Ganancia por cada atributo para identificar los nodos del árbol de acuerdo a los valores de la clase. Una propiedad de este modelo es el procesamiento de la información continua implementando un método de discretización (Tan, Steinbach, and Kumar 2005)

El Clasificador por Mayoría de Votos (CMV) es un método de clasificación ensamblador

básico cuyo esquema está formado por la participación de varios clasificadores, determinando como clase inferida aquella con mayor frecuencia entre los resultados individuales de cada clasificador. Es recomendable utilizar una cantidad de modelos nones cuando la cantidad de clases son 2, debido a la posibilidad de ocurrencia de igualdad entre las clases (Rokach 2010)

3. Materiales y Métodos

3.1 Conjunto de datos

Las bases de datos utilizadas se muestran en la tabla 1, obtenidas del repositorio de datos UCI Machine Learning Repository (Bache and Lichman 2013). La selección se realizó identificando diferentes características como información faltante, variables numéricas, categóricas, cantidad de instancias y de atributos; además de ser biclases con diferente proporción.

ID	Tipo	Instancias	Variables	Valores faltates	Clase
BD1	Num	748	5	N/A	C1 (307) C2 (383)
BD2	Cat/Núm	690	15	Si	C1 (307) C2 (383)
BD3	Cat/Núm	1000	21	No	1 (700) 2 (300)
BD4	Cat/Núm	270	14	No	1 (150) 2 (120)
BD5	Cat	432	7	No	0 (216) 1 (216)
BD6	Núm	1372	5	N/A	0 (762) 1 (610)

Tabla 1: Descripción del conjunto de datos. BD1) Blood Transfusion Service center, BD2) Credit Approval, BD3) German, BD4) Heart, BD5)MONK's Problem, BD6) Bank Authentication Data Set. Cat = Categóricos, Num= Numéricos. Los valores dentro de los paréntesis representa la cantidad de instancias por valor de la clase.

3.2 Metodología

Una vez cargadas las bases de datos en el ambiente de R, se particiona la información original en conjunto de prueba que está conformado por la instancia con el índice i=1, mientras

que el conjunto de entrenamiento es su cumplemento. Este último funge como entrada de información para el aprendizaje de cada uno de los clasificadores: KNN con K=3, DLF, C4.5, IB y CMV, validándose cada uno con la instancia del conjunto de prueba, corroborando que el valor inferido sea igual al original (denominado objetivo).

Una vez realizado el proceso anterior, se contabilizan las concordancias entre los valores inferidos y objetivos, para conformar la matriz de confusión. El valor del índice i inicia en 1 hasta la cantidad de instancias total de la base de datos repitiendo el proceso. Finalmente, se procede a calcular las medidas de rendimiento anteriormente mencionados.

Es importante resaltar que para la discretización de los datos en algunas técnicas se realizó una partición de la variable tomando como umbral su mediana; mientras que por otro lado se convirtieron las variables marginales y nominales a numéricos asignando un entero según al valor identificado. Por último, para la solución del problema de la obtención de una matriz singular durante el proceso del ADF se procedió a tomar los primeros 3 componentes principales de la matriz de datos originales.

4. Resultados

Los resultados muestra en la tabla 2, las cuales de manera general se puede visualizar que ningún clasificador tuvo el mejor rendimiento en todos los conjuntos de datos considerando la cantidad de instancias clasificadas correctamente; aunque cada uno predominó al menos una vez.

Otro punto importante a observar es que cada medida de rendimiento propuesto no necesariamente coincidieron con el mayor porcentaje de instancias correctas, a excepción del ECM, esto se debe a que el cálculo de este valor se realiza con el complemento de la cantidad de correctos.

Además, se percibió que en el cuarto conjunto de datos, el porcentaje de efectividad fue el escenario en donde los modelos tuvieron un desempeño competitivo. Por último se observa que C4.5 pudo clasificar correctamente todos los datos.

	BD1					**BD2**				
	KNN	DLF	C4.5	IB	CMV	KNN	DLF	C4.5	IB	CMV
Correctos	560	494	382	178	479	473	594	391	307	548
Precisión	0.748	0.660	0.510	0.237	0.640	0.685	0.860	0.566	0.444	0.794
Exactitud	0.459	0.388	0.240	0.237	0.376	0.663	0.785	0.513	0.444	0.696
Recuerdo	0.320	0.747	0.488	1	0.780	0.596	0.944	0.488	1	0.951
Valor-F	0.377	0.511	0.322	0.384	0.508	0.627	0.857	0.500	0.615	0.804
ECM	0.251	0.333	0.489	0.762	0.359	0.314	0.139	0.433	0.555	0.205
	BD3					**BD4**				
Correctos	727	571	662	719	728	214	224	199	225	219
Precisión	0.727	0.571	0.662	0.719	0.728	0.792	0.822	0.737	0.833	0.811
Exactitud	0.772	0.846	0.756	0.853	0.811	0.771	0.803	0.699	0.815	0.800
Recuerdo	0.862	0.472	0.762	0.722	0.797	0.758	0.816	0.716	0.808	0.766
Valor-F	0.815	0.606	0.759	0.782	0.804	0.764	0.809	0.707	0.811	0.782
ECM	0.273	0.420	0.338	0.281	0.272	0.207	0.170	0.262	0.166	0.188
	BD5					**BD6**				
Correctos	322	288	432	216	360	1371	1340	683	610	1344
Precisión	0.745	0.666	1	0.5	0.833	0.999	0.976	0.497	0.444	0.979
Exactitud	0.827	0.666	1	0.5	0.75	1	1	0.553	N/A	1
Recuerdo	0.620	0.666	1	1	1	0.998	0.958	0.497	0	0.963
Valor-F	0.708	0.666	1	0.666	0.857	0.999	0.978	0.523	0	0.981
ECM	0.254	0.333	0	0.5	0.166	0.001	0.023	0.523	0.555	0.029

Tabla 2: Resultados de ejecución de las 6 bases de datos. N/A = No Aplica, debido a que el resultado fue propiciado por una división entre 0.

Valor en Riesgo del Índice de Precios y Cotizaciones, 1991-2011*

Genoveva Lorenzo Landa[a]
Especialización en Métodos Estadísticos Universidad Veracruzana

Sergio Fco. Juárez Cerrillo
Facultad de Estadística e Informática, Universidad Veracruzana

Héctor Fco. Coronel Brizio
Facultad de Física, Universidad Veracruzana

Clasificación: Tesis de Licenciatura.
Área: Valores Extremos.
Subárea: Modelación de Valor en Riesgo.
Trabajo presentado en: XXVIII Foro Nacional de Estadística.
Palabras Clave: Bondad de Ajuste, Censura Tipo II, Distribución Pareto, Riesgo de Mercado.

1. Introducción

Los índices de precios se integran por muestras de acciones del mercado que se consideran representativas de éste. En el mercado financiero internacional destacan índices como el Standard & Poors, el Dow Jones y el Nikkei. En México el principal indicador es el Índice de Precios y Cotizaciones (IPC). Dentro de un mercado financiero los instrumentos financieros se valoran de acuerdo a su rendimiento y riesgo de que su valor disminuya al mismo tiempo que los movimientos del mercado. El rendimiento en un período de tiempo $[0, T]$ se determina por $R_T = (S_T - S_0)/S_0$ donde S_T es el precio del instrumento financiero al final del periodo y S_0 es el precio al inicio del periodo. El retorno en el tiempo t es el rendimiento expresado en relación al tiempo anterior, es decir $R_t = (S_t - S_{t-1})/S_{t-1}$. Junto con la integración

* Este trabajo se presentó con financiamiento parcial de la Maestría en Estadística Aplicada de la Universidad Veracruzana
[a] gelorenzo@uv.mx

financiera global, la volatilidad de los mercados financieros se ha convertido en un tema de relevancia para los diferentes agentes económicos. La metodología de Valor en Riesgo (VaR) busca anticipar esta volatilidad y así cuantificar el riesgo calculando las magnitudes de grandes pérdidas así como las probabilidades de que ocurran. Una forma de hacer esto es mediante el cálculo del Valor en Riesgo, el cual se define por $\text{VaR}_p = F^{-1}(p), 0 < p < 1$, donde F es la distribución de las pérdidas. El VaR permite que los reguladores financieros asignen un número a su peor escenario y así planear de acuerdo a ese escenario. En este artículo calculamos el riesgo financiero de los retornos del IPC durante el período 1991-2011. Seguimos la propuesta de Corononel-Brizio y Hernández-Montoya (2010b) de modelar los retornos del IPC con la distribución Pareto para muestras con censura tipo II por la izquierda.

2. El IPC

En la gráfica (a) de la Figura 1 observamos el comportamiento del IPC al cierre desde el 11 de noviembre de 1991 hasta el 29 de noviembre del 2011. Se tienen 5012 observaciones. Sea I_t el valor de cierre del IPC en el tiempo t ($t = 1, ..., 5012$), de modo que los retornos son $R_t = (I_t - I_{t-1})/I_{t-1}$. Las gráficas (b) y (c) de la Figura 1 muestran los retornos estandarizados del IPC. La función de autocorrelación de los retornos estandarizados, gráfica (d), nos permite considerar a los retornos como variables aleatorias independientes.

3. La Distribución Pareto y su Ajuste

La variable aleatoria X sigue la distribución Pareto, Pareto (1897), con parámetros α y θ si su función de distribución es

$$F(x; \alpha, \theta) = 1 - \left(\frac{\theta}{x}\right)^{\alpha}, x > \theta,$$

donde θ es un parámetro positivo de escala y α es un parámetro positivo que se le conoce como índice de Pareto y corresponde al negativo de la pendiente de $\log(1 - F(x; \alpha, \theta))$ vs $\log(x)$. El valor esperado y la varianza de X son $\text{E}(X) = \alpha\theta/(\alpha - 1)$ para $\alpha > 1$ y $\text{Var}(X) = \alpha\theta^2/(\alpha - 1)^2(\alpha - 2)$ para $\alpha > 2$, respectivamente.

En la Figura 2 vemos los histogramas de los retornos estandarizados positivos y negativos

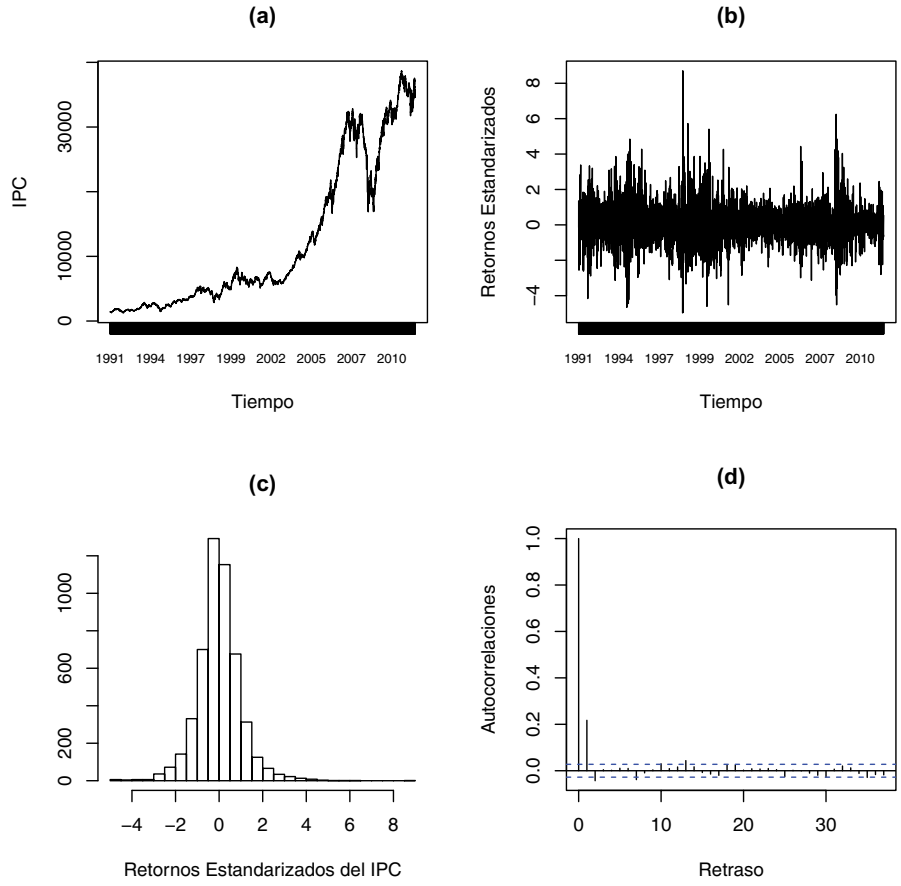


Figura 1: (a) IPC del 8/11/1991 al 29/11/2011, (b) Serie de tiempo de los retornos estandarizados, (c) Histograma de los retornos estandarizados, (d) Función de autocorrelación de los retornos estandarizados.

(en valor absoluto) estandarizados. El comportamiento aproximadamente lineal en la escala log-log proporciona evidencia empírica de que la Pareto es un modelo que describe adecuadamente a los retornos del IPC. Como un procedimiento formal utilizamos el estadístico de Anderson-Darling para probar que las observaciones $x_1, \ldots, x_n$ provienen de una distribución Pareto basándose en las r observaciones más grandes $x_{(n-r+1)} \leq x_{(n-r+2)} \leq \ldots \leq x_{(n-1)} \leq$

	$q = 1 - r/n$	r	$\hat{\alpha}$	$\hat{\theta}$	$A^2_{n,r}$
retornos positivos	$q = 0.85$	389	2.397	0.596	0.22
$n = 2596$	$q = 0.9$	260	2.617	0.664	0.12
retornos negativos	$q = 0.9$	259	3.149	0.737	0.10
$n = 2595$	$q = 0.95$	130	3.973	0.943	0.01

Tabla 1: Parámetros estimados a los retornos del IPC y estadísticos de Anderson-Darling.

$x_{(n)}$. Este estadístico de prueba es

$$\begin{aligned} A^2_{n,r} &= -\frac{1}{n}\sum_{i=1}^{r}(2i-1)\{\log(1-z_{n-i+1}) - \log(z_{n-i+1})\} - 2\sum_{i=1}^{r}\log(z_{n-i+1}) \\ &\quad -\frac{1}{n}\left[(r-n)^2\log(z_{n-r+1}) - r^2\log(1-z_{n-r+1}) + n^2(1-z_{n-r+1})\right], \end{aligned}$$

donde $z_{n-i+1} = F(x_{(n-i+1)}; \hat{\alpha}, \hat{\theta})$, y

$$\hat{\alpha} = r\left[\sum_{i=1}^{r}\log(x_{(n-r+i)}) - r\log(x_{(n-r+1)})\right]^{-1} \quad \text{y} \quad \hat{\theta} = \left(\frac{r}{n}\right)^{1/\hat{\alpha}} x_{(n-r+1)}$$

son los estimadores de máxima verosimilitud de α y θ bajo censura tipo II. Los valores críticos de $A^2_{(n,r)}$ están tabulados en Coronel-Brizio y Hernández-Montoya (2010a).

4. VaR para los Retornos del IPC

La Tabla 1 muestra los resultados de ajustar separadamente la Pareto a los retornos positivos y negativos estandarizados (en valor absoluto) para diferentes niveles de censura tipo II por la izquierda. Los cálculos se hicieron en R. De la inspección de los valores críticos de $A^2_{n,r}$ en Coronel-Brizio y Hernández-Montoya (2010a), tenemos que no rechazamos a la Pareto con un nivel de significancia del 0.15. La Figura 3 muestra el ajuste de las colas derechas de las distribuciones Pareto.

Para calcular el VarR, supongamos que X es Pareto con parámetros α y θ y sea u un valor tal que $P(X > u) = 1 - q$. Una estimación de la cola derecha de X con una proporción

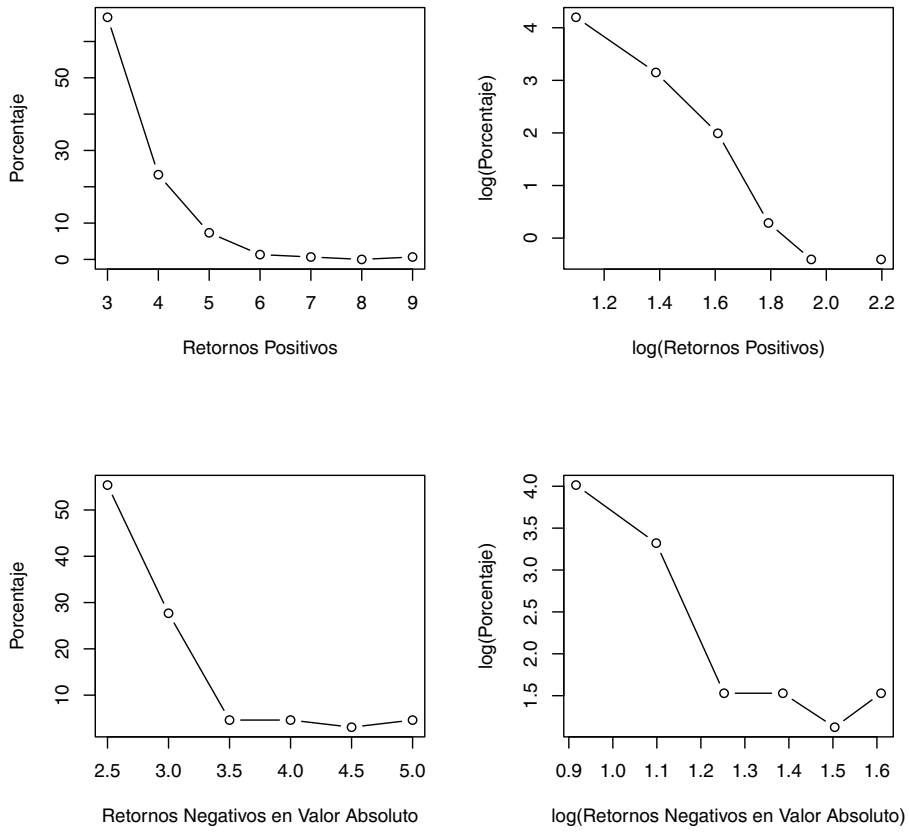


Figura 2: IPC del 8/11/1991 al 29/11/2011. Arriba: Cierres. Abajo: Retornos.

de censura de tipo II de $q = 1 - r/n$ es

$$\hat{P}(X > u + x) = 1 - F(u + x; \hat{\alpha}, \hat{\theta}) = \frac{r}{n}\left(\frac{\hat{\theta}}{x}\right)^{\hat{\alpha}},$$

para $x > \hat{\theta}$. De tal manera que un estimador del cuantil x_p de F se obtiene resolviendo para x_p a la ecuación

$$\hat{P}(X > x_p) = \frac{r}{n}\left(\frac{\hat{\theta}}{x_p - u}\right)^{\hat{\alpha}} = p.$$

Lo que resulta en $\hat{x}_p = u + \hat{\theta}(r/np)^{1/\hat{\alpha}}$. Si $p = 1/m$, entonces $\hat{x}_{1/m}$ es el nivel de retorno con período de retorno de m días, $\hat{x}_{1/m} = u + \hat{\theta}(rm/n)^{1/\hat{\alpha}}$. Para los retornos negativos

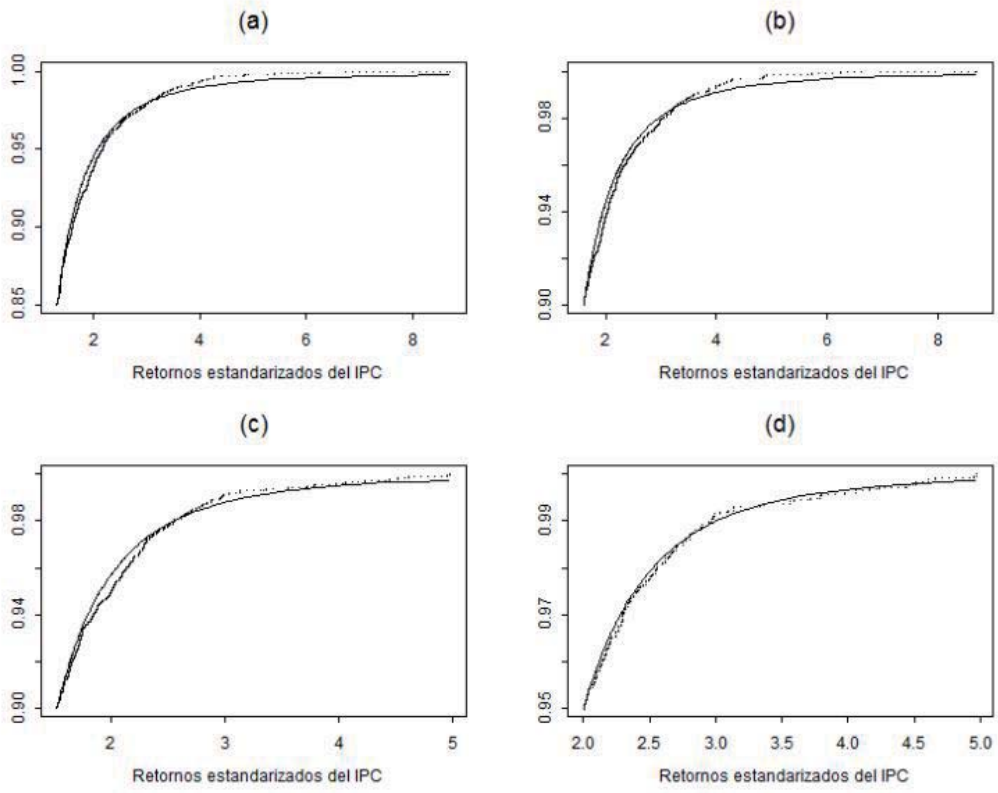


Figura 3: Ajuste de la Pareto a la cola derecha de los retornos positivos con (a) $q = 0.85$ y (b) $q = 0.9$. Ajuste de la Pareto a la cola derecha de los retornos negativos (valor absoluto) con (c) $q = 0.9$ y (d) $q = 0.95$.

estandarizados con $q = 0.9$ tenemos $u = x_{(n-r-1)} = x_{(2595-259-1)}$ =1.53 y, la Tabla 1, el valor en riesgo estimado es

$$\hat{x}_{1/m} = 1.53 + 0.737 \left(\frac{259m}{2595} \right)^{1/3.149}.$$

La Figura 4 muestra las estimaciones de los niveles de retorno de los retornos negativos (en valor absoluto) desestandarizados $\text{Var}_{1/m} = \bar{R} + s\hat{x}_{1/m}$ para un horizonte a 10 años, donde $\bar{R}$ y s son la media y la desviación estándar de los retornos R_t. Así por ejemplo vemos que en los próximos 10 días se espera una caída de aproximadamente el 9 % del valor del IPC.

