UNIVERSIDAD NACIONAL DEL COMAHUE
Maestría en Estadística Aplicada

Estrategias de Análisis Multivariado para tratamiento de variables mixtas con datos faltantes: una aplicación al estudio del efecto de plaguicidas sobre el período perinatal

Autor

Guillermo Sabino

Profesor de Matemáticas

2003

Neuquén, 2011

COMISIÓN ASESORA DE TESIS

Director

Dr. Sergio Jorge Bramardi

Co-directora

Dra. María Gabriela Rovedatti

A Costi y Lola

AGRADECIMIENTOS

A Sergio, por su generosidad, confianza y aliento en lo que emprendo.

A Gabriela, por mostrarme el camino del investigador.

A Luis, Darío y Javier, por los mates compartidos en el GMAC.

A mis compañeros y profesores de la maestría de quienes aprendí cosas muy valiosas.

A mis viejos.

INDICE GENERAL

CAPÍTULO I. INTRODUCCIÓN 1
1.1 REVISIÓN DE ASPECTOS BIOLÓGICOS 1
1.1.1 Problemática ambiental 1
1.1.2 Plaguicidas 3
1.1.3 Placenta humana 4
1.2 ASPECTOS ESTADÍSTICOS 4
1.2.1 Revisión de antecedentes 4
1.2.2 Distancias estadísticas 7
1.2.3 Métodos para la representación de datos multivariados 11
1.2.3.1 Análisis de Componentes Principales 11
1.2.3.2 Análisis de Coordenadas Principales 19
1.2.3.3 Análisis de Procrustes Generalizados 21
1.3 RESEÑA SOBRE SIMULACIÓN 24
1.3.1 Ventajas y desventajas de la simulación 25
1.3.1.1 Modelo analítico 25
1.3.1.2 Modelos de simulación 25
1.3.2 Etapas de la simulación 26
1.3.2.1 Formulación del problema 27
1.3.2.2 Recolección y análisis de datos 27
1.3.2.3 Desarrollo del modelo 27
1.3.2.4 Verificación y validación del modelo 28
1.3.2.5 Experimentación y análisis de las salidas 29
1.3.2.6 Interpretación de los resultados de la simulación 30
1.4 RESEÑA SOBRE DATOS FALTANTES 31
1.4.1 Mecanismo de pérdida 32
1.4.2 Primeros métodos heurísticos para tratamiento de valores faltantes 33
1.4.2.1 Análisis de datos completos 33
1.4.2.2 Análisis de casos disponibles 33
1.4.3 Métodos de imputación simple 33
1.4.3.1 Imputación por la media 33
1.4.3.2 Imputación por regresión 34
1.4.4 Métodos basados en verosimilitudes 34
1.4.4.1 Algoritmo EM 34
1.4.4.2 El método de imputación múltiple 35

CAPÍTULO II. OBJETIVOS E HIPÓTESIS.....36
2.1 EXPOSICIÓN DEL PROBLEMA.....36
2.2 JUSTIFICACIÓN DEL TEMA DE ESTUDIO.....36
2.3 OBJETIVOS E HIPÓTESIS DE TRABAJO.....37
2.3.1 Objetivo general.....37
2.3.2 Objetivos específicos.....38
2.4 PARTES CONSTITUTIVAS DE LA TESIS.....38

CAPÍTULO III. METODOLOGÍA.....39
3.1 ASPECTOS BIOLÓGICOS.....39
3.1.1 Muestreo de la población bajo estudio.....39
3.1.2 Obtención de homogenados de placenta.....41
3.1.3 Actividad de Acetilcolinesterasa.....41
3.1.4 Actividad de Catalasa.....41
3.2 MÉTODOS ESTADÍSTICOS.....41
3.2.1 Fundamentación.....41
3.2.2 Análisis uni y bivariado (análisis clásico).....42
3.2.3 Estrategias multivariadas.....43
3.3 SIMULACIÓN.....44

CAPÍTULO IV. RESULTADOS Y DISCUSIÓN.....46
4.1 MÉTODOS ESTADÍSTICOS.....46
4.1.1 Técnicas clásicas.....46
4.2.2 Estrategias multivariadas aplicadas.....49
4.2.2.1 Estrategia de una Etapa.....49
4.2.2.2 Estrategia Bietápica.....51
4.2.2.3 Comparación de estrategias multivariadas.....52
4.2 SIMULACIÓN.....53
4.2.1 Generador de datos multivariados mixtos.....53
4.2.2 Eficacia y Robustez de métodos alternativos multivariados para el análisis de datos mixtos ante la presencia de valores faltantes.....59
4.2.2.1 Calidad de representación.....59
4.2.2.2 Aplicabilidad de las estrategias.....65

CAPÍTULO V. CONCLUSIONES.....75

REFERENCIAS BIBLIOGRÁFICAS.....79
APÉNDICES.....86

INDICE DE TABLAS

TABLAS

TABLA 3.1: Detalle de las variables utilizadas 40

TABLA 3.2: Combinación de variables utilizadas para la simulación 45

TABLA 4.1: Pruebas de normalidad para variables continuas. 47

TABLA 4.2: Coeficientes de correlación\p-valores de las variables estudiadas. 47

TABLA 4.3: Matriz de datos a partir de la cual se generan los parámetros iniciales 56

TABLA 4.4: Matriz de entrada necesaria para generar datos mixtos. 57

TABLA 4.5: Matriz de Varianza-Covarianza común a todos los cruces. 57

TABLA 4.6: Matriz simulada. 59

TABLA 4.7: Número de variables utilizadas para determinar la probabilidad de hallar matrices NA según estrategia utilizada. 67

TABLA 4.8: Porcentaje de matrices NA según número de individuos, cantidad de variables utilizadas y proporción de faltantes. 68

INDICE DE FIGURAS

FIGURAS

FIGURA 1.1: Descripción del APG 22

FIGURA 1.2: Descripción del APG 22

FIGURA 1.3: Descripción del APG 23

FIGURA 1.4: Etapas de la simulación 26

FIGURA 3.1: Fotografías de una placenta a término 40

FIGURA 4.1: Diagrama de dispersión de las variables estudiadas 48

FIGURA 4.2: Proyección de los individuos en el plano principal, según Época de Parto Estrategia de una Etapa. 50

FIGURA 4.3: Proyección de la Configuración Consenso, según Época de Parto. Estrategia Bietápica. 52

FIGURA 4.4: (5 Cualitativas – 5 Cuantitativas) Promedio de los coeficientes de correlación de Pearson entre configuraciones con faltantes y originales con diferentes proporciones de datos faltantes, según estrategia. 60

FIGURA 4.5: (6 Cualitativas – 4 Cuantitativas) Promedio de los coeficientes de correlación de Pearson entre configuraciones con faltantes y originales con diferentes proporciones de datos faltantes, según estrategia. 61

FIGURA 4.6: (5 Cualitativas – 6 Cuantitativas) Promedio de los coeficientes de correlación de Pearson entre configuraciones con faltantes y originales con diferentes proporciones de datos faltantes, según estrategia. 61

FIGURA 4.7: (6 Cualitativas – 5 Cuantitativas) Promedio de los coeficientes de correlación de Pearson entre configuraciones con faltantes y originales con diferentes proporciones de datos faltantes, según estrategia. 62

FIGURA 4.8: (7 Cualitativas – 4 Cuantitativas) Promedio de los coeficientes de correlación de Pearson entre configuraciones con faltantes y originales con diferentes proporciones de datos faltantes, según estrategia. 62

FIGURA 4.9: (5 Cualitativas – 7 Cuantitativas) Promedio de los coeficientes de correlación de Pearson entre configuraciones con faltantes y originales con diferentes proporciones de datos faltantes, según estrategia. 63

FIGURA 4.10: (6 Cualitativas – 6 Cuantitativas) Promedio de los coeficientes de correlación de Pearson entre configuraciones con faltantes y originales con diferentes proporciones de datos faltantes, según estrategia. 63

FIGURA 4.11: (7 Cualitativas – 5 Cuantitativas) Promedio de los coeficientes de correlación de Pearson entre configuraciones con faltantes y originales con 64

diferentes proporciones de datos faltantes, según estrategia.

FIGURA 4.12: Porcentaje de Matrices NA, según porcentaje de faltantes y número de variables (n=50) 69

FIGURA 4.13: Porcentaje de Matrices NA, según porcentaje de faltantes y número de variables (n=100) 69

FIGURA 4.14: Porcentaje de Matrices NA, según porcentaje de faltantes y número de variables (n=150) 70

FIGURA 4.15: Porcentaje de Matrices NA, según porcentaje de faltantes y número de variables (n=200) 70

FIGURA 4.16: Porcentaje de Matrices NA, según porcentaje de faltantes y número de variables (n=250) 71

ABREVIATURAS

OF	Organofosforados
ACoP	Análisis de Coordenadas Principales
ACP	Análisis de Componentes Principales
ACh	Acetilcolina
AChE	Acetilcolinesterasa
AEB	Alteraciones en el Embarazo
APG	Análisis de Procrustes Generalizados
CAT	Catalasa
CI	Cualitativas
CT	Cuantitativas
E	Expectation
EG	Edad Gestacional
EM	Expectation-Maximization
EMV	Estimadores Máximo Verosímiles
EP	Época de Parto
fuma	Hábito de Fumar
fumih	Fumigación en el hogar
M	Maximization
MAR	Missing at Random
MCAR	Missing Completely at Random
MCMC	Markov Chain Monte Carlo
MDS	Multidimensional Scaling
NA	Not Available
N.A.S.A	Nacional Aeronautics and Space Administration
NIA	Nivel de Instrucción Alcanzado

NMAR	Not Missing at Random
PC	Perímetro Cefálico al Nacer
PCcorr	Perímetro Cefálico al Nacer corregido
Peso	Peso al Nacer
Pesocorr	Peso al Nacer corregido
Pozo	Consumo de agua de pozo
PPla	Peso de la Placenta
Talla	Talla al nacer
Tallacorr	Talla al nacer corregida

RESUMEN

En trabajos de investigación acerca de la incidencia de contaminantes ambientales sobre alteraciones o enfermedades en poblaciones humanas, ocasionalmente se utilizan enfoques multivariados en el análisis de los datos debido al alto número de variables intervinientes. Asimismo, el carácter sociológico-cultural de esta problemática trae aparejada consigo la naturaleza mixta de las variables, ordinales, nominales, continuas, etc.. Otra particularidad a tener en cuenta, es la de encontrarse, en muchas ocasiones, con una importante cantidad de datos faltantes ocasionados por cuestiones fortuitas, olvidos u omisiones.

En esta tesis, las alternativas propuestas para el abordaje multivariado comprendieron dos estrategias a las que llamamos “de una Etapa” y “Bietápica”. Estas consistieron, respectivamente, en hallar una matriz de distancia mediante un coeficiente de similaridad para variables mixtas conocido como coeficiente general de Gower, aplicando posteriormente una técnica factorial de representación; en la segunda se obtuvieron por separado representaciones de las variables cuantitativas y cualitativas, consensuando dichas ordenaciones mediante Análisis de Procrustes Generalizado.

Teniendo en cuenta que los procedimientos suelen ser ineficientes cuando la proporción de faltantes en la base de datos es significativa, se investigó en qué grado se distorsiona la información original, a medida que aumenta la proporción de estos. En este sentido, se estudiaron por separado dos comportamientos complementarios de las estrategias utilizadas: por un lado, la calidad de representación ante la presencia de los datos faltantes; y por otro lado su aplicabilidad.

Para ello se desarrolló una rutina que permitió generar conjuntos de datos que combinaron variables categóricas y continuas, respetando las estructuras de correlación de la base de datos original. Modificando la proporción de faltantes se extrajeron datos de manera aleatoria y se observaron los alejamientos que sufrían las representaciones de las estrategias ante la ausencia de esos valores, como así también la proporción de veces que éstas podían aplicarse.

Los resultados mostraron que la Estrategia Bietápica fue más eficaz en cuanto a que conservó mejor la representación de los datos originales, mientras que la Estrategia de una Etapa fue más robusta en el sentido de aplicabilidad.

PALABRAS CLAVES: Simulación, generador de datos mixtos, distribución normal multivariada.

ABSTRACT

In research papers on the incidence of environmental contaminants on diseases or alternations affecting human populations, multivariate approaches are occasionally used when analyzing data due to a high amount of intervening variables. Likewise, the sociological and cultural side of this issue brings about the mixed nature of variables, ordinal, nominal, continuous, etc.. Another particularity to take into account is that of dealing with a great number of missing data due to random factors, forgetfulness, or omission.

In this research paper, the proposed alternatives for the multivariate approaches involved two types of strategies called “one-stage” and “two-stage” strategies which respectively consisted of the finding of a distance matrix through a coefficient of similarity for mixed variables known as Gower's general coefficient, and applying later on a factorial technique of representation. In the second one, representations from quantitative and qualitative variables were obtained separately, reaching a consensus on these orderings through Generalized Procrustes Analysis.

Taking into account that procedures usually are inefficient when there is a significant amount of missing data in the database, research on the extent to which original information is distorted when dealing with an increasing amount of missing data was carried out. In this regard, two complementary performances of the used strategies were studied separately: on the one hand, the quality of representation in the presence of missing data, and on the other hand, its applicability.

For this purpose, a special routine was developed which generated clusters of data that combined categorical and continuous variables, respecting the structures of the correlation of the original database. By modifying the proportion of missing data, it was possible to extract data randomly and the distancing of the representation of the strategies in the absence of those values was observed, as well as the proportion of times that these representations could apply.

The results showed that the “Two-stage” Strategy was more efficient since it better retained the representation of the original data, whereas the “One-stage” Strategy was more robust in terms of its applicability.

KEY WORDS: Simulation, mixed data generator, multivariate normal distribution.

CAPITULO I

INTRODUCCION

La exposición prenatal a plaguicidas es detrimental para el desarrollo fetal y/o neonatal, sin embargo es escasa la información acerca de los posibles efectos que podría tener la exposición ambiental a pequeñas concentraciones de estos tóxicos. En este sentido el estudio de biomarcadores de efectos tempranos en placentas provenientes de mujeres que viven en la región del norte de la Patagonia, provincia de Río Negro- Argentina, donde se realiza un uso intensivo de plaguicidas asociado a la fruticultura, es de vital importancia (Souza, 2005).

1.1 Revisión de aspectos biológicos

1.1.1 **Problemática ambiental**

La zona del Alto Valle de Río Negro y Neuquén posee aproximadamente 60.000 hectáreas destinadas a la producción frutícola. Como la incidencia del gusano de la pera y la manzana o carpocapsa (*Cydia pomonella*, L.) es un factor crítico en la evolución de dicha producción es que al iniciarse el período de riego en los meses de septiembre u octubre comienza, paralelamente, la pulverización con plaguicidas organofosforados (OF) por un período de cuatro a seis meses. En promedio, las pulverizaciones por temporada varían entre 5 y 8 para los cultivos de pera y de 10 a 12 para los de manzana, según la variedad, la presión de la plaga y la época de cosecha (Sanchez y col., 2003). La detección de residuos de OF en esta región fue reportada por Loewy M. y col. (2003 a y b) en canales de drenaje y por Santagni (2006) en agua destinada al consumo humano, habiéndose también establecido que existe una asociación temporal y espacial entre el uso de OF en agricultura y su presencia en el aire dentro de un radio de 4,8 km debido a su volatilización (Harnly y col., 2005). Un factor adicional de contaminación en las comunidades rurales o en áreas periurbanas es la dispersión aérea de estos compuestos, ya que el período de pulverizaciones coincide con la época de fuertes vientos. Este hecho, sumado a la existencia de conductas de riesgo en la población residente (Bulgaroni, 2008) y a la transferencia de plaguicidas a través de la placenta (Abu-Qare y col., 2001)

sugiere como problemática indiscutible el deterioro de la salud de las personas desde el ambiente intrauterino.

Existen dos tipos de exposición a tóxicos. La exposición directa tiene lugar cuando el individuo está en contacto directo y reconoce la fuente de contaminación o bien la misma es identificable por terceros. Por otra parte, en la exposición indirecta el individuo desconoce la presencia del contaminante. Se produce a través de residuos en alimentos, agua, suelos y aire, afectando, en general, a un segmento mayor de la población. Esto se denomina exposición ambiental exposición ambiental (Magnarelli y col., 2009).

Determinar la exposición a plaguicidas es un proceso complejo y sofisticado. Para lograr datos precisos es necesario efectuar un biomonitoreo de la exposición que se puede llevar a cabo por medio del uso de biomarcadores. *Sensu stricto* se define como **Biomarcador** a la respuesta biológica a un químico o grupo de químicos (Gil y Pla, 2001). No se incluye a la presencia del agente o de sus metabolitos, en cuyo caso se estaría hablando de dosis interna. La determinación de los mismos consiste en medir los cambios, ya sean estos bioquímicos, fisiológicos o morfológicos, que se asocian con la exposición a un tóxico y que indican alteraciones o fallas de procesos, que evidencian una enfermedad subyacente y/o procesos patológicos o toxicológicos (Peña y col., 2001).

La Salud Ambiental abarca un contexto complejo de factores entre los que se incluye la exposición a tóxicos (de Titto y col., 2004). Las consecuencias a largo plazo de la exposición directa laboral a OF han sido extensamente examinadas (Roldán Tapia y Sánchez Santed, 2004; Edwards y Tchounwou, 2005). Una revisión sistemática de la información epidemiológica del período 1992-2004 mostró la existencia de una asociación positiva entre la exposición directa laboral y anomalías congénitas y efectos reproductivos adversos. Sin embargo se encontró, como limitación de muchos de los estudios analizados, la falta de biomarcadores que representaran fehacientemente una medida de la exposición (Sanborn y col., 2004).

Más recientemente ha cobrado relevancia el estudio de la exposición indirecta no-laboral, en particular la de residentes en zonas rurales (ATSDR 1997). Esta situación ha generado una creciente preocupación sobre el riesgo de la

exposición perinatal que tiene lugar en períodos de alta vulnerabilidad a los tóxicos. Los trabajos llevados a cabo para evaluar la exposición prenatal se han abordado fundamentalmente con diseños epidemiológicos, y ellos sugieren una asociación entre la exposición a OF de las embarazadas y abortos espontáneos, neonatos de bajo peso (Anwar 1997), malformaciones (Weselak M. y col., 2008), retardo en el crecimiento intrauterino (Levario Carrillo y col., 2004) y deterioro en el neurodesarrollo (Eskenazi y col., 2007).

La exposición intrauterina resulta, entonces, la primera vía de contacto con estos tóxicos, por la que se puede alterar el delicado equilibrio del sistema madre-placenta-feto. Un embarazo exitoso requiere de una placenta bien desarrollada con una adecuada interacción materno-fetal para el correcto intercambio sanguíneo de gases, nutrientes y desechos metabólicos fetales.

En la actualidad, la evaluación de riesgo a la exposición crónica a plaguicidas requiere de estudios que permitan identificar biomarcadores sensibles ya que representan una medida de la exposición subyacente y posibilitan la detección de efectos tóxicos tempranos antes que se manifiesten los efectos clínicos adversos.

1.1.2 **Plaguicidas**

Los plaguicidas son por definición xenobióticos biocidas, es decir, compuestos tóxicos no sintetizados por los organismos “blanco” (Barnett y Rodgers, 1994). Comprenden un grupo grande y diverso de compuestos que son deliberadamente introducidos en el medio con el propósito de prevenir, reducir o controlar una forma de vida considerada plaga. Es decir, un organismo, en general de alta densidad poblacional, que interfiere negativamente con alguna de las actividades humanas (Matsumura, 1985).

A pesar de la aparente diversidad que existe entre los organismos vivos hay muchos eventos moleculares básicos que son comunes a la mayoría de ellos, incluyendo al hombre. Por esta razón es muy difícil lograr una toxicidad selectiva sobre los organismos plagas y disminuir los efectos negativos en especies benéficas que se desean preservar o "especies no blanco”. Dentro de los plaguicidas, los OF incluyen compuestos de muy alta toxicidad aguda y relativamente bajo efecto residual.

1.1.3 **Placenta humana**

Diversos estudios en animales de experimentación han focalizado su atención en los efectos fetotóxicos de los plaguicidas, sin embargo es escasa la información sobre la toxicidad de estos compuestos en placenta humana.

Desde el punto de vista físico y funcional, la placenta actúa como interfase entre la madre y el embrión en desarrollo. Provee los nutrientes necesarios para el crecimiento del feto y funciona, a su vez, como órgano excretor (Honzik y col., 2006). Además, este órgano participa en el intercambio de gases, la protección física y la síntesis de hormonas y factores reguladores de la respuesta inmune, los cuales estimulan la retención del feto y evitan el rechazo inmunológico (Gilbert, 2005). Hasta hace unas décadas la placenta era, además, considerada como una barrera contra sustancias posiblemente nocivas. Sin embargo, numerosas estudios han comprobado que este órgano está lejos de ser impenetrable (Benjaminov y col, 1992; Abu-Qare, 2001; Whyatt y col., 2005; Stillerman y col., 2008). Es así que la identificación de biomarcadores en placenta permite estudiar la exposición intrauterina a plaguicidas y planificar posteriores acciones de prevención de la salud humana.

1.2 Aspectos estadísticos

1.2.1 **Revisión de antecedentes**

Romero Villafranca (1995) prioriza el objetivo exploratorio en el contexto multivariado sobre otros enfoques más clásicos de naturaleza inferencial. Desde un punto de vista puramente estadístico y de forma algo restrictiva, define el Análisis Multivariado como: “el conjunto de técnicas cuyo objetivo es el análisis descriptivo y/o la realización de inferencias a partir de datos de naturaleza multivariante, es decir, en los que cada observación está constituida por los valores de varias variables interrelacionadas”. Según Cuadras (2007) “El Análisis Multivariante es un conjunto de métodos estadísticos y matemáticos, destinados a describir e interpretar los datos que provienen de la observación de varias variables estadísticas, estudiadas conjuntamente”. Ampliando el concepto afirma que “es la parte de la estadística y del análisis de datos que estudia, analiza, representa e interpreta los datos que resulten de observar un número $p > 1$ de variables estadísticas sobre una muestra de n individuos”. Debe tenerse en

cuenta la importancia de que estas variables estudiadas estén relacionadas, ya que si fueran independientes entre sí, bastaría con utilizar las técnicas estadísticas univariadas clásicas para cada una de ellas. Contemplar la concordancia existente entre las variables permite efectuar con mayor riqueza y profundidad una mirada descriptiva e introductoria a la estructura subyacente del conjunto de datos.

Existe un gran número de técnicas tendientes a caracterizar individuos u objetos cuando se le observaron un elevado número de variables. Por ejemplo si las variables son de tipo cuantitativa podemos recurrir al Análisis de Componentes Principales o al Análisis Canónico de Poblaciones; si las variables son de tipo cualitativo, Análisis Factorial de Correspondencias o Coordenadas Principales. Estas técnicas se encuadran dentro del marco de los Métodos Factoriales de Ordenación, convenientes para visualizar de la mejor manera, en un espacio reducido, lo que ocurre en un hiper-espacio con la menor pérdida de información posible. La otra rama del Análisis Multivariado se compone por los Métodos de Clasificación, donde a través de una matriz de similaridad para variables cualitativas o distancias para variables cuantitativas, se pretende determinar un reagrupamiento de los individuos en el espacio original completo, a estos métodos se los conoce también con el nombre de Análisis de Cluster (Suarez y col., 1999). Otra de las técnicas útiles referenciadas dentro de este marco es el Escalado Multidimensional (Peña, 2003; Bramardi, 2007; Willems, 2005), que es una generalización de la idea de Componentes Principales, cuando en lugar de disponer de una matriz de observaciones por variables, se dispone de una matriz de distancias cualquiera.

Otra particularidad a tener en cuenta en este tipo de investigaciones, es la de encontrarse con una gran cantidad de datos faltantes ocasionados por cuestiones fortuitas como ser cambio de hospital por parte de la madre, ausencia a los controles periódicos y otros; o por olvidos u omisiones como por ejemplo pérdida de planillas, olvido en alguna encuesta, no medición de los parámetros del neonato, etcétera. Ante la presencia de datos faltantes es muy frecuente recurrir a dos diferentes métodos *ad hoc:* 'case deletion' en los cuales se borran los casos con datos faltantes o 'imputation' donde se completan con algún criterio dichas celdas para forzar a un conjunto incompleto de datos a un conjunto de datos rectangulares completo (Schafer, 1997). Estos procedimientos

suelen ser ineficientes cuando la proporción de datos faltantes es importante. Por un lado, omitirlos del análisis puede introducir un sesgo que alteraría las futuras inferencias hacia la población bajo estudio, mientras que "rellenar" las celdas faltantes podría distorsionar la estructura de varianzas-covarianzas original. Cabe destacar que la naturaleza mixta de las variables presentes en este trabajo ocasiona una problemática mayor en cuanto a su tratamiento.

La cantidad y variedad de métodos dentro del Análisis Multivariado puede producir una sensación de confusión al usuario que se introduce por primera vez a estas técnicas, sin embargo una primera distinción está en: a) métodos explicativos o confirmatorios; y b) métodos descriptivos o exploratorios. Los primeros, bajo el nombre de $multivariate\ statistical\ analysis$, suponen generalmente que cuando las variables son cuantitativas, siguen una distribución normal p-dimensional que posibilita realizar estudios de naturaleza inferencial. Entre estos métodos se encuentran, por ejemplo, el análisis multivariante de la varianza, la regresión múltiple y los modelos log-lineales (Bisquerra Alzina, 1989). Entre sus principales referentes se encuentran Galton, Pearson, Spearman, Fisher, Nightingale, Student, Snedecor, Gower, Hotelling, Wilks, Bartlett, etc. En cambio, el $analyse\ des\ données$, nombre con el que se conoce a los métodos exploratorios en la escuela francesa con Benzecri, Lebart, Morineau y Fenelon a la cabeza (Judez Asencio, 1989), realiza el análisis de datos sin hipótesis previas sobre las variables con una finalidad meramente descriptiva. Debe destacarse también los importantes aportes realizados por Mahalanobis, Rao y Roy dentro del Análisis Multivariante por parte de la escuela india.

El desarrollo de los métodos multivariantes se ha ido incrementando exponencialmente en los últimos años gracias al uso de las computadoras y sus aplicaciones se han extendido en todos los campos científicos: comenzaron desarrollándose para resolver problemas de clasificación en Biología, se extendieron para encontrar variables indicadoras y factores en Psicometría, Marketing y Ciencias sociales, alcanzando una gran aplicación en Ingeniería y Ciencias de la computación como herramientas para resumir información y diseñar sistemas de clasificación automática y de reconocimiento de patrones (Peña, 2003). Por lo anterior, el carácter multidisciplinar de la problemática ambiental que se intenta abordar potencia el uso de estas técnicas para su comprensión.

La descripción y/o caracterización multivariada de individuos y variables intervinientes en un problema, que pueden ser cuantitativas, cualitativas o mixtas, es de significativa importancia en este tipo de análisis. Si bien el conjunto de técnicas para resolver la problemática abordada es numeroso, sólo se presentarán con mayor profundidad las que se utilizarán en el presente trabajo. Previamente, se realizará una breve descripción de los distintos tipos de variables y de las medidas de distancia-simililaridad asociadas a cada una de ellas.

1.2.2 **Distancias estadísticas**

Dados dos puntos x_i, x_j pertenecientes a $\mathbb{R}^p$, se define una función distancia *d* si cumple las siguientes propiedades:

1. $d(x_i, x_j) \geq 0$ No negatividad
2. $d(x_i, x_i) = 0, \ \forall i$
3. $d(x_i, x_j) = d(x_j, x_i)$ Simetría
4. $d(x_i, x_j) \leq d(x_i, x_p) + d(x_p, x_j)$ Desigualdad Triangular

Conceptualmente, si las variables son cuantitativas se define a la distancia como el grado de no semejanza entre dos objetos y en caso de ser cualitativas se denominan disimilitudes o similitudes. Este último término es opuesto a distancia. Puede hallarse en la bibliografía otra terminología.

Variables cuantitativas

La distancia de Minkowsky entre dos puntos

$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, $x_j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ de $\mathbb{R}^p$ se define como:

$$d_q(x_i | x_j) = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^q \right)^{1/q}$$

Casos particulares de la distancia d_q son:

1. Distancia de Manhattan (cuando q=1) $d_1(x_i, x_j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$
2. Distancia Euclídea (cuando q=2)

   $d_2(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$

Mahalanobis (1936) presentó una distancia general, perfectamente adecuada para diferenciar individuos o poblaciones mediante variables aleatorias (Cuadras, 1996)

3. Distancia de Mahalanobis

$$d_M(\mathrm{x}_i, \mathrm{x}_j) = \sqrt{\sum_{l=1}^{p}\sum_{k=1}^{p}(x_{ik} - x_{jk})\sigma_{kl}^{-1}(x_{il} - x_{jl})}$$

Donde σ_{kl} es el elemento de la matriz Σ de varianzas-covarianzas entre las *p* variables.