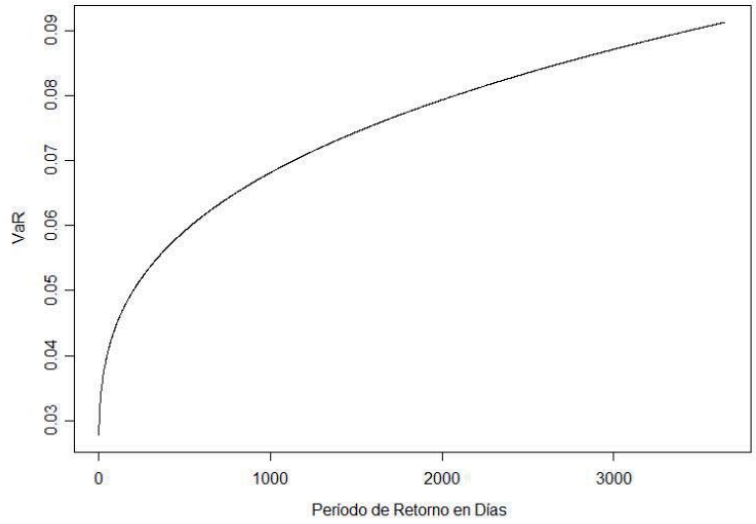


Figura 4: Valor en riesgo estimado de los retornos negativos (en valor absoluto) del IPC.

5. Conclusiones

Hemos visto que la distribución Pareto proporciona un buen ajuste a los retornos máximos y mínimos estandarizados del IPC. Es interesante notar el hecho de que los índices de Pareto estimados para el IPC son muy similares a los estimados para el Dow Jones en Coronel-Brizio y Hernández-Montoya (2010b). El criterio de evaluación de la bondad de ajuste del modelo fue la prueba de Anderson-Darling para muestras con censura tipo II.

Referencias

1. Cano Medina, J.L. (2010). *Valor en Riesgo del IPC de México, 1991-2008.* Tesis de Maestría en Estadística Aplicada, Facultad de Estadística e Informática, Universidad Veracruzana.

2. Coronel-Brizio, H.F., and Hernández-Montoya, A.R. (2010a). *The Anderson-Darling Test of Fit for the Power-Law Distribution from Left-Censored Samples.* Econophysica A 389, 3508-3515.

3. Coronel-Brizio, H.F., and Hernández-Montoya, A.R. (2010b). *On fitting the*

Estudio del uso de las redes sociales dentro de la Universidad Veracruzana

Herminia Domínguez Palmeros[a]
Instituto Tecnológico de la Cuenca del Papaloapan

María Luisa Hernández Maldonado, Antonia Olivia Jarvio Fernández
Universidad Veracruzana

Clasificación: Tesis de Especialización.
Área: Análisis Multivariado.
Subárea: Análisis de Correspondencia.
Trabajo presentado en: XXVIII Foro Nacional de Estadística.

1. Introducción

El uso de computadoras, Internet y celulares han propiciado una nueva modalidad de comunicación escrita. La escritura empleada muestra similitud a la comunicación verbal cotidiana; originándose conversaciones informales como las realizadas en una comunicación real entre amigos, en las que se identifican un propósito básicamente lúdico (Levis, 2007). Una de las herramientas de comunicación más utilizadas a través de los medios mencionados la constituyen las redes sociales. La red social favorita en América Latina es el Facebook, con un promedio de navegación de 3 horas al mes. En México más del 80 % de los usuarios de redes sociales la utilizan, lográndose un incremento de usuarios de internet de febrero del 2009 a febrero del 2010 del 19 % (Sutter, 2010). Jarvio (2011) realiza la encuesta "Prácticas Lectoras en los Nuevos Soportes Digitales de la UV". Esta encuesta da origen a la pregunta abierta objeto del presente estudio: ¿Por qué te gusta formar parte de una red social? El objetivo de este trabajo radica en conocer a través de un análisis estadístico de datos textuales las razones del gusto de la comunidad universitaria por pertenecer a una red social, identificando eventuales similitudes entre las regiones que integran a la Universidad Veracruzana.

[a] hdomp@hotmail.com

2. Marco Teórico

El Análisis Estadístico de Datos Textuales se enfoca en el análisis del número de ocurrencias de las palabras contenidas en el texto correspondiente a las respuestas y su relación con características propias de los encuestados, está relacionado con métodos estadísticos factoriales y de clasificación (Lebart y Salem, 1988 y 1994; Lebart y coautores, 1998; Lebart y otros; 2000. Citados por Fernández y Modroño, 2007). La Tabla Léxica Agregada se forma por el vocabulario depurado y la frecuencia de cada forma gráfica (palabra) para cada parte del texto, como proponen Lebart y colaboradores (2000).

3. Método

Este estudio se basa en las preguntas relacionadas con el uso de las redes sociales, mismas que formaron parte del cuestionario aplicado en la encuesta "Prácticas Lectoras en los Nuevos Soportes Digitales en la UV", dirigida a la comunidad universitaria de las cinco regiones que la integran (Xalapa, Veracruz-Boca del Río, Orizaba-Córdoba, Poza Rica-Tuxpan y Coatzacoalcos-Minatitlán). El periodo de aplicación de la encuesta fue del 27 de julio al 07 de octubre de 2010, fue administrada a través de internet con el apoyo de la Red de Estudios de Opinión. Las preguntas referidas son las siguientes: ¿formas parte de alguna red social?, si contestaste sí, ¿de cuál?, ¿por qué te gusta formar parte de una red social?

Se realizaron invitaciones para participar en la encuesta a través del servicio de correo electrónico, estas invitaciones fueron dirigidas a las personas registradas en la base de datos del módulo de préstamo de libros del sistema Unicornio utilizado en las bibliotecas de la Universidad Veracruzana, dado que a las bibliotecas de la UV acude el 94 % de la comunidad universitaria. El tamaño de muestra fue de 641 individuos, quienes respondieron a la invitación realizada. Para este estudio, se aplicó un filtro a la información obtenida de los 641 miembros de la comunidad universitaria, quedando constituida la matriz de datos exclusivamente por 343 casos, correspondientes a quienes contestaron de forma afirmativa a la pregunta ¿formas parte de alguna red social? (variable dicotómica); se realizó un análisis de frecuencias a las respuestas proporcionadas por los usuarios de las bibliotecas de la UV correspondientes a la pregunta abierta si contestaste sí, ¿de cuál? Se depuró el vocabulario formado por las respuestas libres a la pregunta abierta ¿por qué te gusta formar parte de una

red social? de acuerdo con la metodología propuesta por Lebart y otros (2000) y se aplicó un Análisis Factorial de Correspondencias (AFC) por región, sobre la tabla léxica agregada ***palabras x región.*** El AFC busca detectar diferencias y semejanzas entre la comunidad universitaria en las diversas regiones que integran a la Universidad Veracruzana.

4. Resultados

De los 343 usuarios de las redes sociales que forman parte de la comunidad universitaria, 330 respondieron a la pregunta textual ¿de cuál?; el 69 % son miembros exclusivos de Facebook, el 24 % forman parte de Facebook y otras redes sociales también; de este porcentaje, que corresponde a 80 usuarios, 37 se conectan también a Twitter, 23 a Hi5 y 9 a Sónico y 11 a otras redes sociales (Figura 1). En promedio, la comunidad universitaria en el año 2010 era miembro de 1.4 redes sociales. Con base en el vocabulario seleccionado a partir de la segunda variable textual, se construye una tabla léxica por región de la U.V. (Tabla 1) y se aplica un AFC obteniendo una explicación de la inercia en el primer plano factorial del 55.68 %, que es el plano que se interpreta. Para la formación del primer eje factorial las regiones que más contribución aportan son Veracruz y Xalapa; en la formación del segundo eje factorial, son Coatzacoalcos-Minatitlán y Córdoba-Orizaba. De estas mismas regiones la de mejor calidad de representación en este plano, es Veracruz (70 %), le siguen Xalapa (65 %) y Coatzacoalcos-Minatitlán (65 %) (Tabla 2).

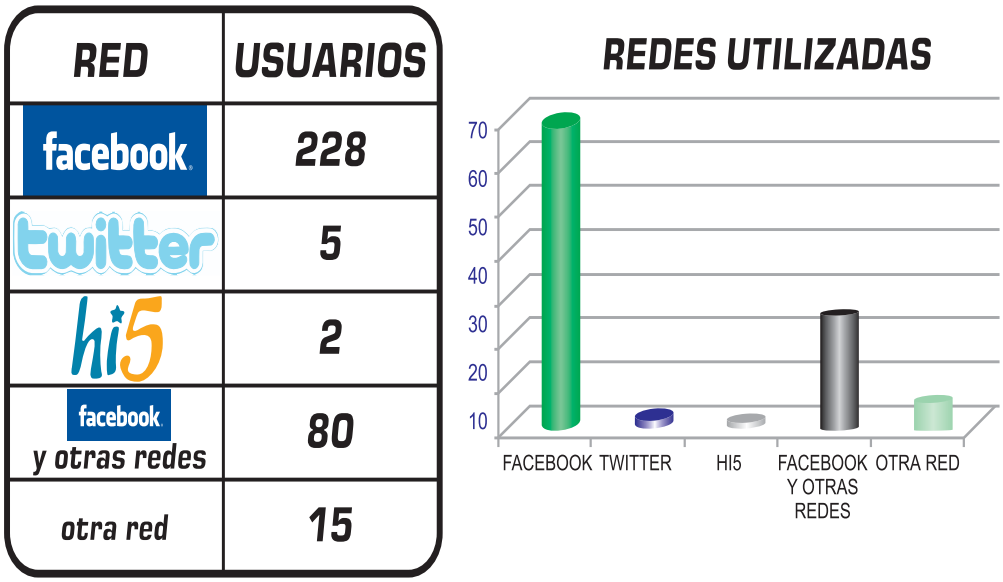


Figura 1: Redes sociales utilizadas por la comunidad universitaria de la UV/2010.

Palabra	Xalapa	Veracruz	Poza Rica-Tuxpan	Córdoba-Orizaba	Coatzacoalcos-Mina
ACTIVIDAD	3	1	0	0	0
ADEMAS	1	5	1	0	0
ALGUN	7	3	1	0	0
ALGUNA	3	0	0	2	0
AMIGO	49	55	14	11	7
AMISTAD	13	3	0	0	1
BUENA	3	1	0	0	0
...	...	...	...	...	...

Tabla 1: Fragmento de la tabla léxica agregada (tabla de contingencia que cruza *palabras x región* a la cual se aplica un Análisis Factorial de Correspondencias

Región	Contribuciones	Cosenos Cuadrados
Xalapa	33.06	0.65
Veracruz	42.00	0.70
Poza Rica-Tuxpan	20.33	0.28
Córdoba-Orizaba	46.25	0.55
Coatzacoalcos-Mina	58.36	0.65

Tabla 2: Contribución y calidad de representación.

El eje factorial 1 caracteriza a los individuos que usan las redes sociales porque están interesados en conocer gente, hacer nuevas amistades, conservar amigos, compartir e intercambiar información, por causas de trabajo y por la familia. Opone a quienes se comunican entre amigos por intercambiar, por diversión y de forma rápida y a aquéllos que se comunican por compartir, hacer nuevas amistades, socializar y aunque menos significativamente también por cuestiones de trabajo. Los primeros se asocian a la región Veracruz y los segundos a la región Xalapa. El eje factorial 2 caracteriza a los individuos por su necesidad de usar las redes sociales para comunicarse por cuestiones académicas y de trabajo principalmente. Opone a quienes se conectan para comunicarse con compañeros y maestros, intercambiar información y cumplir con tareas con quienes declaran significativamente que no les gustan del todo las redes sociales, sin embargo, las utilizan por asuntos de trabajo y también por entretenimiento. Los primeros se asocian a la región Coatzacoalcos-Mina y los segundos a la región Córdoba-Orizaba (Figura 2).

5. Conclusiones

En general, la comunidad universitaria de la Universidad Veracruzana opina tener un gusto por las redes sociales, a excepción de los encuestados que pertenecen a la región Córdoba-Orizaba. Todas las regiones manifiestan que una de las razones por las que les gusta usar las redes sociales es conocer gente; estando esta característica más acentuada en la región Xalapa. De la misma forma, también declaran que parte de su gusto por las redes sociales radica en tener contacto con los amigos y la familia, observándose esta característica más marcada en la región Veracruz. Los contextos identificados para las regiones Xalapa y Veracruz son muy

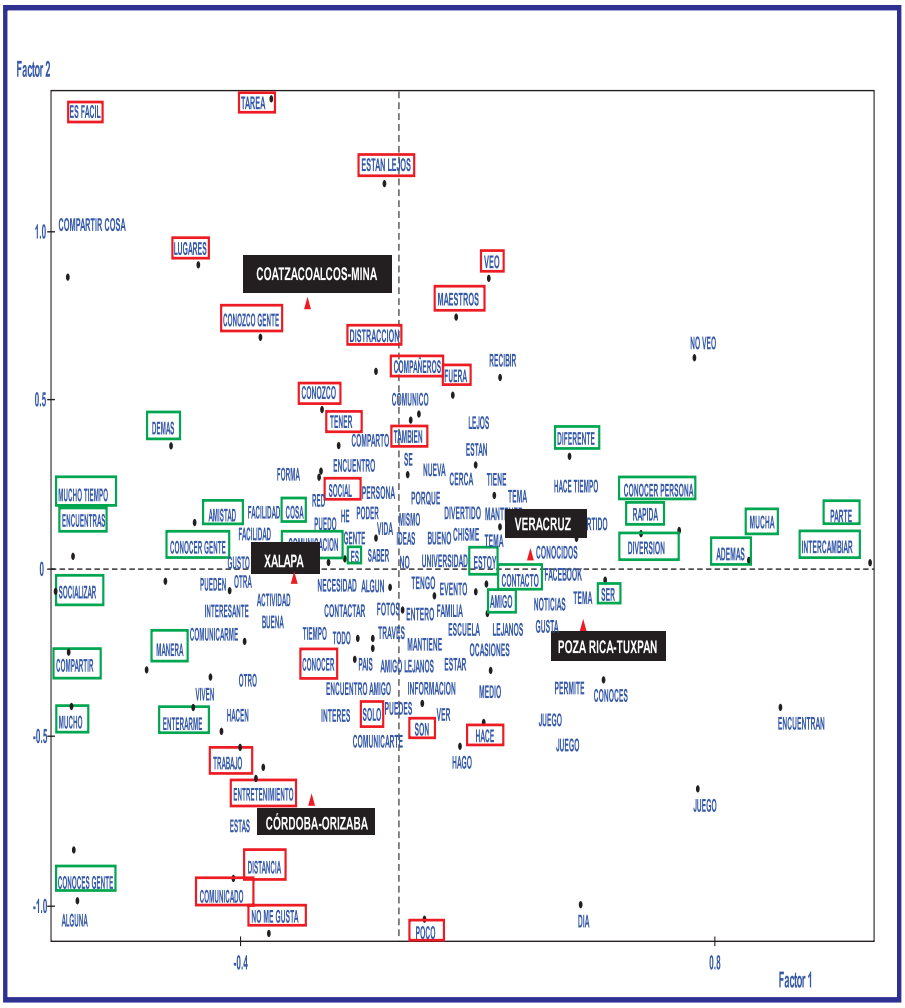


Figura 2: Representación de las palabras y regiones en el 1er. plano factorial. En marco verde las palabras con $\cos^2$ más altos en el eje 1. En marco rojo las palabras con $\cos^2$ más altos en el eje 2. Los segmentos de dos palabras más fueron considerados como variables suplementarias.

similares, en ambas regiones les gusta a los encuestados formar parte de una red social porque conocen gente, por intercambiar información, etc.; sin embargo, en el mapa perceptual del Análisis Factorial de Correspondencias se observa oposición entre ambas regiones, ya que a Veracruz se le percibe además de estar muy asociado a la diversión, estar un poco más apegado a los amigos y a la familia, y a Xalapa se le detecta un uso de las redes sociales más relacionado a socializar, hacer nuevas amistades y también al trabajo. Las regiones Coatzacoalcos-Mina y Córdoba-Orizaba se caracterizan por usar las redes por cuestiones académicas o de trabajo además del entretenimiento y aunque la región Córdoba–Orizaba declara que no le gustan las redes sociales por completo, manifiesta la necesidad de utilizarlas.

6. Bibliografía

Fernández, K.; y J. I. Modroño. (2007). Exploración textual en el contexto del Modelo de Valores en Competencia [En línea]. Estadística Española. Vol. 49, Núm.166, 501-530. Disponible en: <http://www.ine.es/ss/Satellite?L=0
&c=INERevEstad_C&p=1254735226759 &pagename=ProductosYServicios
%2FPYSLayout&_charset_=utf-8&cid=1259924965499&sub m it=Ir>.

Jarvio, A. O. (2011). La lectura digital en el ámbito de la Universidad Veracruzana. España: Eds. Universidad de Salamanca.

Lebart, L.; A. Salem; y M. Bécue. (2000). Análisis Estadístico de Textos. España: Editorial Milenio.

Levis, D. (2007) "Hablar" con el teclado. El habla escrita del chat (y de otros mensajes escritos con computadoras y celulares. Razón y Palabra No. 53 [En línea]. Disponible en:http://www.razonypalabra.org.mx/anteriores/n54
/dlevis.html.

Sutter, J. (2010). Comscore presenta el "Estado de Internet en Latinoamérica" [En línea]. Disponible en: http://www.iabcolombia.com/noticias/comscore-presenta-el-estado-de-internet-en-latinoamerica/.

La Universidad Veracruzana desde la Satisfacción de sus Estudiantes

Ismael Sosa Galindo[a]
Maestría en Estadística Aplicada, Universidad Veracruzana

Sergio Fco. Juárez Cerrillo
Facultad de Estadística e Informática, Universidad Veracruzana

Luis Cruz Kuri
Instituto de Ciencias Básicas, Universidad Veracruzana

Clasificación: Tesis de maestría.
Área: Modelos Lineales.
Subárea: Modelos de Ecuaciones Estructurales.
Trabajo presentado en: XXVIII Foro Nacional de Estadística.
Palabras Clave: Cuadrados Mínimos Parciales, ECSI, Ecuaciones Estructurales, SmartPLS.

1. Introducción

En las instituciones de educación superior (IES) la auto-evaluación permite determinar avances en objetivos y metas, así como valorar el impacto de éstos en términos de sus visiones y misiones. Esto proporciona retroalimentación sobre el diseño de los procesos académicos y administrativos. Retroalimentación que es fundamental para la reorientación y revaloración del rumbo de la IES. Así, la auto-evaluación se convierte en una herramienta orientada a resultados que es instrumental para la mejora continua de la calidad de la educación que brinda. En este contexto, nos enfocamos en la medición de una dimensión de la auto-evaluación: la satisfacción de los estudiantes. Proponemos un indicador para medir el nivel de satisfacción académica que logra la Universidad Veracruzana (UV) en sus estudiantes. El indicador se basa en un modelo del proceso causa-efecto entre la satisfacción y los factores que la causan.

* isoga77@yahoo.com.mx

El indicador se calcula a partir de una adaptación del Índice Europeo de Satisfacción de Clientes (ECSI por sus siglas en inglés), tal y como se presenta en Tenenhaus (2005).

2. El Modelo y el Índice de Satisfacción

En el modelo propuesto, la expectativa del estudiante, la calidad percibida en los aspectos no humanos de la UV, y la calidad percibida en los elementos humanos de la UV, tienen una influencia directa en la percepción del valor del estudiante sobre la educación que está adquiriendo. Esta percepción a la vez afecta directamente a la satisfacción. La expectativa del estudiante y la calidad percibida también impactan directamente a la satisfacción del estudiante. La satisfacción se considera causal de la lealtad, además de que la lealtad se ve afectada directamente por la calidad percibida en los elementos humanos de la UV. La Figura 1 muestra la estructura causa-efecto entre las variables latentes del modelo que usamos. Las Tablas 1 y 2 muestran la operacionalización de estas variables latentes.

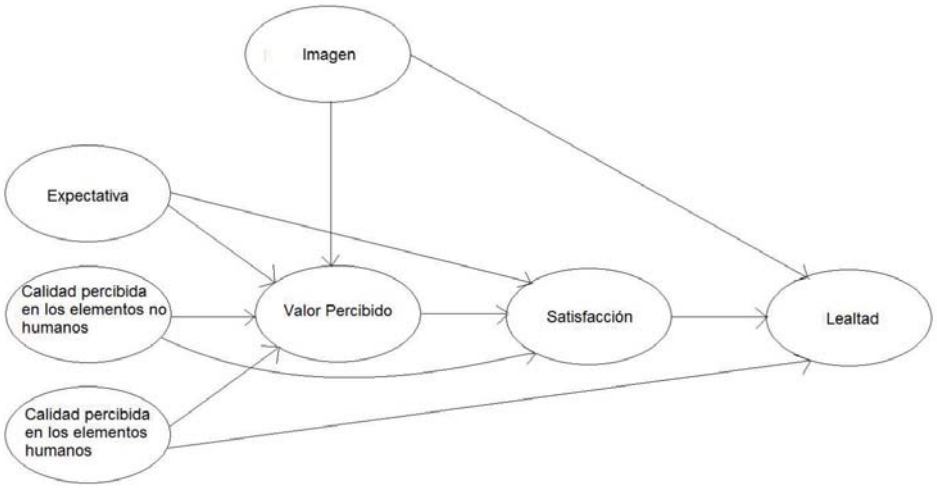


Figura 1: Modelo de satisfacción basado en el ECSI.

Variables latentes	Variables manifiestas
ξ_1: Expectativa al momento de ingresar a la UV	X_1: Calidad del contenido de los cursos X_2: Calidad de las aulas X_3: Calidad de las bibliotecas X_4: Calidad de los centros de cómputo X_5: Calidad del nivel académico de los profesores X_6: Calidad del servicio proporcionado por las autoridades administrativas y secretarias
ξ_2: Calidad percibida en los en los elementos no humanos	X_7: Calidad global del contenido de los cursos X_8: Calidad global de la oferta educativa X_9: Calidad global de los horarios X_{10}: Calidad global de la tutoría X_{11}: Calidad global de las aulas X_{12}: Calidad global de las bibliotecas X_{13}: Calidad global de los centros de cómputo X_{14}: Calidad del Modelo Educativo Integral y Flexible (MEIF) X_{15}: Calidad global del Proyecto AULA
ξ_3: Calidad percibida en los elementos humanos	X_{16}: Calidad global del proceso de enseñanza-aprendizaje por parte de los profesores X_{17}: Calidad global del tutor en su quehacer tutorial X_{18}: Calidad global del servicio prestado por las autoridades académicas (director, secretario, jefe de carrera y/o departamento) X_{19}: Calidad global del servicio prestado por el personal administrativo (administradores, secretarias, encargados de bibliotecas, encargados de centros de cómputo)

Tabla 1: Variables latentes y manifiestas del indicador.

Variables latentes	Variables manifiestas
ξ_4: Imagen	X_{20}: Imagen global de la UV como institución
	X_{21}: Responsabilidad y compromiso social
	X_{21}: Credibilidad y ética
η_1: Valor percibido por el estudiante	Y_1: Valor de la educación que estás adquiriendo en términos de tiempo, dinero y esfuerzo
	Y_2: Valor de la educación que estás adquiriendo en términos del empleo o el posgrado que deseas
η_2: Satisfacción del estudiante	Y_3: Satisfacción global con la UV
	Y_4: La medida en que se han llenado tus expetativas
	Y_5: La UV en comparación con la institución ideal
η_3: Lealtad del estudiante	Y_6: Continuar con un posgrado en la UV
	Y_7: Recomendar a la UV a otros estudiantes
	Y_8: Recomendar la carrera que estudias a otros estudiantes
	Y_9: Elegir a la UV pero diferente carrera
	Y_{10}: Elegir la misma carrera en la UV
	Y_{11}: Recomendarías a la UV a otras personas

Tabla 2: Variables latentes y manifiestas del indicador.

Las ecuaciones del modelo estructural representado en la Figura 1 son

$$\begin{aligned} \eta_1 &= \gamma_{1,1}\xi_1 + \gamma_{1,2}\xi_2 + \gamma_{1,3}\xi_3 + \gamma_{1,4}\xi_4 + \zeta_1, \\ \eta_2 &= \beta_{2,1}\eta_1 + \gamma_{2,1}\xi_1 + \gamma_{2,2}\xi_2 + \zeta_2, \\ \eta_3 &= \beta_{3,2}\eta_2 + \gamma_{3,3}\xi_3 + \gamma_{3,4}\xi_4 + \zeta_3. \end{aligned}$$

las Tablas 1 y 2 determinan las ecuaciones del modelo de medición

$$\begin{aligned} Y_j &= \omega_{j,l}\eta_l + \varepsilon_j, \quad l = 1, \ldots, 3, \\ X_k &= \theta_{k,l}\xi_l + \delta_k, \quad l = 1, \ldots, 4, \end{aligned}$$

los γ's, β's, ω's y θ's son parámetros desconocidos tales que $\omega_{j,1} = 0$ para $j = 3, \ldots, 11$; $\omega_{j,2} = 0$ para $j = 1, 2, 6, \ldots, 11$; $\omega_{j,3} = 0$ para $j = 1, \ldots, 5$; $\theta_{k,1} = 0$ para $k = 7, \ldots, 10$;

$\theta_{k,2} = 0$ para $k = 1, \ldots, 6, 16, \ldots, 20$; y $\theta_{k,3} = 0$ para $k = 1, \ldots, 19$. Los términos de error satisfacen $\mathrm{E}(\zeta_i) = \mathrm{E}(\varepsilon_j) = \mathrm{E}(\delta_k) = 0$, $\mathrm{Var}(\zeta_i) = \sigma^2_\zeta$, $\mathrm{Var}(\varepsilon_j) = \sigma^2_\varepsilon$, $\mathrm{Var}(\delta_k) = \sigma^2_\delta$, $\mathrm{Cov}(\zeta_i, \zeta_{i'}) = \mathrm{Cov}(\varepsilon_j, \varepsilon_{j'}) = \mathrm{Cov}(\delta_k, \delta_{k'}) = \mathrm{Cov}(\zeta_i, \varepsilon_j) = \mathrm{Cov}(\zeta_i, \delta_k) = \mathrm{Cov}(\varepsilon_j, \delta_k) = 0$ para $i, i' = 1, 2, 3; j, j' = 1, \ldots, 11$; y $k, k' = 1, \ldots, 22$. La expresión poblacional del indicador es

$$I = \frac{\mathrm{E}(\eta_2) - \mathrm{mín}(\eta_2)}{\mathrm{máx}(\eta_2) - \mathrm{mín}(\eta_2)} \times 100,$$

donde $\mathrm{E}(\eta_2) = \omega_{3,2}\mathrm{E}(Y_3) + \omega_{4,2}\mathrm{E}(Y_4) + \omega_{5,2}\mathrm{E}(Y_5)$, $\mathrm{mín}(\eta_2) = \omega_{3,2} + \omega_{4,2} + \omega_{5,2}$ y $\mathrm{máx}(\eta_2) = 10(\omega_{3,2} + \omega_{4,2} + \omega_{5,2})$.

3. Resultados

Se diseñó un cuestionario en el que cada variable manifiesta se mide con una pregunta cuya respuesta está en una escala Likert del 1 al 10. Las preguntas se redactaron de tal forma que el 1 indica el nivel más bajo de satisfacción-acuerdo y 10 el más alto. El cuestionario se aplicó a una muestra de 203 estudiantes de las carreras de Publicidad y Relaciones Públicas, Relaciones Industriales, Administración de Negocios Internacionales, Lengua Inglesa, Lengua Francesa, Informática, Biología, Enfermería, Ingeniería Civil, Economía, Adminstración, Agronomía, Derecho, Sociología e Historia. El levantamiento se hizo durante agosto 2012-julio 2013. El modelo se ajustó con el método de cuadrados mínimos parciales (CMP), Tenenhaus et al. (2005). Para ajustar el modelo usamos el software libre SmartPLS, Ringle et al. (2005). Estimamos a I sustituyendo los valores esperados de las Y's por sus respectivos promedios muestrales y a los ω's por sus estimaciones de mínimos cuadrados parciales, de modo que

$$\hat{I} = \frac{\hat{\omega}_{3,2}\bar{Y}_3 + \hat{\omega}_{4,2}\bar{Y}_4 + \hat{\omega}_{5,2}\bar{Y}_5 - (\hat{\omega}_{3,2} + \hat{\omega}_{4,2} + \hat{\omega}_{5,2})}{9(\hat{\omega}_{3,2} + \hat{\omega}_{4,2} + \hat{\omega}_{5,2})} \times 100.$$

Los alfas de Cronbach en la Tabla 3 permiten evaluar la consistencia interna del modelo de medición. Generalmente un alfa mayor que 0.8 se considera indicativo de unidimensionalidad de la variable latente subyacente. En cuanto a la capacidad predictiva del modelo tenemos que las R^2 de las variables latentes endógenas son: lealtad 0.60, satisfacción 0.72 y valor percibido 0.58. El valor del indicador es $\hat{I} = 70.46$. Una detallada inspección de los resultados, siguiendo las recomendaciones en Chin (2010), señaló que las áreas de oportunidad están en la calidad de las aulas, de las autoridades administrativas, de los horarios de clase, de las tutorías, del MEIF, y del proyecto AULA. Los detalles del análisis se pueden ver en Sosa Galindo (2014).

Variable Latente	Alfa de Cronbach
Expectativa	0.88
Calidad percibida en los elementos no humanos	0.89
Imagen	0.88
Lealtad	0.90
Satisfacción	0.94
Calidad percibida en los elementos humanos	0.82
Valor Percibido	0.86

Tabla 3: Alfas de Cronbach de cada variable latente.

4. Conclusiones

En la UV se hacen esfuerzos para alcanzar niveles de excelencia académica. Estos esfuerzos incluyen acciones estratégicas tales como el Proyecto AULA, el Proyecto 3-2-3, la departamentalización, y el MEIF en su segunda generación. Sin embargo estos proyectos no se podrán consolidar completamente sin la participación de los estudiantes. Con nuestra propuesta se espera establecer un puente para que fluya esta participación, el valor de $\hat{I} = 70.46$ indica que hay espacio para la mejora y la opinión de los estudiantes se debe escuchar.

Referencias

1. Chin, W.W. (2010). How to Write Up and Report PLS Analyses. Espocito Vinzi, V. Chin, W.W., Henseler, J., and Wang, H. (eds). *Handbook of Partial Least Squares. Concepts, Methods, and Applications*. Heidelberg: Springer.

2. Ringle, C.M, Wende, S., and Will, A. (2005). *SmartPLS*. Hamburg: Germany.

3. Sosa Galindo, I. (2014). Medición de la Calidad Institucional desde la Perspectiva de la Satisfacción de los Estudiantes. Tesis de la Maestría en Estadística Aplicada. Facultad de Estadística e Informática, Universidad Veracruzana.

4. Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.M., and Lauro, C. (2005). PLS Path Modeling. *Computational Statistics and Data Analysis*, 48, 159-205.

Prospectiva de los Costos Unitarios de Algunas Enfermedades por Grupo de Edad por Sexo

Dora Elena Ledesma Carrión[a], Lidia Hernández Hernández[b], María Teresa Leonor Muciño Porras[c], Maria Esperanza Sainz López[d]
Instituto Nacional de Estadística y Geografía(INEGI)

Clasificación: Análisis de Datos
área: Estadística Aplicada, Econometría, Series de Tiempo.
Subárea: Mínimos Cuadrados Ordinarios(MCO), Modelos ARMA.
Trabajo presentado en: XXVIII Foro Nacional de Estadística

Resumen

Se calculan los costos unitarios de seis enfermedades para derechohabientes del IMSS para todos los grupos de edad y sexo del 2010 a 2050. Se calculan las probabilidades de ingreso *a*, en tratamiento *de* y defunción *por* cada enfermedad: tumor maligno de la mama (CaMa), tumores malignos de los órganos genitourinarios (CaCu), diabetes mellitus (DM), enfermedades hipertensivas (HA), enfermedades renales: Insuficiencia renal, tubulointersticiales, otras enfermedades renales tubulointersticiales y enfermedad renal tubulointersticial, no especificada (IR) y, enfermedad por el virus de la inmunodeficiencia humana (SIDA/VIH). Bases de datos: Instituto Mexicano del Seguro Social (IMSS), Secretaría de Salud (SS), Consejo Nacional de Población (CONAPO), Instituto Nacional de Estadística y Geografía (INEGI), Asociación Mexicana de Instituciones de Seguros (AMIS). Se hace la inferencia a toda la población.

[a] dora.ledesma@inegi.org.mx
[b] lidia.hernandezh@inegi.org.mx
[c] teresa.mucino@inegi.org.mx
[d] maria.sainz@inegi.org.mx

1. Introducción

Los últimos avances en medicina han mostrado que el cambio del azúcar de caña por la fructuosa como endulzante en la dieta de los mexicanos es responsable en gran medida del deterioro físico de la población mexicana respecto a la DM[1] y enfermedades crónico-degenerativas ligadas a ella, HA e IR. Esta afirmación se debe a que la primera se sintetiza en sangre mientras que la fructuosa se degrada en el hígado. Así como la falta de prevención y factores hereditarios son la causa de enfermedades que se controlan más no se curan como el SIDA/VIH, CaMa y CaCu. Por simplicidad se dan los resultados del escenario intermedio denominado escenario II. El costo de estas enfermedades es elevado tanto por su tratamiento como por su duración. Las aseguradoras hoy en día ofrecen seguros para CaMa y no para las otras enfermedades a la población en general ya que las primas resultan onerosas.

2. Metodologías

Se analizan los datos en primera instancia para saber si hay evidencia de seguir alguna tendencia ya sea lineal, potencial, logarítmica o exponencial. Se busca en especial si siguen un comportamiento exponencial ya que seguiría la dinámica poblacional Lotka-Volterra, esto es, que la población se comporta como un modelo presa-depredador:

$$\frac{dN(t)}{dt} = N(\alpha - \beta M) \quad y \quad \frac{dM(t)}{dt} = M(\gamma - \delta N)$$

donde: N, población presa (derechohabiente, población); M, población depredadora (enfermedad); t, tiempo; $\alpha = \delta$ (CONAPO); $\beta = \gamma$ (Nuevos Casos al tiempo t) son parámetros en nuestro caso particular. En el caso del modelo estocástico los parámetros serán las tasas y su dinámica de nuevos casos, permanencia en la enfermedad y defunción por la enfermedad. Se han construido muchos modelos basados en el Lotka-Volterra al estudio de virus y demás de dos especies combinando estos modelos con los estocásticos, como aquí se aplico [3], [5] y [8]. A continuación se esquematiza la dinámica del modelo propuesto:

[1] Se ha encontrado una variante genética entre los mexicanos que aumenta 25% el riesgo de padecerla, ver: http://jama.jamanetwork.com/article.aspx?articleid=1878720#Methods

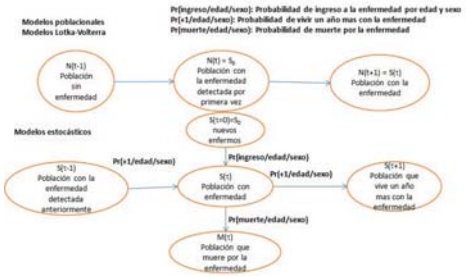


Tabla 1: Tasas por enfermedad por período según prospectiva del IMSS.

Tasas de crecimiento de pacientes en tratamiento por enfermedad por periodo								Tasas de crecimiento de gastos médicos por enfermedad por periodo							
		CaMa	CaCu	DM	HA	IR	SIDA/VIH			CaMa	CaCu	DM	HA	IR	SIDA/VIH
λ_1	2012-2020	2.80 %	-0.90 %	1.50 %	-0.20 %	10.80 %	4.60 %	μ_1	2012-2020	3.50 %	3.90 %	3.80 %	2.30 %	10.00 %	3.90 %
λ_2	2021-2030	2.20 %	-0.50 %	2.20 %	1.10 %	4.80 %	2.50 %	μ_2	2021-2030	2.60 %	3.20 %	3.80 %	3.30 %	6.10 %	3.20 %
λ_3	2031-2040	1.50 %	0.00 %	1.60 %	1.20 %	2.70 %	1.10 %	μ_3	2031-2040	1.60 %	2.50 %	3.30 %	3.50 %	4.60 %	1.80 %
λ_4	2041-2050	1.00 %	0.30 %	1.00 %	0.90 %	1.50 %	0.40 %	μ_4	2041-2050	0.60 %	1.70 %	2.90 %	3.00 %	3.80 %	0.60 %
λ	2012-2050	1.80 %	-0.20 %	1.60 %	0.90 %	4.40 %	1.90 %	μ	2012-2050	2.00 %	2.80 %	3.50 %	3.20 %	5.80 %	2.30 %

Tabla 2: Tasas por enfermedad, sexo de la población y los de 50 años y mas calculadas con datos históricos de la SS. Nuevos casos y defunciones.

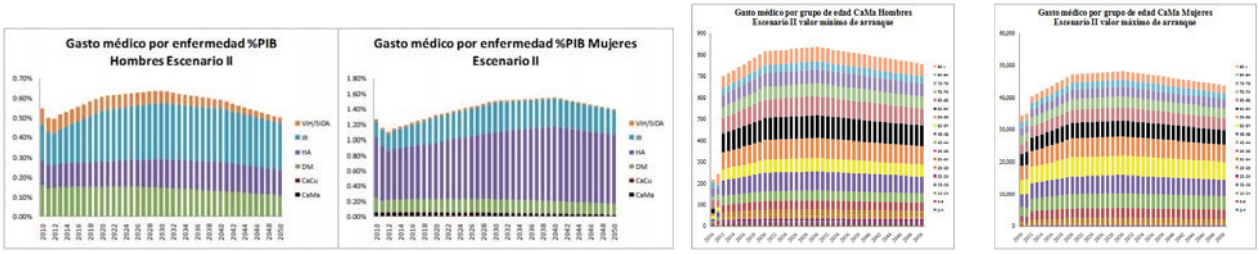
Tasa exponencial de nuevos casos						
	CaMa	CaCu	DM	HA	IR	SIDA/VIH
$\lambda_{nc,h} =$	4.88 %	-0.80 %	6.77 %	2.01 %	4.39 %	10.73 %
$\lambda_{nc,m} =$	8.69 %	-0.78 %	7.74 %	0.12 %	3.93 %	10.43 %
Tasa exponencial de defunción						
Ψ edad	CaMa	CaCu	DM	HA	IR	SIDA/VIH
$\lambda_{def,h} =$	2.80 %	3.48 %	6.01 %	6.69 %	3.79 %	1.38 %
$\lambda_{def,m} =$	3.60 %	0.58 %	4.78 %	5.41 %	2.64 %	3.67 %
Tasa de defunción 50 +						
50+	CaMa	CaCu	DM	HA	IR	SIDA/VIH
$\lambda_{def,h} =$	2.024 %	3.48 %	6.16 %	6.71 %	3.82 %	4.41 %
$\lambda_{def,m} =$	4.45 %	0.92 %	4.88 %	5.53 %	2.83 %	6.42 %

Se toma la población derechohabiente del IMSS como muestra y posteriormente se hace inferencia a la población en general tomando los datos poblacionales de CONAPO[2].
Se manejaron 4 escenarios de proyección del producto interno bruto (PIB): Escenario I: Se analizan los datos trimestrales del PIB de 1996 a 2012 y se convierten a base 2012 para un AR(2)MA(2). Escenario II: Se analizan por series de tiempo los datos trimestrales del PIB de 1980 a 2012 y se convierten a base 2012. Se ajustó un modelo de tendencia sin problemas de heteroscedasticidad. Escenario III (pésimo): Crecimiento entre 1.5 % - 2.5 %. Y el escenario IV (óptimo): Crecimiento del 2.5 % para el 2013 y constante del 4 % del 2014 al 2050.
El método de Runge-Kutta orden 4 resuelve sistemas de ecuaciones diferenciales ordinarias (EDO) dando las condiciones iniciales y los parámetros del sistema, los modelos autorregresivos de promedios móviles (en inglés, ARMA) sirven para proyectar el PIB en el periodo

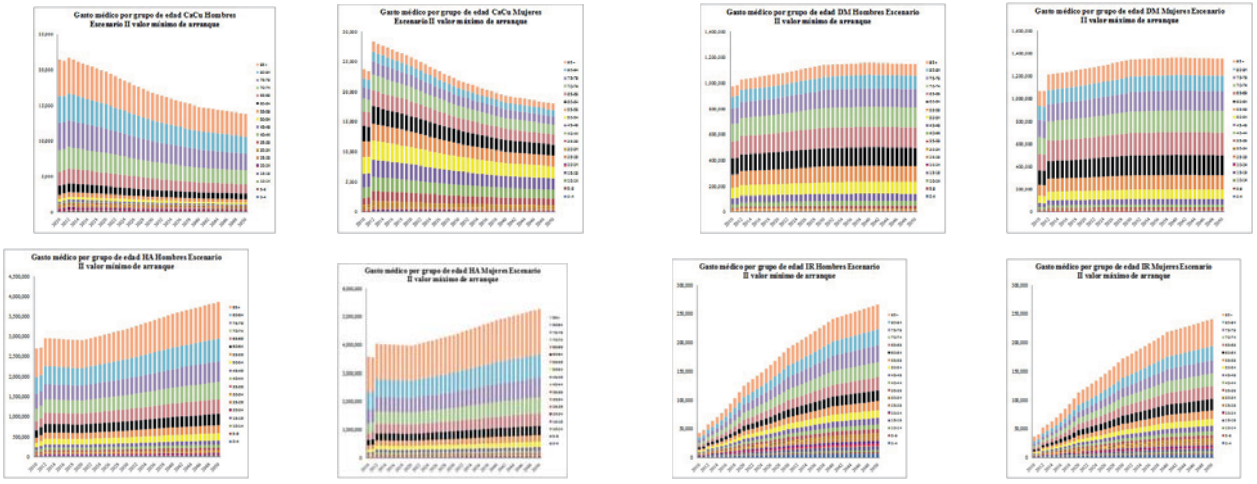
[2] Véase http://www.conapo.gob.mx/es/CONAPO/Proyecciones

en el 2010-2050. Estas metodologías son ampliamente conocidas y pueden ser consultadas en cualquier libro de texto de métodos numéricos, econometría, estadística y matemáticas aplicadas. La Tabla 1 y Tabla 2 muestra los parámetros del sistema.

3. Resultados

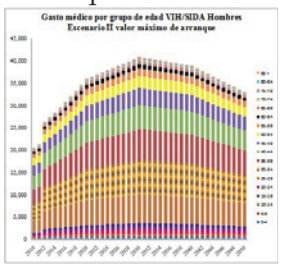


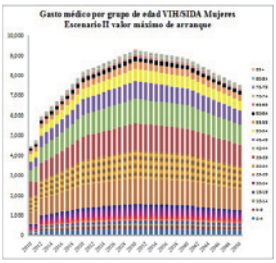
Para hombres la enfermedad más cara sería la IR y para las mujeres HA. La disminución en los casos de SIDA/VIH tanto en hombres como en mujeres obedece a las políticas de prevención. Y mientras disminuye los costos por DM se incrementa los de HA e IR. En el caso del CaMa, las edades de mayor riesgo-costo oscilan entre los 45 y 64 años para los hombres y de los 35 a los 59 años para las mujeres. El CaMa se presenta con mayor incidencia en los grupos de edad de 15 a 24 años y de 60 a 69 años para los hombres. Para las mujeres 15 a 19 y de 50 a 59 años de edad.



En el caso del CaCu, las edades de mayor gasto-incidencia en los hombres se encuentran desde los 60 años de edad y para las mujeres empieza a ser notorio de entre los 30 y 40 años manteniéndose hasta los 74 años. En el caso de la DM tipo I y II, el comportamiento es

similar entre hombres y mujeres siendo más impactante en las mujeres. Es notorio el gran número de casos en adultos jóvenes. Para la HA, en ambos sexos se incrementan los casos años con año. En ambos sexos empiezan a ser notoria la incidencia a partir de los 40 años. En el caso de la IR, crece rápidamente, más que la HA y la DM tanto en hombres como en mujeres. Las edades donde impacta notoriamente son a partir de los 50 años siendo a los 80 y más el mayor incremento. La enfermedad del SIDA/VIH empieza a manifestarse en los hombres entre los 20 y 24 años y las mujeres entre los 15 y 19 años. Siendo su máxima incidencia en ambos casos entre los 25 y 39 años. Los casos tienen un máximo en 2030 cuando se espera se invierta la pirámide poblacional.





4. Conclusiones

De los resultados se observan que la DM, la HA e IR no están controladas, *i.e.*, en 2010 no existían políticas públicas para estas enfermedades, como en el caso del CaMa y SIDA/VIH. En congruencia con especialitas se observa una estrecha relación entre la DM y la HA degenerando en daño renal. A continuación en la Tabla 3 se presenta la proyección para 2025 de la AMIS y la del modelo aquí porpuesto:

Tabla 3: AMIS vs. Modelo 2025

2025	AMIS			Modelo					
	Hombres	Mujeres	Total	Hombres	Mujeres	Total	Hombres(% PIB)	Mujeres(% PIB)	Total(% PIB)
CaMa	6,212	895,992	902,204	3,255	183,038	186,293	0.00027	0.04264	0.04291
CaCu	354,092	119,466	473,558	39,695	92,584	132,280	0.00102	0.01455	0.01557
DM	1,882,277	1,115,012	2,997,289	1,845,402	1,887,260	3,732,661	0.16253	0.18989	0.35241
IR	540,197	537,595	1,077,792	976,087	778,595	1,754,682	0.32782	0.43290	0.76072
HA	1,900,913	1,473,408	3,374,321	6,068,570	5,583,175	11,651,745	0.15605	0.87713	1.03317
SIDA/VIH	155,303	999	156,302	613,439	155,700	769,139	0.06686	0.02432	0.09117

De la Tabla 3 la AMIS sobrestima los casos de Cáncer y subestima IR, HA y SIDA/VIH. Donde hay coincidencia es en DM. En 2014 se han implantado políticas de no sal en las mesas de los restaurantes, no anuncios de "comida chatarra" en los horarios de niños, y el impuesto al azúcar en las golosinas. Queda pendiente recalcular el modelo dentro de algunos

años para evaluar estas políticas públicas.