También tienen interés en las aplicaciones, la distancia normalizada por el rango r_k de la variable *k*, que puede ser poblacional o muestral. Tiene la propiedad de ser invariante al cambio de escala.

4. Distancia de Manhattan normalizada por el Rango R_i de la variable i.

$$d_R(\mathrm{x}_i, \mathrm{x}_j) = \frac{1}{p}\sum_{k=1}^{p}\frac{|x_{ik} - x_{jk}|}{r_k}$$

Otras distancias menos usuales son las Euclídea ponderada, Canberra, Bray-Curtis.

Variables binarias

En ciertos casos, cuando las variables son binarias, lo que se mide es el grado de similaridad entre dos objetos o individuos. Supongamos que tenemos *p* variables binarias $X_1, X_2, \ldots, X_p$, donde cada X_i asume los valores 0 o 1 para ausencia ó presencia, respectivamente. Para cada par de objetos (i, j) queda definida la siguiente tabla de frecuencias:

		Objeto *i*	
		Pres (1)	Aus (0)
Objeto *j*	Pres (1)	a	b
	Aus (0)	c	d

Donde "a" simboliza la frecuencia de presencia simultánea de los atributos en ambos objetos, "d" representa la frecuencia de ausencias simultáneas y "b" y "c" contabilizan las discrepancias entre ambos objetos o individuos. Una vez

construida esta tabla, se pueden definir varias medidas de similitud (s_{ij}) en función de a, b, c y d.

Numerosos autores han propuesto diferentes coeficientes de similaridad, pero podríamos clasificarlos en dos grupos: aquellos coeficientes en los que se establece una simetría entre a y d, donde tanto la ausencia como la presencia simultánea del carácter contribuyen a la semejanza entre las unidades; y aquellos en que no se considera como motivo de aumento de la similaridad, la ausencia simultánea (Bramardi, 2000).

En todos los casos se verifica que $s_{ii} = 1 \geq s_{ij} = s_{ji} \geq 0$, lo que significa que la mayor similaridad resulta igual a 1, que la menor es igual a 0 y que la similaridad es simétrica. Algunos de ellos son:

Simétricos	• Simple Matching (Sokal-Michener, 1958)	$s_{ij} = \frac{a+d}{a+b+c+d}$
Asimétricos	• Jaccard (1908)	$s_{ij} = \frac{a}{a+b+c}$
	• Dice (1945)	$s_{ij} = \frac{2a}{2a+b+c}$

Una vez seleccionado el coeficiente a utilizar, se calculan las $\frac{n(n-1)}{2}$ similaridades entre los n individuos construyendo la $matriz\ de\ similaridad\ entre\ individuos$.

$$\mathbf{S}=\begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & & s_{2n} \\ \vdots & & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{pmatrix}$$

Una cualidad deseada de esta matriz, es que sea semidefinida o definida positiva, para poder representar a los individuos en un espacio euclídeo a través del $Análisis\ de\ Coordenadas\ Principales$. Este método es desarrollado en el punto 1.2.3.2)

Variables cualitativas multiestado

Las variables binarias tienen dos categorías, 0 o 1. Un tratamiento posible para medir la similaridad entre dos objetos cuyas variables son cualitativas

multiestado, es generalizar este concepto a *k* categorías excluyentes $A_1, \dots\dots, A_k$. De esta forma una medida de similaridad entre dos objetos sería la extensión del coeficiente de emparejamiento simple, mencionado en la literatura como Simple Matching:

$$s_{ij} = \frac{\text{número de caracteres coincidentes}}{\text{número total de caracteres}}$$

Variables mixtas

En las aplicaciones a menudo los datos provienen de las observaciones de p_1 variables cuantitativas, p_2 variables binarias y p_3 variables categóricas o cualitativas para más de dos estados. Una manera simple de hallar una medida de similaridad entre dos objetos cuando están caracterizadas por variables de naturaleza mixta, es el coeficiente de similaridad de Gower, que además contempla las situaciones de datos faltantes. La idea básica es promediar las similaridades o distancias entre los objetos, de acuerdo a la naturaleza de los caracteres ya sean éstos cuantitativos, cualitativos o binarios.

Cuando la variable *k* es continua, el aporte a la similaridad entre los objetos (i, j) quedará determinado por $s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{r_k}$, lo que equivale al complemento de la distancia de Manhattan normalizada por el rango r_k, detallado en el contexto de las variables continuas.

Por otra parte, cuando las variables son binarias o categóricas dependerá del investigador utilizar el coeficiente de Simple Matching o el de Jaccard.

Si los objetos se encuentran caracterizados por p_1 variables cuantitativas, p_2 variables dicotómicas y p_3 variables categóricas, los coeficientes propuestos por J. C. Gower dependerán del tratamiento de las variables binarias. A continuación se presentan dos expresiones, de acuerdo se utilice el coeficiente de Simple Matching o Jaccard:

Simple Matching	$s_{ij} = \frac{\sum_{k=1}^{p_1}\left(1 - \frac{\lvert x_{ik} - x_{jk}\rvert}{r_k}\right) + a + d + \alpha}{p_1 + p_2 + p_3}$
Jaccard	$s_{ij} = \frac{\sum_{k=1}^{p_1}\left(1 - \frac{\lvert x_{ik} - x_{jk}\rvert}{r_k}\right) + a + \alpha}{p_1 + (p_2 - d) + p_3}$

donde r_k es el rango de la *k-ésima* variable cuantitativa, a y d son el número de dobles presencias y dobles ausencias de las variables dicotómicas, respectivamente, y α es el número de coincidencias en las variables categóricas. Muchas de las medidas de similaridad y de distancia descritos previamente pueden ser consideradas como un caso particular de este coeficiente.

1.2.3 **Métodos para la representación de datos multivariados**

1.2.3.1 Análisis de Componentes Principales

El Análisis de Componentes Principales (ACP) ofrece diversos puntos de vista. La versión clásica de Hotelling (1933) se presenta como la búsqueda de ejes principales de un elipsoide indicador de una distribución Normal multivariada, siendo esos ejes estimados a partir de una muestra, mientras que en el sentido originario de Pearson (1901) se trata de una técnica de representación de datos, que posee un carácter óptimo según ciertos criterios algebraicos y geométricos, y que se utiliza en general sin referencias a hipótesis estadísticas específicas ni a ningún modelo en particular.

El ACP permite analizar la interdependencia de variables métricas y encontrar una representación gráfica óptima en una dimensión reducida de la variabilidad de los datos de una tabla rectangular de n observaciones y p columnas o variables. Este análisis encuentra un nuevo conjunto de variables no correlacionadas, llamadas componentes principales, que expliquen la estructura de variación en las filas de la tabla de datos.

El desarrollo siguiente corresponde a Peña (2003):

Supongamos que se dispone de los valores de p-variables en n elementos de una población dispuestos en una matriz $\mathbf{X}$ de dimensiones $n \times p$, donde las columnas contienen las variables y las filas los individuos. Si previamente

restamos a cada variable su media, las variables de la nueva matriz centrada $\widetilde{\boldsymbol{X}} = \mathbf{X} - \frac{\mathbf{1}}{\boldsymbol{n}}\mathbf{1}\mathbf{1}'\mathbf{X} = \mathbf{P}\mathbf{X}$ tienen media cero y la matriz de covarianzas $\boldsymbol{\Sigma}$ vendrá dada por $\frac{1}{n}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}$. Para todo el desarrollo de ACP se utilizará esta matriz centrada.

Cálculo de los componentes

Cálculo del primer componente:

El $primer\ componente\ principal$ se define como la combinación lineal de las variables originales que tiene varianza máxima. Los valores en este primer componente de los n individuos se representarán por un vector $\mathbf{z}_1$, dado por

$$\mathbf{z}_1 = \widetilde{\boldsymbol{X}}\mathrm{a}_1$$

Como las variables originales tienen media cero también $\mathbf{z}_1$ tendrá media nula. Su Varianza será:

$$\frac{1}{n}\,\mathbf{z}'_1\mathbf{z}_1 = \frac{1}{n}\mathbf{a}'_1\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\,\mathbf{a}_1 = \mathbf{a}'_1\boldsymbol{\Sigma}\,\mathbf{a}_1$$

Se puede maximizar la varianza sin límite aumentando el módulo del vector $\mathbf{a}_1$. Para que la maximización tenga solución debe imponerse una restricción al módulo del vector $\mathbf{a}_1$, y, sin pérdida de generalidad, se considera que $\mathbf{a}'_1\mathbf{a}_1 = 1$. Esta restricción mediante el multiplicador de Lagrange es:

$$M = \mathbf{a}'_1\boldsymbol{\Sigma}\,\mathbf{a}_1 - \lambda\,(\mathbf{a}'_1\mathbf{a}_1 - 1)$$

Maximizando esta expresión de la forma habitual derivando respecto a los componentes de $\mathbf{a}_1$ e igualando a cero. Entonces

$$\frac{\partial M}{\partial \mathbf{a}_1} = 2\boldsymbol{\Sigma}\mathbf{a}_1 - 2\lambda\mathbf{a}_1 = 0,$$

cuya solución es:

$$\boldsymbol{\Sigma}\mathbf{a}_1 = \lambda\mathbf{a}_1,$$

implica que $\mathbf{a}_1$es un vector propio de la matriz $\boldsymbol{\Sigma}$, y λ su correspondiente valor propio. Para determinar qué valor propio de $\boldsymbol{\Sigma}$ es la solución, multiplicando por la izquierda por $\mathbf{a}'_1$ esta ecuación,

$$\mathbf{a}'_1\boldsymbol{\Sigma}\,\mathbf{a}_1 - \lambda\,\mathbf{a}'_1\mathbf{a}_1 = \lambda$$

Concluyendo que λ es la varianza de $\mathbf{z}_1$. Como es la cantidad a maximizar, λ será el mayor valor propio de la matriz $\boldsymbol{\Sigma}$. Su vector asociado, $\mathbf{a}_1$, define los coeficientes o participación de cada variable en el primer componente principal.

Cálculo del segundo componente:

Aquí se obtiene el mejor plano de proyección de las variables $\widetilde{\boldsymbol{X}}$. Se establece como función objetivo que la suma de las varianzas de $\mathbf{z}_1 = \widetilde{\boldsymbol{X}}\,\mathbf{a}_1$ y $\mathbf{z}_2 = \widetilde{\boldsymbol{X}}\,\mathbf{a}_2$ sea máxima, donde $\mathbf{a}_1$ y $\mathbf{a}_2$ son los vectores que definen el plano. La función objetivo será:

$$\phi = \mathbf{a'}_1\boldsymbol{\Sigma}\,\mathbf{a}_1 + \mathbf{a'}_2\boldsymbol{\Sigma}\,\mathbf{a}_2 - \lambda_1\,(\mathbf{a'}_1\mathbf{a}_1 - 1) - \lambda_2(\mathbf{a'}_2\mathbf{a}_2 - 1)$$

Incorporando las restricciones de que las direcciones deben de tener módulo unitario $(\mathbf{a'}_i\mathbf{a}_i) = 1,\ \ i = 1, 2$. Derivando e igualando a cero:

$$\frac{\partial\phi}{\partial\mathbf{a}_1} = 2\boldsymbol{\Sigma}\mathbf{a}_1 - \ 2\lambda_1\mathbf{a}_1 = 0$$

$$\frac{\partial\phi}{\partial\mathbf{a}_2} = 2\boldsymbol{\Sigma}\mathbf{a}_2 - \ 2\lambda_2\mathbf{a}_2 = 0$$

La solución de este sistema es:

$$\boldsymbol{\Sigma}\mathbf{a}_1 = \lambda_1\mathbf{a}_1,$$

$$\boldsymbol{\Sigma}\mathbf{a}_2 = \lambda_2\mathbf{a}_2$$

Lo que indica que $\mathbf{a}_1$ y $\mathbf{a}_2$ deben ser vectores propios de $\boldsymbol{\Sigma}$. Tomando los vectores propios de norma uno y sustituyendo se obtiene, que en el máximo, la función objetivo es

$$\phi = \lambda_1 + \lambda_2$$

Es claro que λ_1 y λ_2 deben ser los dos máximos autovalores de la matriz $\boldsymbol{\Sigma}$ y $\mathbf{a}_1$ y $\mathbf{a}_2$ sus correspondientes autovectores. Entonces, la covarianza entre $\boldsymbol{z}_1$ y $\boldsymbol{z}_2$ dada por $\mathbf{a'}_1\,\boldsymbol{\Sigma}\,\mathbf{a}_2$ es cero ya que $\mathbf{a'}_1\mathbf{a}_2 = 0$, y las variables $\boldsymbol{z}_1$ y $\boldsymbol{z}_2$ estarán incorreladas.

Generalización:

Puede demostrarse análogamente que el espacio de dimensión r que mejor representa a los puntos viene definido por los vectores propios asociados a los r mayores valores propios de $\boldsymbol{\Sigma}$. Estas direcciones se denominan

$direcciones\ principales$ de los datos y a las nuevas variables por ellas definidas $componentes\ principales$. En general, la matriz $\widetilde{\boldsymbol{X}}$, y por tanto $\boldsymbol{\Sigma}$, tiene rango p, existiendo entonces tantas componentes principales como variables que se obtendrán calculando los valores propios o raíces características, $\lambda_1, \ldots, \lambda_p$, de la matriz de varianzas y covarianzas de las variables, $\boldsymbol{\Sigma}$, mediante:

$$|\boldsymbol{\Sigma} - \lambda \mathbf{I}| = 0$$

Y sus vectores asociados son:

$$(\boldsymbol{\Sigma} - \lambda_i \mathbf{I})\mathbf{a}_i = 0$$

Los términos λ_i son reales, al ser la matriz $\boldsymbol{\Sigma}$ simétrica, y positivos, ya que $\boldsymbol{\Sigma}$ es definida positiva. Por ser $\boldsymbol{\Sigma}$ simétrica si λ_j y λ_h son dos raíces distintas sus vectores asociados son ortogonales. Si $\boldsymbol{\Sigma}$ fuese semidefinida positiva de rango $r < p$, lo que ocurriría si $p - r$ variables fuesen combinación lineal de las demás, habría solamente r raíces características positivas y el resto serían ceros.

Llamando $\mathbf{Z}$ a la matriz cuyas columnas son los valores de los p componentes en los n individuos, estas nuevas variables están relacionadas con la original:

$$\mathbf{Z} = \widetilde{\boldsymbol{X}}\mathbf{A}$$

Donde $\mathbf{A'A} = \mathbf{I}$. Calcular los componentes principales equivale a aplicar una transformación ortogonal $\mathbf{A}$ a las variables $\widetilde{\boldsymbol{X}}$, correspondiente a los ejes originales, para obtener unas nuevas variables $\mathbf{Z}$ incorreladas entre sí.

Propiedades de los componentes

Los componentes principales son nuevas variables con las siguientes propiedades:

1. Conservan la variabilidad inicial: la suma de las varianzas de los componentes es igual a la suma de las varianzas de las variables originales, y la varianza generalizada de los componentes es igual a la original.

   Como $Var\ (z_h) = \lambda_h$ y la suma de los valores propios es la traza de la matriz:

   $$tr(\boldsymbol{\Sigma}) = Var\ (x_1) + \ \ldots + Var\ (x_p) = \lambda_1 + \ \ldots + \ \lambda_p$$

   Por tanto:

$$\sum_{k=1}^{p} Var\,(x_k) = \sum \lambda_k = \sum_{k=1}^{p} Var\,(z_k).$$

Las nuevas variables z_k tienen conjuntamente la misma variabilidad que las variables originales.

Los componentes principales también conservan la $Varianza\ generalizada$, entendida como el determinante de la matriz de covarianzas de las variables. Como el determinante es el producto de los valores propios, llamando $\mathbf{S}_z$ a la matriz de covarianzas de los componentes, que es diagonal con términos λ_k:

$$|\mathbf{\Sigma}_x| = \lambda_1 \ \ldots \ \lambda_p = \prod_{k=1}^{p} Var\,(z_k) = |\mathbf{\Sigma}_z|.$$

2. La proporción de variabilidad explicada por un componente principal es el cociente entre su varianza, el valor propio asociado al vector propio que lo define y la suma de los valores propios de la matriz.

   En efecto, la varianza del componente h es λ_h, y la suma de las varianzas de las variables originales es $\sum_{k=1}^{p} \lambda_k$, igual a la suma de las varianzas de las componentes. La proporción de variabilidad total explicada por el componente h es $\lambda_h / \sum \lambda_k$.

3. Las covarianzas entre cada componente principal y las variables X vienen dadas por el producto de las coordenadas del vector propio que define el componente por su valor propio:

   $$Cov(z_i; x_1, \ldots, x_p) = \lambda_i \mathbf{a}_i = (\lambda_k a_{k1}, \ldots, \lambda_k a_{kp})$$

   donde $\mathbf{a}_k$ es el vector de coeficientes de la componente z_k.

   Para justificar este resultado, vamos a calcular la matriz $p \times p$ de covarianzas entre los componentes y las variables originales. Esta matriz es:

   $$Cov(z, x) = {}^{1}/_{n}\, \mathbf{Z'}\widetilde{\boldsymbol{X}}$$

   y su primera fila proporciona las covarianzas entre la primera componente y las p variables originales. Como $\mathbf{Z}=\widetilde{\boldsymbol{X}}\mathbf{A}$, sustituyendo

   $$Cov(z, x) = {}^{1}/_{n}\, \mathbf{A'}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}=\mathbf{A'\Sigma}=\mathbf{TA'}$$

   donde $\mathbf{A}$ contiene en columnas los vectores propios de $\mathbf{\Sigma}$ y $\mathbf{T}$ es la matriz diagonal de los valores propios. En consecuencia, la covarianza entre, por ejemplo, el primer componente principal y las p variables vendrá dada por la primera fila de $\mathbf{A'\Sigma}$ es decir, $\mathbf{a'}_1\mathbf{\Sigma}$ o también $\lambda_1 \mathbf{a'}_1$, donde $\mathbf{a'}_1$ es el vector de coeficientes de la primer componente principal.

4. Las correlaciones entre un componente principal y una variable X es proporcional al coeficiente de esa variable en la definición del componente, y el coeficiente de proporcionalidad es el cociente entre la desviación típica del componente y la desviación típica de la variable.
   Para comprobarlo:

$$Corr(z_k; x_{k'}) = \frac{Cov(z_k x_{k'})}{\sqrt{Var(z_k)Var(x_{k'})}} = \frac{\lambda_k a_{kk'}}{\sqrt{\lambda_k s_{k'}^2}} = a_{kk'}\frac{\sqrt{\lambda_k}}{s_{k'}}$$

Análisis Normado o con correlaciones

Los componentes principales se obtienen maximizando la varianza de la proyección. En términos de las variables originales esto supone maximizar:

$$M = \sum_{k=1}^{p} a_k^2 s_k^2 + 2\sum_{k=1}^{p}\sum_{k'=k+1}^{p} a_k a_{k'} s_{kk'}$$

Con la restricción $\mathbf{a'a}=1$. Si alguna de las variables, por ejemplo la primera, tiene una varianza s_1^2, mayor que las demás, la manera de aumentar M es incrementar la coordenada a_1 asociada a esta variable. En el límite, si una variable tiene una varianza mucho mayor que las demás, el primer componente principal coincidirá muy aproximadamente con esta variable.

Cuando las variables tienen unidades muy distintas esta propiedad no es conveniente: si disminuimos la escala de medida de una variable cualquiera, de manera que aumenten en magnitud sus valores numéricos, por ejemplo pasando de medir en km a medir en metros, el peso de esa variable en el análisis aumentará, ya que:

1) su varianza será mayor y aumentará su coeficiente en el componente, a_k^2, pues contribuye más a aumentar M;
2) sus covarianzas con todas las variables aumentarán, con el consiguiente efecto de incrementar a_k.

En resumen, cuando las escalas de medida de las variables son muy distintas, la maximización de M dependerá decisivamente de estas escalas de medida y las variables con valores más grandes tendrán más peso en el análisis. Para evitar este problema, conviene estandarizar las variables antes de calcular los

componentes, de manera que las magnitudes de los valores numéricos de las variables X_k sean similares.

La ecuación a maximizar se transforma en:

$$M' = 1 + 2\sum_{k=1}^{p}\sum_{k\prime=k+1}^{p} a_k a_{k\prime} r_{kk\prime}$$

siendo $r_{kk\prime}$ el coeficiente de correlación lineal entre las variables k y k'. En consecuencia la solución depende de las correlaciones y no de las varianzas.

Los $componentes\ principales\ normados$ se obtienen calculando los vectores y valores propios de la matriz $\mathbf{R}$ de correlación. Llamando λ_p^R a las raíces características de esa matriz, que suponemos no singular, se verifica que:

$$\sum_{k=1}^{p}\lambda_k^R = traza(R) = p$$

Las propiedades de los componentes extraídos de $\mathbf{R}$ son:

1) La proporción de variación explicada por λ_h^R será:

$$\frac{\lambda_h^R}{p}$$

2) Las correlaciones entre cada componente $z_{k\prime}$ y las variables X originales vienen dados directamente por $a'_{k\prime}\sqrt{\lambda_{k\prime}}$ siendo $\mathbf{z}_{k\prime} = \widetilde{\boldsymbol{X}}\mathbf{a}_{k\prime}$.

Cuando las variables tienen las mismas unidades, ambas alternativas son posibles. Si las diferencias entre las varianzas de éstas son informativas e interesa tenerlas en cuenta, no deben estandarizarse. Por ejemplo, resulta informativo conservar dos variables con la misma base donde una fluctúa mucho y la otra es casi constante. Para tener en cuenta esta variabilidad no se deben estandarizar las variables, de manera que el índice de mayor variabilidad tenga más peso. Por el contrario, si las diferencias de variabilidad no son relevantes se eliminan con el análisis normado. En caso de duda, conviene realizar ambos análisis, y seleccionar aquél que conduzca a conclusiones más informativas (Peña, 2003).

La representación geométrica de las filas mediante puntos permite visualizar la proximidad entre los individuos. Dos individuos son próximos, en términos de similitud, si sus *p* coordenadas son próximas o similares. Por otra parte las

columnas son representadas mediante vectores, indicando que dos variables estarán relacionadas si miden una misma "cosa" o están ligadas a través de una relación particular (Willems, 2005). La "distancia" entre variables se interpreta en términos del ángulo que forman ambos vectores, cuyo coseno es la correlación entre ambas variables:

$$d^2(x_k, x_{k'}) = \|x_k - x_{k'}\| = \underbrace{\|x_k\|^2}_{1} - \underbrace{\|x_{k'}\|^2}_{1} - 2\|x_k\|^2\|x_{k'}\|^2 \cos\theta_{kk'} \Rightarrow$$

$$d^2(x_k, x_{k'}) = 2(1 - cos\theta_{kk'}) \text{ , luego } r_{kk'} = \cos\theta_{kk'}$$

Para ajustar la nube de variables en el espacio de los individuos (R^n) en un subespacio de s-dimensiones ($s \leq p$) definido por los s primeros autovectores, se debe diagonalizar $\left(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}'}\right)$ de orden $n \times n$.

1.2.3.2 Análisis de Coordenadas Principales

Cómo se desarrolló en el ítem 1.2.3.1, el ACP es una técnica que permite la representación óptima en menor dimensionalidad de una nube de puntos cuando las variables bajo estudio son cuantitativas. En el caso que las variables sean cualitativas, se requiere otro tipo de análisis. Si las variables son dicotómicas, es necesario construir un equivalente a las matrices de covarianza o correlación entre individuos, y estas son las $matrices\ de\ similaridad$ o de $disimilaridad$, que se obtienen teniendo en cuenta la coincidencia o discrepancia con que se dan las características que los describen (Bramardi, 2000).

El desarrollo actual de estas técnicas se debe a las investigaciones de Torgerson (1952, 1958), quien elaboró técnicas conocidas colectivamente como de reescalado multidimensional $(Multidimensional\ Scaling,\ MDS)$, las cuales tienen por objeto producir representaciones de reducida dimensionalidad de colecciones de objetos.

Gower (1966) propuso el $Análisis\ de\ Coordenadas\ Principales$ (ACoP), cuyo objetivo es producir una configuración de puntos, idealmente de muy baja dimensión, cuya distancia euclídea ordinaria reproduzca con la máxima fidelidad las disimilaridades d_{ij} .

El objetivo del ACoP es llegar a una representación geométrica de los n individuos a través de una medida de distancia que respete la estructura de similaridades definida por la matriz de similaridades $\mathbf{S}$. Esto se logra cuando la distancia se define precisamente en función de las similaridades ya establecidas. La distancia cuadrada entre los individuos x_i y x_j será:

$$s_{ij} = \sum_{s=1}^{p} x_{is}x_{js} = \mathrm{x'}_i\mathrm{x'}_j,$$

Por la expresión del producto escalar, $s_{ij} = |\mathrm{x}_i||\mathrm{x}_j|\cos\theta_{ij}$, si los dos elementos tienen coordenadas similares, $\cos\theta_{ij} \cong 1$ y s_{ij} será grande. Caso contrario, si los dos elementos son muy distintos, $\cos\theta_{ij} \cong 0$ y s_{ij} será pequeño. En este sentido podemos interpretar $\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}'$ como la matriz de similitud entre elementos (Peña, 2003).

Las distancias entre las observaciones se deducen inmediatamente de esta matriz de similitud. La distancia euclídea al cuadrado entre dos elementos es:

$$d_{ij}^2 = \sum_{s=1}^{p} \left(x_{is} - x_{js}\right)^2 = \sum_{s=1}^{p} x_{is}^2 + \sum_{s=1}^{p} x_{js}^2 - 2\sum_{s=1}^{p} x_{is}x_{js}$$

Que puede calcularse en función de los términos de la matriz **S**, por la expresión

$$d_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}$$

Luego si s_{ij} es una similaridad, $0 \leq s_{ij} \leq 1$, entonces $d_{ij}^2 = 2\left(1 - s_{ij}\right)$y verifica $d_{ij}^2 = 0$ si $s_{ij} = 1$ (similaridad total), $d_{ij}^2 = 2$ (máxima distancia) si $s_{ij} = 0$ (disimilaridad total).

El ACoP no es más que un análisis de componentes principales aplicado a una tabla de datos compatibles con la distancia $d_{ij}^2 = s_{ii} + s_{jj} - 2s_{ij}$. Pero a diferencia de este, aquí el nuevo sistema de ejes no representa direcciones dado el carácter cualitativo de las variables originales.

Tussel (2005) propone el siguiente algoritmo para el cálculo de las Coordenadas Principales:

1. Obtener una matriz de disimilaridades.
2. $A \leftarrow \left[-\frac{1}{2}d_{ij}^2\right]$
3. $B \leftarrow \left(I - \frac{1}{n}11'\right)A\left(I - \frac{1}{n}11'\right)$
4. Diagonalizar B:

$$B = V'\Lambda V$$

Si no fuera semidefinida positiva, añadir una constante a las disimilaridades no diagonales, y recalcular; alternativamente, prescindir de los valores propios no positivos de$B$.

5. Obtener la configuración de puntos Y:

$$Y \leftarrow V'\Lambda^{\frac{1}{2}}$$

y retener el número de columnas deseado, normalmente dos.

1.2.3.3 Análisis de Procrustes Generalizado

Las configuraciones geométricas obtenidas mediante componentes principales, coordenadas principales u otras técnicas similares, ofrecen una de las maneras más clásicas de representar la estructura y relación empírica de un conjunto de elementos o individuos a los cuales se les ha observado simultáneamente una serie de atributos. En muchos casos, la orientación de las dimensiones es arbitraria, y cuando se han obtenido varias configuraciones sobre la misma muestra de elementos ya sea porque se realizaron en diferentes momentos o por distintos observadores o técnicas, no es posible compararlas con un sentido completo sin antes rotarlas y adecuarlas en forma óptima (Bramardi, 2007).

El $Análisis\ de\ Procrustes$ tiene por objeto examinar en qué medida dos configuraciones de puntos en el espacio euclídeo son similares. Para ello una de las matrices se somete a rotaciones y dilataciones de forma tal que pueda coincidir de la mejor manera posible con la $matriz\ objetivo$.

En 1975 Gower propone el $Análisis\ de\ Procrustes\ Generalizados$ (APG) para m configuraciones $(m > 2)$, el cual consiste en hallar una matriz consenso de todas las configuraciones, y utilizar la misma como nueva matriz objetivo. La técnica de cálculo propone la armonización de las configuraciones individuales a través de una serie de pasos iterativos por transformación de éstas.

Suponiendo que $P_i^{(m)}(m = 1, 2, \dots, q,\ i = 1, 2, \dots, n)$ ofrece la ubicación de los qn puntos en un espacio $p_m - dimensional$, los movimientos de traslación, rotación, simetría y escalado de estas q configuraciones se efectúan bajo dos criterios: que se mantengan las distancias entre individuos de las configuraciones individuales y que se minimice la suma de cuadrados entre puntos análogos, es decir correspondientes al mismo elemento, y su centroide. Esto es minimizar $\sum_{m=1}^{q}\sum_{i=1}^{n}\Delta^2\left(P_i^{(m)}G_i\right)$, donde G_i es el centroide de los q puntos $P_i^{(m)}(i = 1,2 \dots, q)$. Puede expresarse esta transformación mediante:

$$X_m \rightarrow \rho_m X_m H_m + T_m$$

donde ρ_m es un factor de escala, H_m es una matriz ortogonal de rotación y T_m es una matriz de traslación.

Siguiendo un ejemplo de Dijksterhuis y Gower (1991), en el cual existen $q = 3$ configuraciones, cada una de las cuales provee información sobre $n = 4$ objetos

en $p = 2$ dimensiones. Llamando G al centroide de todo el conjunto de puntos y tomando la distancia entre dos puntos A y B como $\overline{AB}$, puede utilizarse la fórmula de Huygen (1629-1695) para hallar la suma total de cuadrados dada por:

$$\sum_{m=1}^{q}\sum_{i=1}^{n}\overline{GA}_{im}{}^{2} = \sum_{m=1}^{q}\sum_{i=1}^{n}\overline{G_mA}_{im}{}^{2} + n\sum_{i=1}^{n}\overline{GG}_{m}{}^{2}$$

donde A_{im} es el punto que representa el $i-ésimo$ individuo de la configuración m. G_1, G_2 y G_3 representan los centroides de cada configuración.

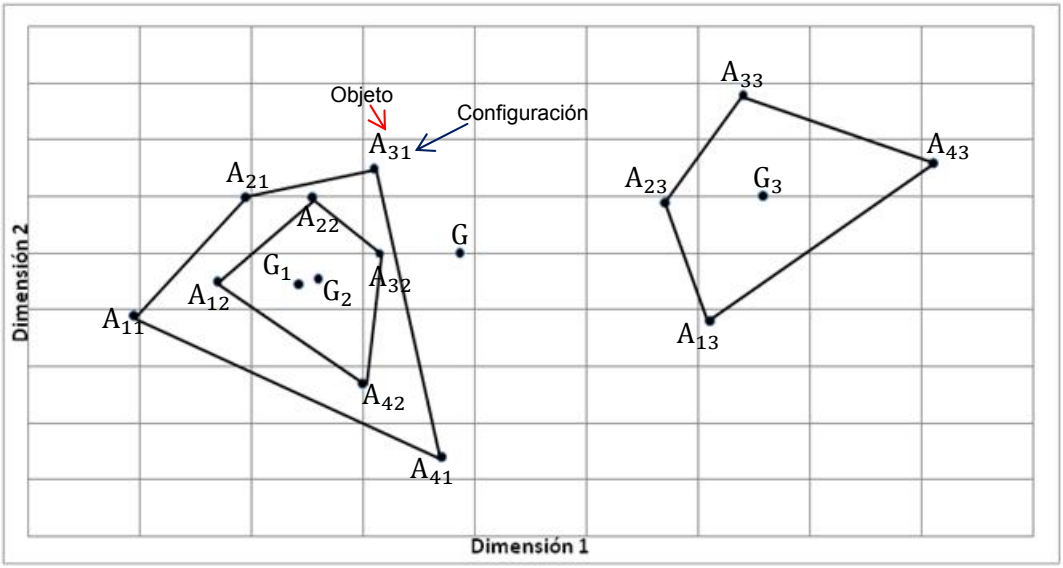


Figura 1.1: A_{im} es el punto que representa el $i-ésimo$ objeto de la configuración m. G_1, G_2 y G_3 representan los centroides de cada configuración.

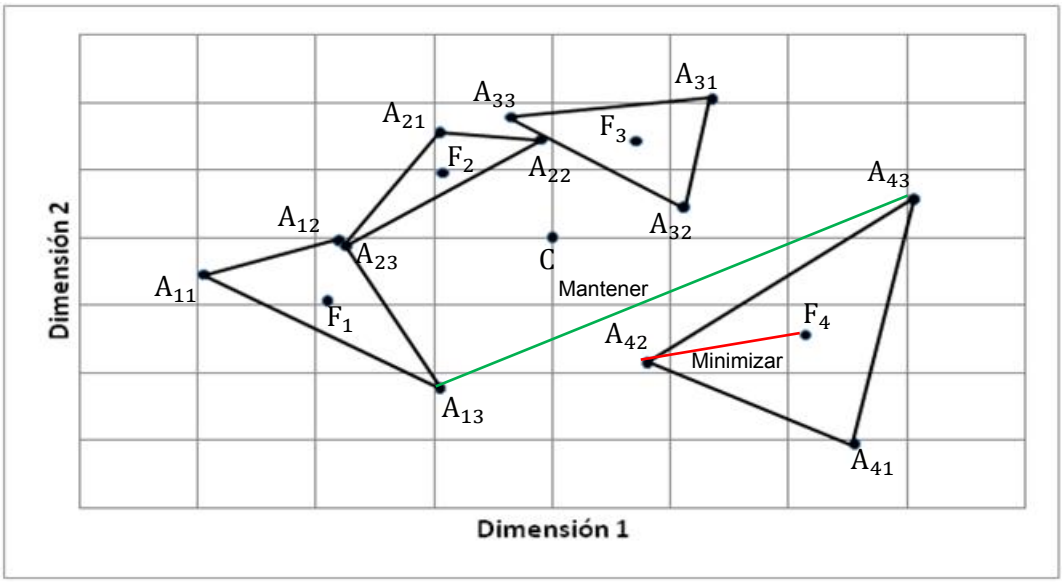


Figura 1.2: Configuraciones de la Figura 1.1 trasladadas a un centroide común C, sin cambios de orientación. Los puntos $F_1, F_2, F_3\ y\ F_4$ representan el centroide de cada uno de los objetos en las distintas configuraciones.

La posición relativa de los centroides sólo tienen sentido cuando las dimensiones utilizadas se refieren a las mismas variables de cada configuración. Cuando las matrices surgen de escalados multidimensionales, la posición de los centroides es arbitraria. Por lo tanto es válido superponer las configuraciones para tener un centroide común C, donde G, G_1, G_2 y G_3 coinciden con C (Figura 1.2). Este paso es el de traslación.

Una vez rotadas y escaladas las configuraciones para minimizar las sumas de cuadrados, la representación formada por los centroides $F_1, \dots, F_4$ asociadas con los objetos, provee un valioso resumen de cómo las configuraciones ven la relación entre los objetos y sus promedios. En 1975 Gower llamó a esta configuración $consenso$ y más adelante Dijksterhuis y Gower llamaron a esta representación $grupo\ promedio$.

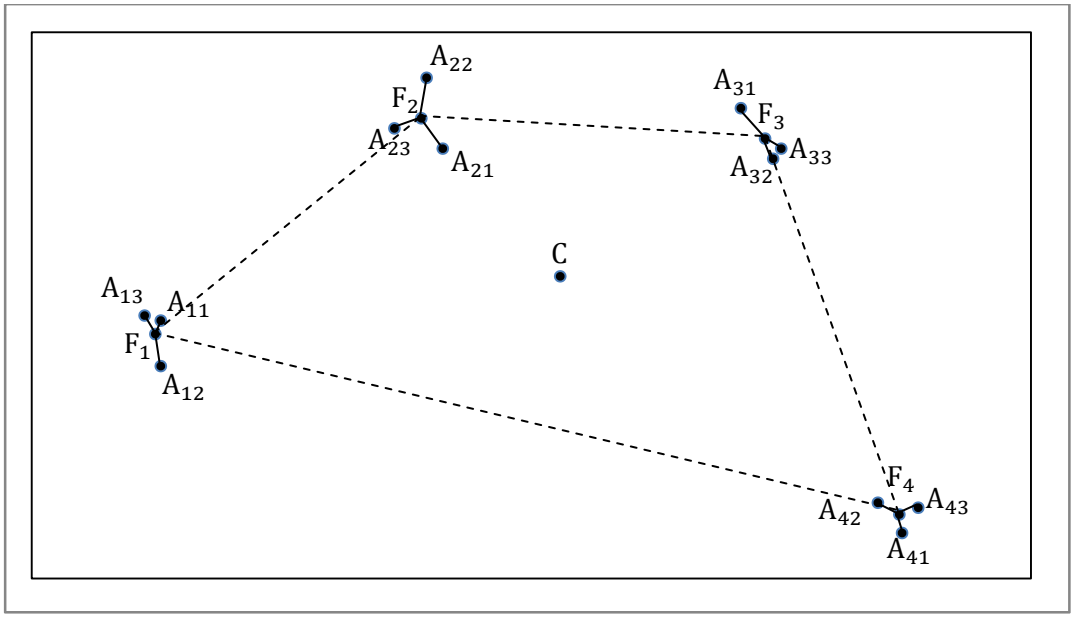


Figura 1.3: Las configuraciones han sido rotadas y escaladas para minimizar la suma de cuadrados entre los puntos objeto con sus respectivos centroides $F_1, F_2, F_3\ y\ F_4$. El $grupo\ promedio$ del APG está indicado con líneas punteadas.

1.3 Reseña sobre Simulación

La simulación es uno de los procesos cuantitativos más ampliamente utilizados en la toma de decisiones y resolución de problemas; sirve para aprender lo relacionado con un sistema real mediante la experimentación con el modelo que lo representa. Las primeras referencias sobre simulación se encuentran hacia el año 1947 cuando Ulam y Von Neumann trabajaron sobre la simulación del flujo de neutrones para la construcción de la bomba atómica en el proyecto "Montecarlo". A partir de ese momento se conocieron las técnicas de simulación como procesos Montecarlo, aunque en la actualidad se diferencian ambas, siendo los segundos un tipo particular de simulación. También se realizó un proceso de simulación para el proyecto APOLLO dentro del plan espacial de la N.A.S.A (Nacional Aeronautics and Space Administration), acerca del movimiento dentro del ambiente lunar, entendiéndose el mismo como una región de moléculas muy delgada, al cual podría llamársele atmósfera.

Actualmente, la simulación es una poderosa técnica para la resolución de problemas. Sus orígenes se encuentran en la teoría de muestreo estadístico y análisis de sistemas físicos probabilísticos complejos. El aspecto común de ambos es el uso de números y muestras aleatorias para aproximar soluciones.

Los investigadores suelen utilizar técnicas para imitar o simular el comportamiento de sistemas del mundo real. Para estudiar científicamente estos sistemas, a menudo se han de hacer una serie de suposiciones acerca de cómo trabaja. Estas suposiciones que usualmente toman la forma de relaciones matemáticas o lógicas, constituyen un modelo que va a ser usado para intentar comprender el comportamiento del sistema correspondiente. (Barrera, 2009)

En este punto, es conveniente plantear las siguientes definiciones:

- **Sistema:** Conjunto de objetos o ideas que están interrelacionados entre sí como una unidad para la consecución de un fin (Shannon, 1988).
- **Modelo:** Representación de un sistema desarrollado con el propósito de estudiar dicho sistema.
- **Simulación:** Proceso de diseñar un modelo de un sistema real y llevar a cabo experiencias con él, con la finalidad de entender el comportamiento del sistema o de evaluar diversas estrategias para el funcionamiento del sistema (Shannon, 1988).

Si las relaciones que componen el modelo son suficientemente simples, es posible usar métodos matemáticos, tales como álgebra, cálculo o teoría de la probabilidad, para obtener una información exacta de las cuestiones de interés; comúnmente denominada *solución analítica*. Sin embargo, la mayoría de los sistemas del mundo real son demasiado complejos y normalmente los modelos realistas de los mismos no pueden evaluarse analíticamente. Lo que se puede hacer es estudiar dichos modelos mediante simulación. En una simulación se utiliza el soporte computacional para experimentar con un modelo de un modo numérico, de forma que con los resultados obtenidos se haga una estimación de las características del sistema. (Barrera, 2009).

1.3.1 **Ventajas y desventajas de la simulación**

1.3.1.1 Modelo analítico

Ventajas:

- Conciso en la descripción del problema.
- Conjunto de soluciones cerrado.
- Permiten evaluar fácilmente el impacto producido por cambios en las entradas sobre las medidas de salida.
- Posibilidad de llegar a una solución óptima.

Inconvenientes:

- Las suposiciones hechas para describir el sistema pueden ser poco realistas.
- Las fórmulas matemáticas pueden ser muy complicadas impidiendo llegar a una solución.

1.3.1.2 Modelos de simulación

Ventajas:

- Pueden describir sistemas que sean muy complejos.
- Pueden ser usados para experimentar con sistemas que todavía no existan, o para experimentar con sistemas existentes sin que éstos se alteren. Esto también los pueden hacer los métodos analíticos siempre y cuando el sistema no sea muy complejo.

Inconvenientes:

- No existe un conjunto de soluciones cerrado.
- Cada cambio en las variables de entrada requiere una solución separada o conjunto de ejecuciones.
- Los modelos de simulación complejos pueden requerir mucho tiempo de construcción y ejecución.
- Puede resultar dificultoso establecer la validez del modelo, es decir, la correspondencia con el sistema real.

1.3.2 **Etapas de la simulación**

Aunque se van a presentar una serie de pasos en forma secuencial, realmente es un proceso iterativo, como se muestra en la Figura 1.4.

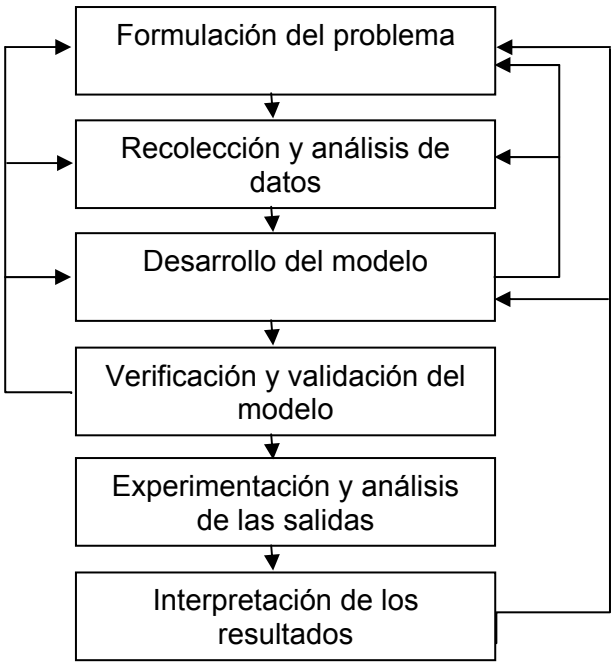


Figura 1.4: Etapas de la simulación

1.3.2.1 Formulación del problema

Debe quedar establecido el objeto de la simulación. Se deberá acordar detalladamente los siguientes factores: resultados que se esperan del simulador, plan de experimentación, tiempo disponible, variables de interés, tipo de perturbaciones a estudiar, tratamiento estadístico de los resultados, complejidad de la interfaz del simulador, etc. (Banks, 1996). En esta etapa se definen cuestiones para las que se buscan respuestas, variables implicadas y medidas de ejecución a utilizar (Barrera, 2009).

1.3.2.2 Recolección y análisis de datos

Aunque la recopilación de datos emerge como el segundo paso, es bastante posible que se haya tenido que recoger datos para la formulación del problema. Sin embargo, durante este paso se recoge, reduce y analiza el mayor volumen de datos (Barrera, 2009).

La naturaleza y cantidad de datos necesarios están determinadas por la formulación del problema y del modelo. Los datos pueden ser provistos por registros históricos, experimentos de laboratorios o mediciones realizadas en el sistema real. Los mismos deberán ser procesados adecuadamente para darles el formato exigido por el modelo (Banks, 1996).

Una vez realizado el muestreo, los datos se analizan e introducen en el modelo. La información usada para definir el modelo puede ser:

- Determinística: Datos conocidos con certeza. Pueden introducirse fácilmente en el modelo.
- Probabilística: Existen dos formas de incluirlos en el modelo:
    - Usar la muestra de datos recogida para representar la distribución de probabilidades.
    - Determinar una distribución probabilística teórica que se comporte como la muestra y usarla en el modelo. Esto permite tener una mejor comprensión o generalización.

1.3.2.3 Desarrollo del modelo

Esta etapa incluye la construcción y depuración del modelo del sistema real, selección de un lenguaje de programación y codificación del modelo. Se divide en dos partes:

- **Comprensión del sistema:** Una de las tareas más difíciles es adquirir el conocimiento suficiente del sistema para desarrollar un modelo apropiado. Para conocer el comportamiento del sistema, una técnica comúnmente usada es la aproximación de flujo físico. Se identifican las entidades o grupos considerados como unidades, cuyo procesamiento o transformación constituye el propósito principal del sistema. Estas entidades pueden tomar diferentes caminos en el sistema, las rutas que siguen se determinan mediante reglas de decisión. Un diagrama de flujo de entidad con los elementos de procesamiento del sistema permitirá la representación del mismo.
- **Construcción del modelo:** Comienza con el desarrollo de un modelo simple que captura los aspectos relevantes del sistema real, los cuales dependen de la formulación del problema. El mismo se irá enriqueciendo como resultado de varias iteraciones. Aquí resulta de vital importancia el lenguaje de programación seleccionado, puesto que puede influir en la forma exacta del diagrama de flujo. Existe un creciente número de lenguajes de programación disponibles para la implementación de modelos de simulación, mientras que otros de uso más general, resultan adecuados para este propósito, entre otros FORTRAN, PASCAL, R, etc. El lenguaje seleccionado puede influir en la generación de números y variables aleatorias, que serán necesarios para representar valores de entrada probabilística. Utilizando estos números aleatorios podemos obtener valores de variables aleatorias que sigan ciertas distribuciones de probabilidad.

1.3.2.4 Verificación y validación del modelo

Por un lado, la verificación consiste en determinar la consistencia interna del modelo. Se comprueba que no se hayan cometidos errores durante su implementación. Para ello, se utilizan las herramientas de *debugging* o depuración provistas por el entorno de programación.

Por otro, la validación asegura la correspondencia entre el sistema real y el modelo. Un buen método para estos fines es realizar un test para conocer cómo el modelo predice el comportamiento del sistema ante determinadas entradas.

La verificación y validación se realiza en todos los niveles de modelización: conceptual, lógico y computacional.

1.3.2.5 Experimentación y análisis de las salidas

Deben diseñarse los experimentos a llevar a cabo sobre el modelo y luego analizar las salidas obtenidas, de forma que permita responder a las cuestiones que se plantearon.

- **Experimentación:** El propósito es obtener información acerca del comportamiento del sistema al momento de tomar decisiones. Durante la ejecución de un sistema se desea conocer cómo se comporta en sentido absoluto, o bien, para poder contrastar varias configuraciones alternativas del sistema. En esta etapa se deciden las características de los experimentos a realizar: tiempo de arranque, tiempo de simulación y número de simulaciones.
- **Análisis de las Salidas:** Las salidas del modelo se consideran muestras. Los principales requisitos en la obtención de estimaciones útiles a partir de muestras son: que la muestra sea representativa del comportamiento del sistema y que el tamaño de la muestra sea lo suficientemente grande para que las estimaciones de las medidas de ejecución alcancen un buen nivel de precisión. Aunque el tamaño de la muestra esté bien definido, la representatividad del comportamiento del sistema depende de cuestiones que tienen que ser contestadas por el modelo.

  Se pueden realizar dos tipos de análisis con un modelo de simulación:

  - Análisis para sistemas con final definido: la ejecución del modelo finaliza cuando ocurre un evento específico. Se toma una muestra por ejecución.
  - Análisis para sistemas con final no definido: el interés reside en el promedio de las medidas de comportamiento en ejecuciones largas. Las medidas en estado estacionario se pueden definir como el valor de las medidas en el límite cuando la longitud de la ejecución tiende a infinito. El tamaño de la muestra es importante ya que la precisión de las estimaciones depende de la varianza de la media muestral y a su vez la varianza cambia de forma inversamente proporcional al tamaño de la muestra. Es decir que

si se cuadriplica el tamaño de la muestra la desviación estándar se reduce a la mitad.

La definición del tamaño de muestra depende del tipo de análisis que se haya hecho. Para sistemas con final definido se puede reproducir el periodo de interés un determinado número de veces hasta conseguir la precisión deseada de la estimación, en condiciones iniciales apropiadas. En cada ejecución de obtiene un elemento de la muestra. Con un análisis en estado estacionario el tamaño de la muestra se encuentra estrechamente enlazado con el tamaño de la ejecución del modelo o cantidad de tiempo de simulación (Barrera, 2009).

1.3.2.6 Interpretación de los resultados de la simulación

Se analiza la sensibilidad del modelo con respecto a los parámetros que tienen asociados la mayor incertidumbre. Se arriba a conclusiones a partir de los resultados obtenidos en la simulación y en caso de ser necesario, se debe recolectar datos adicionales para refinar la estimación de los parámetros críticos.

1.4 Reseña sobre Datos Faltantes

La mayoría de los métodos estadísticos mencionados fueron desarrollados en sus comienzos con matrices rectangulares de datos completos. Al realizarse trabajos de experimentación de campo es habitual encontrar falta de información, ocasionando una pérdida en la precisión de las estimaciones, conforme aumenta la proporción de faltantes. Actualmente, este problema es fundamental para la inferencia por su presencia permanente, hecho que continúa motivando la aparición de diferentes tratamientos para su solución. La elección de los mismos se debe realizar teniendo en cuenta el comportamiento de las pérdidas, tipo de variables afectadas y procedimiento de análisis que se desea aplicar (Badler, 2004).

Los métodos utilizados habitualmente para tratar el problema de los datos perdidos eran: eliminación del caso, sustitución/imputación de éstos por valores plausibles como la media de la variable o predicción obtenida mediante regresión sobre las demás variables del vector, en el caso de tratar con información cuantitativa. Éstos no suelen tener una base teórica sólida, aunque fáciles de implementar y adecuados en situaciones concretas, presentan en general importantes inconvenientes y carencias, especialmente en contextos multivariantes. Los principales problemas inferenciales asociados son: ineficiencia, sesgos, distorsión de la estructura de covarianzas; además de no incorporar la incertidumbre asociada a los datos faltantes.

Frente a estos métodos clásicos, se han ido desarrollando en los últimos años y de forma paralela a la formalización del problema de los datos faltantes, métodos con una base teórica más sólida. Así, en Dempster, Laird y Rubin (1977) establecen una formulación general y rigurosa para la inferencia en presencia de datos faltantes mediante el algoritmo EM (Expectation – Maximization). Por otro lado, Rubin (1987) desarrolla una nueva metodología de propósito general, flexible y fundamentada que denomina imputación múltiple, y que salva muchos de los inconvenientes asociados al tratamiento tradicional de los datos faltantes (Gómez García, 2006).

1.4.1 **Mecanismo de pérdida**

El proceso que conduce a la pérdida de información en un relevamiento o experimento es denominado mecanismo de pérdida. Este mecanismo intenta responder la pregunta si existe relación entre la pérdida y la información subyacente en el conjunto original de datos, es decir, si ésta pérdida es aleatoria o está asociada a ciertos niveles o valores de una o varias variables.

Los mecanismos de pérdida son cruciales dado que las propiedades de los métodos de datos faltantes dependen fuertemente de la naturaleza de las dependencias de estos mecanismos. El rol crucial de los mecanismos en el análisis de datos con valores faltantes ha sido ignorado durante mucho tiempo hasta que el concepto fue formalizado en la teoría de Rubin (1976), a través de un simple sistema de tratamiento de indicadores de datos faltantes como variables aleatorias a las cuales se les asigna una distribución (Little y Rubin, 2002).

La información incompleta en una variable puede presentarse en forma aleatoria o en categorías de valores dentro de la propia variable, determinando en este último caso que los valores no observados sean diferentes a los observados. Dichos mecanismos de pérdida se pueden clasificar a partir de las características de la probabilidad de respuesta de los datos faltantes:

- Los datos están perdidos completamente al azar (*missing completely at random* MCAR): si la probabilidad de respuesta de los faltantes es independiente de las variables observadas y de las no observadas.
- Los datos están perdidos al azar (*missing at random* MAR): si la probabilidad de respuesta de los faltantes es independiente a las variables no observadas y no de las observadas.
- Los datos no están perdidos al azar (*not missing at random* NMAR): si la probabilidad de respuesta de los faltantes no es independiente de las variables no observadas y posiblemente, tampoco, de las observadas.

1.4.2 Primeros métodos heurísticos para tratamiento de valores faltantes

Aunque intuitivamente parecen soluciones razonables y funcionan bien cuando la cantidad de información perdida es pequeña, no son aceptables en general.

1.4.2.1 Análisis de datos completos

Consiste en utilizar aquellos elementos en los cuáles se encuentra información completa de todas las variables, o sea filas completas. Como ventajas se pueden destacar su simplicidad y el hecho de que todos los estadísticos son calculados con el mismo tamaño muestral, lo que permite su comparación. Sin embargo, la principal desventaja ante grandes pérdidas de información, es el sesgo y la falta de precisión en las estimaciones de los parámetros, si los faltantes no son una muestra aleatoria de la muestra completa.

1.4.2.2 Análisis de casos disponibles

Incluye todos los casos que son observados en cada variable. Es decir, todos los valores disponibles de cada columna. Su principal problema es que esto implica en general trabajar con distintos tamaños muestrales e incluso combinarlos en la estimación de un mismo estadístico. Por ejemplo, los promedios y variancias se calculan para los casos disponibles en cada variable, mientras que las covariancias y correlaciones en base a todos los casos que no presenten datos faltantes para el par de variables implicadas.

1.4.3 Métodos de imputación simple

En ellos se resuelve el problema de los datos faltantes sustituyendo a los mismos por valores estimados en base a la información existente en la muestra, cuando las variables son cuantitativas. De esta forma se construye una matriz rectangular completa de datos sobre la que se realizan los análisis. La forma de estimar o predecir los valores perdidos permite diferenciar los distintos métodos.

1.4.3.1 Imputación por la media

Reemplaza los valores faltantes de una determinada variable por la media calculada en la muestra de esa variable. Aunque sencilla e intuitivamente satisfactoria, esta estrategia presenta un importante defecto: tiende a subestimar

la variabilidad real de la muestra al sustituir los faltantes por valores centrales de la distribución.

1.4.3.2 Imputación por regresión

Aquí los valores faltantes son estimados por regresión lineal múltiple. Se estima la regresión de la variable X_k sobre las variables $X_{k\prime}$, $\forall\, k \neq k'$, a partir de los casos completos, imputando cada valor perdido con la predicción dada por la ecuación de regresión estimada. Esto es, si para el caso I no se observa el valor x_{Ik}, entonces se imputa mediante:

$$\hat{x}_{Ik} = \hat{\beta}_{O.obs} + \sum_{k \neq k\prime} \hat{\beta}_{k\prime.obs}.\, x_{Ik\prime}$$

Donde $\hat{\beta}_{O.obs}$ y $\hat{\beta}_{k\prime.obs}$,$k' \neq k$, representan los coeficientes de la regresión de X_k sobre $X_{k\prime}$, $\forall k \neq k'$, basadas en las observaciones completas. A diferencia del método anterior, éste incorpora la información que contiene el resto de las variables.

Al imputar mediante regresión se reemplaza el valor perdido por una media condicionada, por lo que también se tiende a subestimar sistemáticamente la variabilidad. Una sencilla alternativa para atenuar este efecto consiste en añadir al valor predicho por la regresión una perturbación aleatoria determinada. Así se obtiene una realización de la distribución predictiva de los valores perdidos condicionada a los valores observados. A esta estrategia se la denomina $imputación\ por\ regresión\ estocástica$.

1.4.4 Métodos basados en verosimilitudes

Se basan en funciones de verosimilitud, por lo tanto son métodos bajo los que subyace un modelo probabilístico. Los dos más utilizados son:

1.4.4.1 Algoritmo EM

Dempster, Laird y Rubin en 1977 crean un algoritmo iterativo que consiste en un paso E ($expectation$) y un paso M ($maximization$) diseñado para la obtención de estimadores máximo-verosímiles (EMV) en problemas con muestras incompletas. En primer lugar, el paso E consiste en calcular la esperanza condicional de los datos faltantes dados los datos observados y la estimación de los parámetros. Luego estas esperanzas sustituyen a los datos faltantes. Luego,

el paso M realiza la estimación máximo-verosímil del parámetro de interés como si no existieran datos faltantes. En síntesis, se reemplazan los valores perdidos por los valores estimados; se estiman los parámetros; luego se re-estiman los valores perdidos asumiendo que son correctas las nuevas estimaciones de los parámetros y nuevamente se vuelven a estimar. Un criterio de convergencia habitual en la práctica consiste en detener el proceso cuando la diferencia entre dos estimaciones sucesivas de los parámetros sea suficientemente pequeña. Esta convergencia puede ser lenta cuando existe una gran proporción de datos faltantes. Su aplicación requiere que los datos sean perdidos al azar (*MAR*).