Referencias

1. CEPAL(2008), *"Directrices para la elaboración de módulos sobre envejecimiento en las encuestas de hogares Centro Latinoamericano y Caribeño de Demografía"*,ONU, ISBN: 978-92-1-323244-6 LC/L.2969-P, Santiago de Chile, noviembre de 2008.

2. Comisión Nacional de Arbitraje Médico CONAMED(2001), *"Recomendaciones para mejorar la práctica en la atención del paciente con cáncer"*, ISBN 970-721-215-2, México.

3. Nguyen Huu Du, Vu Hai Sam(2006), *"Dynamics of a stochastic Lotka-Volterra model perturbed by white noise"*, Journal of Mathematical Analysis and Applications, 324, 82-97.

4. Prajneshu(1980), *"Diffusion approximations for models of population growth with logarithmic interactions"*, Stochastic Processes and their application Vol. 10 pp. 87-99, North-Holland Publishing Company.

5. Rui Xu, Chaplain M.A.J., Davidson F.A.(2006), *"A Lotka-Volterra type food chain model with stage structure and time delays"*, Journal of Mathematical Analysis and Applications, 315, 90-105.

6. SS(2006), *"La mortalidad en México, 2000-2004. Muertes Evitables: magnitud, distribución y tendencias"*, México.

7. Villarreal Ríos Enrique, Campos Esparza Maribel, Galicia Rodríguez Liliana, Martínez González Lidia, Vargas Daza Emma Rosa; Torres Labra Guadalupe, Patiño Vega Adolfo, Rivera Martínez María Teresa, Aparicio Rojas Raúl, Juárez Durán Martín(2011), *"Costo anual per cápita en primer nivel de atención por género"*, Unidad de Investigación Epidemiológica y en Servicios de Salud Querétaro,IMSS, 16(3):1961,1968.

8. Zhu C., Yinb G.(2009), *"On competitive Lotka-Volterra model in random environments"*, Journal of Mathematical Analysis and Applications, 357, 154-170.

Bayesian Approach for Modeling Speed and Direction of Wind

José Martín Cadena Barajas[a]
Doctorado en Matemáticas, Universidad Veracruzana

Sergio Fco. Juárez Cerrillo
Facultad de Estadística e Informática, Universidad Veracruzana

David A. Stephens
Department of Mathematics and Statistics, McGill University

Clasificación: Tesis de Doctorado.
Área: Inferencia Bayesiana.
Subárea: Datos Circulares.
Trabajo presentado en: XXVIII Foro Nacional de Estadística.

1. Introduction

This work is motivated for modeling wind data recorded in the nuclear power plant Laguna Verde, located in Veracruz state. The full data base consists of hourly observations of speed and direction from 2000 to 2007. Speed is measured in meters/second and direction is recorded in degrees. The observations are recorded at 10 mts and 60 mts from the ground. In this work we analyze the data corresponding to 2007 at 10 mts from the ground, see Figure 1. The time series has 8,760 observations; 52 were calm wind (a calm wind is defined as a wind speed below the measurement threshold of the wind instrument); and 40 values were missing. The calm winds and the missing values were ignored of the analysis.

[a] martincadenab@gmail.com

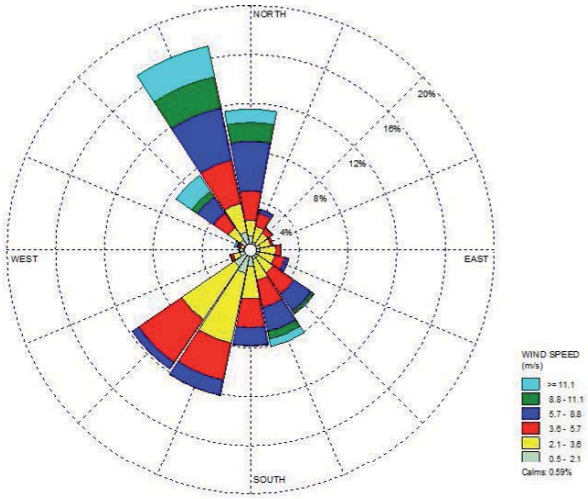


Figura 1: Wind data modeled.

2. The Model

Let (S_t, D_t), $t = 1, \ldots, n$, be the speed-direction wind time series. For the speed S we choose the log-normal distribution. It is known that this distribution provides good fit to wind speed, García Torres et al. (1998) and Luna and Church (1974). The direction D is an angular variable as in Mardia and Jupp (1999). Based on the directional approach it is possible to fit circular distributions and to construct joint distributions for the speed and direction as in Carta et al. (2008a) and Carta et al. (2008b) or to use auto-regressive models as in Erdem and Shi (2011). However we treat the direction as a discrete variable with support values on the 16 cardinal directions shown in Figure 1, $\{1 = \text{N}, 2 = \text{NNE}, 3 = \text{NE}, 4 = \text{ENE}, 5 = \text{E}, 6 = \text{ESE}, 7 = \text{SE}, 8 = \text{SSE}, 9 = \text{S}, 10 = \text{SSW}, 11 = \text{SW}, 12 = \text{WSW}, 13 = \text{W}, 14 = \text{WNW}, 15 = \text{NW}, 16 = \text{NNW}\}$, and use the multinomial distribution as the model for direction.

We explored several dependence structures for (S_t, D_t) using the deviance as a model selection criterion. We ended up with speed at time $t + 1$ depending on i) the direction at time $t+1$ and ii) the speed and direction at time t; while direction at time $t+1$ depends only on ii). The influence diagram in Figure 2 represents the selected model. The joint probability distribution of the probabiistic structure in Figure 2 is

$$f_{S,D}(s, d|\beta, \sigma^2, \gamma) = g_D(d|\gamma)h_{S|D}(s|d, \beta, \sigma^2),$$

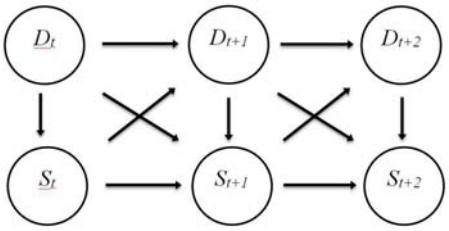


Figura 2: Dependence structure used for speed and direction.

where g_D is the multinomial probability mass function, $h_{S|D}$ is the normal density function, the argument s is the logarithm of the speed, and β and γ are regression hyperparameters that will be explained bellow. The posterior distribution for the unknown parameters is

$$\pi(\beta, \gamma, \sigma^2|s, d) \propto f_{S,D}(s, d|\beta, \gamma, \sigma^2)p(\beta, \gamma, \sigma^2)$$

where $\beta = (D_{t+1}, S_t, D_t)$, σ^2 is the variance of $h_{S|D}$ and $\gamma = (D_t, S_t)$. The term $f_{S,D}$ is proportional to the likelihood of our model and it reflects the observed data while $p(\beta, \gamma, \sigma^2)$ is the prior probability and can include the meteorological knowledge about the system under study. Assuming that the data values are obtained independently, the likelihood function is given by the following conditional independence structure

$$L(\beta, \gamma, \sigma^2) = \prod_{t=1}^{n} \quad g_D(d_{t+1}|\gamma) \quad h_{S|D}(S_{t+1}|\beta, \sigma^2) \tag{1}$$

where $h_{S|D}$ is the normal density function with mean $\mu(\beta) = \beta_0 + \beta_1 S_t + \beta_2 D + \beta_3 D_t$ and variance σ^2 and g_D is the multinomial probability mass function with parameter $\lambda(\gamma) = (\lambda_1(\gamma), \ldots, \lambda_{16}(\gamma))$ where

$$\lambda_k(\gamma) = \frac{\exp(Z_k\gamma)}{1 + \sum_{l=1}^{15} \exp(Z_l\gamma)}, \quad k = 1, \ldots, 15,$$

and

$$\lambda_{16}(\gamma) = \frac{1}{1 + \sum_{l=1}^{15} \exp(Z_l\gamma)},$$

where $\lambda_k(\gamma) = P(Z_{tk} = 1)$ with $Z_{tk} = 1$ if direction at time t, D_t, falls into direction k and $Z_{tk} = 0$ otherwise; and $Z_k\gamma = \gamma_{0k} + \gamma_1 D_{tk} + \gamma_2 S_{tk}, \quad k = 1, \ldots, 15$. Then, the likelihood (1) has the form

$$L(\beta, \gamma, \sigma^2) = \prod_{k=1}^{16} \lambda_k(\gamma)^{z_{tk}} \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(s_{t+1} - \mu(\beta))^2}{2\sigma^2}\right)$$

Note that $\mu(\beta)$ and $\lambda(\gamma)$ reflect the dependence structure in Figure 2. Suppose the wind blows on direction j, then $\sum_{k=1}^{16} \log\left(\lambda_k(\gamma)^{z_{tk}}\right) = z_{tj} \log\ \lambda_j(\gamma)$, thus the log-likelihood can be writen as follows:

$$\log(L(\beta, \gamma, \sigma^2)) = \sum_{t=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(s_t - \mu(\beta))^2}{2\sigma^2}\right)\right) + z_{tj} \log \lambda_j(\gamma).$$

We assume flat prior distributions for β and γ and an informative prior for σ^2. The prior proposed for σ^2 is an inverse gamma with parameters a and b. Thus the posterior distribution has the form:

$$\pi(\beta, \gamma, \sigma^2 | s, d) \propto \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(s_{t+1} - \mu(\beta))^2}{2\sigma^2}\right) \prod_{k=1}^{16} \lambda_k(\gamma)^{z_{ik}} \quad \frac{b^a}{\Gamma(a)} (\sigma^2)^{-a-1} \exp\left(\frac{-b}{\sigma^2}\right).$$

There is no full conditional for β, and γ in a standard form so we use the general Metropolis-Hastings update. Take $\mu(\beta) = \mu$, for sampling σ^2 note that

$$\begin{aligned} \pi(\sigma^2 | \beta, \gamma, s, d) \quad &\propto \quad \left\{\prod_{t=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(s_{t+1} - \mu)^2}{2\sigma^2}\right)\right\} \times \frac{b^a}{\Gamma(a)} (\sigma^2)^{-a-1} \exp\left(\frac{-b}{\sigma^2}\right) \\ &= \quad \frac{1}{(2\pi)^{n/2}} \frac{b^a}{\Gamma(a)} \left(\sigma^2\right)^{-\left(\frac{n}{2} + a\right)-1} \times \exp\left(\frac{-\left(\frac{1}{2}\sum_{t=1}^{n}(s_{t+1} - \mu)^2 + b\right)}{\sigma^2}\right) \end{aligned}$$

Then

$$\pi(\sigma^2 | \beta, \gamma, s, d) \propto \mathcal{IG}\left(\frac{n}{2} + a\ ,\ \frac{1}{2}\sum_{t=1}^{n}(s_{t+1} - \mu)^2 + b\right)$$

We are now in a position to run the Metropolis-Hastings algorithm as follows:

1. Take the regression coefficients of the models fitted for $\mu(\beta)$ and $\lambda(\gamma)$ and the estimation of σ^2 as starting values for the Markov chain.

2. For $t = 1, \ldots, n-1$, update each variable in turn:

   a) Sample $\gamma^{(t+1)} \sim \pi(\gamma^{(t)}, \beta^{(t)}, \sigma^{2(t)})$. This requires a Metropolis-Hastings update:

   - Propose a new value for γ, say γ^*, according to a multivariate normal distribution $q(\cdot|\gamma^{(t)})$.
   - Compute the accept-reject ratio $\alpha(\gamma, \gamma^*) = \text{mín}\left(\frac{\pi(\gamma^*)}{\pi(\gamma^{(t)})}\frac{q(\gamma^{(t)})}{q(\gamma^*)}, 1\right)$.
   - Accept the new value γ^* with probability $\alpha(\gamma, \gamma^*)$, otherwise take the next value of γ the same as before.

   b) Sample $\beta^{(t+1)} \sim \pi(\gamma^{(t+1)}, \beta^{(t)}, \sigma^{2(t)})$ with the Metropolis-Hastings update:

   - Propose a new value for β, say β^*, according to a multivariate normal proposal distribution $h(\cdot|\beta^{(t)})$.
   - Compute the accept- reject ratio $\alpha(\beta, \beta^*) = \text{mín}\left(\frac{\pi(\beta^*)}{\pi(\beta^{(t)})}\frac{h(\beta^{(t)})}{h(\beta^*)}, 1\right)$.
   - Accept the new value β^* with probability $\alpha(\beta, \beta^*)$, otherwise take the next value of β the same as before.

   c) Sample $\sigma^{2(t+1)} \sim \pi(\gamma^{(t+1)}, \beta^{(t+1)}, \sigma^{2(t)})$ from $\mathcal{IG}\left(\frac{n}{2} + a\ ,\ \frac{1}{2}\sum_{t=1}^{n}(s_{t+1} - \mu)^2 + b\right)$.

3. Store the values $(\gamma^{(t+1)}, \beta^{(t+1)}, \sigma^{2(t+1)})$.

3. Results

We wrote R code for the posterior probability function and the Metropolis-Hastings algorithm presented in the previous section. We also tested the effectiveness of two variants of the acceptance-rejection method (Metropolis and Metropolis Hastings) and two ways for generating candidates (independence and random walk). We found that the combination that works best is Metropolis with independent generation. Traceplots and density plots for 9,000 draws with a burn-in of 1,000 iterations for γ_i,$i = 1, \ldots, 4$ are shown in Figure 3.

4. Further Research

Currently we are doing formal convergence diagnostics in order to use the model for short-term forecast.

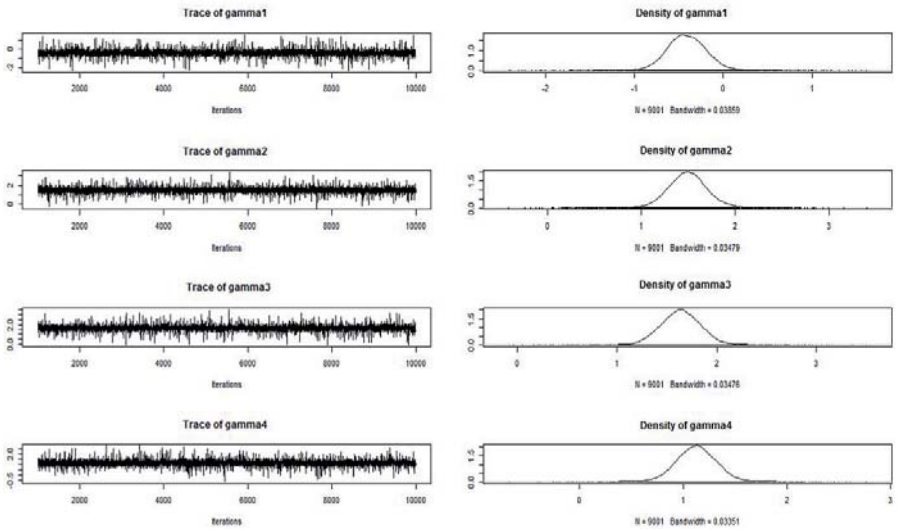


Figura 3: Traceplots and density plots of γ_1, γ_2, γ_3 and γ_4.

Referencias

1. Carta, J.A., Bueno, C., and Ramírez, P. (2008). Statistical Modelling of Directional Wind Speeds using Mixtures of von Mises Distributions: Case study, Energy Conversion and Management, Volume 49, Issue 5, May 2008, pp. 897-907

2. Carta, J.A., Ramírez, P., and Bueno, C. (2008). A Joint Probability Density Function of Wind Speed and Direction for Wind Energy Analysis, Energy Conversion and Management, Vol. 49, Issue 6, pp. 1309-1320

3. Erdem, E. and Shi, J. (2011) ARMA based Approaches for Forecasting the Tuple of Wind Speed and Direction, Applied Energy, Volume 88, Issue 4, pp. 1405-1414

4. A. Garcia, A., Torres, J.L., Prieto, E., and de Francisco, A. (1998). Fitting Wind Speed Distributions: A Case Study, Solar Energy, Volume 62, Issue 2 , pp. 139-144

5. Luna, R., Church, H. (1974). Estimation of Long-Term Concentrations Using a Universal Wind Speed Distribution, Journal of Applied Meteorology, 13, pp. 910-916.

6. Mardia, K. V. and Jupp, P. E. (1999). Directional Statistics. New York: Wiley.

Análisis estadístico del portafolio de evidencias como estrategia didáctica para el desarrollo de competencias genéricas

L.C.yT.E. Milady Lucia Ruiz Mendoza, L.E. Araceli Pineda Moreno[a]
Universidad Veracruzana

Clasificación: Tesis de Licenciatura.
Área: Análisis Multivariado.
Subárea: Correspondencia, Comparaciones pareadas para medidas repetidas (intervalos simultáneos y de Bonferroni).
Trabajo presentado en: XXVIII Foro Nacional de Estadística.

1. Introducción

El presente trabajo es una investigación que tiene por objetivo analizar estadísticamente el desarrollo de competencias genéricas: Instrumentales, Interpersonales y Sistémicas, a partir del portafolio de evidencias, como instrumento de recopilación de productos en la evolución del aprendizaje de los alumnos del cuarto semestre de la asignatura ?Estructura Socioeconómica de México? del Colegio Preparatorio de Xalapa, aplicado al enfoque educativo basado en competencias. De igual manera, se busca contribuir a la investigación educativa, mostrando de forma sistemática el tratamiento estadístico propuesto para datos emanados del pre-test y post-test, mediados a partir de la utilización del portafolio de evidencias, para evaluar el desarrollo de competencias genéricas en la población bajo estudio y estimar la efectividad de la herramienta pedagógica como instrumento de apoyo.

El *portafolio de evidencias* es un instrumento que permite la compilación de todos los trabajos realizados por los estudiantes durante un curso o ciclo escolar, pueden ser agrupados datos de vistas técnicas, resúmenes de textos, proyectos, informes, anotaciones diversas, en

[a] mlucia_ruizm@gmail.com

fin, todos los productos que el profesor requirió del estudiante como muestra de aprendizaje. El portafolio de evidencias, también compila, las pruebas y las autoevaluaciones de los alumnos; la finalidad de este instrumento es auxiliar al estudiante a desarrollar la capacidad de evaluar su propio trabajo, reflexionando sobre él, mejorando su producto y al profesor, ofrece la oportunidad de trazar referencias de la clase como un todo, a partir del análisis individual, conociendo la evaluación de los alumnos a lo largo del proceso de la enseñanza y del aprendizaje.

En torno a las *competencias* y en el entendido de que el termino es polisémico y tal vez hasta problemático porque su origen no es único, sino múltiple, y porque llegó al lenguaje educativo proveniente del mundo laboral; se dice que una persona tiene una determinada competencia cuando muestra desempeños eficientes y eficaces, en un campo específico, en el desarrollo de tareas concretas y distinguidas, con respuestas o soluciones variadas y pertinentes, con recursos propios y externos, que vistos desde criterios objetivos y válidos permiten concluir la existencia de una determinada competencia.

Es entonces cuando para una formación integral se necesita una herramienta de evaluación integral y el portafolio de evidencias resulta ser un instrumento de evaluación del razonamiento reflexivo, propicia oportunidades para documentar, registrar y estructurar los procedimientos y el propio aprendizaje, y es por esa reflexión que el estudiante puede, con ayuda del profesor, verificar lo que necesita mejorar en su desempeño. Por otro lado, el portafolio permite al profesor conocer mejor a su alumno, sus ideas, sus expectativas, su concepción de mundo, se convierte en un instrumento de diálogo entre el profesor y el alumno, de manera que se espera que posibilite nuevas formas de ver e interpretar un problema y solucionarlo, he aquí la utilidad de la estadística.

2. Marco Teórico

El término competencia en su sentido más simple y general tiene dos significados. Tal como lo describe el diccionario electrónico etimológico: es un sustantivo femenino que pertenece a la familia léxica de dos verbos diferentes, competer y competir.

Competer que proviene del latín *competentĭa*; cf. *Competente*, que por un lado denota incumbencia, y por otro poseer la pericia, aptitud, idoneidad para hacer algo o intervenir en un asunto determinado. (Real Academia Española (22.a ed.)., 2001)

Competir que proviene del latín *competentĭa*; cf. *Competir*, que denota una disputa, oposición rivalidad entre dos o más personas sobre algo, para obtener una misma cosa. (Real Academia Española (22.a ed.)., 2001)

Con el propósito de explicar el término de competencia así como su definición, se pueden tomar en cuenta las siguientes preguntas propuestas por Zabala y Arnau (2007) Presentadas en la Tabla 1.

Tabla 1. Preguntas guía sobre el concepto de competencias

Preguntas	**Respuestas**
¿Qué es?	Una capacidad.
¿Para qué?	Para resolver situaciones, conflictos y demandas.
¿Cómo?	De forma idónea.
¿Dónde?	En un contexto determinado.
¿Con qué?	La movilización de saberes, actitudes, habilidades y aptitudes.
¿Cuándo?	Al mismo tiempo y de manera interrelacionada.

De acuerdo a los anterior y como bien lo cita Fuentes, D. M. (2012) es necesario que el estudiante aprenda a desarrollar todos los componentes en un mismo contexto y tiempo ya que de nada sirve si sabe, sabe hacerlo, sabe estar, pero no quiere hacerlo; esto se traduciría como una competencia no desarrollada, o en otras palabras que no será llevada a la práctica.

El desarrollo del concepto de competencias ha llevado a que se distingan o se reconozcan diferentes tipos de competencias y que en general corresponden a las competencias básicas, genéricas o transferibles y competencias técnicas o específicas (Gutierrez O. A., 2005). Sin embargo, el presente documento solo analiza el desarrollo de competencias genéricas o transferibles mismas que se describen a continuación según Villa y Poblete (2010).

Tabla 2. Organización de competencias, sub-competencias y características o habilidades a desarrollar por el alumno.

Competencias	**Sub-competencias**	**Habilidades o actitudes a desarrollar**
Instrumentales	Cognitivas	Pensamiento; Analítico, Sistémico, Crítico, Reflexivo, Lógico, Analógico, Practico, Colegiado, Creativo y Deliberativo.
	Metodológicas	Gestión de tiempo Resolución de problemas Toma de decisiones Orientación al aprendizaje (en el marco pedagógico, estrategias de aprendizaje)
	Tecnológicas	Uso de las TIC Utilización de base de datos
	Lingüísticas	Comunicación verbal Comunicación escrita Manejo de idioma extranjero
Interpersonales	Individuales	Automatización Diversidad e interculturalidad Resistencia y adaptación al entorno Sentido ético
	Sociales	Comunicación interpersonal Trabajo en equipo Tratamiento de conflictos y negociación
Sistémicas	Organización	Gestión por objetos Gestión de proyectos Orientación a la calidad
	Capacidad	Creatividad

3. Conclusiones

Con base a los resultados obtenidos y de acuerdo al análisis estadístico desarrollado, se puede concluir que existen diferencias significativas entre el diseño curricular en procesos: propósitos y temas (modelo anterior) y el diseño curricular por competencias. Adicionalmente y con respecto a los resultados obtenidos del análisis de muestras pareadas para medidas repetidas se afirma que la estrategia didáctica "portafolio de evidencias.es una herramienta que contribuye al desarrollo de competencias genéricas individuales, interpersonales y sistémicas tal y como lo menciona Villa y Poblete (2010). Específicamente en las sub competencias: *metodológicas* que corresponde a la competencia genérica individual, las sub competencias de *organización* y *capacidad emprendedora* que se encuentran en el apartado de competencias sistémicas.

Como se puede apreciar los alumnos del Colegio Preparatorio de Xalapa de la asignatura Estructura Socioeconómica de México, desarrollaron la sub competencia *metodológica* por lo cual desenvolvieron habilidades y actitudes como: gestión de tiempo, resolución de problemas, toma de decisiones así como nuevas técnicas de aprendizaje. Mientras que en la sub competencia de *organización* aumento la gestión de tiempo, de proyectos y orientación a la calidad. Sin embargo, a partir del desarrollo de la sub competencia de *capacidad emprendedora* los alumnos incrementaron su creatividad, espíritu emprendedor e innovación como se muestran en la Tabla 2, según Villa y Poblete (2010) y Aguerrondo (1999).

Hay que hacer notar que las sub competencias de la competencia interpersonal (individuales y sociales). Así como las sub competencias (cognitivas, tecnológicas, lingüísticas) pertenecientes a la competencia individual y la sub competencia (liderazgo) para la competencia sistémica tuvieron el mismo efecto en el diseño curricular en procesos: propósitos y temas (modelo anterior) y el diseño curricular en competencias. Es importante hacer notar en esta conclusión que el diseño del portafolio de evidencias esta puntualizado en el desarrollo de ciertas competencias; y es algo que se ve reflejado entre el análisis, los resultados obtenidos y los objetivos de esta investigación.

Referencias

1. Gutierrez, D. O. (Mayo-Junio de 2005). Desarrollo de Competencias y Habilidades. México. Recuperado el 17 de Abril de 2013, de *www.uvmnet.edu/ praxis/ presenta/*

competencias.ppt

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

3. Richard A. Johnson, Dean W. Wichern. (2007). *Applied Multivariate Statistical Analysis* (Sixth Edition ed.). USA: Pearson.

4. Sánchez, M. Y. (Agosto-Diciembre de 2011). Cap.6 Medidas repetidas. Universidad Veracruzana. Xalapa, Veracruz, México.

Análisis de sensibilidad de proyecciones de población a pequeños cambios de la Tasa Global de Fecundidad

Milenka Linneth Argote Cusi[a]
Consultora-Investigadora de Corpotalentos sede Colombia

1. Introducción

Actualmente la demografía se encuentra en un periodo de reflexión y de análisis de los métodos tradicionales hasta ahora aplicados para el estudio y modelación de la dinámica poblacional (Gilbert y Michel, 2005). Dos aspectos favorecen este salto, por un lado la disponibilidad actual de diversos tipos de información con mejor calidad y el uso intensivo de las computadoras en el procesamiento, búsqueda, selección y clasificación de esa información para la toma de decisiones.

Considerando el comportamiento no lineal de las variables demográficas (Caswell, 2008 y 2009) y la existencia de incertidumbre ligada a estos sistemas el problema que orienta la presente investigación es: ¿Cuál es el impacto de pequeños cambios en los valores del estimador de la TGF en un ejercicio de proyección de población? En 1963, Edward Lorenz se preguntó del ¿por qué pequeños cambios en los parámetros de un modelo de predicción, daban resultados tan diferentes? La respuesta a esta pregunta se llamó "el *efecto mariposa*" para explicar cómo pequeños errores provocan grandes errores (amplificación del error). En 2011 Argote encontró que a pequeños cambios de la TGF, indicador resumen la fecundidad de una población de mujeres, los nacimientos en los diferentes grupos quinquenales de edad variaban ampliamente (20 % aproximadamente) presentándose un efecto amplificador en algunos grupos. En un proceso inverso ¿Cómo varían las estimaciones producto de un ejercicio

[a] milenqita@hotmail.com

de proyección de población si se realizan pequeños cambios a uno de los parámetros (en este caso la TGF)?

Es decir, sea $T\bar{G}F_i^x$ la media de TGF_i^x una variable aleatoria que representa la distribución por muestreo de la Tasa Global de Fecundidad en la iteración i del año x, donde $i = 1, 2, ..., 1000$, $x = 1, ..., n$ años. Además TGF_i^x es un estimador de razón definido por:

$$TEF_{x,x+5}^{t,t+1} = \frac{Nac_{x,x+5}^{t,t+1}}{Temujeres_{x,x+5}^{t,t+1}} \tag{1}$$

$$TGF_i^x = 5 * \sum_{i=1}^{7} TEF_{x,x+5}^{t,t+1} \tag{2}$$

Donde, $Nac_{x,x+5}^{t,t+1}$ son los nacimientos entre el periodo $t, t+1$ y las edades $x, x+5$ y $Temujeres_{x,x+5}^{t,t+1}$ es el tiempo de exposición de las mujeres a experimentar el evento de tener un hijo en el periodo $t, t+1$y las edades $x, x+5$, expresado en años-persona.

Y sea $P_i'^x$ la población estimada en un momento especifico x acorde a su distribución por muestreo, $\bar{P}_i^x$ la media de la distribución por muestreo de la población proyectada, α_i^x la diferencia entre las dos anteriores y w_i^x la diferencia entre un valor puntual $TGF_i'^x$ de la distribución estadística por muestreo de la TGF_i^x y la media $T\bar{G}F_i^x$:

$$TGF_i'^x - T\bar{G}F_i^x = \alpha_i^x * 100$$

$$P_i'^x - \bar{P}_i^x = w_i^x * 100$$

$$\alpha_i^x > A * w_i^x \tag{3}$$

Se espera encontrar que a pequeñas variaciones de la tasa global de fecundidad, la diferencia entre los valores de la correspondiente población proyectada $P_i'^x$ presentan el efecto mariposa de tal manera que la variación α_i^x es mayor a w_i^x en A veces.

2. La complejidad de los indicadores demográficos

De acuerdo a Escobedo (2007) los datos son elementos complejos que se construyen a partir de filosofía y metodología que provienen del investigador u observador. Entonces es ilógico considerar a los datos como simples símbolos, cuando son función de una o más variables lineales o no lineales.

En este sentido, muchas controversias emergen de los indicadores demográficos debido a su dificultosa tarea de representar el estado de los sistemas sociales y mayor es la discusión acerca de su precisión y el manejo del error (Gottlieb,2001; Bagajewicz, 2005). Por ejemplo la Tasa Global de fecundidad es un indicador resumen que representa el número de hijos por mujer al final de su vida reproductiva bajo el supuesto que a lo largo de su vida tendrá la fecundidad presente (supuesto de cohorte ficticia); la complejidad de este estimador de razón se observa en su definición formal (ver introducción).

La demografía nos provee de un marco teórico y de métodos para la estimación de indicadores que representan la dinámica de la población. Debido a que la dinámica poblacional involucra un conjunto de variables interrelacionadas y los métodos para la captura de dicha información, las estimaciones están sujetas a determinado margen de error entre los cuales podemos mencionar, los errores por muestreo, errores humanos, errores de trabajo de campo, etc. Sin embargo, los ajustes y correcciones parecen una tarea de nunca acabar cuando nos queremos acercar a la estimación más próxima a la realidad. En este sentido es frecuente la corrección de datos poblacionales por cientos o miles de personas, que podrían ser beneficiarios o no de determinada política de salud o de mitigación de la pobreza.

Siempre me he preguntado cuando estimábamos la Tasa Global de Fecundidad (TGF) u otros indicadores demográficos por diferentes métodos en mis clases de posgrado, ¿Cuál estimación elegimos? ¿Se puede redondear fácilmente el valor de las estimaciones que tienen varios decimales?, ¿Qué representan estos decimales? ¿Es posible mejorar las estimaciones y errar menos?. En su esencia estas preguntas tienen origen matemático más que demográfico y la verdad, es que el tema de la precisión o de cuan exactamente podemos hacer estimaciones, ha tenido origen en las matemáticas hace cientos de miles de años. Por ejemplo, ¿cómo midió Eratóstenes el tamaño de la tierra alrededor de 250 a.c.? Estimo un valor de 39.250 kms y la cifra moderna es de 39.840 kms (Stewart, 2007), con cual estimación nos quedamos?.

Existen investigaciones que han profundizado en el tema de la precisión de los indicadores y de la interpretación compleja de los datos. Griffiths et. al. (2000) analizaron por qué razón el índice de masculinidad se ha mantenido a favor de los hombres a través del tiempo en la India, en relación a las fuentes de datos y variables culturales que pueden estar influyendo. Por otro lado, en España debido a las bajas tasas de fecundidad que están por debajo del nivel de reemplazo, se han realizado análisis mas detallados desde las fuentes de información, el efecto

de la edad, la infertilidad, etc. para verificar si realmente la fecundidad esta disminuyendo y cuando se estabilizaría (Ortega et. al., 2000).

Alho y Spencer (2005) dedican varios apartados al análisis de la precisión y formalización de la función de error en el modelado de los eventos demográficos. El análisis de la precisión nos lleva de manera natural al análisis de sensibilidad de lo datos, es decir, si consideramos a los indicadores como variables aleatorias con infinitos valores posibles en un espacio muestral determinado, el análisis de sensibilidad busca determinar cuales son los parámetros más sensibles a pequeñas variaciones de los parámetros en un modelo matemático.

3. Análisis de sensibilidad

Considerando la incertidumbre atada a los escenarios futuros y que la información de proyecciones de población se utiliza para la toma de decisiones en política pública, el análisis de sensibilidad es un tema transcendental en la demografía actual. Partiendo de la pregunta: ¿Cuál es el impacto de pequeñas variaciones de la TGF en un ejercicio de proyección de población?, el análisis de sensibilidad se constituye en la herramienta científica con la cual podemos responder a la pregunta.

Este no es el primer estudio que realiza análisis de sensibilidad en el campo demográfico. Para Caswell (2008) la importancia de los análisis de sensibilidad radica en su aplicación en política pública y en teoría del muestreo ya que los parámetros que son más sensibles son los que deberían ser estimados de forma más precisa. Este mismo autor ha venido profundizando en el modelado de los sistemas estocásticos dentro los cuales considera la "*individualidad estocástica*" en el ámbito demográfico. Este trabajo ha desarrollado el modelado de varios tipos de población mediante matrices estocásticas, las cuales le han permitido realizar el análisis de la perturbación que incluye la sensibilidad y elasticidad (Caswell, 2009).

Un ejemplo clásico sobre sensibilidad, son las investigaciones de Edward Lorenz (1963) sobre el modelado de datos atmosféricos para predecir el clima. Uno de los hallazgos mas importantes, ha sido llamado el "efecto mariposa" con el cual se explica el hecho de que, a pequeñas variaciones en los parámetros de un modelo de predicción se observan resultados muy diferentes. Este concepto ha dado origen a la "teoría del caos" que trata con sistemas que se comportan de manera estable e inestable en determinados periodos (Stewart, 2007).

Entre otras investigaciones Argote ha realizado un estudio de la precisión e incertidumbre

alrededor de la Tasa Global de Fecundidad, como indicador resumen de la fecundidad de las mujeres bolivianas en 1998 y 2003 y como entrada en un ejercicio de proyección de la población (Argote 2007 y 2009). La metodología utilizada le ha permitido realizar un análisis de sensibilidad de los nacimientos por grupos quinquenales de edad a pequeños cambios de la Tasa Global de fecundidad (Argote, 2011). El contar con la distribución estadística por muestreo de los nacimientos por grupo quinquenales de edad, se pudo analizar estos efectos, donde se encontró que a pequeños cambios de la TGF se presenta un efecto amplificador en los nacimientos. Esto implica que cualquier variación de la TGF es significativa a nivel desagregado y que puede ser representada a través del nivel de variación de los nacimientos.

4. Proyecciones de Población

La inquietud de proyectar para conocer el futuro es intrínseca a todas las disciplinas. El ejemplo más conocido es la predicción del clima. En demografía es de mucho interés la proyección de la población y su distribución etárea con el objetivo de planificar la gestión, recursos y servicios para la población.

El método más utilizado en la proyección de población es el método por componentes (CEPAL, 2006) que considera la siguiente ecuación general:

$$P_t + 1 = P_t + N_t - D_t + I_t - E_t \tag{4}$$

Donde N_t son los nacimientos en t, D_t las defunciones, I_t la inmigración y E_t la emigración, las cuales balancean la .ecuación compensadora" para conocer la población en $t + 1$. Los cuatro componentes demográficos se representan por sus tasas de frecuencia anual (tasas específicas de fecundidad, tasa de mortalidad, tasa neta de migración) los cuales se proyectan bajo diferentes hipótesis de comportamiento (CEPAL, 2006:29-33). Este método se utiliza de manera estándar y se adapta a diferentes contextos sociales sobre todo en los cuales la disponibilidad y calidad de la información puede ser deficiente.

Alho et. al. (2006) innovan en su trabajo de proyección de la población de 18 países europeos: "cuantifican la incertidumbre demográfica". No solo estiman un número sino una distribución probabilística de la proyección con base a lo cual pueden hacer afirmaciones del tipo .El valor Y_i tiene una probabilidad P_i de ocurrir". Esta investigación utiliza el método "*socalled scaled model for error*" para cuantificar la incertidumbre relacionada a la proyección

de población. El método de proyección estocástica tiene la ventaja de proyectar la población futura incluyendo un intervalo de proyección probabilística. Encuentran que a diferencia de los datos oficiales es probable que la población en general crezca y su descenso se retrase más tiempo debido a la alta esperanza de vida e incremento de la migración.

La importancia de la incertidumbre en proyecciones de población, fue considerada indirectamente por Welpthon en 1947. Fue el primero en desarrollar el método por componentes para las proyecciones de población y utilizó la función logística para representar el comportamiento de la fecundidad. Sin embargo sus modelos, ni los de ningún otro, pudieron predecir en su momento, el "*baby boom*" que hizo que las tasas de fecundidad que iban en descenso se dispararan. De esta manera hizo análisis exhaustivos de las tendencias de la fecundidad en varios países utilizando diversas fuentes y concluye: "A largo plazo, la tendencia de descenso de la fecundidad es una regla universal. Aunque raramente puede ocurrir que ascienda de forma relativa y corta" (Alho y Spencer, 2005).

5. Datos y métodos

Un razonamiento inductivo y deductivo, de composición y descomposición, que permite el método utilizado, se aplica para observar los cambios en los componentes de la TGF y en un ejercicio de proyección de población. La distribución estadística de la TGF generada mediante remuestreo (Argote, 2009) es utilizada como dominio para la selección de diferentes valores del estimador de la TGF en el análisis de sensibilidad por objetivo, de un ejercicio de proyección de población considerando la mortalidad y migración constantes.

En cuanto al ejercicio de proyección de población, en esta primera fase de la investigación, se construye el algoritmo para la proyección de 2001 a 2006 por el método de componentes, con la información disponible de Bolivia. Para la proyección de población de 2001 a 2006 se parte de la población base según el censo de 1992 y 2001 disponible en línea (INE-Bolivia). Esta población es depurada mediante el método de un dieciséisavo y ajustada a mediados de año. Se toman las tablas de mortalidad utilizadas por el INE para el periodo 2000-2005, así mismo se toma la migración registrada a través de las tasas netas de migración para el componente de migración.

Para la proyección de 2001 a 2006 lo primero que se proyectó fueron los nacimientos de 0 a 4 años que componen el flujo de entrada para el año proyectado. Estos nacimientos se

calcularon tomando en cuenta la estructura de fecundidad que nos provee la ENDSA 2003. Luego este grupo es afectado por la tasas de mortalidad y de migración (Tablas del INE 2000-2005) con lo cual obtenemos el grupo de nacimientos de 0 a 4 años en 2006. Para los grupos entre 5 y 79 años de edad se reconstruyen las cohortes y se afectan con la mortalidad y migración correspondiente de su grupo quinquenal de edad. El grupo de 80 y más en 2006 queda conformado por la unión de los grupos 75-79 y 80+ en el año base afectados por la mortalidad y la migración.

6. Resultados

6.1 Distribución por muestreo de los nacimientos y de la población total proyectada

Se estima a 2006 una población total boliviana de 9.072.364 habitantes (4.596.295 mujeres y 4.476.069 hombres). El intervalo de confianza al 95 % es [9.028.20, 9.117.231] esto implica que la proyección puede variar en 44,161 habitantes respecto el límite inferior y en 44,867 respecto el límite superior (véase tabla 2). Estos datos son diferentes de las estimaciones oficiales del INE de Bolivia que publican 9.627.269 millones de habitantes en 2006 (4.799.178 hombres y 4.828.091 mujeres) que no se han conciliado con conteos u otros censos, únicamente son estimaciones provenientes de encuestas.

De las Gráficas 1 y 2 y de las Tablas 1 y 2, se observa que las variaciones del estimador de la TGF alrededor de su media son a lo sumo de 2 % aproximadamente y de la población proyectada de 0.3 %. Se hace necesario ampliar el espectro de análisis considerando un conjunto de valores del estimador de la TGF y ver de esta manera la presencia o no del efecto mariposa con relación a los resultados de la proyección.

El análisis de sensibilidad implica, evaluar los cambios en los resultados, al realizar cambios en determinados parámetros del modelo. En este caso se realizan pequeñas variaciones a la TGF, se ejecuta el modelo de proyección y se observan los cambios en las estimaciones de la población total. En este sentido se parte de variaciones de 0.5 %, 1 % y 2 % por encima y por debajo de la media de la TGF_i^x. Acorde a ello la probabilidad de que ocurran estos valores es muy alta ya que se encuentran en el área de "mayor peso" de la distribución.

Contrario a lo que se esperaba, a variaciones del 0.5 %, 1 % y 2 % de la TGF_i por encima

y debajo del valor medio, la población proyectada P_i no varía ampliamente alrededor de la media (véase gráfica 3). Al parecer a medida que se incorpora mayor información en una proyección de población partiendo de los nacimientos, entonces los efectos se atenúan. (ver, gráfica 3)

Para analizar la relación entre cambios en la TGF_i y la población proyectada P_i, ampliamos a 30 la muestra de valores por debajo de la media de la TGF (subestimación) y tomamos otra muestra de 30 valores en la cola superior de la distribución (sobreestimación), estos son valores extremos que podría adoptar el estimador.

El análisis de sensibilidad considerando la información de las colas de la distribución por muestreo de la TGF muestra un comportamiento heterogéneo en la relación estudiada, ya que en algunos puntos se observa un efecto reductor y en otros de leve amplificación (ver, gráfica 4 y 5).

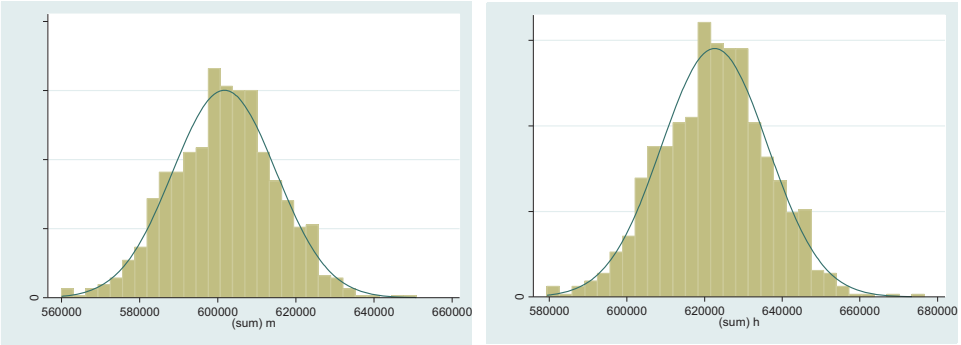
7. Discusión

El método aplicado nos brinda información coherente y de fácil interpretación para la evaluación de nuestras hipótesis de investigación. Según la estimación del INE de Bolivia, la población total proyectada a 2006 es de 9.627.269 millones de habitantes de los cuales 4.799.178 son hombres y 4.828.091 son mujeres. En esta investigación se obtuvo una población media de 9.072.364 (4.476.069 hombres y 4.596295 mujeres) y un intervalo de confianza de [9.028.203, 9.117.231]. Nuestras estimaciones están por debajo de las estimaciones oficiales en un 6 % del total. Así mismo el método permite también estimar los intervalos de confianza de la población proyectada por grupos quinquenales de edad, lo cual aporta en el análisis del error e incertidumbre de las estimaciones realizadas (Argote, 2011).