1.4.4.2 El método de imputación múltiple

Mediante imputación múltiple se reemplaza cada valor perdido por un conjunto de valores simulados con el fin de incorporar a la estimación la incertidumbre debida a la presencia de datos faltantes. La referencia básica sobre imputación múltiple es Rubin (1987), aunque se puede encontrar una variedad de trabajos relevantes como por ejemplo Rubin (1996), Schafer (1997), Little y Rubin (2002) o Zhang (2003).

Esta metodología permaneció durante algunos años en un segundo plano por su limitada aplicabilidad. Principalmente debido a la inexistencia de herramientas computacionales adecuadas para poder crear las imputaciones. El desarrollo tecnológico de las últimas décadas permitió la implementación de algoritmos y procedimientos de cálculo computacionalmente intensivos necesarios para dar solución a problemas intratables analíticamente. Durante los años 90 se popularizaron los algoritmos de Markov Chain Monte Carlo (MCMC) que permiten una modelización estadística más compleja, al tiempo que realista. Este tipo de algoritmos también encontraron su aplicación en el ámbito de los datos faltantes, en concreto, su incorporación al contexto de la imputación múltiple (Schafer, 1997) ha convertido este procedimiento en un destacado método para el análisis de datos incompletos. (Gómez García, 2006).

CAPITULO II

OBJETIVOS E HIPÓTESIS

2.1 Exposición del problema

El objetivo del proyecto de Investigación "Evaluación toxicológica de poblaciones de riesgo en la exposición crónica a plaguicidas", en el cual se enmarca esta Tesis, fue identificar si existían alteraciones tempranas en poblaciones expuestas directa e indirectamente a plaguicidas y detectar biomarcadores de toxicidad que permitan predecir riesgo. Para ello se llevó a cabo un estudio epidemiológico de tipo prospectivo en una población de embarazadas que asistieron al programa de atención prenatal en el Hospital Área Programa Cinco Saltos, Provincia de Río Negro. Se relevaron datos sobre las características de la población y sus hábitos por medio de entrevistas. La información sobre biomarcadores se obtuvo por análisis de laboratorio de muestras de sangre de las embarazadas y de placenta a término. Los datos del neonato se registraron en base a la planilla de control del recién nacido.

Este proyecto de investigación puede enmarcarse dentro de la llamada "epidemiología molecular", abordaje en el que se aplican herramientas tanto de la epidemiología clásica como de la biología molecular.

El carácter sociológico-cultural de esta problemática trae aparejada consigo la naturaleza mixta de las variables: ordinales, nominales, continuas, etcétera, propia de la temática que circunscribe este tipo de estudios. Esto se debe a la diversidad de métodos para la obtención de los datos, como análisis de laboratorio, entrevistas, historias clínicas, etc.

2.2 Justificación del tema de estudio

En el extenso bagaje de técnicas dentro del análisis multivariado se encuentra la riqueza que permite contemplar la naturaleza de las variables, aplicando un tratamiento diferenciado de acuerdo a que sean cualitativas o cuantitativas. El uso de los métodos multivariados en la epidemiología clásica incluye modelos de regresión múltiple para datos categóricos, interacción y factores de confusión

(Holford, 2002). Asimismo, estudios de la importancia de los plaguicidas sobre la incidencia de alteraciones o enfermedades en poblaciones humanas recurren también a un enfoque multivariado en el análisis de los datos (Hardell y Eriksson, 1999; Harnly y col. 2005). Dado que la realidad estudiada se manifiesta de manera compleja en cuanto a la diversificación de las variables, se requiere de distintas técnicas capaces de transformar los datos recogidos a campo, en información útil para el investigador idóneo de la disciplina.

Por otra parte, y continuando con la problemática de los datos faltantes, más allá de la metodología de rellenado utilizada cabe preguntarse en qué grado se distorsiona la información original, teniendo en cuenta que los procedimientos suelen tornarse ineficientes cuando la proporción de faltantes en la base de datos aumenta. Se presume que a partir de cierta proporción de faltantes algunas estrategias multivariadas no podrían aplicarse, ya que no sería posible hallar matrices de similaridad entre individuos. Entonces surge el siguiente interrogante: si bien algunas técnicas son más eficaces que otras en cuanto a la calidad de representación de matrices con faltantes, ¿es posible determinar cuáles son más robustas — en el sentido de aplicabilidad — a la presencia de datos faltantes?

2.3 Objetivos e hipótesis de trabajo

2.3.1 **Objetivo general**

El objetivo del presente trabajo de tesis fue analizar e indagar la factibilidad y adecuación de las técnicas del Análisis Multivariado para caracterizar poblaciones de embarazadas expuestas a plaguicidas a fin de detectar biomarcadores de toxicidad que permitan predecir riesgo.

Para alcanzar este objetivo se partió de la hipótesis:

“Existen técnicas estadísticas capaces de analizar de manera eficiente los datos del proyecto ‘*Evaluación toxicológica de poblaciones de riesgo en la exposición crónica a plaguicidas*’, salvando el problema de los datos faltantes”

2.3.2 **Objetivos específicos**

- Estudiar en forma comparativa técnicas y estrategias en la caracterización y agrupamiento de mujeres embarazadas potencialmente expuestas a plaguicidas caracterizadas por variables mixtas.
- Determinar cuáles son los variables independientes más importantes en las alteraciones del período perinatal, en relación a la exposición a plaguicidas.
- Evaluar el impacto de los valores faltantes en los análisis estadísticos propuestos.

2.4 Partes constitutivas de la tesis

En el Capítulo 1 se planteó el problema y se desarrolló la revisión de literatura necesaria para comprender el trabajo de tesis. Allí se encuentran los aspectos biológicos necesarios para interpretar los efectos de los plaguicidas sobre los seres humanos. Luego, un raconto de estadística multivariada permite al lector interiorizarse en las estrategias utilizadas para la caracterización de poblaciones de riesgo. Una síntesis pormenorizada sobre simulación de datos y estudio de datos faltantes presentó los aspectos más relevantes en los que se fundamenta el presente trabajo. En el Capítulo 2 se presentó el caso a estudiar y se expusieron los objetivos e hipótesis de la tesis.

El desarrollo de los Materiales y Métodos experimentales y los Métodos estadísticos ocupa el lugar principal del Capítulo 3, explicando el diseño del estudio epidemiológico, las técnicas bioquímicas utilizadas, las técnicas estadísticas clásicas necesarias para cualquier abordaje estadístico, y también los métodos vinculados con la simulación y el tratamiento de datos faltantes. La caracterización de las variables contempladas en las simulaciones, la aplicabilidad de las estrategias y su calidad de representación resulta de vital importancia para la interpretación y discusión de los resultados presentados en el Capítulo 4, el cual conforma la parte central del trabajo.

Las conclusiones desarrolladas en el Capítulo 5 conjugan la tarea de síntesis y jerarquización de los hallazgos, identificando explícitamente los elementos de originalidad, y las proyecciones hacia el futuro en materia de aplicación y/o investigación.

CAPITULO III

METODOLOGÍA

3.1 Aspectos biológicos

3.1.1 **Muestreo de la población bajo estudio**

Para evaluar el grado de exposición a plaguicidas y sus posibles efectos se estudiaron 306 embarazadas atendidas durante la gestación y parto en el servicio de Ginecología y Obstetricia del Hospital de Cinco Saltos entre junio de 2002 y junio de 2003. Se incluyeron madres sanas de entre 17 y 35 años, que no recibieron medicación potencialmente teratogénica, ni padecieron rubéola o toxoplasmosis durante el embarazo. Además se excluyeron las madres diabéticas, hipertensas y madres expuestas a rayos X en el primer trimestre de embarazo. El personal médico solicitó un consentimiento informado a las madres para la toma de muestras y realizó una entrevista. Se obtuvieron registros de la madre, la placenta y el neonato. De la madre se registró mediante la mencionada entrevista: lugar de residencia, estado nutricional, nivel de instrucción y hábitos como el uso de plaguicidas en el hogar — entre otros — de acuerdo con la planilla que se encuentra en el Apéndice 1. Las muestras de placenta fueron obtenidas dentro de la media hora posterior al parto para evitar alteraciones en su metabolismo en concordancia con Serkova y col. (2003). Dado que es un órgano muy heterogéneo y la expresión de los distintos componentes puede variar de acuerdo a su localización dentro de la placenta (Pidoux y col., 2004), las muestras se tomaron siempre en la misma ubicación (Figura 3.1), se congelaron inmediatamente a -20ºC y se procesaron dentro de los 2 meses posteriores. En placenta se determinó: peso, actividad de la enzima Acetilcolinesterasa (AChE), biomarcador de exposición a OF y de relevancia en el metabolismo de acetilcolina (ACh) placentaria. En un subgrupo de placentas se determinó, adicionalmente, la actividad de Catalasa (CAT), enzima involucrada en la defensa antioxidante. A partir de las historias clínicas se obtuvieron los datos del neonato: sexo, peso, talla y perímetro cefálico al nacer. En la Tabla 3.1 se detallan sólo las variables consideradas en el estudio.

Tabla 3.1: Detalle de las variables utilizadas

Nº	Descripción de las variables (abreviatura)	Tipo	Categorías
	Relacionadas con la madre		
1	Hábito de fumar (fuma)	cl	(3 categorías) No Fumadora Fumadora Pasiva / Fumadora Activa
2	Nivel de Instrucción Alcanzado (NIA)	cl	(3 categorías) Primario Completo Secundario Completo / Universitario
3	Consumo de agua de pozo (pozo)	D	Si/No
4	Fumigación en el hogar (fumih)	D	Si/No
5	Alteraciones en el embarazo (AEB)	D	Si/No
6	Época de parto (EP)	D	Pulverización/Receso
	Relacionados con el Neonato		
7	Peso al nacer (Peso)	CT	
8	Talla al nacer (Talla)	CT	
9	Perímetro Cefálico al nacer (PC)	CT	
10	Sexo (Sexo)	D	Femenino/Masculino
11	Edad Gestacional (EG)	CT	
	Relacionados con la Placenta		
12	Peso de la Placenta (PPla)	CT	
13	Acetilcolinesterasa (AChE)	CT	
14	Catalasa (CAT)	CT	

cl: cualitativa D:dicotómica CT: cuantitativa

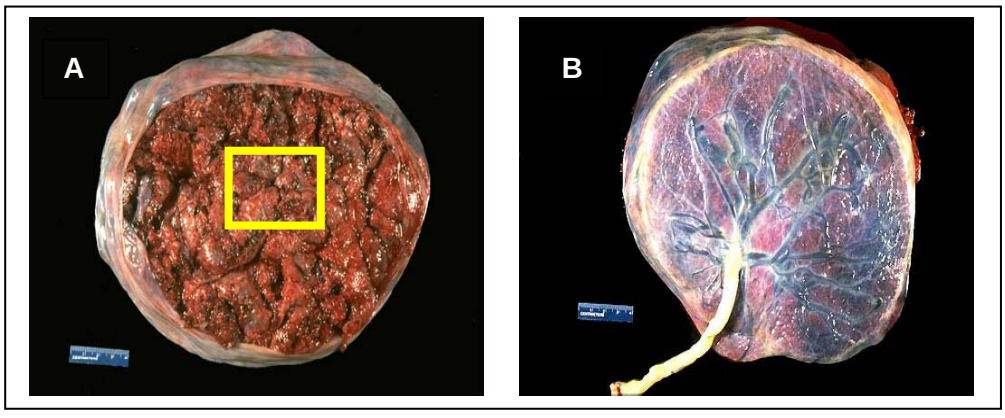


Figura 3.1: Fotografías de una placenta a término. La regla que sirve de referencia tiene una longitud total de 4 cm. En **A** se muestra la cara materna, el cuadrado amarillo indica la localización del área de muestreo y en **B** se muestra la cara fetal de la placenta y el cordón umbilical.

3.1.2 **Obtención de homogenados de placenta**

Las muestras fueron tomadas de la porción central de la "cara materna" de la placenta (Figura 3.1). Se cortó en trozos la porción extraída, se lavó con solución fisiológica fría para eliminar restos de sangre, se secó con papel de filtro y se pesó. El homogenado fue realizado en amortiguadores adecuados para cada determinación y luego filtrado en una doble capa de muselina.

3.1.3 **Actividad de Acetilcolinesterasa**

La actividad enzimática de AChE se determinó en homogenados de vellosidades placentarias de acuerdo al método de Ellman y col. (1960). La medición cinética fue realizada en solución amortiguadora pH=8 usando ioduro de acetilcolina como sustrato. El grupo tiol, producto de su degradación, fue detectado en su reacción con DTNB a 412 nm durante 2 minutos a 30ºC. La actividad específica se expresó como micromoles de acetiltiocolina hidrolizada por minuto y por miligramo de proteína del tejido.

3.1.4 **Actividad de Catalasa**

Los homogenados placentarios obtenidos para la medición de esta enzima se realizaron en solución amortiguadora pH=7. Se registró espectrofotométricamente, de manera continua, la disminución de la absorbancia del peróxido de hidrógeno a una longitud de onda de 240nm a 25 ºC en presencia de la muestra (Beers y Sizer, 1952). La actividad específica de CAT se expresó como UI/mg proteínas.

3.2 Métodos Estadísticos

3.2.1 **Fundamentación**

Para lograr el objetivo general del Proyecto de Investigación es necesario que sus conclusiones estén fundamentadas desde el marco teórico de la estadística. Luego de un pormenorizado análisis uni y bivariado de la tabla de datos, se incursionó en técnicas del Análisis Multivariado con un fin descriptivo preliminar para discriminar grupos de madres de particularidades similares, considerando que en la base de datos conviven variables mixtas. Para ello se empleó ACP para las variables cuantitativas y ACoP para las variables cualitativas. El coeficiente de similaridad de Gower (1971) fue una opción para una

caracterización conjunta en función de variables cuanti y cualitativas, junto con APG (Gower y Dijksterhuis, 2004) para contrastar diferencias entre las madres originadas por distintos grupos de variables. Se analizó en profundidad las bondades y desventajas en el tratamiento de variables mixtas de las distintas técnicas, teniendo en cuenta la calidad de representación de las mismas, la robustez ante la presencia de datos faltantes, medidas de pérdida de información, entre otras.

Por otra parte, se recurrió a diferentes técnicas de simulación capaces de captar la información contenida en la matriz de datos observados y, a partir de esta estructura, obtener un archivo rectangular de datos completos. Una vez obtenida esta nueva matriz, se emplearon estrategias de eliminación de celdas y sobre estas nuevas situaciones se aplicaron los métodos arriba descriptos, estudiando la robustez de los mismos ante la presencia de distintas magnitudes de datos faltantes.

Con el objetivo de captar la estructura de variación conjunta entre ambos tipos de variables, cuali y cuantitativas, se programó un macro que permitió simular de manera eficiente la base de datos originales.

3.2.2 **Análisis uni y bivariado (análisis clásico)**

Se aplicaron técnicas univariadas como histogramas, box-plot y diagramas de dispersión para describir el comportamiento de las variables cuantitativas, mientras que se calcularon los porcentajes correspondientes a las categorías de las variables cualitativas. Los análisis en los que intervinieron alguna de las variables antropométricas del neonato, es decir, peso, talla y perímetro cefálico, fueron corregidas previamente por sexo y edad gestacional (Lejarraga 1986), utilizando los z scores –Pesocorr, Tallacorr, PCcorr–. Se verificó el supuesto de normalidad para las variables continuas, mediante la prueba de Shapiro-Wilks modificado por Mahibbur y Govindarajulu (1997). Se calcularon los coeficientes de correlación entre las distintas variables cuantitativas. Se realizaron pruebas Chi cuadrado de independencia para determinar asociación entre las variables cualitativas. En el caso en el que las variables categóricas tuvieron sólo dos categorías, se presentaron los cocientes de chance , referido en la literatura en inglés como “odds ratios”. Para relacionar variables cualitativas de dos niveles con cuantitativas se utilizaron pruebas t para muestras independientes para

establecer diferencia de medias, previa prueba de homogeneidad de varianzas. Asimismo, se realizaron Análisis de la Varianza de un Factor para relacionar variables multiestado con cuantitativas. Se utilizó la prueba a posteriori de LSD Fisher con el objetivo de detectar diferencias significativas dentro de los niveles de un factor. Cuando la variable respuesta resultó ser dicotómica, se utilizaron Modelos Lineales Generalizados. Estas técnicas fueron empleadas mediante el Software Infostat (Versión 2008) y R (versión 2.10).

3.2.3 **Estrategias Multivariadas**

En muchas aplicaciones prácticas de caracterización nos encontramos que algunas variables son cualitativas y otras cuantitativas (variables mixtas). Dado que generalmente es de interés del investigador tratar en forma simultánea toda la información en forma conjunta, se pueden seguir dos caminos alternativos: (i) por un lado, buscar una medida de distancia y/o similaridad que permita cuantificar la asociación entre individuos en función de todas las variables estudiadas y luego sobre ella aplicar algún método de ordenación o clasificación apropiado, o bien (ii) asumir que se cuenta con variables diferentes que requieren tanto del uso de medidas de similitud como técnicas de agrupamiento particulares y obtener una configuración o clasificación de los individuos u objetos bajo estudio para cada tipo de variable para luego proceder a 'consensuar' ambas caracterizaciones.

Se consideraron dos estrategias multivariadas al momento de estudiar los individuos: la primera consistió en hallar una matriz de distancia, mediante un coeficiente de similaridad para variables mixtas, aplicando posteriormente una técnica factorial de representación; en la segunda se obtuvieron por separado representaciones de las variables cuantitativas y cualitativas, consensuando dichas ordenaciones mediante algún método. A continuación se detallan las dos estrategias multivariadas comparadas:

Estrategia de una Etapa (EduE):

Hallar la matriz de distancias a partir del coeficiente de similaridad de Gower Simple Matching (1966). Luego, aplicar ACoP a la matriz hallada.

Estrategia Bietápica (EB):

Cuando las variables en estudio son cualitativas, calcular la matriz de similaridad a partir del coeficiente de similaridad Simple Matching y posteriormente aplicar el ACoP. Análogamente, cuando las variables en estudio son cuantitativas, obtener la matriz de correlaciones entre variables y luego utilizar ACP. Aplicar APG para hallar la configuración consenso, resultante de las dos configuraciones halladas previamente. Cabe destacar que al momento de realizar ACP, se estimaron los faltantes mediante imputación por la media de la variable.

3.3 Simulación

Se elaboró un programa con el paquete estadístico R para generar matrices de datos de naturaleza mixta a partir de la estructura de la base de datos empírica. Esto permitió que los datos recabados en el proyecto de investigación puedan adaptarse a diferentes situaciones tales como número de individuos, proporción de faltantes, etc., posibilitando generalizar y arribar a conclusiones acerca del funcionamiento de las estrategias mencionadas anteriormente.

Teniendo en cuenta que el número de individuos, número de variables y proporción de datos faltantes podía afectar los resultados posteriores, se contemplaron distintos valores para estos factores. En el caso del número de individuos, se decidió analizar 5 situaciones diferentes: $n = 50, 100, 150, 200, 250$; debido, entre otras cosas, al tamaño de muestra de la base original, estudios de similares características y limitaciones computacionales. También se utilizó el mismo procedimiento con el número de variables (p) a trabajar, tomando las variables cualitativas más importantes y aquellas cuantitativas que en el estudio univariado no presentaron un alejamiento significativo de la distribución normal. Se trabajó con $p = 10, 11\ y\ 12$, combinándose las cantidades de variables cuanti y cualitativas que se detallan en la Tabla 3.2. De esta manera cada simulación quedó determinada por el número de individuos y las diferentes combinaciones de variables, obteniéndose 40 situaciones distintas. Cada una de estas situaciones se repitió 1000 veces.

Tabla 3.2: Combinación de variables utilizadas para la simulación

Combinación nº	Tipos de variables	Cantidad de variables p
1	5 cualitativas – 5 cuantitativas	10
2	6 cualitativas – 4 cuantitativas	
3	5 cualitativas – 6 cuantitativas	11
4	6 cualitativas – 5 cuantitativas	
5	7 cualitativas – 4 cuantitativas	
6	5 cualitativas – 7 cuantitativas	12
7	6 cualitativas – 6 cuantitativas	
8	7 cualitativas – 5 cuantitativas	

Posteriormente, se decidió el porcentaje de faltantes a extraer en cada una de esas combinaciones. Suponiendo que los datos faltantes eran MCAR (*missing completely at random*), se procedió a la extracción al azar datos de la base simulada. Dado que la proporción de faltantes de la base original era de aproximadamente el 17,9%, se eligieron los siguientes porcentajes de faltantes $q = 1, 5, 10, 15, 20, 25\ y\ 30\%$. Cuando la proporción de faltantes de una matriz no resultó ser un número entero de datos a extraer, se tomó el siguiente. Al número de datos faltantes se lo denominó con la letra f.

Una vez obtenidas las matrices, se observó el comportamiento de las estrategias mencionadas ante la presencia de los datos faltantes mediante correlaciones entre las matrices de distancias de las configuraciones obtenidas para los datos completos y las matrices de distancia de las configuraciones correspondientes a los datos con faltantes. De esta forma se observa como inciden los faltantes en la calidad de representación de cada una de las estrategias utilizadas.

Paralelamente y de forma complementaria, se realizaron perfiles multivariados para determinar el comportamiento de matrices a las que no se les pudo aplicar estrategias multivariadas, teniendo en cuenta que la quita de datos puede provocar la imposibilidad de encontrar distancias entre dos individuos. De esta forma se encuentran $Matrices\ Not\ Available$, o Matrices NA. De ahí surge el estudio de la proporción de matrices de distancia con elementos indeterminados que se expone en el apartado 4.2.2.2.

CAPITULO IV

RESULTADOS Y DISCUSION

Como se detalló en el Capítulo 3, la metodología estadística aplicada se dividió en dos partes. La primera tendiente a explicar el fenómeno estudiado mediante técnicas estadísticas habitualmente utilizadas o clásicas, y la segunda orientada a la caracterización multivariada de las madres incorporadas dentro del muestreo.

Por otra parte, y con el objetivo de generalizar el comportamiento de estas estrategias ante la presencia de datos faltantes en situaciones diferentes a las encontradas en la base de datos original, se trabajó con simulaciones que permitan responder a dichos objetivos.

4.1 Métodos estadísticos

4.1.1 **Técnicas clásicas**

El Apéndice 2A presenta medidas de resumen descriptivas, histogramas y box-plot de las variables cuantitativas estudiadas. Se debe tener en cuenta que los distintos tamaños muestrales quedaron determinados por la gran cantidad de faltantes en la base de datos. A su vez, el comportamiento de las variables antropométricas estandarizadas (z scores), peso de la placenta y biomarcadores AChE y CAT tuvieron una distribución aproximadamente simétrica, sin registrarse valores atípicos; en tanto que la EG de los neonatos, resultó ser una distribución con presencia de algunos valores atípicos a la izquierda.

En cuanto a las variables cualitativas, el 36% de las madres no eran fumadoras, 44% fumadoras pasivas y 20% fumadoras activas, mientras que el máximo nivel de instrucción alcanzado fue hasta primario completo 50%, hasta secundario completo 44% y Universitario o Terciario 6%. El 16% de las mujeres consumía agua de pozo, 45% fumigaban en el hogar y un 42% dio a luz en época de pulverizaciones con OF. Por último, el 48% de los neonatos registrados fue de sexo femenino (Apéndice 2B).

Continuando el desarrollo desde una perspectiva inferencial (ver Apéndice 3), se detectaron leves asociaciones, pero no significativas al 5%, entre el nivel de instrucción alcanzado y el hábito de fumar (p=0,07), como así también con el hábito de fumigar en el hogar (p=0,09). Asimismo, la persona que tomó agua de pozo tuvo aproximadamente 4 veces más chances de haber fumigado en su hogar que otra que no lo hace.

Prosiguiendo con el estudio de las variables continuas, la Tabla 4.1 muestra que no hubieron alejamientos muy significativos de la normalidad (p=0,01), salvo para la edad gestacional (p<0,0001). En este sentido, múltiples factores fisiológicos se combinan para desencadenar el parto en forma natural. Sin embargo, existen otros determinantes de la edad gestacional del neonato relacionados con intervenciones médicas, como es la inducción del parto con la hormona oxitocina, que podrían haber modificado la distribución normal de los valores. Por ello, esta última no se incluyó en el proceso de simulación, ya que el macro se sustenta en el supuesto de que las variables siguen una distribución normal multivariada.

Tabla 4.1: Pruebas de normalidad para variables continuas.

Variable	n	Media	D.E.	W*	P
Pesocorr	253	0,23	0,88	0,99	0,4594
Tallacorr	253	-0,65	1,11	0,98	0,0216
PCcorr	253	0,28	1,19	0,99	0,7031
AChE	111	12,13	3,94	0,97	0,0847
PPla	248	602,26	134,77	0,98	0,0200
CAT	27	0,15	0,04	0,95	0,4925
EG	260	39,10	1,45	0,81	<0,0001

A continuación se presentan los coeficientes de correlación de Pearson bajo la diagonal principal con sus respectivos p-values por encima de la misma (Tabla 4.2) y un matrix-plot (Figura 4.1), con las relaciones entre las variables continuas.

Tabla 4.2: Coeficientes de correlación de Pearson de las variables estudiadas por debajo de la diagonal – p-valores por encima de la diagonal.

	Pesocorr	Tallacorr	PCcorr	PPla	AChE	CAT
Pesocorr	-	0,0000	0,0000	0,0000	0,7478	0,3923
Tallacorr	0,5494	-	1,1E-06	0,0003	0,5370	0,4143
PCcorr	0,6038	0,3006	-	2,4E-08	0,0665	0,0639
PPla	0,5823	0,2519	0,3768	-	0,3220	0,4449
AChE	-0,0332	-0,063	-0,1881	0,1068	-	0,0213
CAT	0,1919	0,1833	0,4016	0,1811	0,4877	-

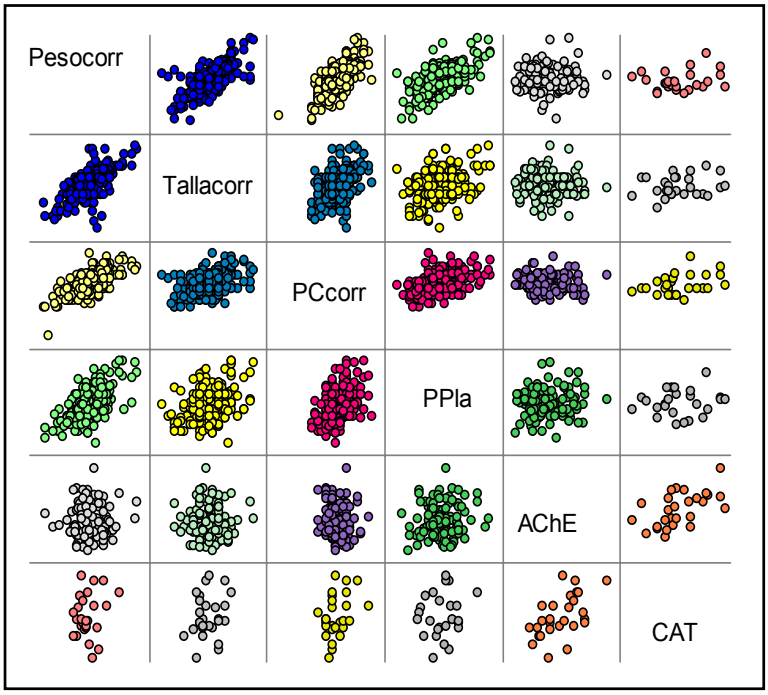


Figura 4.1: Diagrama de dispersión de las variables estudiadas.

Se pueden apreciar correlaciones lineales directas significativas entre Pesocorr, Tallacorr, PCcorr y PPla. En este sentido cabe mencionar que la asociación entre el peso de la placenta y el peso y la talla del neonato es un resultado esperado dado que el peso de la placenta es un importante predictor del peso del neonato (Sanin y col., 2001). Estudios en recién nacidos con bajo peso mostraron placentas a nivel del área de los vasos y del área de las vellosidades coriales libres significativamente diferentes de las placentas de neonatos de peso adecuado para la edad gestacional (Prieto Gómez y col., 2008). Por otra parte se observa una correlación positiva entre la actividad de la AChE y la actividad de CAT, enzima relacionada con la defensa antioxidante, resultado que es coherente con la información que los OF causan estrés oxidativo (Mansour y Mossa, 2009; Bayrami y col. 2011). Por el contrario, no se hallaron correlaciones significativas entre los biomarcadores y los parámetros antropométricos.

En el Apéndice 4 se muestran los cambios en biomarcadores y parámetros antropométricos para diferentes niveles de las variables categóricas. Se observó que los perímetros cefálicos corregidos en niños de madres que tomaban agua de pozo fueron mayores de las que no consumían ($p<0,01$), como así también en aquellas que fumigaban dentro del hogar ($p<0,01$). El resto de las variables no fueron afectadas significativamente por esos factores. Se observó que tanto

la AChE como CAT mostraron aumentos altamente significativos cuando la Época de parto correspondió al período de pulverizaciones (p<0,001).

En madres fumadoras se registró que los neonatos obtuvieron pesos corregidos significativamente más bajos que los de aquellos cuyas madres que no fumaban o lo hacían de forma pasiva (p=0,01). En el resto de las variables continuas no se produjeron cambios. Por último, el nivel de instrucción no alteró el comportamiento de las variables antropométricas ni el de los biomarcadores.

Los modelos lineales generalizados logit no detectaron relaciones significativas entre las variables, por lo que no se pudo arribar a una función de riesgo que permita estimar la probabilidad de una Alteración en el embarazo en función de las variables Hábito de fumar, Nivel de instrucción alcanzado, Consumo de agua de pozo, Fumigación en el hogar y Época de parto. La alta proporción de datos faltantes en la base de datos generó un gran número de celdas vacías, lo que dificultó el estudio.

4.1.2 Estrategias multivariadas aplicadas

Los siguientes procedimientos se realizaron sobre 306 individuos, tomando al individuo como el conjunto madre-placenta-feto, que presentaron como mínimo un dato en alguna de las siete variables cuantitativas.

4.1.2.1 Estrategia de una Etapa

Gower propuso un coeficiente de similaridad de forma general para tipos mixtos de variables que además contempla las situaciones de valores faltantes. Este coeficiente aplicado sobre los individuos de la base de datos original a lo largo de sus variables cualitativas: fuma, NIA, pozo, fumih, AEB, EP; y cuantitativas: Peso, Talla, PC, Sexo, EG, PPla, AChE, CAT, simultáneamente, permitió conformar una matriz de similaridad teniendo en cuenta la coincidencia o discrepancia con que se dan las características que los describen. Se aplicó ACoP con el objetivo de llegar a una representación geométrica en dos dimensiones de estos 306 individuos que respete la estructura de similaridades definida por la matriz obtenida anteriormente.

Posteriormente, para determinar cuál o cuáles variables cualitativas permitían discriminar en distintos grupos homogéneos la nube de puntos obtenida, se pintaron los individuos de diferentes colores para cada una de las categorías. De

todas las variables cualitativas estudiadas, la Época de parto resultó la que mejor discriminó a los individuos (Figura 4.2). Quizás esto pueda responder a los resultados observados anteriormente mediante las técnicas inferenciales clásicas, donde la variable categórica que produjo diferencias altamente significativas ($p<0,001$) en los biomarcadores de placenta estudiados fue la Época de parto. El hecho de que la aplicación del plaguicida haya sido realizada en diferentes épocas del año incide en estos biomarcadores, produciendo una posterior discriminación en dos grupos. Cabe destacar que existen individuos "de transición" entre un grupo y otro, producto del carácter multivariante de la disciplina estudiada. Para el resto de las variables cualitativas, los individuos de diferentes colores se ubicaban en el gráfico de manera aleatoria, sin lograr grupos de manera efectiva. Incluso en aquellos hábitos que habían producido modificaciones en las variables antropométricas. De la misma forma se pintaron de colores diferentes a los individuos con valores mayores y menores a la media en las variables cuantitativas, no encontrándose grupos homogéneos.

La desventaja que tiene esta estrategia multivariada, es que resulta muy difícil establecer gradientes dentro de las variables cuantitativas. La siguiente figura no identifica a cada individuo por separado, ya que no resulta de importancia para el estudio y dificulta la observación del gráfico.

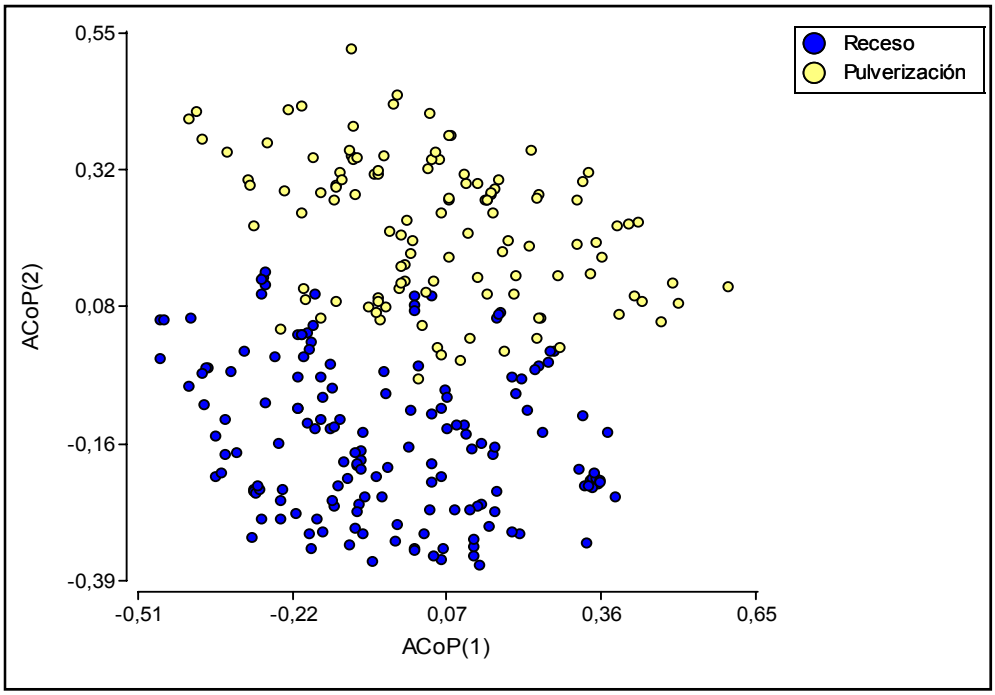


Figura 4.2: Proyección de los individuos en el plano principal, según Época de parto. Estrategia de una Etapa

4.1.2.2 Estrategia Bietápica

Para este análisis se llevó a cabo una primera etapa en búsqueda de la representación en dos dimensiones de los individuos; por un lado para las variables cualitativas, y por otro las cuantitativas. Luego, la segunda etapa consistió en consensuar o armonizar ambas configuraciones para el mismo conjunto de individuos a través de una serie de pasos iterativos.

En la primera instancia se realizó un ACoP para las siete variables categóricas: fuma, NIA, pozo, fumih, AEB, EP; y se guardaron los siete ejes resultantes. Del mismo modo para el ACP con las variables cuantitativas: Peso, Talla, PC, Sexo, EG, PPla, AChE, CAT; donde el primer eje, asociado a las variables antropométricas y peso de la placenta, retuvo el 41,75% de la variabilidad total; mientras que el segundo eje, vinculado a los biomarcadores AChE y CAT, retuvo un 21,25% de la información. Esto concuerda con lo observado en el matrix plot y la matriz de correlaciones, donde se apreciaba una relación lineal directa dentro de los parámetros antropométricos y de los biomarcadores, pero no así entre ellos.

En la segunda etapa se aplicó APG para encontrar la configuración de consenso obtenida como la media de ambas configuraciones individuales transformadas. Nuevamente la variable categórica que mejor discriminó a los individuos fue la Época de parto, no hallándose grupos discriminados para el resto de las variables cualitativas. Nuevamente se clasificó a los individuos que tengan valores superiores e inferiores a la media en las variables cuantitativas y se pintaron de diferentes colores. Al igual que en la Estrategia de una Etapa, esta categorización no produjo una discriminación efectiva en la nube de puntos.

Esta estrategia asigna igual importancia a ambas configuraciones a mezclar, por lo tanto en este caso particular se asignó igual importancia a las siete variables cualitativas respecto de las siete cuantitativas.

La configuración consenso final se representa a continuación:

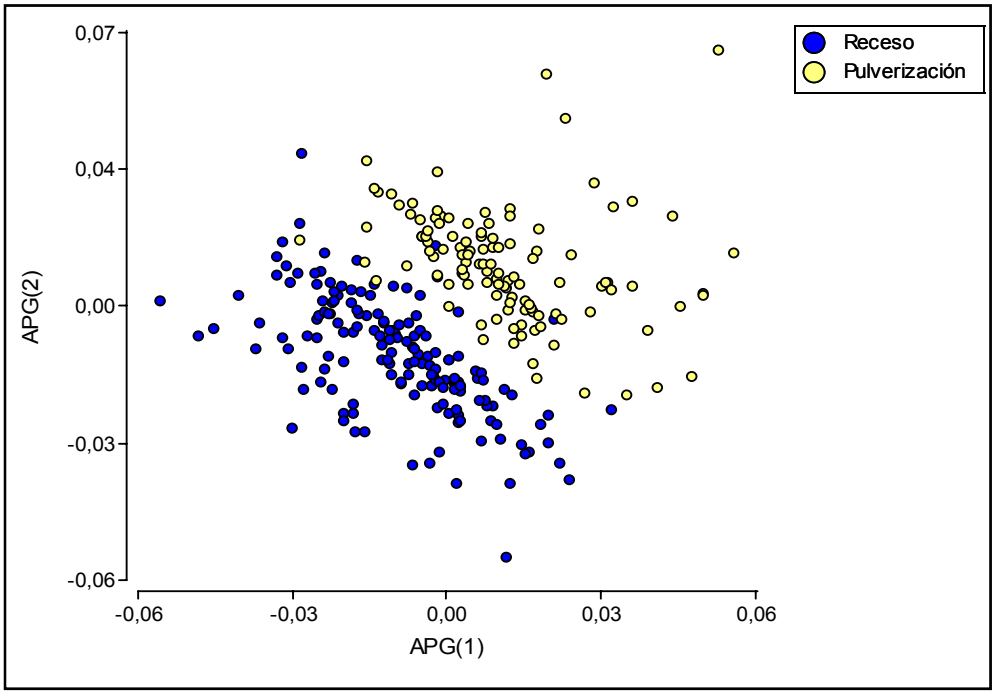


Figura 4.3: Proyección de la Configuración Consenso, según Época de parto. Estrategia Bietápica

4.1.2.3 Comparación de estrategias multivariadas

Una vez obtenidas las dos configuraciones de los individuos mediante la Estrategia de una Etapa y la Bietápica en el plano principal, se hallaron las distancias entre individuos en ambas representaciones y se utilizó el coeficiente de correlación de Pearson para cuantificar la concordancia entre ellas. El valor hallado fue $r = 0{,}2456$, indicando una baja correlación en cuanto a la información que brindan sendas caracterizaciones.

Para verificar este resultado y sus posibles implicancias, se simularon 100 matrices sin datos faltantes. Estas conservaron la estructura de correlación de la base de datos original, el número de individuos, que fue 306 y la cantidad de variables cuali y cuantitativas, que en este caso ambas fueron siete. Luego de aplicar ambas estrategias, la correlación promedio de las cien simulaciones arrojó un $r = 0{,}3707$. Se puede apreciar que la correlación sigue siendo baja, incluso trabajando sin datos faltantes. Esto se debe a que la estructura de las variables cuantitativas no tiene, en general, una asociación significativa con las variables cualitativas, como se mostró en 4.1 mediante pruebas clásicas. Luego, dentro de la Estrategia Bietápica se obtienen configuraciones muy diferentes mediante ACP para variables cuantitativas con respecto al ACoP en el caso de

las variables cualitativas. Por lo tanto, el consenso obtenido entre ambas configuraciones resultantes no posee una calidad de representación óptima, distando, lógicamente, de la representación de la Estrategia de una Etapa. Debe tenerse en cuenta que APG tiene sustento cuando, además de tener igual peso ambas configuraciones, subyace una razón experimental por la que se constituyen los grupos de variables a consensuar, situación que aquí no se presenta, siendo el único motivo de constitución de ambas configuraciones el tipo de variable involucrada.

Por otra parte, la disminución observada entre el coeficiente de correlación promedio de las simulaciones versus el empírico obtenido de la base de datos original, de 0,37 a 0,24, quizás se deba a la gran cantidad de faltantes presentes.

4.2 Simulación

Con el fin de precisar el comportamiento de las dos estrategias multivariadas ante la presencia de datos faltantes se realizaron simulaciones modificando paulatinamente los distintos factores que pudieran incidir en los resultados. Para ello se desarrolló un macro de dos etapas. La primera consistió en la creación de un generador de datos multivariados mixtos, ya que no se encuentra disponible en los paquetes estadísticos. En la segunda, se estudió la $eficacia$ y $robustez$ de las estrategias multivariados para el análisis de esos datos mixtos ante la presencia de valores faltantes.

4.2.1 **Generador de datos multivariados mixtos**

La rutina propuesta genera variables aleatorias multinomiales multivariadas, a partir de un vector de probabilidades correspondientes a la distribución conjunta. Por otra parte para cada combinación de las variables multinomiales, se crean vectores normales multivariados con esperanza especificada para cada combinación de multinomiales y matriz de varianza-covarianza común (Sabino y otros, 2008a). De esta forma pueden generarse conjuntos de datos que combinan variables categóricas y continuas, respetando las estructuras de correlación más relevantes de la base de datos original. A continuación se detalla el procedimiento para la obtención de los parámetros iniciales, posteriormente se describe el programa y por último se presenta un pequeño conjunto de datos para ejemplificar ambos procedimientos.

- ✓ ***Parámetros de entrada***

Para poder utilizar el generador de datos mixtos se necesitan tres matrices y un vector de entrada, todas ellas estimadas a partir de los datos muestrales. La matriz **Cruce** y el vector **mult** dependerán de la sumatoria de categorías de las variables cualitativas, mientras que las matrices **Medias** y **Varcov** dependerán de las variables cuantitativas. A continuación se describe cada uno de los parámetros:

Cruce: Matriz de orden $r \times k_c$, donde k_c es el número de variables cualitativas y r es el número de combinaciones posibles de las categorías de las variables cualitativas ($r = l_1 \times l_2 \times \ldots \times l_{k_c}$) donde l_i representa la cantidad de categorías de la $i-ésima$ variable. En las filas de la matriz se encuentran cada una de las posibles combinaciones de las variables multinomiales.

mult: Vector de probabilidades de dimensión r, cuyas componentes contienen la frecuencia relativa empírica de cada una de las combinaciones halladas en la matriz **Cruce**. A las combinaciones que no hayan sido observadas en la base de datos, se les asignará una probabilidad despreciable.

Medias: Matriz de orden $r \times k_q$, donde k_q es el número de variables cuantitativas. El usuario, para cada combinación de niveles de las variables categóricas, debe especificar la media de una distribución normal por cada una de las variables cuantitativas. Es decir que la $i-ésima$ fila contiene un vector de esperanzas normal multivariado correspondiente a la $i-ésima$ fila de la matriz **Cruce**. El valor m_{ij} de esta matriz quedará conformado por el promedio de la $j-ésima$ variable de aquellos individuos cuyos atributos coincidan con la $i-ésima$ combinación. En caso de que no se registren individuos con esa combinación, se les asignará un vector correspondiente al promedio de todos los individuos.

Varcov: De orden $k_q \times k_q$, es la matriz de Varianza-Covarianza empírica común a todos los individuos, utilizando el método de observaciones completas de a pares $pairwise$.

✓ ***Procedimiento para la generación de Datos Mixtos***

Una vez logrados los parámetros de entrada se continúa con los siguientes pasos hasta la obtención de la matriz de datos mixtos simulada (Apéndice 5).

Paso 1: Multiplicar el vector **mult** por un número entero t $(t \geq 1000)$ y tomar la parte entera. Esta multiplicación por un escalar tiene como objetivo transformar las componentes de un vector de probabilidades a un vector de frecuencias empíricas. Luego, se obtiene un vector $\vec{v}$ de tamaño r (r = cantidad de cruces), como sigue:

$$\vec{v} = (v_1, v_2, \dots, v_r)$$

Paso 2: Establecer un nuevo vector $\vec{u}$ con números consecutivos de 1 a r, donde el número 1 aparecerá v_1 veces; el número 2 aparecerá v_2 veces, etc. Es decir:

$$\vec{u} = \left(\underbrace{\overbrace{1, 1, 1, \dots, 1}^{v_1}, \ \overbrace{2, 2, \dots 2}^{v_2}, \dots \dots \dots \dots, \ \overbrace{r, r, \dots, r}^{v_r}}_{h} \right)$$

donde h es igual a la suma de los elementos del vector $\vec{v}$.

Paso 3: Generar un número aleatorio de una distribución uniforme $(0; h)$: Tomar su parte entera más uno y asignársela al número "a". Dentro del vector $\vec{u}$, el elemento que se encuentra en la posición a, determinará el número de combinación de variables multinomiales correspondiente.

Paso 4: Generar números aleatorios normales multivariados, para lo cual se necesita el vector de medias y la matriz de varianza-covarianza correspondiente a la combinación seleccionada.

Paso 5: Generar datos mixtos mediante la combinación de las variables cualitativas con los vectores normales simulados. Es decir, se unen los vectores de variables cualitativas y cuantitativas formando un nuevo vector de dimensión $p = k_c + k_q$.

Paso 6: Repetir este último procedimiento las n veces que el usuario lo especifique, hasta completar una matriz de $n \times p$.

Al simular 1000 matrices por cada situación especificada –5 tamaños de muestra diferentes y 8 combinaciones distintas de variables cuali y cuantitativas– se generaron 40.000 matrices de datos mixtos.

✓ ***Ejemplo***

A continuación se ejemplifica el cálculo de los parámetros de entrada a partir de un pequeño conjunto de datos La matriz consta de dos variables cualitativas. Fuma, con tres niveles: no fuma=0, fuma pasivo=1, fuma activo=2 y Sexo con 2 niveles: femenino=0, masculino=1. Además, tres variables cuantitativas del neonato: Peso en gramos, Talla y Perímetro Cefálico en centímetros. Esta información fue registrada para doce unidades experimentales, en este caso un par madre-hijo, obteniendo una matriz de 12×5. En la primer columna de la matriz se detalla el número de caso. En aquellas variables en las que hubo un faltante se registró mediante un punto. Cabe destacar que en esta base de datos no se registró ningún neonato de sexo femenino cuya madre fume activamente. Es decir, no se registraron casos que cumplan la combinación 2-0. En la última fila se encuentra el promedio de las variables cuantitativas, sin importar el hábito de fumar de la madre ni el sexo del neonato. Los individuos de iguales características cualitativas se encuentran pintadas con el mismo tono.

Tabla 4.3: Matriz de datos a partir de la cual se generan los parámetros iniciales.

Individuo	Fuma	Sexo	Peso	Talla	Perímetro Cefálico
1	0	0	3300	**.**	34
2	0	0	3350	48	33
3	0	0	3250	46	35
4	0	0	3500	49	36
5	0	0	3400	49	36
6	0	1	3300	50	35
7	0	1	3500	50	**.**
8	1	0	**.**	48	34,5
9	1	1	**.**	50	35,5
10	1	1	3350	48	37
11	1	1	3450	**.**	**.**
12	2	1	3220	48,5	35
Promedio Total			**3362**	**48,65**	**35,1**

Para hallar la matriz **Cruce**, se deben calcular todas las combinaciones posibles de las variables multinomiales. En este caso se tendrán $2 \times 3 = 6$ filas y 2 columnas.

Para el cálculo del vector **mult**, se halló la frecuencia relativa de aparición de cada combinación, es decir, el cociente entre cantidad de individuos que cumplen con las categorías de las variables cualitativas y el total observado. La combinación 0-0, no fuma y femenino, se presenta 2 veces en un total de 12. Por lo tanto, $f_r = \frac{2}{12} \cong 0{,}166666$. Teniendo en cuenta que la combinación 2-0 no fue registrada en la base de datos, se le asignará una probabilidad despreciable. La sumatoria de las componentes del vector **mult** debe ser igual a uno.

Para hallar la matriz **Medias** del ejemplo se promediaron los pesos, tallas y perímetros cefálicos de los neonatos de los 2 individuos correspondientes a la combinación 0-0, obteniendo el vector multivariado $\vec{v}_{00} = (3400; 50; 35)$. En el caso de que una combinación no tenga información sobre alguna de las variables cuantitativas, se le asignará el promedio de todos los individuos para esa variable. De esta forma a la combinación 1-0 se le asignará el vector multivariado $\vec{v}_{10} = (3362; 48; 34{,}5)$. Análogamente para la combinación 2-0 quedará asignado $\vec{v}_{22} = (3362; 48{,}65; 35{,}1)$. En la Tabla 4.4 se presenta la información conjunta de las matrices **Cruce** y **Medias**, como así también el vector **mult**.

Varcov es la matriz de Varianza-Covarianza empírica común a los 12 individuos estudiados, como se muestra en la Tabla 4.5.

Tabla 4.4: Matriz de entrada necesaria para generar datos mixtos.

	Cruce		**mult**	**Medias**		
Nº	Fuma	Sexo	Probabilidades	Peso	Talla	Perímetro Cefálico
1	0	0	0,416666	3360	48,00	34,80
2	0	1	0,166666	3400	50,00	35,00
3	1	0	0,083333	3362	48,00	34,50
4	1	1	0,250000	3400	49,00	36,25
5	2	0	0,000002	3362	48,65	35,10
6	2	1	0,083333	3220	48,50	35,00

Tabla 4.5: Matriz de Varianza-Covarianza común a todos los cruces.

Varcov			
	Peso	**Talla**	**Perímetro Cefálico**
Peso	9773,333	70,804	38,036
Talla	70,804	1,558	0,250
Perímetro Cefálico	38,036	0,250	1,267

Una vez finalizada la etapa inicial referente a la conformación de parámetros iniciales, se continúa con la etapa de la generación de datos mixtos para los datos del ejemplo:

Paso 1: Al multiplicar el vector **mult** por un número entero $t \geq 1000$ –en este caso $t = 1000000$– y tomar la parte entera, se obtiene el vector $\vec{v}$ de longitud 6:

$$\vec{v} = (416666, 166666, 83333, 250000, 2, 83333)$$

Paso 2: El nuevo vector $\vec{u}$ contará con números consecutivos del 1 al 6, donde el número 1 aparecerá 416666 veces; el número 2 aparecerá 166666 veces, etc. Es decir:

$$\vec{u} = \left(\underbrace{\overbrace{1, 1, 1, \dots, 1}^{416666}, \overbrace{2, 2, \dots 2}^{166666}, \dots\dots\dots\dots, \overbrace{6, 6, \dots, 6}^{83333}}_{1000000} \right)$$

Paso 3: Se genera un número aleatorio de la distribución uniforme $(0; 1000000)$ y resulta ser el $989025{,}2$. Luego se debe seleccionar el elemento que se encuentra en la posición 989026 del vector $\vec{u}$, que en este caso resulta ser 6. La combinación de multinomiales que se encuentra en la sexta fila de las columnas **Cruce** (Tabla 4.4) es la composición 2-1, Fuma activo y sexo masculino.

Paso 4: El vector normal multivariado generado tendrá como esperanza el vector que se encuentra en la sexta fila de la matriz **Medias**

$$(3220; 48{,}5; 35)$$

correspondiente a peso, talla y perímetro cefálico promedio de los neonatos varones, cuyas madres fuman. En tanto que la matriz **Varcov** es siempre la misma, no importa cuál sea la combinación de multinomiales seleccionada. Un vector generado posible es $(3100{,}061;\ 48{,}41239;\ 33{,}35477)$.

Paso 5: En consecuencia, el vector de datos mixtos generados sería:

$$(2; 1;\ 3100{,}061;\ 48{,}41239;\ 33{,}35477)$$

Paso 6: Tomando $n = 10$, quedaría determinada una posible matriz de datos mixtos como sigue:

Tabla 4.6: Matriz simulada.

Individuo	Fuma	Sexo	Peso	Talla	Perímetro Cefálico
1	2	1	3100	48,41	33,35
2	0	0	3310	46,06	35,64
3	1	1	3393	49,56	35,99
4	1	1	3460	49,80	34,76
5	1	0	3266	47,59	35,15
6	1	1	3409	48,02	36,58
7	0	0	3472	47,88	35,64
8	0	1	3409	49,73	36,36
9	0	0	3236	45,99	34,76
10	0	1	3476	48,63	34,45

4.2.2 Eficacia y Robustez de métodos alternativos multivariados para el análisis de datos mixtos ante la presencia de valores faltantes

En este apartado se estudiaron por separado dos comportamientos complementarios de las estrategias multivariadas utilizadas: por un lado, calidad de representación ante la presencia de los datos faltantes o *eficacia.* Y por otro su aplicabilidad, de aquí en más *robustez*.

En el Apéndice 6 se encuentran las sentencias necesarias en R para calcular las distancias entre individuos de las configuraciones resultantes en ambas estrategias. El Apéndice 7 detalla el algoritmo para la extracción de datos y posterior cálculo de correlaciones entre las distancias de las configuraciones originales y su respectiva con faltantes.

4.2.2.1 Calidad de representación

En total se generaron 40.000 matrices de **datos completos**, resultantes de las 1000 simulaciones correspondientes al producto cantidad diferente de individuos y número de combinaciones de variables cuali y cuantitativas, que en este caso fueron cinco y ocho, respectivamente (Tabla 3.2). Posteriormente se extrajeron de forma gradual datos de esas matrices completas generando siete proporciones distintas, lo que implicó que se conformen 280.000 matrices con **datos faltantes**. El cálculo de la correlación de la distancia entre individuos para cada una de las matrices con faltantes con su respectiva distancia original

generó 280.000 correlaciones por estrategia. Al estudiar dos estrategias, EduE y EB, se concluyó con un total de 560.000 correlaciones.

Los siguientes gráficos, permitieron comparar la calidad de las representaciones de las distintas matrices mediante el promedio obtenido en los coeficientes de correlación de Pearson entre configuraciones con faltantes y las completas u originales para cada una de las estrategias (Sabino y otros, 2008b). Cada punto indica el promedio. En las referencias de los gráficos se encuentran en siglas las 2 estrategias empleadas – Estrategia de una Etapa y Estrategia Bietápica –, como así también se indica entre paréntesis el número de individuos utilizados.

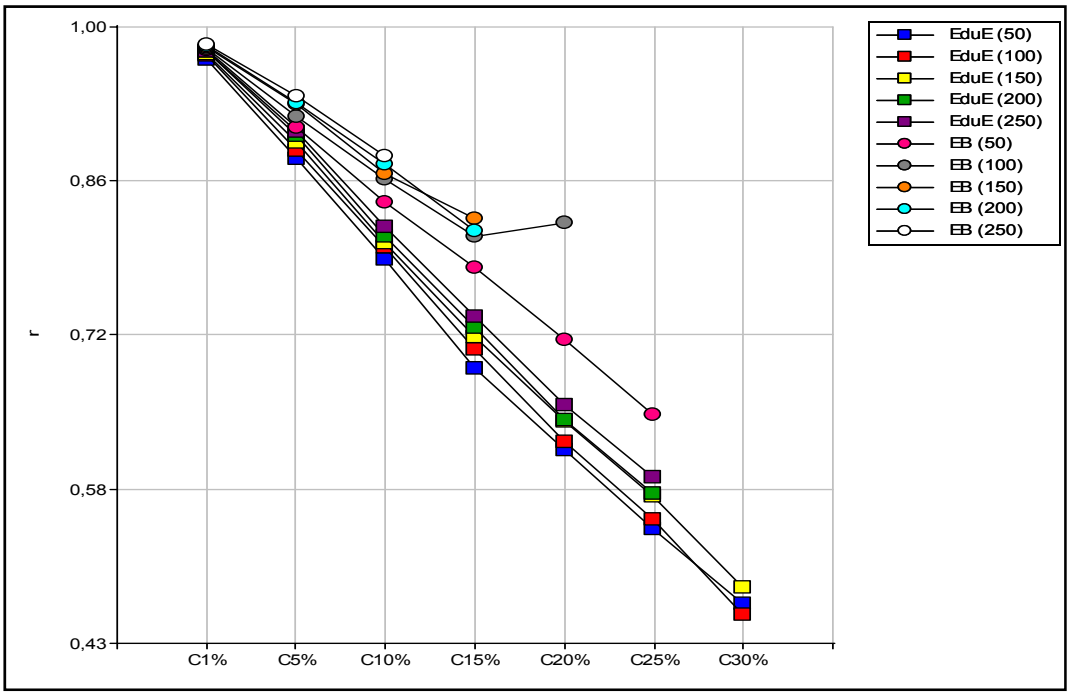


Figura 4.4: Cada punto corresponde al promedio de mil coeficientes de correlación de Pearson entre configuraciones con datos faltantes y originales para **5 variables Cualitativas y 5 Cuantitativas**. Se utilizaron diferentes proporciones de datos faltantes, según Estrategia de una Etapa (EduE) y Estrategia Bietápica (EB). El número de individuos utilizados se encuentra entre paréntesis.