Con base a la evidencia proveniente de las distribuciones de la TGF y de la Población proyectada a 2006 para los grupos quinquenales de edad entre 0 y 80+ años, no se podría afirmar que existe un efecto de amplificación del error entre estas dos distribuciones (véanse gráfica 5 y 6). En términos relativos, pequeñas variaciones de la TGF no repercute en grandes cambios en los resultados de proyección de la población boliviana en 2006, esto se puede atribuir a un efecto de cancelación de efectos debido a la agregación de los datos y a las características del método de proyección aplicado. Sin embargo el análisis de sensibilidad y su construcción ha permitido evidenciar la aleatoriedad y no linealidad de las relaciones

Gráfica 1. Distribución estadística por muestreo de los nacimientos del grupo 0-4 por sexo, Bolivia 2006

a) Mujeres **b) Hombres**



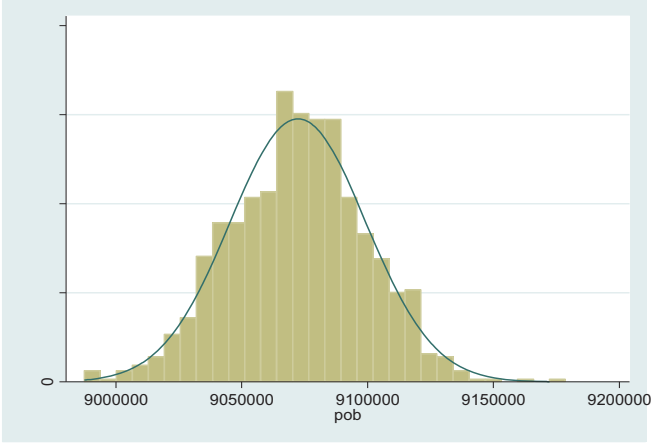
Fuente: Elaboración propia

Tabla 1. Estadísticos de la distribución por muestreo de la población del grupo 0 a 4 años por sexo, Bolivia 2006

	Mujeres	Hombres	Total
Media	601,725	622,612	1,224,337
Desviación standard	13,273.46	13,734.20	27,008
Skewness	-0.0431838	-0.0431866	-0.04318
Kurtosis	3.05670	3.05661	3.05661
Percentil 50%	601,933.3	622,827.4	1,224,761.0
Limite inferior	580,021	600,155	1,180,175
Limite superior	623,776	645,428	1,269,204

Fuente: Elaboración propia, estadísticos generados en STATA

Gráfica 2. Distribución por muestreo de la población proyectada a 2006



Fuente: Elaboración propia

Tabla 2. Estadísticos de la distribución por muestreo de la población total proyectada (2006) y de la Tasa Global de Fecundidad (2003) de Bolivia

	Mujeres	Hombres	Total	TGF
Media	4,596,295	4,476,069	9,072,364	3.8348
Desviación standard	13,273.46	13,734.20	27,008	0.0896
Percentil 50%	4,596,503.0	4,476,285.0	9,072,788	3.8388
Limite inferior	4,574,591	4,453,612	9,028,203	3.6932
Limite superior	4,618,346	4,498,885	9,117,231	3.9841

Fuente: Elaboración propia

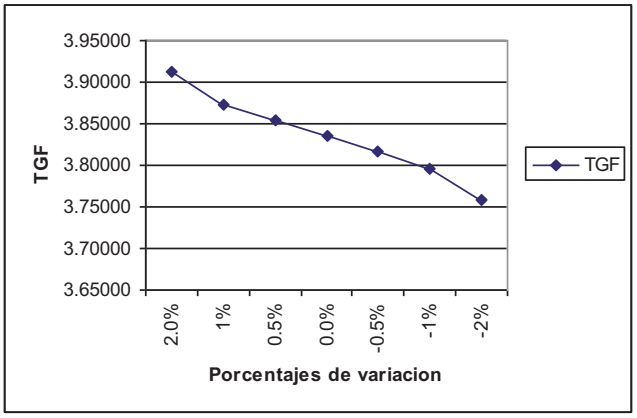
Tabla 3. Estimaciones puntuales de la proyección de las población total y por sexo, 2001-2006

año	mujeres	hombres	total
2001	4,141,063	4,104,158	8,245,221
2006	4,596,295	4,476,069	9,072,364
r	10.99	9.06	10.03

r: tasa de crecimiento

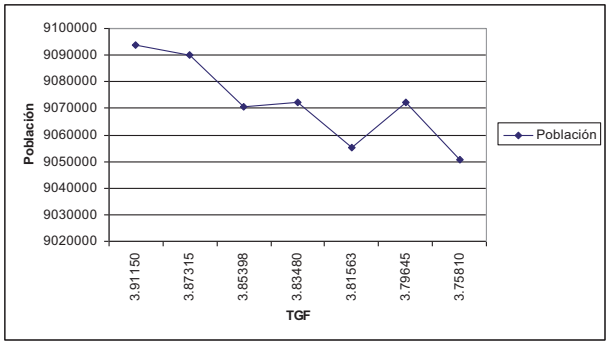
Fuente: Elaboración propia con base a información del INE para el año 2001

Gráfica 3. Valores de la Tasa Global de fecundad a Δ_i porcentaje de variación



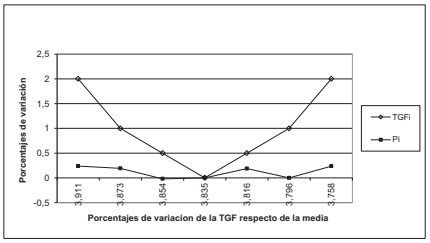
Fuente: Cálculos propios

Gráfica 4. Población proyectada P_i correspondiente a cada valor x_i de la TGF



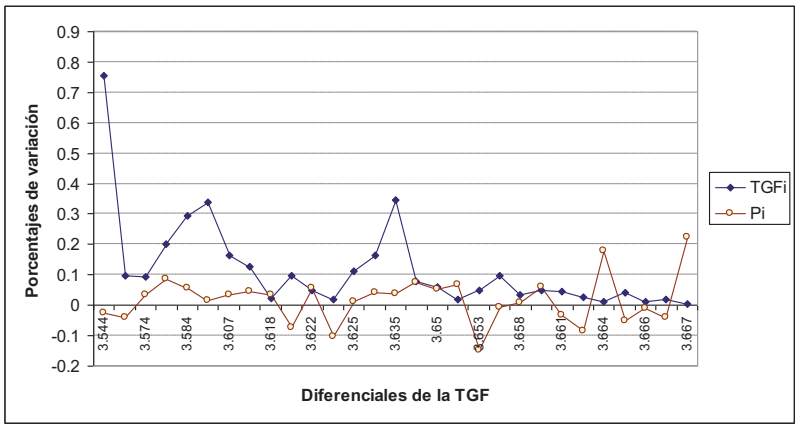
Fuente: Cálculos propios

Gráfica 5. Sensibilidad de P_i a variaciones de la TGF_i, por sexo



Fuente: Cálculos propios

Gráfica 6. Porcentajes de variación entre TGF_{i-1} y TGF_i versus P_{i-1} y P_i en el intervalo [3.543646, 3.667286]



Fuente: Cálculos propios

entre los componentes de la proyección y el error asociado a estas estimaciones.

En términos absolutos se encontró que el intervalo de confianza en que se puede encontrar la población boliviana total proyectada a 2006 varia entre más o menos 44,000 que significan un 0.4 % del total de la población. ¿Qué tan significativo es considerar o no estos nacimientos para una política publica orientada a los recién nacidos, por ejemplo?. La planificación por escenarios nos daría una respuesta.

Referencias

1. Alho Juha, Alders Maarten, Cruijsen Harri, Keilman Nico, Nikander Timo, Dinh Quang Pham (2006). New forecast: Population decline postponed in Europe. Statistical Journal of the United Nations ECE 23 (2006) 1–10 1, IOS Press.

2. Alho, Juha (2005) . Statistical Demography and Forecasting. Springer Series in Statistics.

3. Argote, M. L. (2007). Estimación de la distribución estadística de la tasa global de fecundidad. Papeles de Población, 13(54), 87-113.

4. Argote, M. L. (2009). Comparación y evaluación de la distribución del estimador de la tasa global de fecundidad de Bolivia. Papeles de población, 15(62), 201-222.

5. Argote, M. L. (2012). Análisis de sensibilidad de los nacimientos respecto a la Tasa Global de Fecundidad. Papeles de población, 18(72), 85-112.

6. Bagajewicz Miguel J. (2005). On a new definition of a stochastic-based accuracy concept of data reconciliation-based estimators. *Computer Aided Chemical Engineering*, *Volume 20, Part 2*, *2005*, *Pages 1135-1140*M.

7. Caswell, Hall (2008). Perturbation analysis of nonlinear matrix population models. Demographic Research, Volume 18, Article 2, Pages 27–83

8. Caswell, Hall (2009). Stage, age and individual stochasticity in demography. Oikos 118: 1763_1782, 2009. The Author. Journal compilation # 2009 Oikos, Subject Editor: Per Lundberg. Accepted 16 April 2009.

9. CEPAL (2006). "Chile: estimaciones y proyecciones de la población. 1950-2050", serie OI Nº208, Santiago de Chile, Instituto Nacional de Estadísticas/Centro Latinoamericano y Caribeño de Demografía. Las cifras fueron corroboradas en 2006 por el Centro Latinoamericano y Caribeño de Demografía (CELADE) - División de Población de la CEPAL, sobre la base de las nuevas fuentes de información disponibles.

10. Escobedo Rivera, José (2007). El dato en la investigación demográfica: una visión epistemológica. En Papeles de Población Nueva Época Año 13 Nro 54, Octubre-diciembre de 2007.

11. Gottlieb, A. D. (2001). Asymptotic accuracy of the jackknife variance estimator for certain smooth statistics (preprint), `http://lanl.arxiv.org/abs/math.PR/0109002`.

12. Griffiths, Paula; Matthews, Zoe y Hinde, Andrew (2000). "Understanding the sex ratio in India: a simulation approach". Demography, Volume 37, number 4, November 2000: 477-488.

13. Lorenz, Edward (1963). Deterministic Noperiodic Flow. Massachussets Institute Of Technology. Journal of the atmospheric Sciences, Volume 20, 130-141, March 1963.

14. Ortega Osona, José Antonio y Kohler, Hans-Peter (2000). "¿Está cayendo realmente la fecundidad española? Separación de los efectos de intensidad, calendario y varianza en el Índice Sintético de Fecundidad". Revista Española de Investigaciones Sociológicas, num. 96, 2001, pp. 95-122, Centro de investigaciones sociológicas, España.

15. Ritschard, Gilbert, y Oris, Michel (2005). Life course data in demography and social sciences: statistical and data-mining approaches. Advances in life course research, 10(Complete), 283-314.

16. Stewart, Ian (2007). Taming the infinite. The history of mathematics. Published by arrangement with Quercus Publishing PLC (UK), 2007.

Análisis Shift-Share del Crecimiento Regional del Empleo Manufacturero del Estado de Veracruz

Mónica Pérez García[a]
Especialización en Métodos Estadísticos, Universidad Veracruzana

Alejandro J. Juárez Gómez
Instituto de Investigaciones y Estudios Superiores Económicos y Sociales, Universidad Veracruzana

Sergio Fco. Juárez Cerrillo
Facultad de Estadística e Informática, Universidad Veracruzana

Clasificación: Tesina de Especialización.
Área: Econometría.
Subárea: Modelos econométricos de desarrollo regional.
Trabajo presentado en: XXVIII Foro Nacional de Estadística.
Palabras Clave: Análisis Regional, Cambio Económico, Modelos Lineales.

1. Introducción

Perloff et al.(1960) introducen el concepto y modelo de análisis shift-share con el objetivo de cuantificar los sesgos geográficos de la actividad económica. Dicho modelo parte de la hipótesis de que los cambios en la estructura del crecimiento regional son producto de las decisiones acerca de la localización y producción regional que toman las empresas en términos de los accesos input-outputs. En el análisis shift-share se han utilizado exitosamente distintas metodologías para la estimación conjunta de los efectos nacional, sectorial y regional; vease por ejemplo Buck y Aktins (1983), Berzeg (1984), Nazara y Hewings (2004). Entre las técnicas más utilizadas se encuentran el análisis de varianza, los modelos de regresión y

[a] moni_15_24@hotmail.com

las matrices de vecindad. En este trabajo empleamos el modelo de regresión, conocido como shift-share estocástico, para hacer el análisis del empleo manufacturero del Estado de Veracruz. Entendemos por industria manufacturera la actividad económica que transforma una gran diversidad de materias primas en diferentes productos manufactureros para el consumo final o intermedio.

2. El Modelo de Shift-Share

En el planteamiento clásico del análisis shift-share se considera el cambio de una variable económica entre dos instantes de tiempo. En nuestro caso la variable económica es el empleo manufacturero en cada una de las diez regiones de Veracruz determinadas por el CONAPO. Identificamos tres componentes: efecto nacional, efecto sectorial y efecto regional. Sea $E_{i,j,t}$ el empleo manufacturero en el sector i $(i = 1, ..., S)$ en la región j $(j = 1, ..., R)$ en el tiempo t y sea $E_{i,j,t-1}$ la misma variable en el tiempo $t-1$. El cambio que experimenta la variable queda expresado por

$$E_{i,j,t} - E_{i,j,t-1} = E_{i,j,t-1}r + E_{i,j,t-1}(r_i - r) + E_{i,j,t-1}(r_{i,j} - r_i) \tag{1}$$

donde

$$r = \frac{\sum_{i=1}^{S}\sum_{j=1}^{R}(E_{i,j,t} - E_{i,j,t-1})}{\sum_{i=1}^{S}\sum_{j=1}^{R}E_{i,j,t}} \tag{2}$$

$$\begin{aligned} r_i &= \frac{\sum_{j=1}^{R}(E_{i,j,t} - E_{i,j,t-1})}{\sum_{j=1}^{R}E_{i,j,t-1}}, \\ r_{i,j} &= \frac{(E_{i,j,t} - E_{i,j,t-1})}{E_{i,j,t-1}}. \end{aligned}$$

Berzec (1979) demostró que la identidad del análisis shift-share puede expresarse como un modelo lineal de la forma

$$r_{i,j,t} = \alpha_{0,i,t} + \alpha_{1,i}\beta_{i,t} + \alpha_{2,i}G_{j,t} + \epsilon_{i,j,t}, \tag{3}$$

donde el término independiente $\alpha_{0,i,t}$ es la tasa de crecimiento nacional (2), la magnitud $\beta_{i,t}$ es la diferencia entre la tasa de crecimiento del sector i y la tasa media del crecimiento nacional, $r_i - r$, y $G_{j,t}$ es la diferencia entre las tasas de crecimiento regional y nacional, $r_j - r$. La diferencia entre la tasa de crecimiento regional j y la tasa nacional de crecimiento del sector i, $r_{i,j} - r_i$, queda determinada por el término de error $\epsilon_{i,j,t}$ el cual se considera una variable aleatoria, con esperanza cero y varianza constante. Sean $\hat{\alpha}_{0,i,t}$, $\hat{\alpha}_{1,i}$ y $\hat{\alpha}_{2,i}$ los estimadores de mínimos cuadrados ordinarios de los coeficientes del modelo (3), los efectos nacional, sectorial y regional se estiman por $EN_{ij,t} = \hat{\alpha}_{0,i,t}(E_{ij,t-1})$, $ES_{ij,t} = \hat{\alpha}_{1,i}(r_i - r)E_{ij,t-1}$, y $ER_{ij,t} = \hat{\alpha}_{2,i}(r_{ij} - r_i)E_{ij,t-1}$.

3. Datos de Veracruz

Los datos de empleo manufacturero fueron tomados de los censos industriales de 1993 y 2009 levantados por el INEGI. Los 210 municipios de Veracruz (existentes al 2009) los agrupamos en las 10 regiones socioeconómicas que determina el CONAPO en base al Índice de Marginación. Las regiones se muestran en la Tabla (2). Utilizamos las 21 ramas manufactureras de la división sectorial de INEGI, las cuales podemos ver en la Tabla (1). La variación se estudia para los años 1993 y 2009.

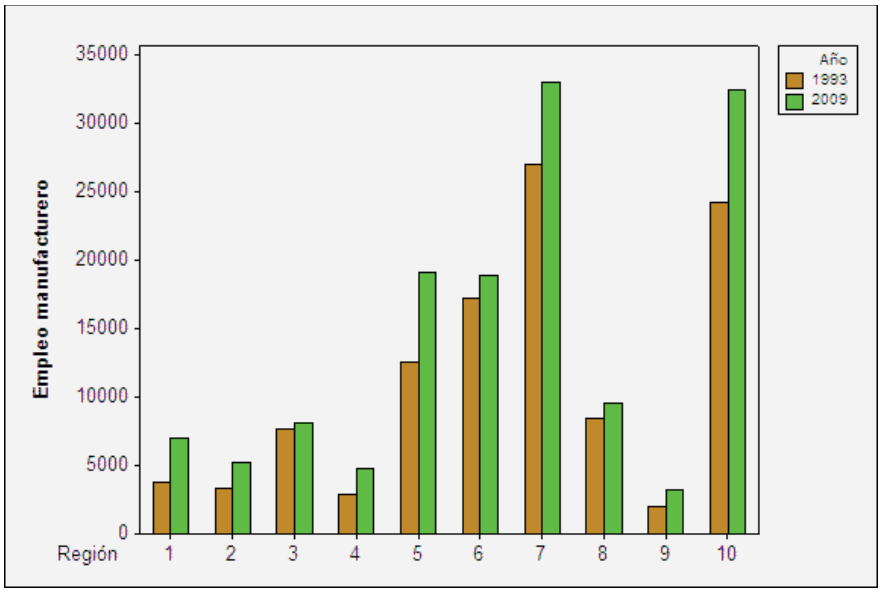


Figura 1: Empleo manufactero por región.

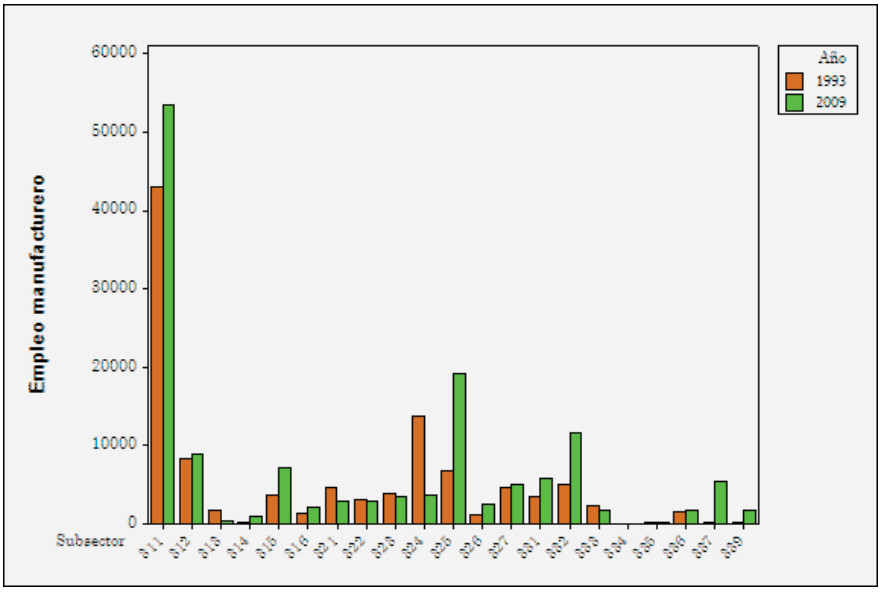


Figura 2: Empleo manufacturero por sector.

4. Resultados

El modelo estimado se muestra en las Tablas (1) y (2). Los resultados manifiestan que el crecimiento económico del empleo manufacturero en Veracruz provoca una inercia positiva en la generación de empleos. Esta inercia se debe al efecto de arrastre de la economía nacional, el cual compensa cualquier negatividad que pudiesen tener los efectos sectorial y regional. Así mísmo, el análisis residual demostró la razonabilidad de una distribución normal.

5. Conclusiones

La variación experimentada por una magnitud regional permite analizar su evolución a partir de distintas fuentes de crecimiento. Esto es lo que se hizo en este trabajo, estimar a nivel regional para estudiar la estructura productiva y así identificar las ventajas comparativas de determinados sectores así como la inercia de la economía estatal en las regiones. Estos elementos podrían indicar qué factores determinan el crecimiento del empleo regional en la entidad, información relevante para el diseño y la planeación regional de una política industrial.

Subsectores	Estimaciones
Intercepto	5103.25
Industria de las bebidas y del tabaco	−4428.79
Fabricación de insumos textiles y acabado de textiles	−5272.19
Fabricación de productos textiles, excepto prendas de vestir	−5224.69
Fabricación de prendas de vestir	−4598.09
Curtido y acabado de cuero y piel, y fabricación de productos de cuero, piel y materiales sucedáneos	−5108.19
Industria de la madera	−5031.29
Industria del papel	−5034.59
Impresión e industrias conexas	−4969.19
Fabricación de productos derivados del petróleo y del carbón	−4940.59
Industria química	−3392.69
Industria del plástico y del hule	−5061.39
Fabricación de productos a base de minerales no metálicos	−4813.49
Industrias metálicas básicas	−4751.19
Fabricación de productos metálicos	−4147.49
Fabricación de maquinaria y equipo	−5151.29
Fabricación de equipo de computación, comunicación, medición y otros equipos, componentes y accesorios electrónicos	−5313.49
Fabricación de accesorios, aparatos eléctricos y equipo de generación de energía eléctrica	−5297.69
Fabricación de equipo de transporte	−5139.99
Fabricación de muebles, colchones y persianas	−4947.59
Otras industrias manufactureras	−5144.39

Tabla 1: Estimaciones de los efectos para los subsectores.

Regiones	Estimaciones
Huasteca baja	−88.62
Totonaca	−114.38
Nautla	73.53
Capital	117.19
Sotavento	73.34
De las montañas	546.76
Papalopan	121.38
Los Tuxtlas	482.29
Olmeca	919.91

Tabla 2: Estimaciones de los efectos para las regiones.

Referencias

1. Berzeg, K. (1979). The Error Components Model: Conditions for the Existence of the Maximum Likelihood Estimates. Journal of Econometrics, 10, p. 99-102.

2. Berzeg, K. (1984), A Note on Statistical Approaches to Shift-Share Analysis. Journal of Regional Science, 24 p. 277-285.

3. Buck, T. and Aktins, M. (1983). Regional Policies in Retrospect: An Application of Analysis of Variance. Regional Studies, Vol. 17, Issue 3.

4. Knudsen, D.C. (2000). Shift-Share Analysis: Further Examination of Models for the Description of Economic Change. Socio-Economic Planning Sciences, Vol. 34, p. 177-198.

5. Nazara, S. and Hewings, G.J. (2004). Spatial Structure and Taxonomy of Decomposition in Shift-Share Analysis. Growth and Change. Gatton College of Business and Economics, University of Kentucky, Vol. 35, p. 476-490.

6. Perloff, H.S., Dunn, E.S., Lampard, E.E., Muth, R.F. (1960). Regions, Resources, and Economic Growth. Johns Hopkins Press for Resources for the Future: Baltimore. xxv. 716.

ANESBA 1.0: Un software para el análisis de la inferencia bayesiana

Norma Edith Alamilla López[a], Reyle Mar Sarao
Universidad Politécnica del Centro.

Alan M. Hernández-Solano
El Colegio de México.

1. Introducción

La estadística es una disciplina que en los últimos años ha tenido una gran injerencia tanto en los diversos campos de la ciencia como en la vida real.

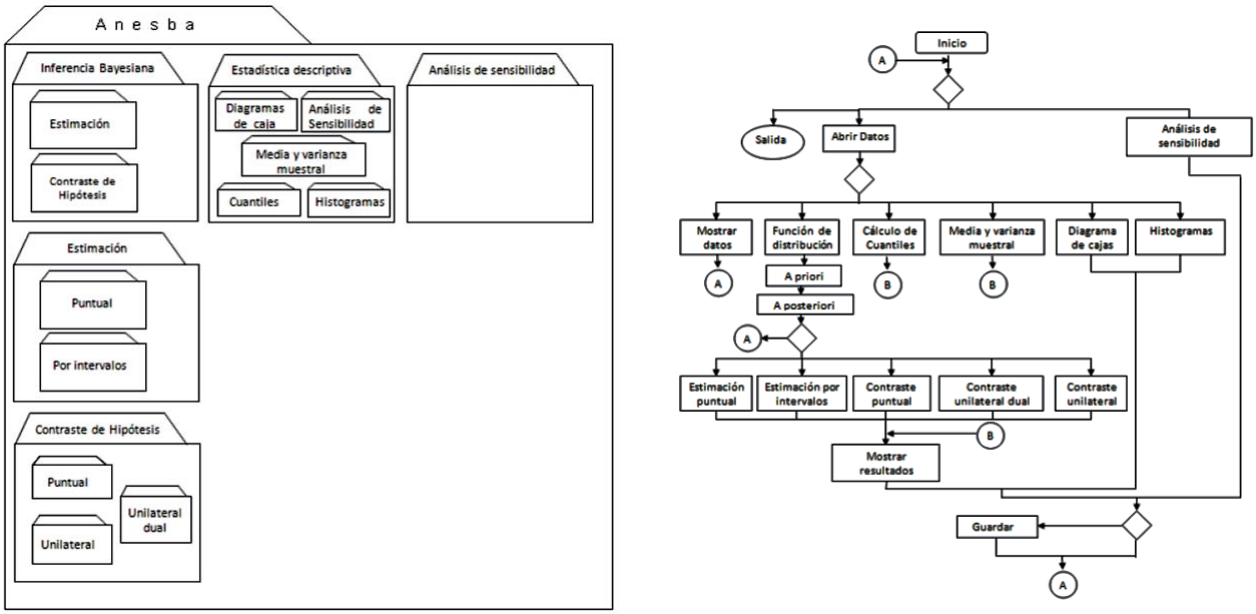
Es basta la literatura que se ha encontrado relacionado con los temas que aborda la Estadística Bayesiana, sin embargo no se ha encontrado o se ha encontrado poco, acerca de la existencia de algún software en español que pudiera apoyar a los cálculos de la inferencia bayesiana.

En este trabajo se muestra un software desarrollado en el lenguaje de programación Visual Basic Net; el cuál es una herramienta, que ayuda en la resolución de los cálculos que aparecen cuando se lleva a cabo las inferencias bayesiana, como estimación y contraste de hipótesis; además de contar con un apartado para realizar análisis de sensibilidad, así como un apartado donde se realiza análisis descriptivo de datos.

2. Diseño e Implementación.

El nombre de este software es ANESBA, cuyo nombre surgió a raíz de utilizar las primeras letras de cada uno de las palabras de: ANálisis EStadístico BAyesiano. Las funciones principales que realiza el software se pueden presentar de forma muy general mediante un

[a] norma.alamilla@gmail.com



(a) Diagrama de carpetas **(b)** Diagrama de flujo

Tabla 1: Diseño de software ANESBA

diagrama de carpetas, tal como se muestra en la Figura 1*a*. A su vez mediante el uso de un diagrama de flujo como en la Figura 1*b* se muestra como se interrelacionan las funciones principales del software.

En el cuadro 1, se muestran los diagramas de carpeta y de flujo, en donde se puede apreciar que en el diagrama de carpetas se presentan las funciones principales que realiza el software ANESBA, mientras que el diagrama de flujo nos muestra las secuencias que se realizan en ANESBA.

Estadística Descriptiva

Otro de los apartados que tiene el software ANESBA, es la Estadística descriptiva en donde se realizan cálculos de los siguientes medidas y gráficas: Cálculo de Cuantiles, Media, Varianza, Histograma y Diagrama de caja.

Análisis de Sensibilidad

Por otro lado, también se consideró incluir un apartado para realizar un análisis de sensibilidad, mediante gráficas en donde se puede ir variando los valores de los parámetros de las distintas distribuciones que se consideran en el software, y así poder decidir cual sería la distribución a priori, a considerar; en el caso en el que se seleccionara una distribución a priori conjugada.

3. Metodología

En esta sección se desarrollará un ejemplo en donde se mostrará el funcionamiento del software ANESBA.

Ejemplo 3.1 (Test De Inteligencia.)**.** *Sea* $X \sim N(\theta, \sigma)$ *y* $\theta \sim N(\mu, \tau)$*. Donde* μ, τ, σ *son conocidos.*

Entonces

$$\pi(\theta|x) = N(\mu(x), \rho^{-1}),$$

donde

$$\mu(x) = x - \frac{\sigma^2(x-\mu)}{\sigma^2+\tau^2}, \rho = \frac{\sigma^2+\tau^2}{\sigma^2\tau^2}.$$

Además supongase que

- $X :=$ *Puntuación que un niño adquiere en un test de IQ.*
- $\theta :=$ *Verdadero IQ del niño*

$$\sigma = \mu = 100, \tau = 225$$

entonces

$$\pi(\theta|x) = N(\frac{400+9x}{13}, 69.23).$$

Es claro que

$$V^{\pi}(x) = \rho^{-1} = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}.$$

Por tanto, en el ejemplo de el test de inteligencia, el niño con $X = 115$ *podría estimarse que tiene un IQ de* $\mu^{\pi}(115) = 110.39$*, con un error estándar*

$$\sqrt{V^{\pi}(115)} = \sqrt{69.23} = 8.32.$$

La estimación clásica de θ *para este problema es* $\delta = x$ *y tenemos que*

$$\begin{aligned} V_{\delta}^{\pi} &= V^{\pi}(x) + (\mu^{\pi}(x) - \delta)^2 \\ &= V^{\pi}(x) + (\frac{\sigma^2\mu^2}{\sigma^2+\tau^2} + \frac{\sigma^2 x}{\sigma^2+\tau^2} - x)^2 \\ &= V^{\pi}(x) + \frac{\sigma^4}{(\sigma^2+\tau^2)}(\mu - x)^2 \end{aligned}$$

Note que el estimador clásico $\delta = x = 115$ *tendría un error estándar de:*

$$\sqrt{V_{115}^{\pi}(115)} = [69.23 + \frac{(100)^2(10-115)}{325}]^{\frac{1}{2}} = \sqrt{90.48} = 9.49.$$

Como la densidad a posterior de θ *es* $N(\mu(x), \rho^{-1})$ *la cual es unimodal y simétrica cerca de* $\mu(x)$*, entonces el Conjunto Creible HPD al* $100(1-\alpha)\,\%$*, es:*

$$C = (\mu(x) + z(\frac{\alpha}{2})\rho^{-1/2}, \mu(x) - z(\frac{\alpha}{2})\rho^{-1/2}),$$

donde $z(\alpha)$ *es el* $\alpha - cuantil$ *de una* $N(0,1)$*.*

Cuando el niño obtiene una puntuación de 115 *en el test de inteligencia, se tiene como distribución a posteriori* $N(110.39, 69.23)$ *para* θ*, luego un conjunto creíble HPD al* $95\,\%$ *para* θ *es*

$$(110.39 + (-1.96)(69.23^{1/2}), 110.39 + (1.96)(69.23^{1/2})) =$$

$$(94.08, 126.70).$$

Ejemplo 3.2. *Supóngase que se desea contrastar la hipótesis* $H_0 : \theta \leq 100$ *frente a* $H_1 : \theta > 100$*.*

Supóngase que la puntuación del niño fue de 115*. Como* $\pi(\theta|x) = N(110.39, 69.23)$*. Se obtiene que:*

$$\alpha_0 = \pi(\theta \leq 100) = 0.106 \text{ y } \alpha_1 = \pi(\theta \geq 100) = 0.894.$$

Por tanto

$$\frac{\alpha_0}{\alpha_1} = 1/8.44.$$

Es decir H_1 *es* 8.44 *veces mas probable que* H_0. *Además puesto que la distribución a priori es* $N(100, 225)$. *Se obtiene que:*

$$\pi_0 = \pi(\theta \leq 100) = 0.5 \text{ y } \pi_1 = \pi(\theta \geq 100) = 0.5.$$

El factor de Bayes es:

$$B = \frac{\alpha_0 \pi_1}{\alpha_1 \pi_0} = 1/8.44.$$

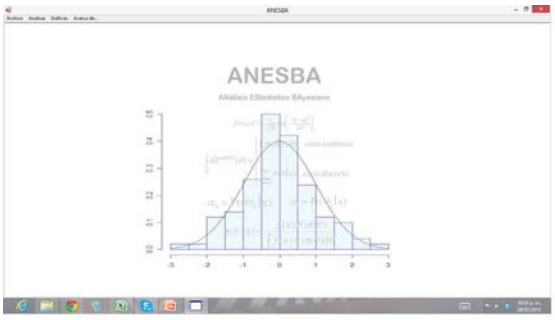
Asi H_1 *es* 8.44 *veces mas probable que* H_0.

En el cuadro 2, se muestran todas las funciones que se realizan en el software ANESBA, en donde se muestra paso a paso lo realizado en el ejemplo desarrollado.

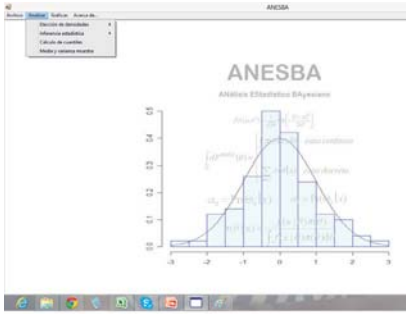
En el cuadro 3, se muestran todos los resultados obtenidos mediante el software ANESBA, después de haber ejecutado todas las funciones, que se muestran durante el desarrollo del ejemplo anterior.

Conclusión

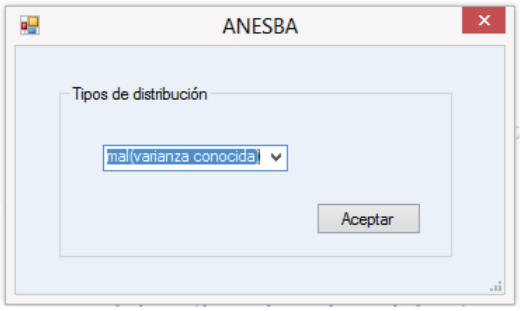
En este artículo se analizan algunos aspectos importantes de la teoría bayesiana, en donde la finalidad era crear un software que permitiera realizar algunos de los métodos principales de la inferencia bayesiana. Así como también al programa se le agregaron otras herramientas de la estadística descriptiva. Las principales funciones que presenta el software ANESBA son: Estimación puntual y por intervalo, contraste de hipótesis, estadística descriptiva (media, varianza, cálculo de cuantiles, así como diagramas de caja e histograma). Los resultados obtenidos por el software ANESBA son los mismos, que los mostrados durante el desarrollo del ejemplo, por lo cual puede apreciarse el buen funcionamiento de ANESBA, y también su fácil manejo en la inferencia bayesiana, con lo cual se cumple las expectativas de crear un software en español, visual y confiable.



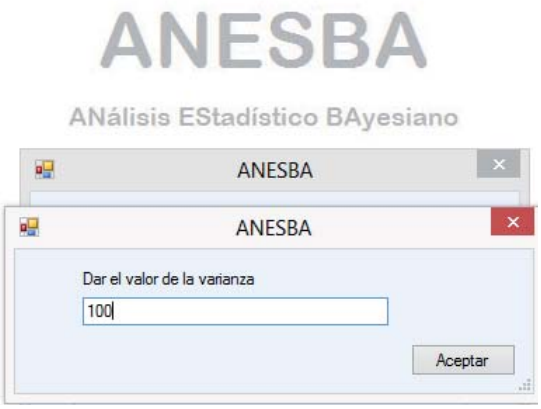
(a) Presentación de software ANESBA.



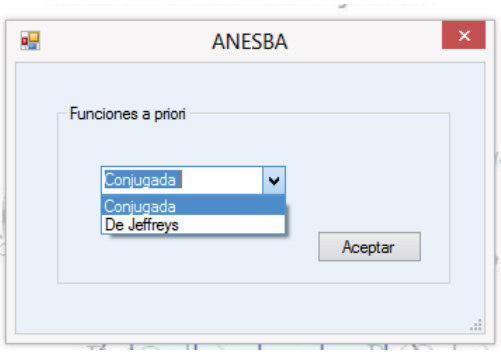
(b) Elección de la funciónn de densidad para la muestra aleatoria.



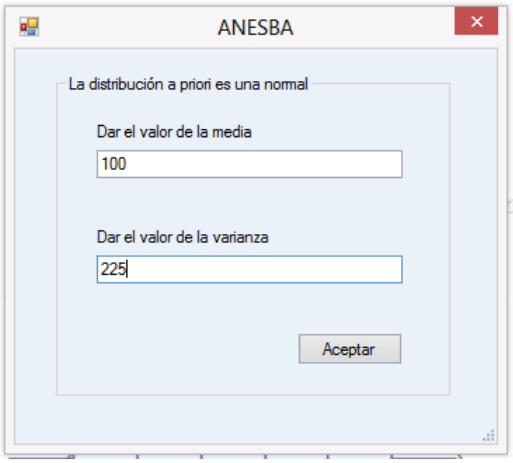
(c) Selección del tipo de distribución de la muestra aleatoria.



(d) Se introduce el valor de la varianza para la distribución normal.

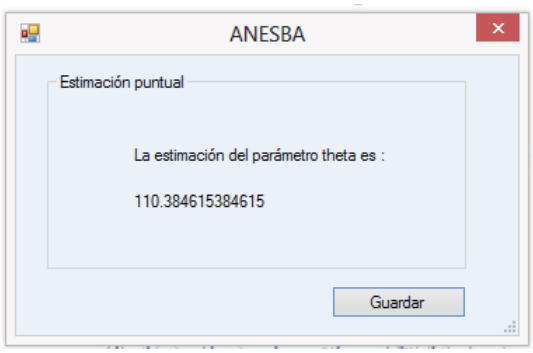


(e) Menú para elegir la distribución a priori.

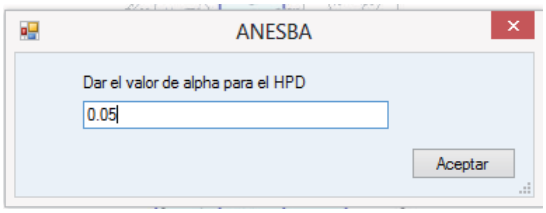


(f) Media y varianza para la distribución a priori elegida.

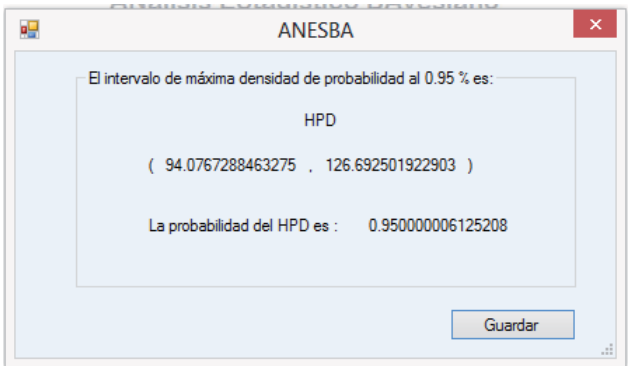
Tabla 2: Funciones que realiza ANESBA



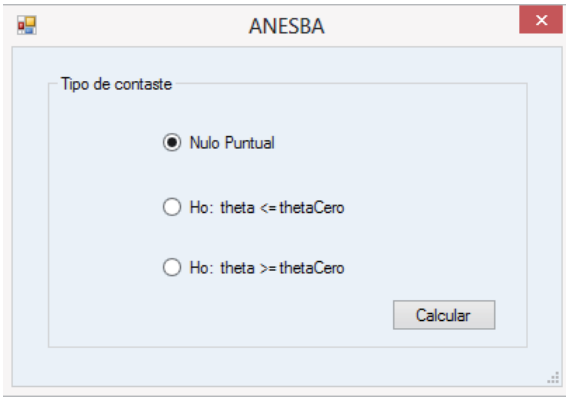
(a) Estimación puntual del parámetro lambda.



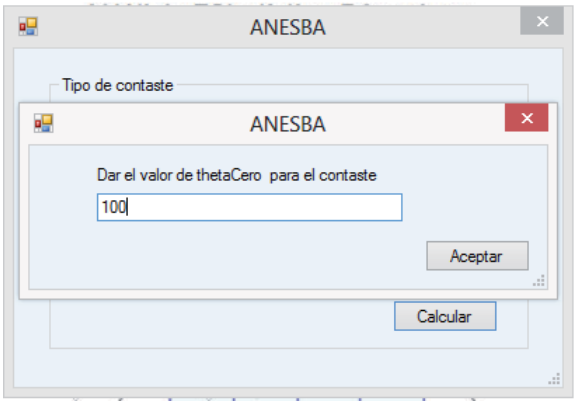
(b) Elección de nivel de significancia para calcular el HPD.



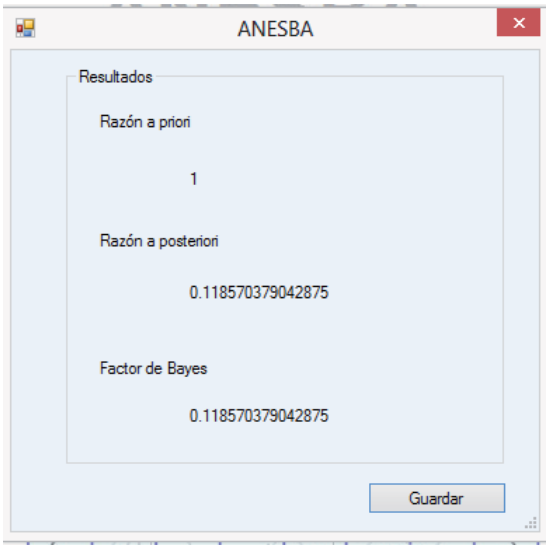
(c) Se muestra el intervalo de HPD y el área que determina.



(d) Opciones de los tipos de contraste de hipótesis.



(e) Ingreso del valor supuesto para realizar el contraste de hipótesis.



(f) Resultados del contraste del hipótesis unilateral realizado.

Tabla 3: Resultados de la ejecución de ANESBA

Referencias

Berger, J.O. 1985. *Statistical Decision Theory and Bayesian Analysis*. Springer.

Carlin, B.P., and T.A. Louis. 2000. *Bayes and Empirical Bayes Methods for Data Analysis, Second Edition*. Taylor & Francis.

Chen, M.H., Q.M. Shao, and J.G. Ibrahim. 2000. *Monte Carlo Methods in Bayesian Computation: With 32 Illustrations*. Springer-Verlag.

Ojeda, F.C. 2005. *Programacion con Visual Basic 2005 / Programming with Visual Basic 2005*. Grupo Anaya Comercial.

Wolpert, R.L. 1990. *Monte Carlo Integration in Bayesian Statistical Analysis*. Institute of Statistics and Decision Sciences, Duke University.

DEMANDA DE IMPORTACIONES DE MANGO MEXICANO EN EL MERCADO DE ESTADOS UNIDOS (1991-2009)*

Plácido Salomón Álvarez-López [a], Elizabeth Trujillo Ubaldo[b], Juan Hernández Ortiz[c]
Universidad Autónoma Chapingo-DICEA

Clasificación: Trabajo de Investigación.
Área: Análisis de Datos
Subárea: Modelos Lineales Generalizados
Trabajo presentado en: XXVIII Foro Nacional de Estadística

1. Introducción

Los principales países productores de mango son: India, China, Indonesia, México y Tailandia, ocupando México el cuarto lugar (FAO, 2012). La producción de mango aumentó en forma sostenida en las ultimas dos décadas; posicionó a México en el primer lugar de exportación (USDA, 2012). Estados Unidos es uno de los mercados más grandes de fruta fresca, con un consumo per cápita de 45 kg en el año 2002. En el periodo 2006-2009 México presentó un comportamiento estacionario en su volumen de exportación hacia EE. UU., a diferencia de sus principales países competidores, que presentaron un pequeño repunte; esta tendencia compromete las futuras exportaciones mexicanas de mango, especialmente por los tratados comerciales firmados con Estados Unidos y Canadá (TLCAN). Los principales destinos del mango mexicano de exportación en 2007 fueron: Estados Unidos (83.5 %); Canadá (13 %); Japón (2 %), Holanda (1 %) y otros (0.5 %). En los últimos 4 años las exportaciones

* Este trabajo fue realizado con el auspicio de la Universidad Autónoma Chapingo
[a] salomones141@gmail.com
[b] elizabeth.trubal@gmail.com
[c] jhdzo@yahoo.com.mx

mexicanas mostraron una caída en las tasas de crecimiento, pasando de 9.7 % a 1.5 %, siendo una de las más bajas comparadas con el periodo anterior. Esto se debió a que países como Brasil, Ecuador y Perú penetraron fuertemente en el mercado de Estados Unidos.

En este contexto surgieron interrogantes sobre la demanda de mango mexicano en Estados Unidos. Una de ellas fue, ¿cuál es la participación de los distintos países en las importaciones de mango de Estados Unidos, a cambios en los precios relativos y el gasto total en importaciones de mango? Otra interrogante fue la magnitud de la sustitución entre mangos provenientes de distintos países por parte de los consumidores estadounidenses. Una forma de abordar estas interrogantes es con la teoría del consumidor, haciendo uso de las preferencias reveladas.

Esta investigación tuvo como objetivo estimar la sensibilidad de la demanda de Estados Unidos por importaciones de mango provenientes de distintos países exportadores. Ésta se evidencia a través de estimaciones de elasticidades precio y gasto (o ingreso), las que a su vez fueron usadas para determinar la competitividad relativa de las exportaciones mexicanas comparada con la de los otros países exportadores. La hipótesis de la investigación fue: los consumidores no son capaces de diferenciar a nivel minorista entre mangos provenientes de México y de los demás países exportadores, siendo más bien productos sustitutos en consumo y bienes necesarios en gasto.

2. Metodología

2.1 Sistema de Demanda Casi Ideal (AIDS)

Para el análisis de la demanda de importaciones de mango a Estados Unidos, se seleccionó el modelo AIDS (Almost Ideal Demand System). La derivación de este modelo se encuentra en el trabajo de Deaton y Muellbauer (1980). Las ecuaciones de demanda se expresan en términos de la participación (w_i) del país exportador en el gasto total del país importador:

$$w_i = \alpha_i + \beta_i Ln(X/P^*) + \sum_{j=1}^{n} \gamma_{ij} Ln(p_j) + \epsilon_i \tag{1}$$

donde los subíndices i y j corresponden a un indice asociado a cada país exportador (México, Brasil, Ecuador, Perú y Resto del Mundo), en este caso los indices i y j van de 1 a 5; α_i, β_i y γ_{ij} son los parámetros del sistema a estimar; ϵ_i es el error de predicción; y w_i es la

proporción de participación en dinero del país exportador i en el valor total de importaciones de mango del país importador (Estados Unidos), de esta manera se tiene que $w_i = \frac{p_i * q_i}{X}$

donde q_i es la cantidad importada (cantidad de mango exportada por el país i), p_i es el precio por kilogramo de mango importado, y $X = \sum_{i=1}^{n} p_i q_i$ es el valor total de las importaciones de mango. El índice de precios $P^* = Ln(P)$, se expresa como: $P^* = \sum_{i=1}^{n} \overline{w}_i * Ln(p_i)$

donde $\overline{w}_i$ es la participación promedio del exportador i en el total de las importaciones.

Debido a que en la presente investigación se usó el índice de precios lineal Stone, se genera una aproximación lineal del modelo original (1), conocido como *Linear Approximation of an Almost Ideal Demand System* (LA/AIDS).

Bajo las siguientes restricciones paramétricas el modelo propuesto, satisface los supuestos de la teoría de la demanda: *Aditividad:* $\sum_{i=1}^{n} \alpha_i = 1; \sum_{i=1}^{n} \beta_i = 0; \sum_{i=1}^{n} \gamma_{ij} = 0$; *Homogeneidad:* $\sum_{j=1}^{n} \gamma_{ij} = 0$; *Simetría:* $\gamma_{ij} = \gamma_{ji}$

Consideradas las restricciones, la ecuación (1) representa un sistema de funciones de demandas que agregan de forma total el gasto, son homogéneas de grado cero en precio y en gasto total, y satisface la simetría de Slutsky. Para la estimación de los parámetros del modelo AIDS se hace uso del método de Regresiones Aparentemente no Relacionadas, imponiendo las condiciones iniciales de aditividad, homogeneidad y simetría. El sistema de ecuaciones esta formado de n ecuaciones, de la cuales $n-1$ son independientes, por lo tanto se eliminó la última ecuación; dadas las restricciones de agregación los resultados de estimación son invariantes a la ecuación que se eliminó (Molina, 1993).