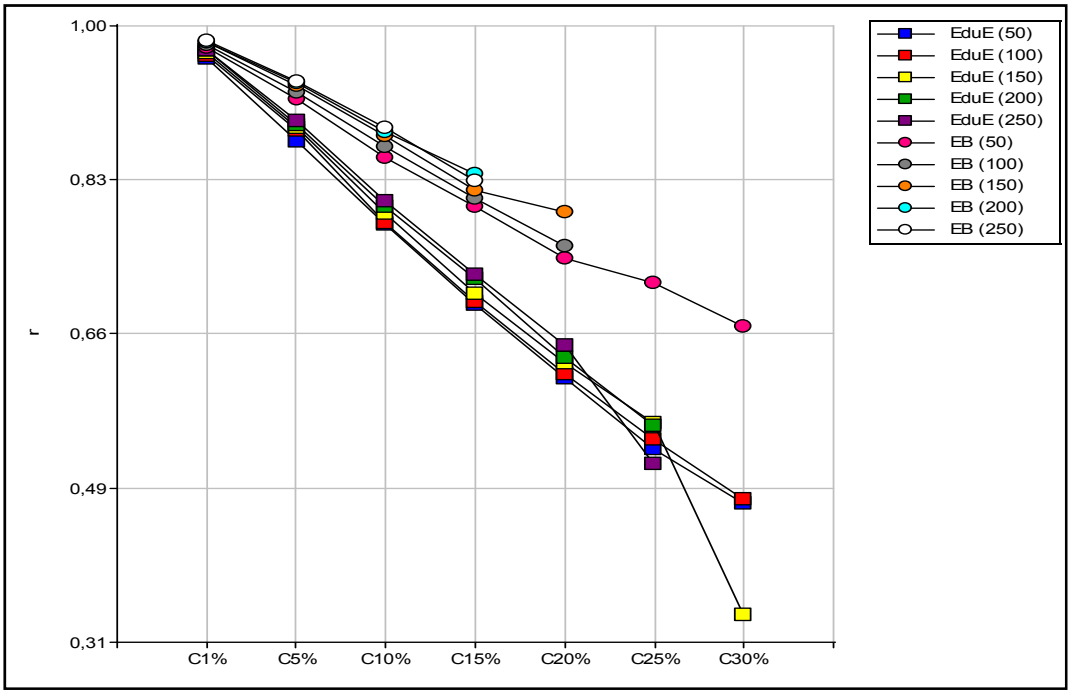


Figura 4.5: Cada punto corresponde al promedio de mil coeficientes de correlación de Pearson entre configuraciones con datos faltantes y originales para **6 variables Cualitativas y 4 Cuantitativas**. Se utilizaron diferentes proporciones de datos faltantes, según Estrategia de una Etapa (EduE) y Estrategia Bietápica (EB). El número de individuos utilizados se encuentra entre paréntesis.

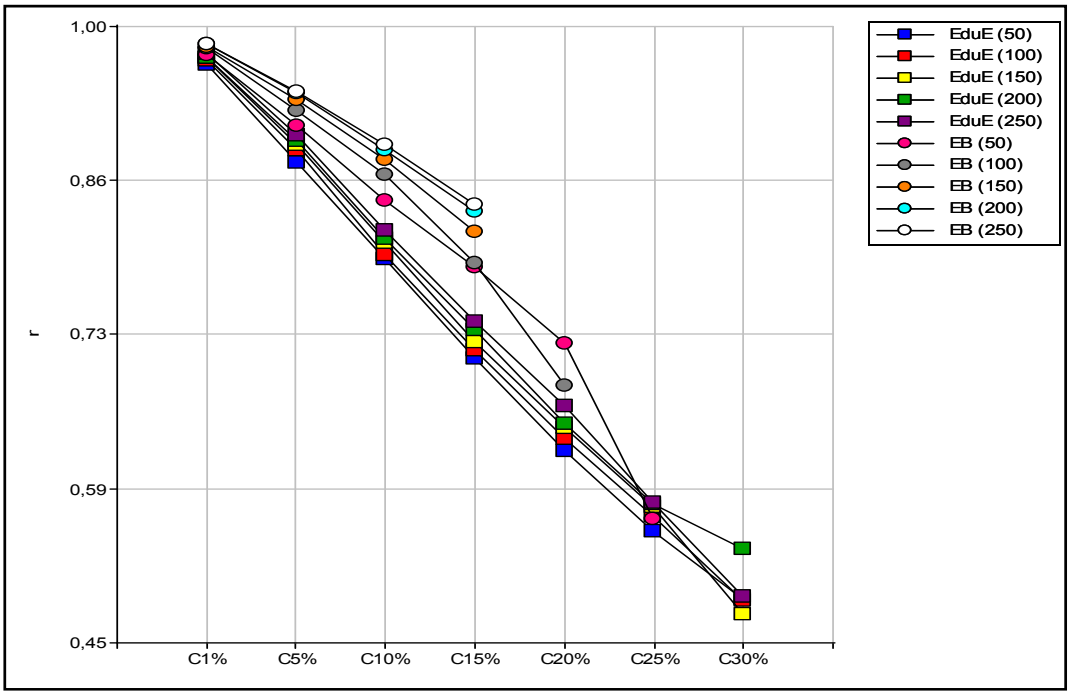


Figura 4.6: Cada punto corresponde al promedio de mil coeficientes de correlación de Pearson entre configuraciones con datos faltantes y originales para **5 variables Cualitativas y 6 Cuantitativas**. Se utilizaron diferentes proporciones de datos faltantes, según Estrategia de una Etapa (EduE) y Estrategia Bietápica (EB). El número de individuos utilizados se encuentra entre paréntesis.

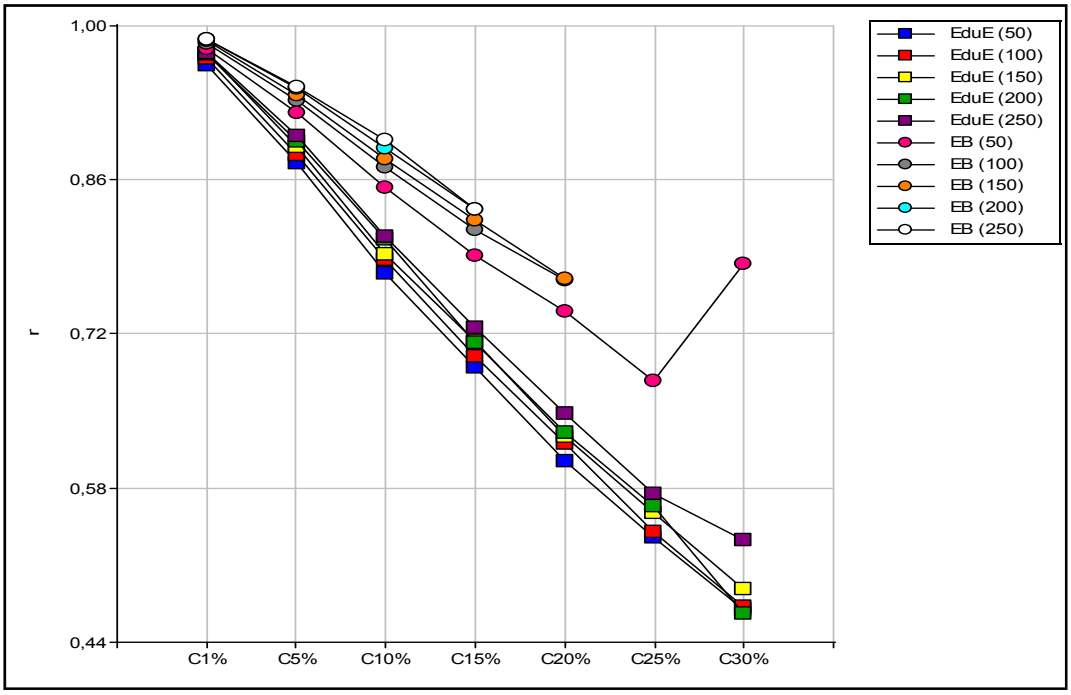


Figura 4.7: Cada punto corresponde al promedio de mil coeficientes de correlación de Pearson entre configuraciones con datos faltantes y originales para **6 variables Cualitativas y 5 Cuantitativas**. Se utilizaron diferentes proporciones de datos faltantes, según Estrategia de una Etapa (EduE) y Estrategia Bietápica (EB). El número de individuos utilizados se encuentra entre paréntesis.

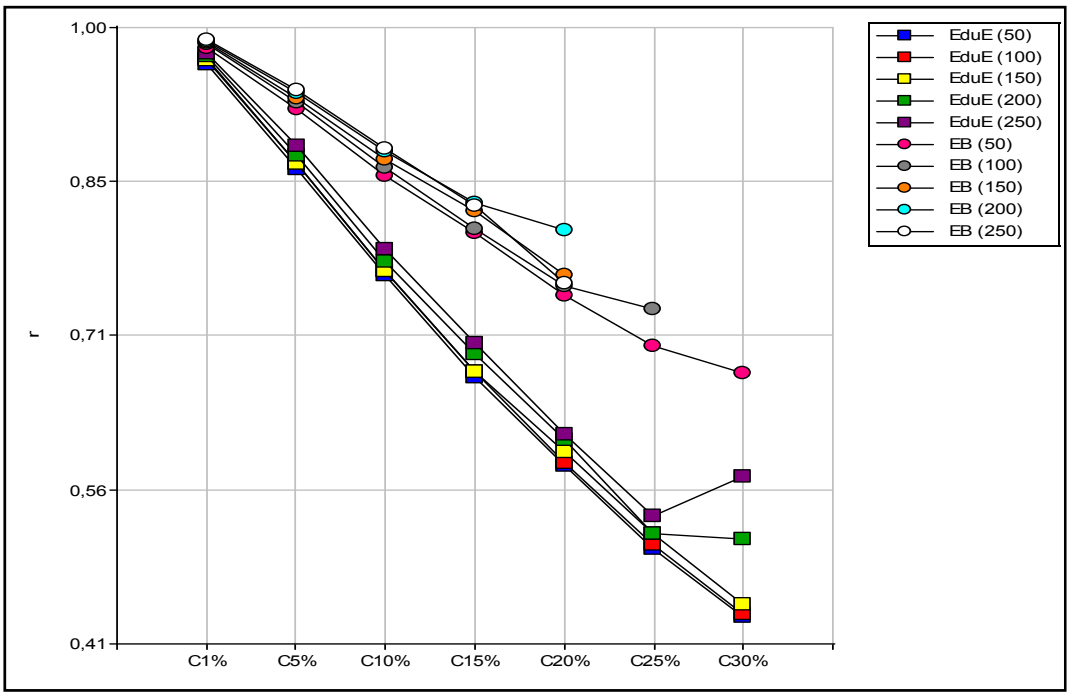


Figura 4.8: Cada punto corresponde al promedio de mil coeficientes de correlación de Pearson entre configuraciones con datos faltantes y originales para **7 variables Cualitativas y 4 Cuantitativas**. Se utilizaron diferentes proporciones de datos faltantes, según Estrategia de una Etapa (EduE) y Estrategia Bietápica (EB). El número de individuos utilizados se encuentra entre paréntesis.

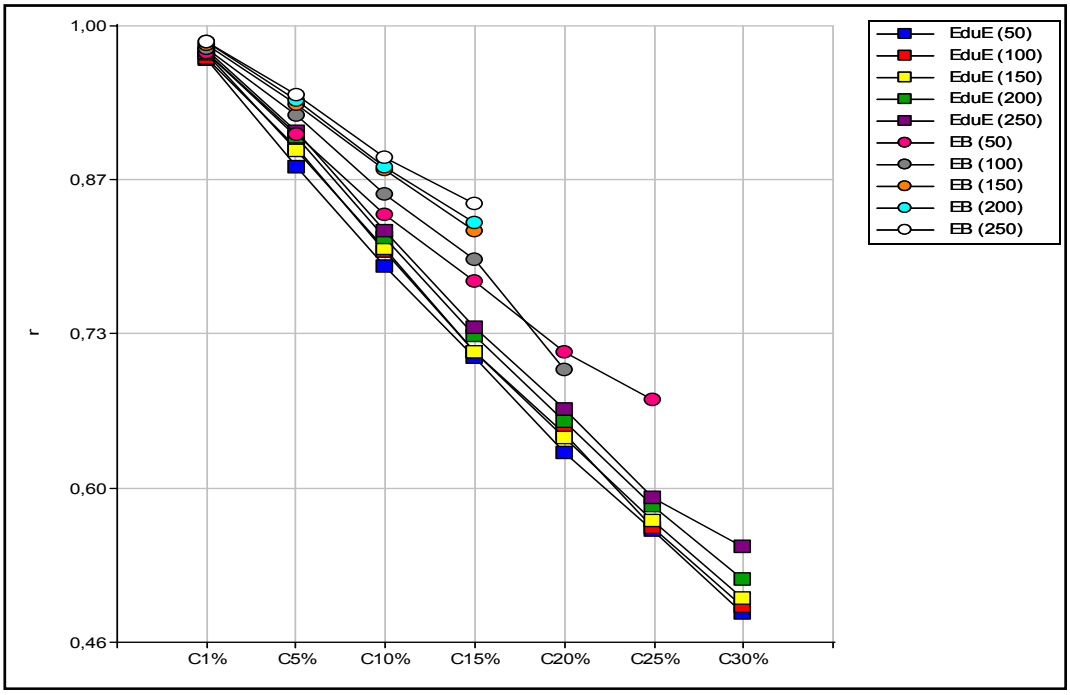


Figura 4.9: Cada punto corresponde al promedio de mil coeficientes de correlación de Pearson entre configuraciones con datos faltantes y originales para **5 variables Cualitativas y 7 Cuantitativas**. Se utilizaron diferentes proporciones de datos faltantes, según Estrategia de una Etapa (EduE) y Estrategia Bietápica (EB). El número de individuos utilizados se encuentra entre paréntesis.

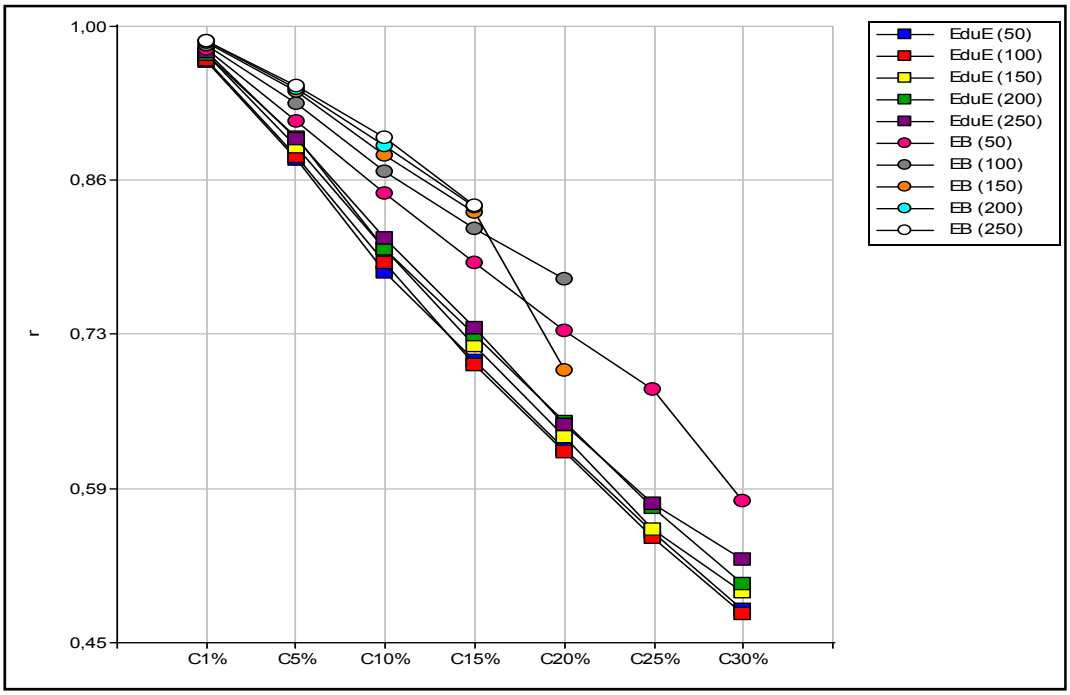


Figura 4.10: Cada punto corresponde al promedio de mil coeficientes de correlación de Pearson entre configuraciones con datos faltantes y originales para **6 variables Cualitativas y 6 Cuantitativas**. Se utilizaron diferentes proporciones de datos faltantes, según Estrategia de una Etapa (EduE) y Estrategia Bietápica (EB). El número de individuos utilizados se encuentra entre paréntesis.

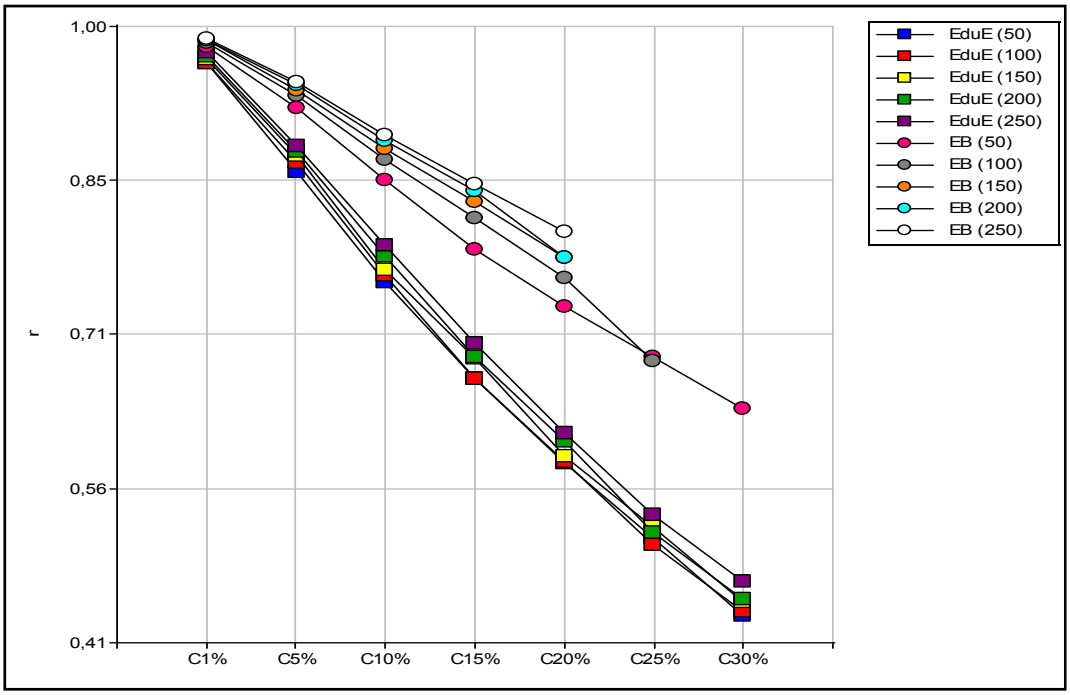


Figura 4.11: Cada punto corresponde al promedio de mil coeficientes de correlación de Pearson entre configuraciones con datos faltantes y originales para **7 variables Cualitativas y 5 Cuantitativas**. Se utilizaron diferentes proporciones de datos faltantes, según Estrategia de una Etapa (EduE) y Estrategia Bietápica (EB). El número de individuos utilizados se encuentra entre paréntesis.

La combinación de variables utilizadas no alteró la calidad de representación. Se pudo apreciar un comportamiento lineal decreciente en los promedios de los coeficientes de Pearson a medida que se incrementaron los faltantes. Tanto la Estrategia de una Etapa como la Bietápica no se vieron afectadas cuando sólo falta el 1% de los datos, pero se observaron cambios a medida que se incrementó la cantidad de faltantes.

La extracción gradual de datos repercutió del mismo modo pero con distinta intensidad en las dos técnicas empleadas. Si bien ambas técnicas decayeron en su calidad de representación a medida que la proporción de datos faltantes aumentó, ésta fue sensiblemente menor cuando se utilizó la Estrategia de una Etapa.

Considerando como calidad de representación “aceptable” aquella cuya correlación fuera de aproximadamente 0,70 o superior, se pudo determinar que el porcentaje de faltantes tolerable para realizar la Estrategia de una Etapa fue del 15%, mientras que para la Bietápica fue del 20 al 25%.

4.2.2.2 Aplicabilidad de las estrategias

La extracción gradual de observaciones en las bases simuladas manifestó una particularidad en cuanto al número de matrices de distancia en las cuales no se pudo establecer al menos la relación entre un par de individuos (matrices de distancia con algún elemento indeterminado), imposibilitando llevar adelante el desarrollo de las estrategias de Análisis Multivariado correspondiente (Sabino y otros, 2009).

A continuación se desarrolla un ejemplo para comprender el problema:

Se supone una matriz de dimensiones $n \times p$, donde n representa los individuos y p las variables. Se pueden calcular $\binom{n}{2} = \frac{n.(n-1)}{2}$ distancias entre individuos.

Tomando una matriz de 5×4, donde la aparición del número 1 quiere decir que allí hay un valor o dato válido, se pueden computar 10 distancias.

$$\text{INDIVIDUOS}\left\{\begin{array}{c} \overbrace{\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}}^{\text{VARIABLES}} \end{array}\right.$$

A través de la simulación se extraen datos de forma aleatoria —representando datos faltantes— y se quiere determinar cuándo no se puede calcular, como mínimo, una de esas distancias.

Es decir, si el porcentaje de faltantes (q) es del 30%, la cantidad de datos faltantes ($f = n.p.q$) es igual a 6. Entonces una matriz posible sería:

$$A = \begin{pmatrix} \blacksquare & 1 & 1 & \blacksquare \\ 1 & \blacksquare & \blacksquare & 1 \\ 1 & 1 & \blacksquare & 1 \\ 1 & 1 & 1 & 1 \\ 1 & \blacksquare & 1 & 1 \end{pmatrix}$$

Donde ■ representa un dato faltante. En este caso se computan 9 distancias ya que no es posible hallar la distancia entre el individuo 1 y 2, dado que se necesita que en al menos una variable, los 2 individuos tengan el valor 1.

Otro ejemplo:

$$B = \begin{pmatrix} ■ & 1 & 1 & ■ \\ 1 & ■ & 1 & 1 \\ 1 & 1 & ■ & 1 \\ ■ & 1 & 1 & 1 \\ 1 & ■ & 1 & 1 \end{pmatrix}$$

En cambio, aquí sí se pueden calcular todas las distancias, ya que en al menos una variable todo par de individuos tiene datos válidos.

Se llamará **matriz NA** $(not\ available)$ cuando no se pueda calcular, como mínimo, una distancia entre dos individuos. Las simulaciones permiten describir el comportamiento de estas matrices, encontrando la estrategia multivariada más $robusta$ para este problema.

Es importante destacar que al utilizar la Estrategia de una Etapa, el número de variables utilizadas para el cálculo de matrices de similaridad es igual a la suma de variables cualitativas y cuantitativas $(k_c + k_q)$.

En cambio, al aplicar la EB, se calcula una matriz de distancia o similaridad para cada tipo de variable. Cuando las variables son cuantitativas, si existen faltantes en la base de datos al momento de aplicar ACP, los individuos habitualmente se descartan o se completan mediante imputación. La eliminación de filas es una solución del problema que escapa al presente trabajo, por lo que se completaron las celdas vacías mediante imputación por la media. En consecuencia, el número de variables utilizadas para estudiar el comportamiento de las matrices NA cuando se aplica la EB, quedó definido por el número de variables cualitativas (k_c) .

Sintetizando, la Tabla 4.7 presenta los 8 casos estudiados y el número de variables utilizadas para cada estrategia:

Tabla 4.7: Número de variables utilizadas para determinar la probabilidad de hallar matrices NA según estrategia utilizada.

Cantidad de variables (p)	**Tipos de variables**	**Utilizadas en EduE $(k_c + k_q)$**	**Utilizadas en EB (k_c)**
10	5 cualitativas – 5 cuantitativas	10	5
	6 cualitativas – 4 cuantitativas	10	6
11	5 cualitativas – 6 cuantitativas	11	5
	6 cualitativas – 5 cuantitativas	11	6
	7 cualitativas – 4 cuantitativas	11	7
12	5 cualitativas – 7 cuantitativas	12	5
	6 cualitativas – 6 cuantitativas	12	6
	7 cualitativas – 5 cuantitativas	12	7

Por otro lado, en la Tabla 4.8 se encuentran tabulados los porcentajes de matrices NA según cantidad de individuos $(n = 50, 100, 150, 200, 250)$, número de variables consideradas en cada estrategia $(p = 10, 11, 12;\ k_c = 5, 6, 7)$ y porcentaje de faltantes $(q = 1, 5, 10, 15, 20, 25, 30)$. Estos resultados corresponden a aquellas 280.000 matrices simuladas a las que no fue posible aplicar una estrategia.

En principio existen dos resultados que no contradicen la intuición: el primero es que a medida que se incrementa el número de faltantes f, la probabilidad de que no se pueda calcular una distancia entre dos individuos, como mínimo, también aumenta. Asimismo, la probabilidad de hallar una matriz NA disminuirá en tanto sea mayor el número de variables, a número de individuos constante.

Sin embargo, cuanto mayor es el número de individuos n, mayor es la probabilidad de hallar matrices de distancia con algún elemento indeterminado. Si bien parece contrario a la percepción, este resultado se sustenta en que también aumenta el número de faltantes f, posibilitando que un mayor número de individuos comparables no obtengan una distancia entre sí.

Las situaciones planteadas no son fácilmente apreciables en la Tabla 4.8. Por este motivo, se presentan los gráficos correspondientes al porcentaje de matrices NA para cada situación, dependiendo del número de individuos, porcentaje de faltantes y número de variables utilizadas para calcular matrices de similaridad.

Tabla 4.8: Porcentaje de matrices NA según número de individuos, cantidad de variables utilizadas y proporción de faltantes.

		Porcentaje de faltantes q						
n	**k**	**1%**	**5%**	**10%**	**15%**	**20%**	**25%**	**30%**
50	**5**	0,00	0,43	12,93	53,46	88,90	99,20	99,96
	6	0,00	0,03	3,13	24,00	65,40	93,83	99,66
	7	0,00	0,00	0,56	8,60	37,80	79,10	97,66
	10	0,00	0,00	0,00	0,15	3,45	17,65	54,80
	11	0,00	0,00	0,00	0,06	0,80	8,33	35,60
	12	0,00	0,00	0,00	0,00	0,40	3,50	22,30
100	**5**	0,00	2,33	38,10	88,96	99,66	100,00	100,00
	6	0,00	0,13	12,03	59,76	95,40	99,93	100,00
	7	0,00	0,03	2,60	27,03	78,33	98,73	100,00
	10	0,00	0,00	0,00	1,05	10,70	49,90	93,90
	11	0,00	0,00	0,00	0,20	4,33	27,83	78,80
	12	0,00	0,00	0,00	0,07	1,90	14,52	58,27
150	**5**	0,00	4,53	57,53	97,46	100,00	100,00	100,00
	6	0,00	0,70	22,00	80,63	99,70	100,00	100,00
	7	0,00	0,06	5,86	45,76	94,23	100,00	100,00
	10	0,00	0,00	0,05	2,00	21,90	76,10	99,45
	11	0,00	0,00	0,00	0,66	8,73	50,10	95,13
	12	0,00	0,00	0,00	0,15	3,77	28,75	83,50
200	**5**	0,03	7,73	73,90	99,53	100,00	100,00	100,00
	6	0,00	1,60	32,00	90,50	99,96	100,00	100,00
	7	0,00	0,06	9,63	64,00	98,56	100,00	100,00
	10	0,00	0,00	0,20	3,65	34,55	89,50	100,00
	11	0,00	0,00	0,00	1,33	15,60	69,16	99,20
	12	0,00	0,00	0,00	0,45	7,20	44,80	94,50
250	**5**	0,00	11,83	85,43	99,86	100,00	100,00	100,00
	6	0,00	1,46	43,26	96,13	100,00	100,00	100,00
	7	0,00	0,10	12,60	77,40	99,85	100,00	100,00
	10	0,00	0,00	0,25	5,70	45,15	96,75	100,00
	11	0,00	0,00	0,03	1,63	23,76	82,13	99,96
	12	0,00	0,00	0,00	0,46	9,80	58,76	93,10

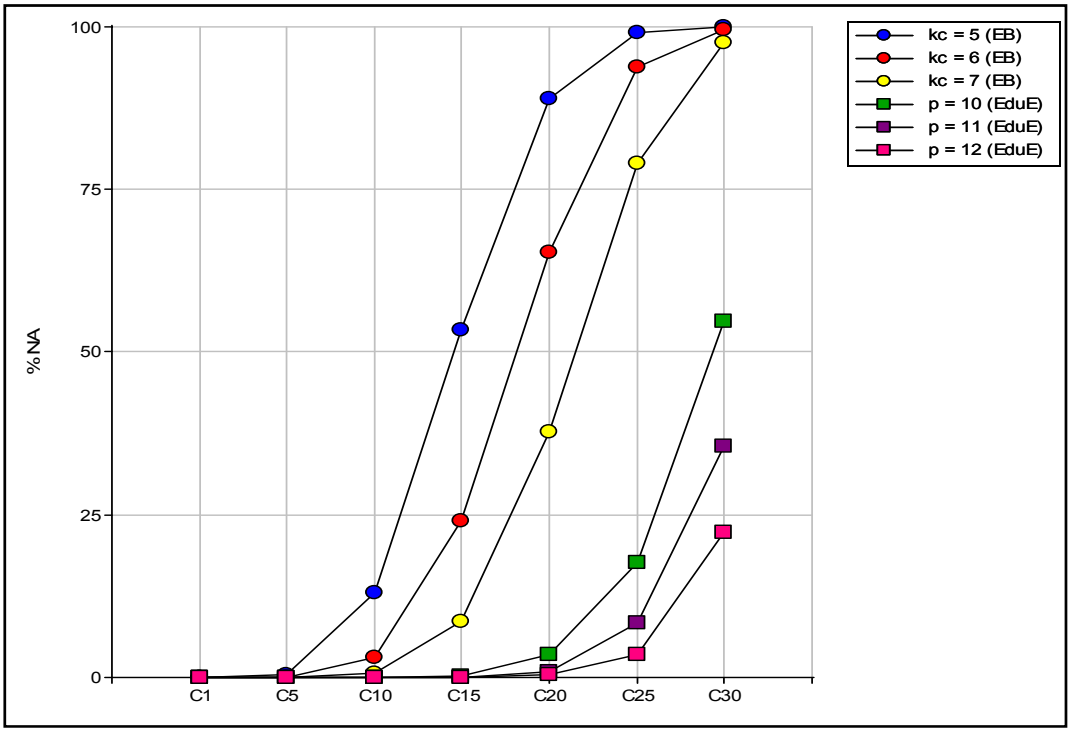


Figura 4.12: Porcentaje de Matrices NA, según porcentaje de faltantes y número de variables (n=50)

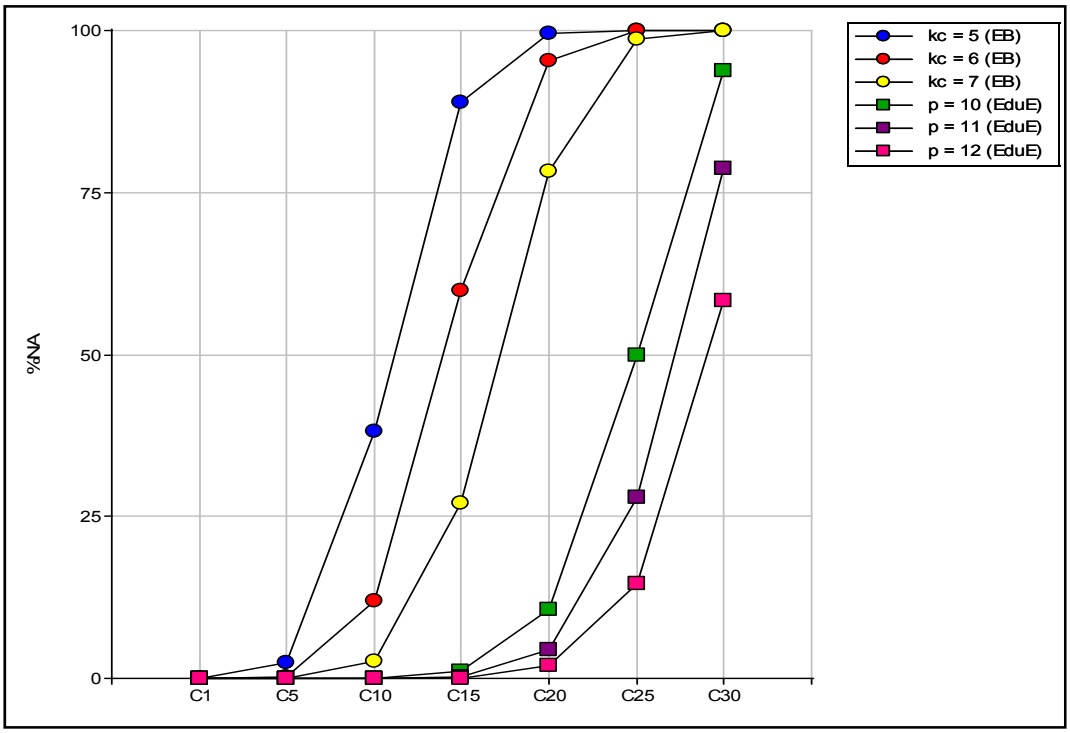


Figura 4.13: Porcentaje de Matrices NA, según porcentaje de faltantes y número de variables(n=100)

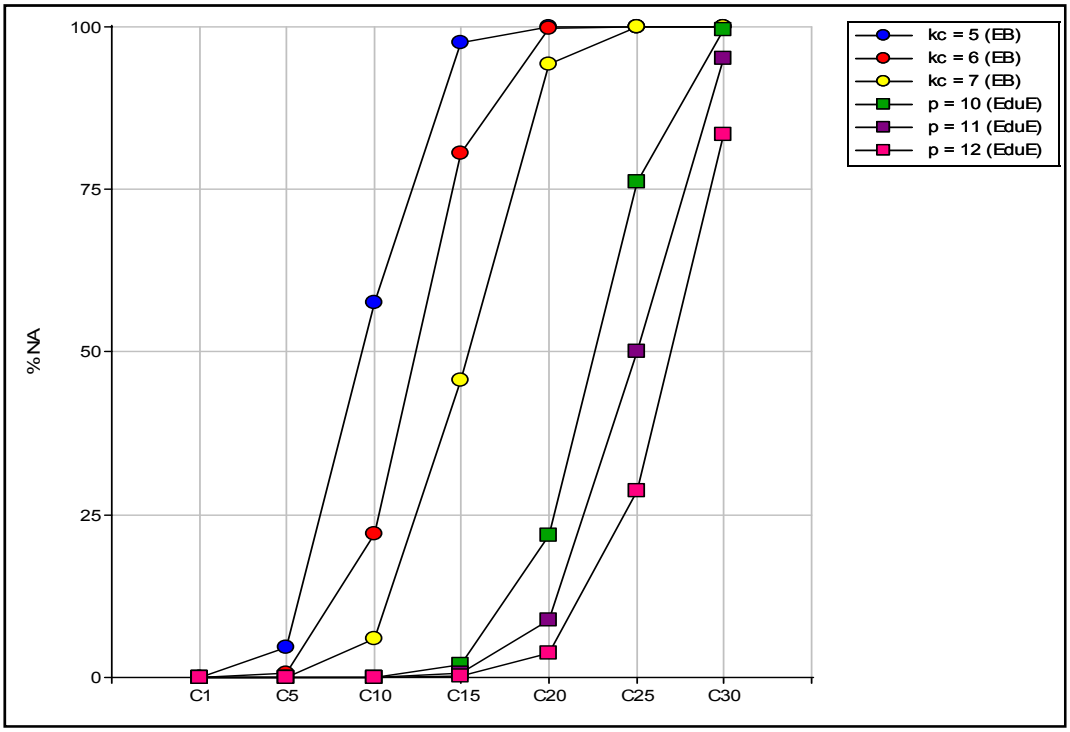


Figura 4.14: Porcentaje de Matrices NA, según porcentaje de faltantes y número de variables (n=150)

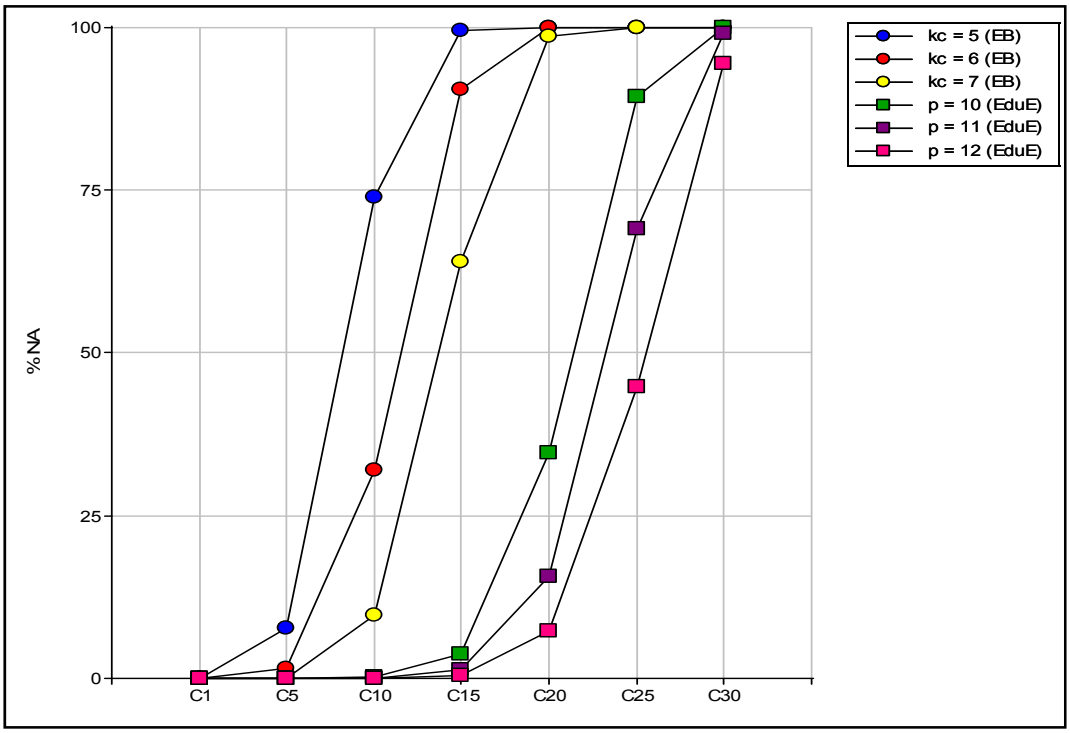


Figura 4.15: Porcentaje de Matrices NA, según porcentaje de faltantes y número de variables (n=200)

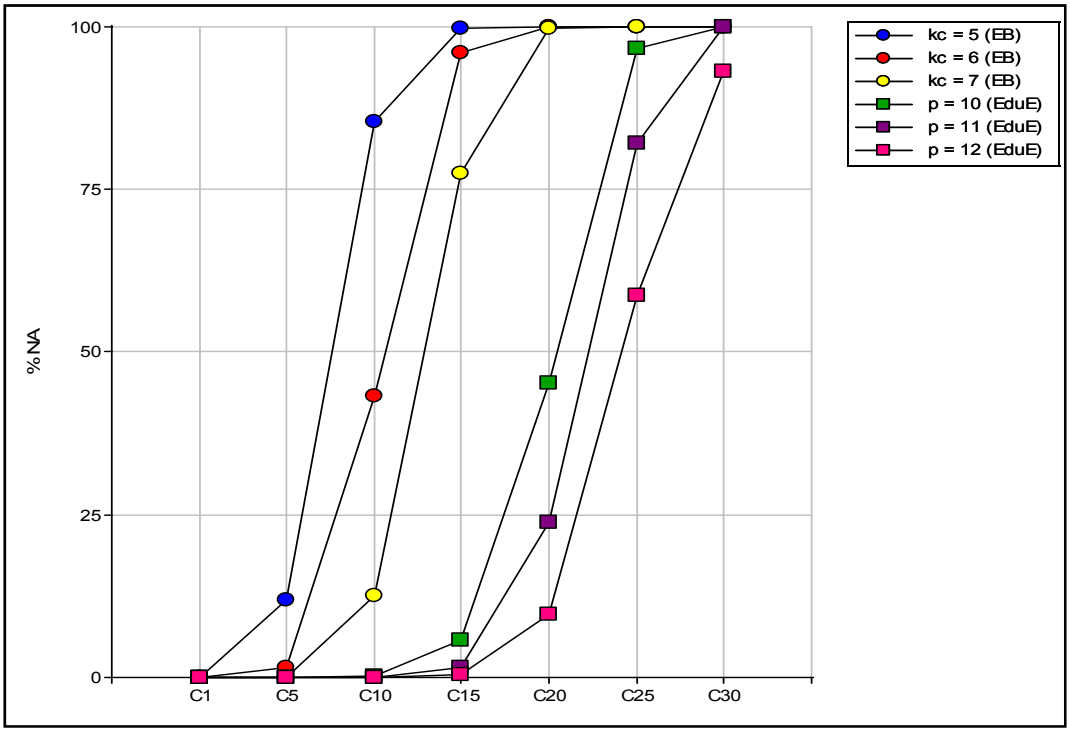


Figura 4.16: Porcentaje de Matrices NA, según porcentaje de faltantes y número de variables (n=250)

Como se mencionó anteriormente, la probabilidad de obtener matrices de distancia con algún elemento indeterminado depende del número de individuos (n), del número de variables (p), y porcentaje de faltantes (q). Dado que a lo largo de las simulaciones se observó un comportamiento particular en cuanto al número de matrices NA, se prosiguió con el desarrollo de una fórmula general a fin de estimar la proporción de matrices de distancia con elementos indeterminados para cualquier $n, p\ y\ q$. El desarrollo teórico de este comportamiento permitió no sólo conocer de forma aproximada la probabilidad de aplicar exitosamente una estrategia para una base de datos particular, sino que también permitió controlar el correcto funcionamiento del simulador.

Desarrollo teórico:

Una vez que los valores n, p y q ya están fijos, y en consecuencia $f = n.p.q$, **el problema consiste en hallar la probabilidad de que una matriz sea NA**.

Pensando en la definición clásica de probabilidad:

$$P(NA) = \frac{casos\ favorables}{casos\ posibles} = \frac{cantidad\ de\ matrices\ NA}{cantidad\ de\ matrices\ posibles}$$

Existen $\binom{np}{f} = \frac{np!}{(np-f)!f!}$ casos posibles, restando calcular la cantidad de matrices NA, es decir, los casos favorables.

Es lógico pensar en distintas expresiones para la probabilidad buscada, dependiendo de los valores que asuma f (cantidad de faltantes).

A continuación se expresa una función por partes de la probabilidad de obtener una matriz NA:

$$P(NA) = \begin{cases} 0 & si\ f < p \\ n + \binom{n}{2}(2^p - 2) & si\ f = p \\ ¿? & si\ p < f \leq n(p-1) \\ 1 & si\ f > n(p-1) \end{cases}$$

Es decir:

- Si la cantidad de faltantes es inferior al número de variables, es imposible que no se pueda calcular distancia entre dos individuos. Esto se debe a que en por lo menos una variable, ambos tienen información presente.
- Si $f = p$ entonces se pueden dar dos situaciones 1) que todos los faltantes se encuentren en el mismo individuo $(n\ casos)$; 2) que los faltantes se encuentren distribuidos sólo en dos individuos de forma tal que en todas las variables se encuentre un faltante para alguno de ellos $\left(\binom{n}{2}(2^p - 2)\ casos\right)$.
- El caso en que $f = n(p-1)$, es el máximo número de faltantes posible previo a que una matriz sea necesariamente NA. Esto viene dado por el caso particular en que $p-1$ variables tienen datos faltantes en todas sus filas y la última columna tiene información en todas sus filas. Al incrementar en uno el número de faltantes una de las filas se anula, convirtiéndose en matriz NA.
- Ante la dificultad de encontrar una función exacta que exprese el número de matrices NA cuando $p < f \leq n(p-1)$, se trabajó en fórmulas para acotar dicha función que permitieron evaluar con mayor profundidad el proceso de simulación y descubrir el umbral de faltantes tolerable en diferentes situaciones empíricas.

En este sentido, se tuvo en cuenta la probabilidad que tiene un elemento cualquiera de ser borrado. En la primer extracción se determinó que esa probabilidad era igual a $f/(n.p) = q$. A medida que se eliminen más elementos,

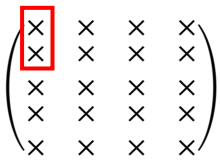
dicha probabilidad irá disminuyendo. Es decir, al quitar el segundo elemento, aquella es $(f-1)/(n.p-1) < q$, el tercer elemento tendrá una probabilidad de ser seleccionado de $(f-2)/(n.p-2) < q$ y así continuando hasta la extracción del último elemento posible, el cual tendrá una probabilidad de $1/(n.p-f+1)$ de ser seleccionado. En estas condiciones se consideró que si la probabilidad de eliminar un elemento en cualquier momento es constante e igual a q, se podría hallar una cota superior, ya que la probabilidad de tener datos faltantes es mayor, como así también la de hallar una matriz NA. Ahora bien, si aquella se mantiene constante, las extracciones son independientes, por lo que se puede hallar una fórmula mediante el desarrollo de variables con distribución binomial.

Desarrollo de la cota superior

Para una variable en los dos primeros individuos, podemos definir:

x: $cantidad\ de\ datos\ faltantes\ en\ \boldsymbol{una}\ variable\ para\ \boldsymbol{dos}\ individuos.$

Entonces $x \sim B_i(2; q)$ Gráficamente:



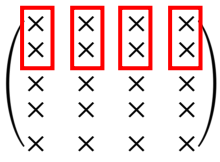
Luego, la probabilidad de calcular una distancia entre dos individuos con una variable o que no haya faltantes en ninguno de los dos individuos es:

$$P(x = 0) = (1-q)^2$$

Por lo tanto, la probabilidad de obtener al menos un dato faltante en una variable para dos individuos es igual a $1-(1-q)^2$.

Si esta situación se repite a lo largo de todas las variables de los dos individuos, puede definirse entonces:

y: $cantidad\ de\ variables\ con\ al\ menos\ un\ dato\ faltante\ (entre\ dos\ individuos)$



$y \sim B_i(p; 1-(1-q)^2)$

Para que no sea posible calcular distancia entre dos individuos a lo largo de todas las variables, esta situación debe repetirse p veces.

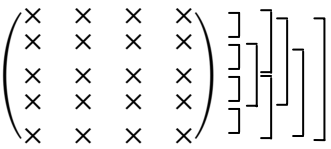
$$P(y = p) = [1 - (1 - q)^2]^p$$

Así, la probabilidad de hallar distancia entre 2 individuos es:

$$1 - [1 - (1 - q)^2]^p$$

Para poder hallar la cota superior para la probabilidad de matriz NA, se computa como primer paso que no lo sea. Para ello, se debe calcular la distancia para todo par de individuos.

Considerando n individuos, existen $\binom{n}{2} = \frac{n(n-1)}{2}$ distancias.



Entonces, la probabilidad de que todas ellas sean calculables es:

$$\{1 - [1 - (1 - q)^2]^p\}^{\frac{n(n-1)}{2}}$$

De esta forma:

$$P(NA) < 1 - \{1 - [1 - (1 - q)^2]^p\}^{\frac{n(n-1)}{2}}$$

En el Apéndice 8 se encuentran los gráficos de los porcentajes de matrices NA encontradas en la simulación, con su respectiva cota superior.

CAPITULO V

CONCLUSIONES

En primer lugar y a partir del desarrollo de las técnicas estadísticas clásicas, se observó que ciertas características maternas influyeron en el normal desarrollo del neonato. Entre ellas se puede destacar que cuando la Época de parto coincidió con las pulverizaciones con organofosforados se alteró de manera altamente significativa la actividad de los biomarcadores AChE y CAT. Por otra parte, y con una menor intensidad, se detectaron hábitos que incidieron en los parámetros antropométricos de los neonatos, como tomar agua de pozo, fumigar en el hogar y el hábito de fumar. Debe subrayarse que estos son prácticas fácilmente modificables lo cual, desde el punto de vista de la salud ambiental, alerta sobre la necesidad de fortalecer las acciones de prevención en esta población de riesgo.

Continuando con el desarrollo del Análisis Multivariado, la Estrategia de una Etapa, consistente en realizar ACoP a partir del coeficiente de similaridad de Gower, y Bietápica ,utilizar APG para hallar la configuración consenso de las dos configuraciones halladas para las variables cuanti y cualitativas, mostraron que la variable Época de parto resultó importante para la discriminación en dos grupos –pulverización, receso– para ambas caracterizaciones. Las otras variables no permitieron dividir de forma clara el grupo de madres.

Cuantificando la concordancia entre las distancias de los individuos entre las configuraciones halladas se obtuvo un $r = 0{,}24$. Esto muestra que ambas estrategias no presentaron la información del mismo modo, dada la baja relación existente entre las configuraciones obtenidas a partir de las matrices de datos cuali y cuantitativos y quizás también atribuible al alto porcentaje de faltantes en la base de datos, que es del 17,9%.

Los resultados mencionados en el párrafo anterior dieron inicio a uno de los objetivos específicos de esta tesis: estudiar los comportamientos de ciertas estrategias multivariadas ante la presencia de datos faltantes. Para poder determinar dichos comportamientos se desarrolló una sentencia que permitiera generar datos de naturaleza mixta en el paquete estadístico R. La función que se

utilizó generó aleatoriamente una matriz de datos que combinó variables categóricas y continuas, de forma tal que permitió simular de manera eficiente a partir de información experimental obtenida empíricamente. Habitualmente se requiere simular datos que conserven estructuras propias de un conjunto de información original para poder recrear determinadas situaciones y estudiar los cambios que surjan. En distintos lenguajes de programación estadísticos se encuentran generadores de números aleatorios de la distribución normal multivariada, como así también de variables aleatorias binomiales multivariadas correlacionadas, pero hasta la presente tesis no se había desarrollado un paquete para la simulación conjunta de variables cuantitativas y cualitativas. Con el desarrollo de esta sentencia se generaron conjuntos de datos que combinaron variables categóricas y continuas, respetando las estructuras de correlación más relevantes en los datos originales.

Mediante la generación de distintas situaciones se estudió cómo la proporción de faltantes altera los resultados de una base de datos, permitiendo elegir el método más $eficaz$ para esta situación.

Se alcanzaron las siguientes conclusiones:

- A medida que aumenta la proporción de faltantes, la Estrategia Bietápica resulta ser la más apropiada para la representación, ya que conserva de manera más $eficaz$ la información de la base de datos original que la Estrategia de una Etapa.
- En ambas estrategias, la pérdida de información es menor cuando se trabaja con mayor cantidad de individuos.
- Considerando aceptable una representación del 70%, la Estrategia de una Etapa tolera un faltante del 15%, mientras que la Bietápica resiste hasta un 25%.

Por otra parte, se estudió en qué momento las estrategias podían ser utilizadas, ya que en algunas oportunidades no es posible determinar matrices de distancia/similaridad, es decir matrices NA, imposibilitando aplicar las técnicas usadas dentro de cada estrategia. En este sentido se puede decir que:

- A medida que aumenta el número de variables (p), disminuye el porcentaje de matrices NA.

- A medida que aumenta el número de individuos (n), aumenta el porcentaje de matrices NA.
- A medida que aumenta la proporción de faltantes (q), aumenta el número de datos faltantes (f), por lo tanto aumenta el porcentaje de matrices NA.

Una vez que las condiciones de simulación están establecidas, cantidad de faltantes, número de variables y cantidad de individuos, se debe calcular la matriz de distancia. Éstas no siempre pueden calcularse, y depende en gran medida del número de variables que cada matriz posea. Más variables implican un incremento en la proporción de veces que se puede calcular la matriz de distancia. Teniendo en cuenta que al realizar ACP, en la EB se reemplazan los faltantes mediante imputación por la media, la cantidad de variables con datos faltantes se reduce sólo a las variables cualitativas (k_c). Mientras que en el caso de la EduE se calcula la similaridad entre individuos mediante el coeficiente de Gower utilizando todas las variables $(k_c + k_q = p)$. Esto hace que la Estrategia de una Etapa sea más $robusta$ que la Bietápica, ya que en la primera se utiliza un mayor número de variables al momento de calcular las matrices de distancia.

Mediante el desarrollo de variables binomiales se elaboró una cota superior genérica para el comportamiento de las matrices NA:

$$P(NA) < 1-\{1-[1-(1-q)^2]^p\}^{\frac{n(n-1)}{2}}$$

Donde:

q = proporción de faltantes en la matriz de datos originales.

p = cantidad de variables utilizadas.

n = número de individuos.

Para futuras investigaciones se recomienda indagar los siguientes aspectos:

- Extensión de la simulación a un número mayor de individuos y variables, lo cual se vio limitado en el desarrollo de la presente tesis por motivos computacionales.
- Utilización de algún otro tipo de mecanismo de pérdida de datos que permita abarcar diferentes situaciones no contempladas en el presente trabajo.

- Extensión de generación de variables cuantitativas a situaciones con distribución no normal.
- Construcción de una función que permita acotar inferiormente la probabilidad de hallar una matriz NA. O su fórmula exacta.
- Estudiar cómo se comportan distintos métodos de imputación.

Los objetivos alcanzados con este trabajo de tesis:

- Se determinó el comportamiento de dos estrategias multivariadas ante la presencia de valores faltantes, según las características de la base de datos original.
- Se analizaron los datos del Proyecto de Investigación "Evaluación toxicológica de poblaciones de riesgo en la exposición crónica a plaguicidas", detectando las variables más influyentes, en cuanto al efecto de los plaguicidas sobre la placenta y los neonatos.
- Se elaboró una sentencia en el lenguaje R que permite simular datos de naturaleza mixta.
- Se desarrolló un mecanismo para la simulación de faltantes (MCAR)
- Se acotó de forma parcial la probabilidad de poder hallar matrices de distancia ante la presencia de faltantes.
- Se caracterizaron las madres mediante técnicas estadísticas multivariadas.

REFERENCIAS BIBLIOGRAFICAS

Abu Qare, A. W., and Abou Donia, M. B. (2001). Inhibition and recovery of maternal and fetal cholinesterase enzyme activity following a single cutaneous dose of methyl parathion and diazinon alone and in combination, in pregnant rats. *Journal of Applied Toxicology* **21**, 307–316.

Anwar, W. A. (1997). Biomarkers of human exposure to pesticides. *Environmental Health Perspectives* **105**, 801–806.

ATSDR. 1997. Toxicological Profile. Metylparathion. CRC Press, Inc.

Badler, C. E., Alsina, S. M., Puigsubirá, C. B., Vitelleschi, M. S. (2004). Tratamiento de bases de datos con información faltante según análisis de las pérdidas con SPSS. *Novenas Jornadas Investigaciones en la Facultad de Ciencias Económicas y Estadística. Universidad Nacional de Rosario.*

Bayrami, M., Hashemi, T., Malekirad, A.A., Ashayeri, H., Faraji F., Abdollahi, M. (2011). Electroencephalogram, cognitive state, psychological disorders, clinical symptom, and oxidative stress in horticulture farmers exposed to organophosphate pesticides. *Toxicol Ind. Health*, en prensa.

Banks, J., Carson, J. S., Nelson, B. (1996). *Discrete − event system simulation. 2º edition*. Prentice-Hall, New Jersey.

Barnett, J. B., Rodgers, K. E. (1994). *Inmunotoxicology and Inmunopharmacology*. Chapter 11. "Pesticides" 2º edition. Raven Press. New York.

Barrera, G. (2009). Introducción a la simulación. Universidad Autónoma de Bucaramanga. Facultad de Ingeniería de Sistemas. http://fis.unab.edu.co/docentes/gbarrera/Introduccion_Simulacion.pdf

Beers, R. F., Sizer, I. W. (1952). A spectrophotometric method for measuring the breakdown of hydrogen peroxide by catalase. *Journal of Biological Chemistry* **195,** 133–140.

de Titto, E., Benítez, R., Derlindatti, A., Domingo, E., Sandlak, J., Álvarez, J., Rivero, S. y Eiman Grossi, M. (2004) Más salud ambiental por más salud. Ingeniería Sanitaria y Ambiental **74**: 42-45.

Benjaminov, O. E., Hoffer, U., Aiteman, J., Urbach, D., Brandes, J. M. (1992). Parathion transfer and acetylcholinesterase activity *in vivo* and *in vitro* perfused term human placenta. *Veterinary and Human Toxicology* **1**, 10–12.

Bisquerra Alzina, R. (1989). *Introducción conceptual al Análisis Multivariable. Vol I y II*. Barcelona: Promociones y Publicaciones Universitarias, S.A.

Bramardi, S. J. (2000). Tesis Doctoral: *Estrategias del Análisis Multivariado para la caracterización de Recursos Fitogenéticos*. Universidad de Valencia

Bramardi, S. J. (2007). *Métodos Multivariados*. Apuntes asignatura Métodos Estadísticos II. Maestría en Estadística Aplicada, Universidad Nacional del Comahue: 1-10 pp.

Bulgaroni, V. (2008). Tesina de grado: *Evaluación de la exposición residencial en niños de comunidades rurales*. Escuela Superior de Salud y Ambiente, Universidad Nacional del Comahue. 122 pp.

Cuadras, C. M. (1996). *Métodos de análisis multivariante*. Barcelona: EUB, S.L.

Cuadras, C. M. (2007). *Nuevos métodos de análisis multivariantes*. CMC Editions.

Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discusion). *Journal of the Royal Statistical Society* **39**, 1–38.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology* **26**, 297–302.

Dijksterhuis, G. B., and Gower, J. C. (1991). The interpretation of Generalized Procrustes Analysis and allied methods. *Food Quality and Preference* **3**, 67-87.

Edwards, F.L. y Tchounwou, P.B. (2005). Environmental toxicology and health effects associated with methyl parathion exposure. A scientific review. *Int. J. Environ. Res. Public Health* **2**, 430-441.