2.2 Cálculo de las elasticidades

Para el cálculo de las elasticidades gasto o ingreso (η) y precio Marshalliana (ϵ), se utilizaron las formulas derivadas por Hayes et. al. (1990), citadas por Ortiz y Martínez (2003):

Elasticidades precio propias o directas Marshallianas: $\epsilon_{ii} = \gamma_{ii}/\overline{w}_i - \beta_i - 1$

Elasticidades precio cruzadas Marshallianas: $\epsilon_{ij} = \gamma_{ii}/\overline{w}_i - \beta_i\,(\overline{w}_j/\overline{w}_i)$

Elasticidades del gasto: $\eta_i = 1 + \beta_i/\overline{w}_i$

2.3 Cálculo de intervalos de confianza para las elasticidades

Dado que las elasticidades son variables aleatorias, se pueden construir intervalos de confianza para las elasticidades propias Marshallianas (no compensadas), y del gasto. Las varianzas de

las elasticidades se calcularon con las siguientes expresiones:

Marshallianas $Var(\epsilon_{ii}) = \left(\frac{1}{\overline{w}_i}\right)^2 Var(\widehat{\gamma_{ii}}) + Var(\widehat{\beta}_i) - 2\left(\frac{1}{\overline{w}_i}\right) Cov(\widehat{\gamma_{ii}}\widehat{\beta}_i)$

Gasto $Var(\eta) = \left(\frac{1}{\overline{w}_i}\right)^2 Var(\widehat{\beta}_i)$

Por lo tanto el intervalo de confianza al nivel de confiabilidad deseado es:

$$P\left[\widehat{\phi}_i - t_{\alpha/2,qg.l.}\sqrt{Var(\widehat{\phi}_i)} \leq \ \widehat{\phi}_i \leq \ \widehat{\phi}_i + t_{\alpha/2,qg.l.}\sqrt{Var(\widehat{\phi}_i)}\right] = 1 - \alpha \qquad (2)$$

Los datos de importación de mango de EE. UU., se tomaron de la base de datos de la FAO-stat (mayo de 2012), la serie de datos se consideró de 1991 al 2009, los datos de cantidad están expresados en miles de toneladas y los precios en miles de dólares. La estimación de las funciones de demanda casi ideal, se hizo con el paquete estadístico SAS System 9.0, mediante el método de Regresiones Aparentemente no Relacionadas.

3. Resultados

México es el país que concentra la mayor cantidad de las importaciones de Estados Unidos de 1991 al 2009, presentando un promedio total de participación de 64.27 %, el resto de países presentan una participación mínima comparada con la de México, pero han mostrado aumentos considerables en sus importaciones los últimos 4 años (ver Tabla 1).

Tabla 1: Participación promedio de los países exportadores (1991-2009)

País	Media	Desviación Estándar	Mínimo	Máximo
Brasil	0.0951240	0.0547832	0.0199727	0.2039676
Ecuador	0.0572732	0.0436870	0.0019738	0.1184192
México	0.6427127	0.1218812	0.5080853	0.8652851
Perú	0.0864482	0.0483886	0.0056256	0.1583574
Resto del Mundo	0.1184420	0.0382322	0.0472588	0.2115741

Fuente: Elaboración propias con datos de la salida de SAS

Las elasticidades fueron calculadas con las participaciones promedio de las exportaciones de los países ($\overline{w}_i$) (ver Tabla 1). En la Tabla 2 se presentan las elasticidades precio propias y cruzadas Marshallianas y del gasto. Las elasticidades propias Marshallianas de Brasil, Ecuador, México y Resto del Mundo son inelásticas (responde poco ante cambios en los precios), Perú presenta comportamiento elástico (es muy sensible ante cambios en los precios), las elasticidades cruzadas nos indican como son vistas las exportaciones de mango de cada país ante cambios en los precios de otro país exportador, son bienes sustitutos los que presentan elasticidad positiva y bienes complementarios los de elasticidades negativas. La elasticidad gasto fue positiva en todos los casos, lo cual indica que son bienes normales o superiores, de las cuales las elasticidades de México y el RM muestran evidencia de ser bienes necesarios y las restantes bienes de lujo.

Tabla 2: Elasticidades precio propias y cruzadas Marshallianas y del gasto

	Brasil	Ecuador	México	Perú	RM	Gasto
Brasil	-0.8152	-0.5772	-0.6637	0.2764	-0.4432	2.2228
Ecuador	-0.9822	-0.7876	-1.1952	-0.4884	0.9827	2.4706
México	0.0564	0.0008	-0.6031	0.0954	-0.1469	0.5975
Perú	0.3160	-0.3023	-0.2558	-1.9109	0.0544	2.0986
RM	-0.2101	0.5772	-0.8561	0.1616	-0.3618	0.6893

Fuente: Elaboración propias con datos de la salida de SAS

En la Tabla 3 se presentan los valores puntuales de las elasticidades precio propias Marshallianas y del gasto, y sus respectivos intervalos de confianza al 95 %.

4. Conclusiones

Con base en los resultados obtenidos, se concluye que México es el principal país exportador de mango hacia Estados Unidos, teniendo una participación promedio de 64.27 % de 1991 al 2009. Asimismo, México presenta una elasticidad precio propia inelástica (Marshalliana), lo que nos indica que la demanda de importaciones responde poco ante cambios en los precio, es decir, las exportaciones de mango hacia Estados Unidos no están condicionadas en gran

Tabla 3: Intervalos de confianza para las elasticidades propias Marshallianas y Gasto

Marshallianas			Gasto		
ε_{ii}	Lim. Inf	Lim. Sup	η_i	Lim. Inf	Lim. Sup
-0.815185	-1.461514	-0.168856	2.222804	2.010283	2.435325
-0.787577	-1.174187	-0.400967	2.470583	2.198864	2.742302
-0.603093	-0.767112	-0.439074	0.597472	0.540214	0.654729
-1.910900	-2.467936	-1.353864	2.098554	1.877619	2.319489
-0.361841	-0.924707	0.201024	0.689291	0.545663	0.832919

Fuente: Elaboración propias con datos de la salida de SAS

medida a cambios en los precios.

La elasticidad gasto para México presenta comportamiento de un bien necesario. Para Brasil, Ecuador y Perú se comporta como bien de lujo, con lo que se concluye que estos países compiten con México vendiendo productos diferenciados (mayor valor agregado). A su vez existe evidencia estadística para afirmar que los consumidores estadounidenses diferencian entre el mango procedente de los distintos países exportadores; hay evidencia que la elasticidad gasto para México se comporta como bien necesario, con lo que se aprecia que los mangos provenientes de otros países están satisfaciendo la demandas de estratos específicos en Estados Unidos.

Referencias

1. Deaton, A.; J. Muellbauer. (1980). An almost ideal demand system. American Economic Association. Vol. 70, No. 3. Pp 312-326.

2. FAO, Food and Agricultural Organization. (2012). Base de datos estadisticos. http:// www.fao.org (marzo de 2012).

3. Molina, J. A. (1993). Evolución de la demanda en los productos alimenticios en los países mediterráneos, Estimación de AIDS. Investigación Agraria Economía. Volumen 8. España.

4. Ortiz, J.; M. A. Martinez. (2003). Estimación de un sistema AIDS y elasticidades para cinco hortalizas en México. Comunicaciones en Socioeconomía, Estadística e Informática. Vol. 7, No. 2. Pp 13-24.

5. USDA, United States Department of Agriculture. (2012). Foreign Agricultural Service On line. USA. http://www.fas.usda.gov (Marzo 2012).

Muestreo Probabilístico para la Recuperación de los Microdatos del Censo General de Población de 1930*

Dr. Francisco J. Zamudio S., M.C. Roxana I. Arana O.,
Lic. Javier Jiménez M., Lic. Carlos Minutti M., Lic. Javier Santibáñez C.
Universidad Autónoma Chapingo

Dr. Robert McCaa
Universidad de Minnesota

Clasificación: Trabajo de divulgación
Área: Muestreo
Trabajo presentado: XVIII Foro Nacional de Estadística

1. Introducción

El Censo General de Población de 1930 fue el quinto de su tipo; se instrumentó con el fin de contar el número de habitantes en la República Mexicana y las principales características socio-demográficas del país.

Como estrategia para el levantamiento de información, se instrumentó una campaña de comunicación para conseguir una adecuada sensibilización de los informantes, con el fin de obtener información de calidad (INEGI 1996). La mayor innovación de este censo fue que por primera vez se consideró el levantamiento de información sobre la gente que habitualmente residía en la vivienda, en lugar de la que se encontraba en ella en el momento censal, se puede decir que se trata del primer censo de buena calidad levantado en México (INEGI 2011).

El censo y la publicación de resultados estuvieron a cargo de la entonces Dirección General de Estadística, la cual publicó un volumen con las principales características demográficas y

* Este trabajo fue realizado con el auspicio del Fondo CONACyT-INEGI.

geográficas por cada Entidad Federativa. Tanto esta información, como los tabulados básicos, fueron publicados en 1932 y actualmente son la única información digital disponible en línea en el portal del INEGI.

El fondo sectorial de investigación constituido por el CONACYT y el INEGI publicó en su convocatoria 2012, la necesidad de recuperar los microdatos del Censo General de Población de 1930. El objetivo general del proyecto de investigación desarrollado es utilizar un diseño de muestreo probabilístico que, con un tamaño de muestra del 10 % de los habitantes, permita obtener estimaciones con precisión medible a escala municipal. La Universidad Autónoma Chapingo, a través de la Licenciatura en Estadística, fue seleccionada para atender esta demanda, y es quien desarrolla el presente documento.

El trabajo de investigación que se expone ha representado una oportunidad de incorporar a estudiantes y egresados de la Licenciatura en Estadística a la solución de un problema práctico, de carácter nacional y además, propio de su formación académica.

2. Materiales y Métodos

Situación clave en la recuperación de la información, es que ésta, se encuentra en documentos que actualmente tienen un importante valor histórico, y están resguardados en el Archivo General de la Nación (AGN), lo que dificultaba la captura masiva de los datos *in-situ*. Debido a esto, se decidió digitalizar las boletas seleccionadas en la muestra, por medio de fotografías de alta calidad, para su posterior captura.

Una vez seleccionada la muestra, el trabajo consistió en realizar la captura de la información, validarla, codificarla y expandirla. Finalmente, se realizará el tratamiento necesario para entregar al INEGI una base de datos para su publicación en línea.

2.1 Muestreo Aleatorio Estratificado

Para obtener una muestra representativa de las boletas censales se usó un Muestreo Estratificado por Conglomerados en una etapa con selección sistemática, este diseño fue propuesto por el Dr. Robert McCaa de la Universidad de Minnesota, quien coordinó la captura del 1 % del censo y emitió la recomendación de realizar la captura del 10 %; esta metodología ha sido usada por IPUMS (International Census Microdata Harmonization Projects), en la

recuperación de microdatos.

Los estratos son los municipios y los conglomerados son las boletas censales. Para seleccionar a los conglomerados, dentro de los estratos, se realizó un esquema de muestreo sistemático con inicio aleatorio. Lo anterior asegura en la muestra, aproximadamente, el 10 % de los registros de cada municipio, de todos aquellos para los cuales las boletas se encuentran disponibles[1].

Las unidades muestrales fueron las boletas del censo, las cuales tienen 50 líneas de captura por el haz y 50 por el envés. En promedio cada cara contiene 40 registros y cada registro contiene la información de 34 variables de una persona. Las unidades fueron seleccionadas sistemáticamente cada 10 elementos, comenzando por el folio 4, por lo que la muestra es llamada "Muestra 4", esto facilitó el manejo de los documentos y aseguró la calidad de la información levantada.

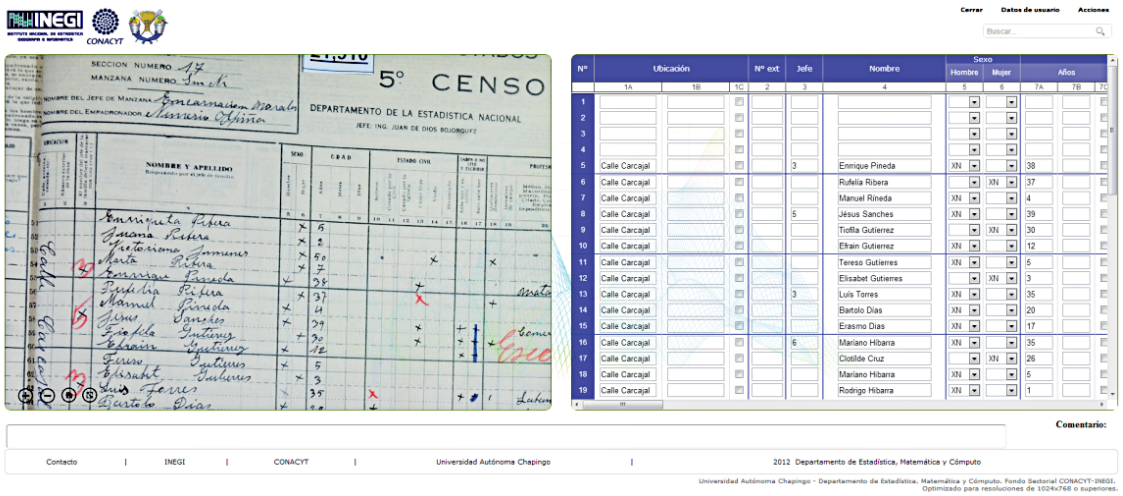
2.2 Captura

- **Captura en Excel©:** Se decidió empezar a capturar en una plantilla de datos en Excel ya que era de fácil manejo para los capturistas, lo que permitió que la llevaran consigo y realizar su trabajo en casa. De esta fase se obtuvieron diferentes aprendizajes con lo que se perfeccionó la captura de la muestra, y se calibró la Plataforma en Línea.

- **Captura en Plataforma en Línea:** La plataforma en línea es un desarrollo tecnológico innovador en el ámbito de captura de datos que, además de cumplir con el objetivo de recuperación de datos de la muestra, brinda una oportunidad de trabajo a larga distancia e ingresos a sectores de la población que no se pueden incorporar de tiempo completo a una actividad productiva, como amas de casa, estudiantes o personas que se encuentran lejos de las oficinas del proyecto. La página web está disponible en la siguiente liga: www.demyc.net.

[1] Alrededor del 20% de las boletas del Censo, incluyendo el 100% de las boletas del Distrito Federal, están extraviadas

Figura 1: Plataforma en Línea: www.demyc.net

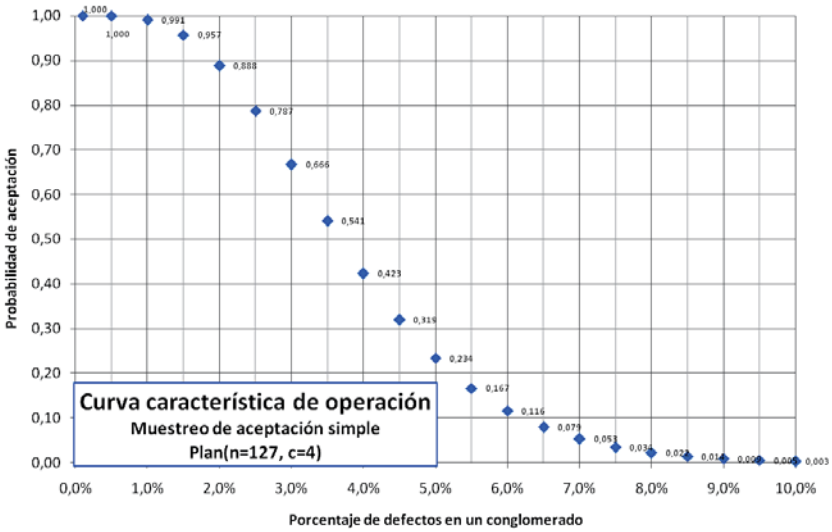


3. Validación de la captura de datos

Para asegurar la calidad de captura, se instrumentó un proceso de validación sobre las **36,127** boletas pertenecientes a la muestra.

1. Muestreo de Aceptación Simple
   - Se tomaron 5,000 registros para calcular el porcentaje de error promedio de un capturista en cada variable.
   - Se utilizó el nomograma de (Montgomery 2009) para determinar los parámetros del plan de muestreo considerando 3,000 campos de captura.
   - Se determinó que al revisar 127 casillas, es plausible aceptar o rechazar una captura con un 95 % de confianza.
2. Los supervisores de validación revisaron el 15 % de las imágenes aceptadas con el mismo esquema de aceptación.
3. Todas las boletas rechazadas al capturista, fueron corregidas por él mismo, para posteriormente volver a ser evaluadas.

Figura 2: Curva característica de operación. Muestreo de aceptación simple. Plan(n=127, c=4)



α: La probabilidad de aceptar un conglomerado con un porcentaje bajo de error (<1 %), es 95 %.
β: La probabilidad de aceptar un conglomerado con un porcentaje de error alto (>6 %), es 10 %.

4. Expansión de la muestra

Con los tabulados básicos publicados por INEGI y las varianzas obtenidas de los microdatos, serán calculados factores de expansión para la realización de inferencias sobre la población. Los estimados para los municipios faltantes, se calcularán con procesos de extrapolación.

5. Resultados preliminares

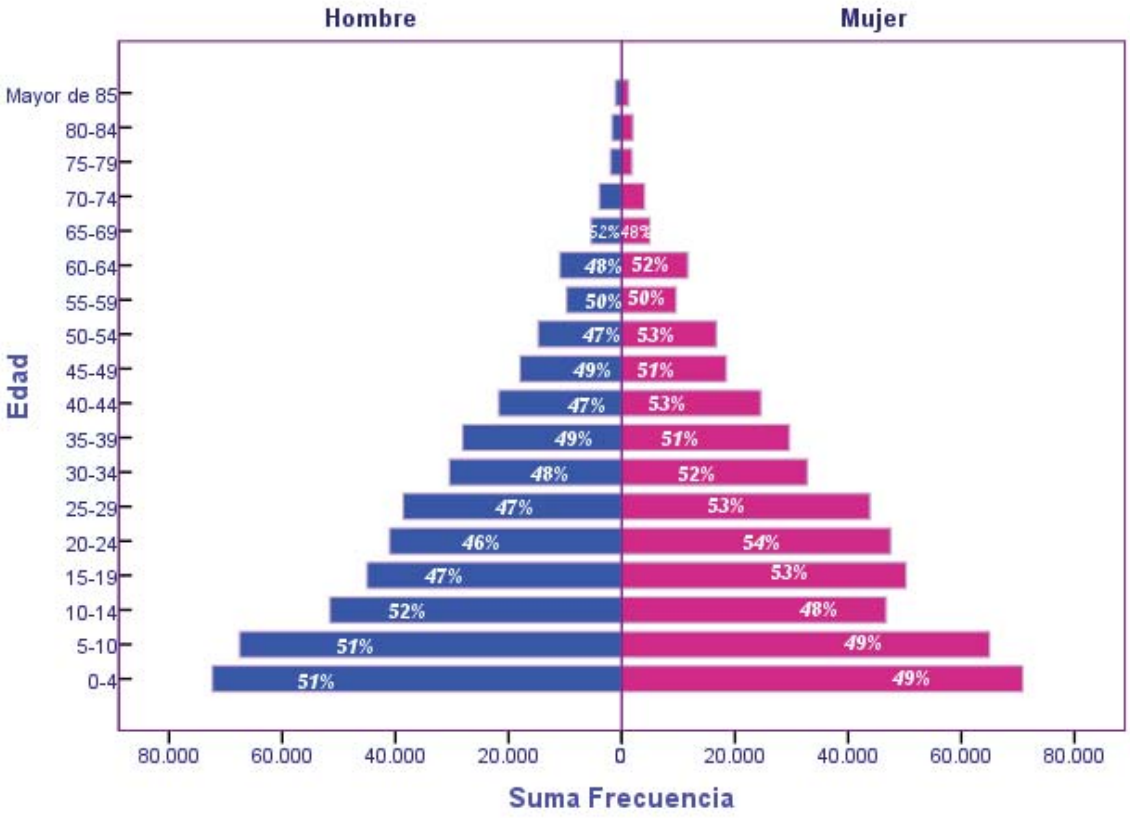
5.1 Edad

Se puede observar que la población de 1930 era joven, además se logra ver que en las categorías de menor edad, hay mayor porcentaje de hombres que de mujeres, en cambio, a partir de 15 años, es mayor la población femenina, esto refleja los estragos de la Revolución Mexicana.

5.2 Estado civil

Se observa que el porcentaje de viudas es significativamente mayor que el de viudos, reflejo también del movimiento armado, otro aspecto interesante es que había más personas que se

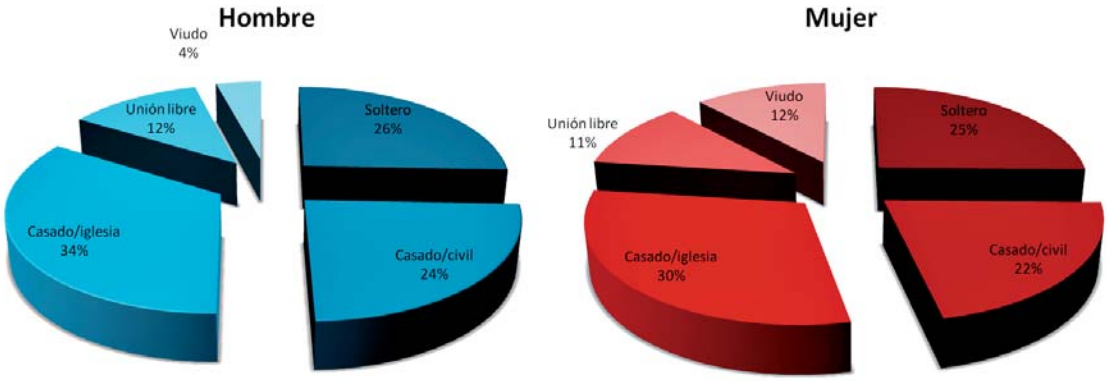
Figura 3: Pirámide poblacional de México en 1930



Datos preliminares de la muestra.

casaban por la iglesia que por lo civil. Finalmente, se aprecia que el porcentaje de hombres casados es mayor que el de las mujeres.

Figura 4: Estado civil de hombres y mujeres en 1930



6. Conclusión

El proyecto de la recuperación de los microdatos del censo es una importante iniciativa del Instituto Nacional de Geografía que aportará información relevante a investigadores del país para el análisis de los principales aspectos socio-demográficos de la población de 1930, y servirá como punto de partida para el análisis longitudinal de los aspectos demográficos de la población mexicana. Análisis más complejos serán presentados una vez que se termine de validar la base de microdatos.

Referencias

INEGI. 1996. *Estados Unidos Mexicanos. Cien Años de Censos de Población.* Aguascalientes, México: Instituto Nacional de Estadística y Geografía (INEGI).

———. 2011. *Censos y Conteos de Población y Vivienda.* Aguascalientes, México: http://www.inegi.org.mx/est/contenidos/proyectos/ccpv/cpv1930/default.aspx.

Montgomery, Douglas C. 2009. *Statistical quality control.* Wiley Hoboken, N.J.