Ellman, G., Courtney, D., Valentino, A., Feathersstone, R. (1960). A new and rapid colorimetric determination of acetylcholinesterase activity. *Biochemical Pharmacology* **7**, 88-95.

Eskenazi, B., Marks, A. R., Bradman, A., Harley, K., Barr, D. B., Johnson, C., Morga, N., Jewell, N. P. (2007). Organophosphate pesticide exposure and neurodevelopment in young mexican-american children. *Environmental Health Perspectives*. **115**, 792-798.

Gil, F., Pla, A. (2001). Biomarkers as biological indicators of xenobiotic exposure-Review. *Journal of Applied Toxicology* **21**, 245-255.

Gilbert, S. F. (2005). *Biología del desarrollo*. 7º Edición, 1º reimpresión. Editorial Médica Panamericana, Buenos Aires. 902 pp.

Gómez García, J., Palarea Albaladejo, J., Martín Fernández, J. A. (2006). Métodos de inferencia estadística con datos faltantes. Estudio de simulación sobre los efectos en las estimaciones. *Estadística Española*. **48**, 241- 270.

Gower, J. C. (1966). Some distance properties of latent root and vector methods in multivariate analysis. *Biometrika* **53**, 315-328.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* **27**, 857-874.

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika* **40**, 33-51.

Gower, J. C., Dijksterhuis, G. B. (2004). *Procrustes problems*. Oxford University Press. 233 pp.

Hardell, L., Eriksson, M. (1999). A case-control study of non-Hodgkin lymphoma and exposure to pesticides. *American Cancer Society* **85**, 1353-1360.

Harnly, M., McLaughlin, R., Bradman, A., Anderson, M., Gunier, R. (2005). Correlating agricultural use of organophosphates with outdoor air concentrations: a particular concern for children. *Environmental Health Perspectives*. **113**, 1184–1189.

Holford, T. R. (2002). *Multivariate Methods in Epidemiology*. Oxford University Press. 440 pp.

Honzik, T., Dragota, Z., Böhm, M., Jeŝina P., Mráĉek T., Paul, J., Zeman, J., Houŝtêk, J. (2006). Specific properties of heavy fraction of mitochondria from human term placenta glycerophosohate dependen hydrogen peroxide production. *Placenta* **27**, 348-356.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24**, 417-441, 498-520.

InfoStat (2008). *InfoStat versión 2008*. Grupo InfoStat, FCA, Universidad Nacional de Córdoba. Córdoba, Argentina.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.* **44**: 223-270.

Judez Asencio, L. (1989). Técnicas de análisis de datos multidimensionales. Madrid: *Centro de Publicaciones, Ministerio de Agricultura, Pesca y Alimentación.*

Lejarraga, H., and Fustiñana, C. (1986). Estándares de peso, longitud corporal y perímetro cefálico desde las 26 hasta las 92 semanas de edad postmenstrual. *Archivo Argentino de Pediatría* **84**, 210-214.

Levario Carrillo, M., Ostrosky Wegman, M., Amato, D., Ostrosky-Wegman, P., González Horta, C., Corona, Y., Sanin, L. H. (2004). Relation between pesticide exposure and intrauterine growth retardation. *Chemosphere* **55**, 1421-1427.

Little, R. J. A., Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley & Sons.

Loewy R. M., Carvajal L. G., Pechen de D'Angelo A. M. (2003a). Residuos de plaguicidas en efluentes de industrias agroalimentarias y aguas superficiales. *Toxicología y Química Ambiental*. Contribuciones para un Desarrollo Sustentable SETAC LA, p.p. 193-195.

Loewy, R. M., Carvajal, L. G., Pechen de D´Angelo, A. M. (2003b). Effect of pesticide use in fruit production orchards on shallow ground water. *Journal* of *Environmental Science* and *Health* **38**, 317-325.

Magnarelli, G. G., Rovedatti, M. G., Pechén de D'Angelo, A. M. (2009). *Plaguicidas y salud humana*. Capítulo 6. En: Anguiano O. L., Montagna, C. M. Clasificación y toxicología de plaguicidas. EDUCO - Universidad Nacional del Comahue, ISBN: 978-987-604-154-6, 390 pp.

Mahalanobis, P. C. (1936). On the generalized distance in statistic. *In Proceedings National Institute of Science, India* **2**, 49-55.

Mahibbur, R. M., and Govindarajulu, Z. (1997). A modification of the test of Shapiro and Wilks for normality. *Journal of Applied Statistics* **24**, 219-235.

Mansour, S., Mossa, A. (2009). Lipid peroxidation and oxidative stress in rat erythrocyte induced by chlopyrifos and the protective effect of zinc. *Pesticide Biochemistry and Physiology* **93**, 34-39.

Matsumura, F. (1985). *Toxicology of Insecticides*. 2º edition. Plenum Press. New York.

Pearson, K. (1901). On lines planes of closest fit to systems of points in space. *Philosophical Magazine*, **6**, 559-572.

Peña, C. E., Carter, D. E., Ayala Fierro, F. (2001). *Toxicología ambiental*: evaluación de riesgo y restauración ambiental. http://superfund.pharmacy.arizona.edu/toxoamb/

Peña, D. (2003). *Análisis de datos Multivariantes*. Mc Graw Hill.

Pidoux, G., Gerbaud, P., Laurendeau, I., Guibourdenche, J., Bertin, G., Vidaud, M., Evain-Brion, D., Frendo, J. (2004). Large variability of trophoblast gene expression within and between human normal term placentae. *Placenta* **25**, 469-473.

Prieto Gómez, R., Matamala, F., Mariana Rojas, M. (2008). Características morfológicas y morfométricas de la placenta de término, en recién nacidos pequeños para la edad gestacional (PEG) en la Ciudad de Temuco-Chile. *Int. J. Morphol.* **26** (3), 615-621.

R Development Core Team. (2008). The R manuals.

Roldán Tapia, L. y Sánchez Santed, F. (2004). Secuelas neuropsicológicas de las intoxicaciones agudas por plaguicidas inhibidores de las colinesterasas. *Rev. Neurol.* **38** (6): 591-597.

Romero Villafranca, R. (1995). *Curso de introducción a los métodos de análisis estadístico multivariante*. Valencia: Servicio de Publicaciones.SP.UPV.95-606.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581-592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley & Sons.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal American Statistical Association* **91**, 473-489.

Sabino, G., Bramardi, S., Di Rienzo, J. (2008a). Generador de datos multivariados mixtos. *XIII Reunión Científica del Grupo Argentino de Biometría*. San Miguel de Tucumán, Argentina. 17 al 19 de Septiembre.

Sabino, G., Rovedatti, M. G., Magnarelli, G. G., Souza, S., Bramardi, S. (2008b). Eficiencia de métodos alternativos para el análisis de datos mixtos ante la presencia de valores faltantes. *VIII Congreso Latinoamericano de Sociedades de Estadística*. Montevideo, Uruguay. 7 al 10 de Octubre.

Sabino, G., Boché, S., Bramardi, S. (2009). Proporción aproximada de matrices de distancia con elementos indeterminados. *XIV Reunión Científica del Grupo Argentino de Biometría*. Trelew, Argentina. 9 al 11 de Septiembre.

Sanborn, M., Colle, D., Kerr, K., Vakiill, C., Saniin, L. H., Bassill, K. (2004). Systematic review of pesticide human health effects. *Pesticides Literature Review* **11**, 167-169.

Sanchez, R. (2003). Evaluación técnico - económica del impacto de Carpocapsa en la producción de fruta de pepita de la norpatagonia argentina. *Comisión de Sanidad Vegetal FUNBAPA*.

Sanin, L.H., López, S.R., Olivares, E.T., Terrazas, M.C., Silva, M.A., Carrillo, M.L. (2001). Relation between birth weight and placenta weight. *Biol. Neonate* **80** (2),113-7.

Santagni, A. (2006). Tesis de grado: Contaminación con plaguicidas del agua subterránea para consumo humano en Campo Grande y Cinco Saltos (Río Negro)". Escuela de Salud y Ambiente, Univ. Nacional del Comahue. 67 pp.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall.15 pp.

Shannon, R. E., (1988). *Simulación de Sistemas. Diseño, desarrollo e implementación*. Trillas, México.

Serkova, N., Bendrick-Peart, J., Alexander, B., Tissot van Patot, M. C. (2003). Metabolite concentrations in human term placentae and their changes due to delayed collection after delivery. *Placenta* **24**, 227-235.

Sokal, R. R., Michener, C. D. (1958). A statistical method for evaluating systematic relationships. University of Kansas. *Science Bulletin* **38**, 1409-1438.

Souza, S. (2005). Tesis de Doctorado: *Identificación de biomarcadores de exposición a plaguicidas en placenta humana.* Centro Regional Universitario Bariloche. Universidad Nacional del Comahue.

Stillerman, K. P., Mattison, D. R., Giudice, L. C., Woodruff, T. J. (2008). Environmental exposures and adverse pregnancy outcomes: a review of the science. *Reproductive Sciences* **15**, 631-650.

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika* **17**, 401-419.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York:John. Wiley.

Tussel, F. (2005). Análisis Multivariante. Apunte asignatura de Estadística: Análisis Multivariante.

Ulam, S., von Neumann, J. On the group of homeomorphisms of the surface of

the sphere, *Bull. Amer. Math. Soc.* vol. 53 (1947) p. 506.

Weselak, M., Arbuckle, T. E., Wigle, D. T., Krewski, D. (2008). Pre and post conception pesticide exposure and the risk of birth defects in an Ontario farm. Reprod Toxicol **25** (4): 472-480.

Willems, P. (2005). *Análisis General*. Apuntes asignatura Análisis Multivariado. Maestría en Estadística Aplicada, Universidad Nacional Comahue 1-15 pp.

Whyatt, R. M., Camann, T. D., Perera, F. P., Rauh, V. A., Tang, D., Kinney, P. L., Garfinke, R., Andrews, H., Hoepner, L., Barr, D. B. (2005). Biomarkers in assessing residential insecticide exposures during pregnancy and effects on fetal growth. *Toxicology and Applied Pharmacology* **206**, 246– 254.

Zhang, P. (2003). Multiple imputation: theory and method. *International Statistical Review* **71**, 581-592.

APÉNDICES

Apéndice 1: Planilla tipo.

Planilla Tipo

NOMBRE DE LA PACIENTE: FECHA DE PARTO:
Historia clínica nº:
Edad: Nivel de Instrucción: A/P/S/U

Domicilio: Nº Muestra:

Completar o marcar lo que corresponda

Vive o vivió en chacra?: nº de años:

Nºhijos previos: Nº hijos previos amamantados: Consume agua de pozo? No/Sí

¿Se fumiga dentro de su hogar? No/Sí

¿Utiliza insecticidas dentro de su hogar? No/Sí

¿Fuma o fumó estando embarazada? No/Sí Fuma su esposo? No/Sí

¿Se fuma en su trabajo? No/Sí

¿Consumo de alcohol durante el embarazo? Frecuencia: cuantos vasos por día /semana?.....
¿Que consume?
Vino/cerveza/otro

Consumo de drogas de adicción: No/Sí

Medicación: No/Sí Tipo ---------

Estado nutricional: Bueno/Malo (en base al body mass index)
Perdida de peso durante el embarazo: No/Sí

Alteraciones en el embarazo

Tipo	Semana de embarazo	Fecha
Amenaza de aborto		
Amenaza de parto prematuro		
Parto prematuro		
Otra		

Datos del neonato:

Peso	Talla:	Perímetro cefálico	Edad gestacional

Datos de la placenta:

Peso	Observaciones:

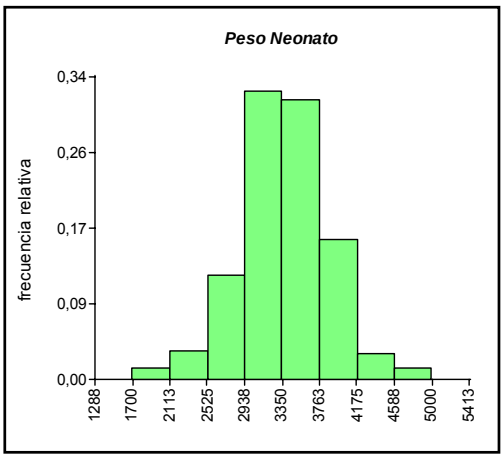
Apéndice 2A: Medidas de Resumen, Histogramas y Box plot de las variables cuantitativas.

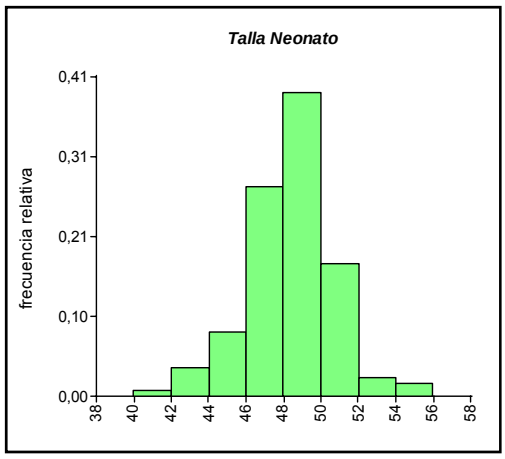
Medidas de Resumen

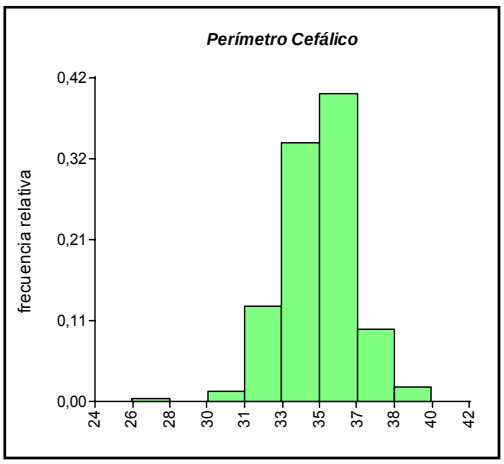
Estadística descriptiva

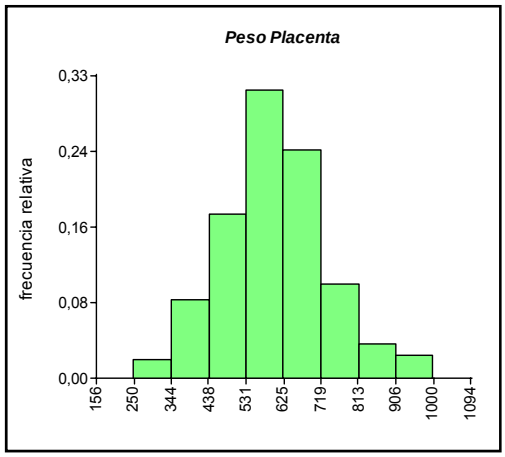
Variable	n	Media	D.E.	Mín	Máx	Mediana
PESO	306	3377,81	503,66	1770,00	4970,00	3395,00
TALLA	303	48,77	2,30	41,00	55,50	49,00
PC	304	34,88	1,63	26,50	39,50	35,00
PPLA	248	602,26	134,77	250,00	1000,00	600,00
ACHE	111	12,13	3,94	4,83	24,83	11,63
CAT	27	0,15	0,04	0,07	0,22	0,14
EG	260	39,10	1,45	28,00	42,00	39,00

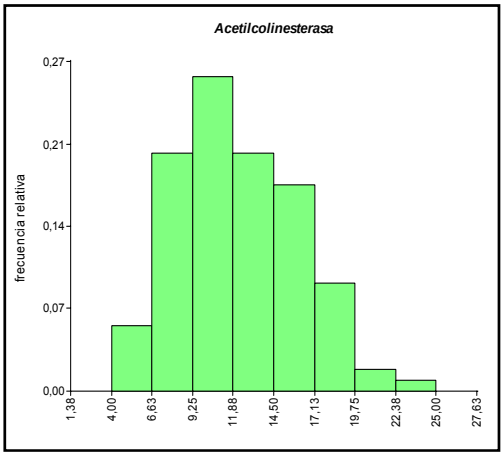
Histogramas

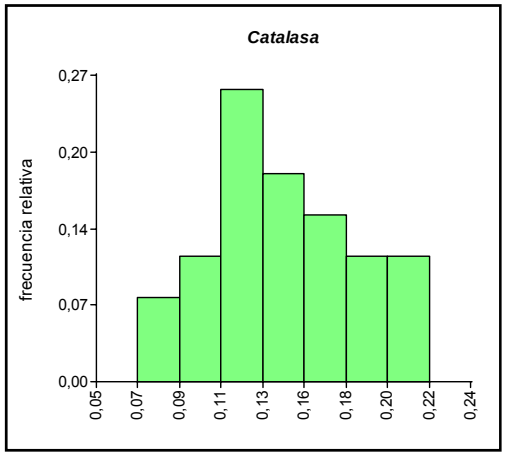


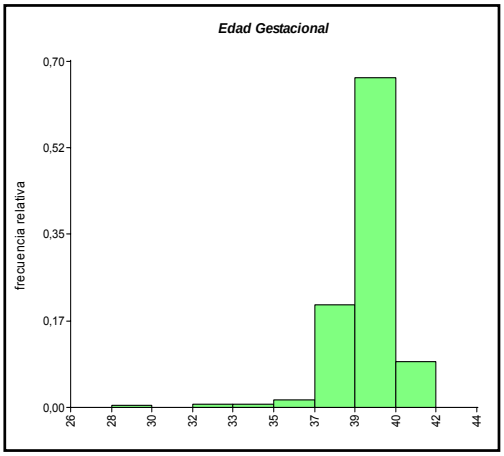




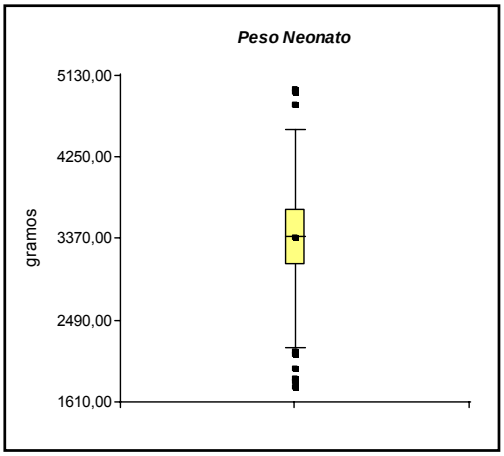


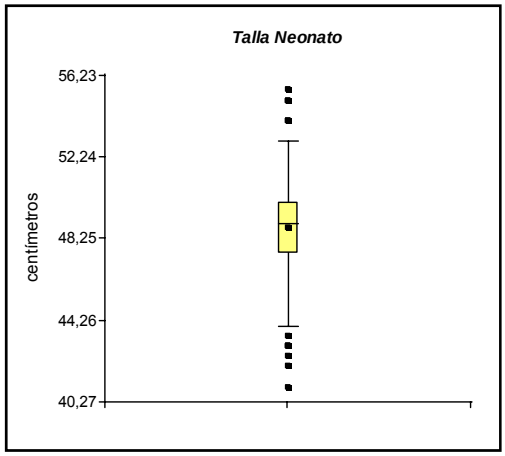

Acetilcolinesterasa
frecuencia relativa
0,27
0,21
0,14
0,07
0,00
1,38
4,00
6,63
9,25
11,88
14,50
17,13
19,75
22,38
25,00
27,63

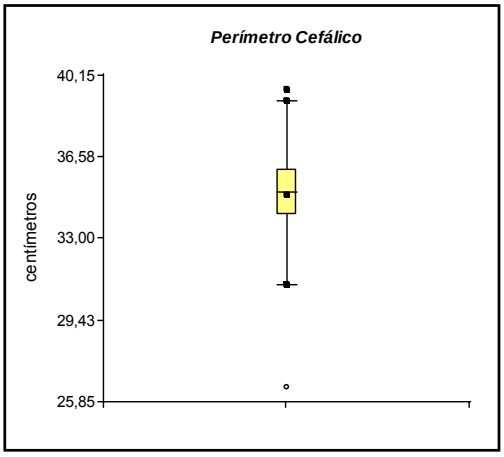

Catalasa
frecuencia relativa
0,27
0,20
0,14
0,07
0,00
0,05
0,07
0,09
0,11
0,13
0,16
0,18
0,20
0,22
0,24

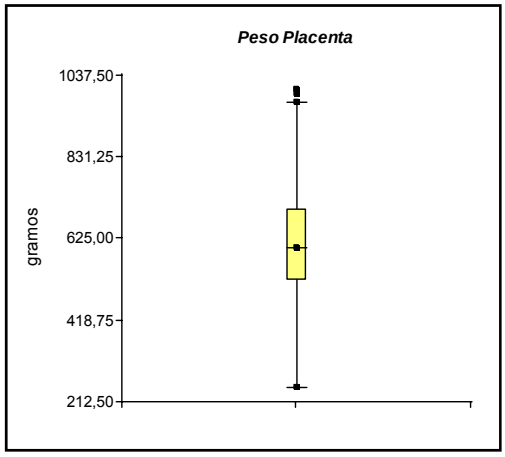

Edad Gestacional
frecuencia relativa
0,70
0,52
0,35
0,17
0,00
26
28
30
32
33
35
37
39
40
42
44

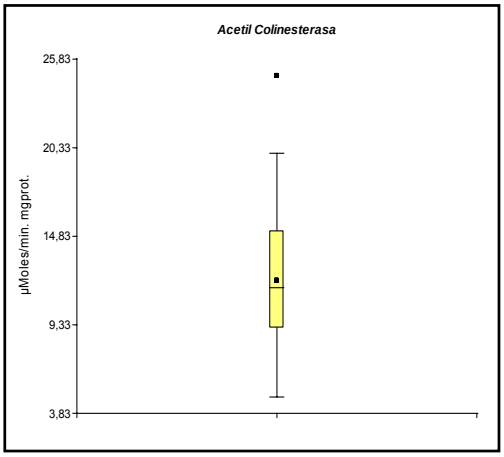
Box plot

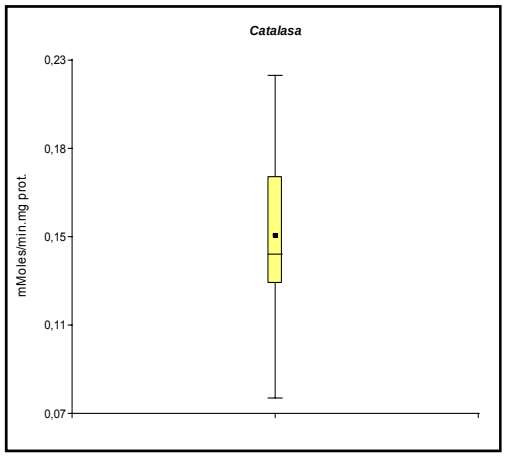

Peso Neonato
gramos
5130,00
4250,00
3370,00
2490,00
1610,00

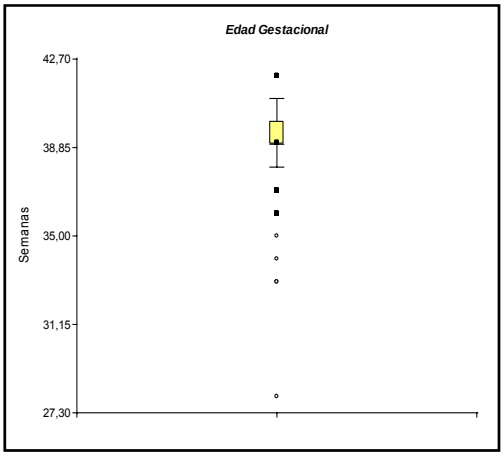

Talla Neonato
centímetros
56,23
52,24
48,25
44,26
40,27


Perímetro Cefálico
centímetros
40,15
36,58
33,00
29,43
25,85

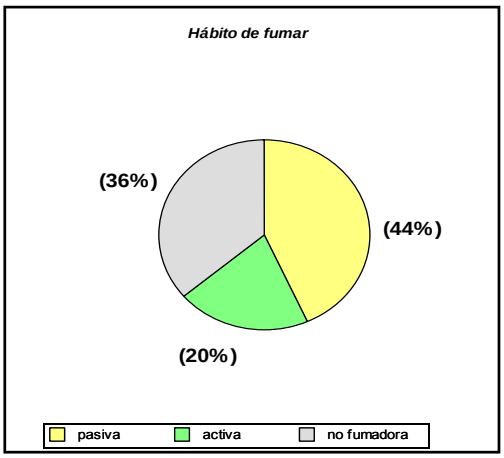

Peso Placenta
gramos
1037,50
831,25
625,00
418,75
212,50

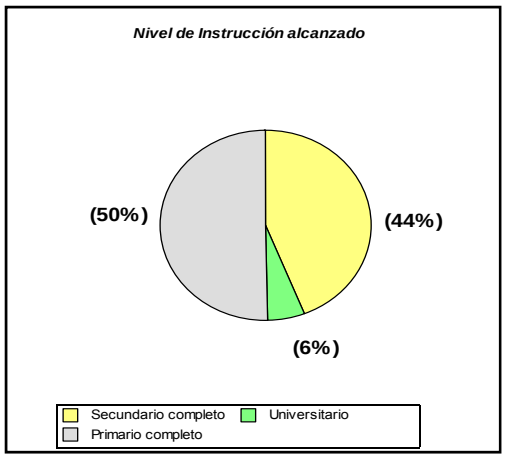

Acetil Colinesterasa
µMoles/min. mgprot.
25,83
20,33
14,83
9,33
3,83

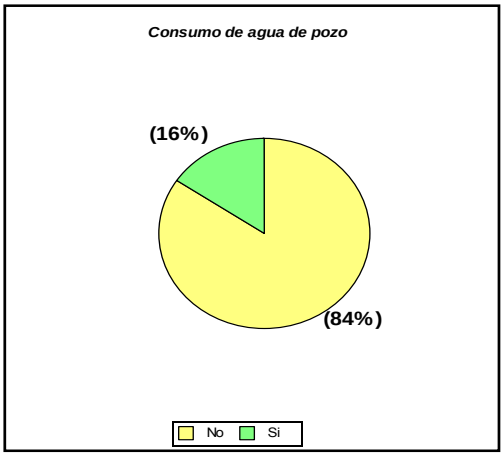

Catalasa
mMoles/min.mg prot.
0,23
0,18
0,15
0,11
0,07

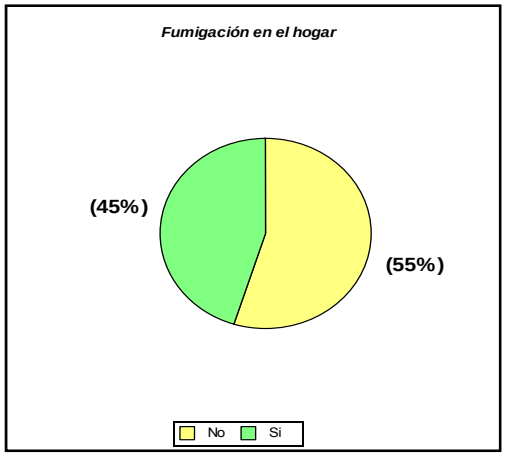

Edad Gestacional
Semanas
42,70
38,85
35,00
31,15
27,30

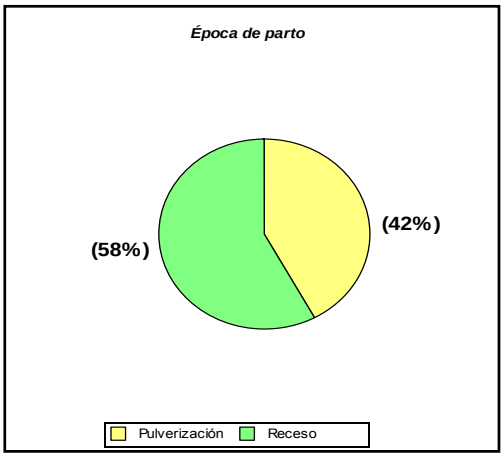
Apéndice 2B: Gráficos de sectores de las variables cualitativas.

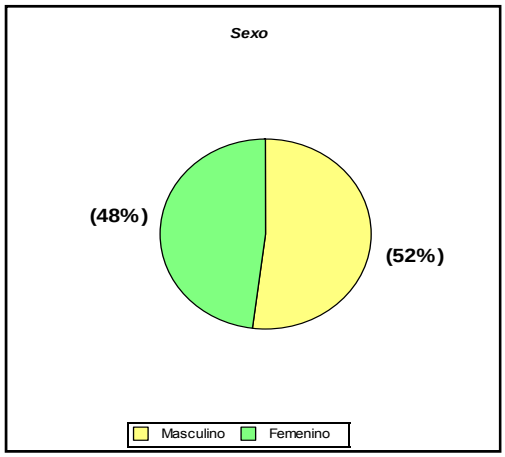












Apéndice 3: Pruebas Chi-cuadrado para determinar asociación entre variables categóricas.

Nivel de Instrucción alcanzado - Hábito de Fumar

Estadístico	Valor	gl	p
Chi Cuadrado Pearson	8,55	4	0,0733
Chi Cuadrado MV-G2	8,97	4	0,0619
Coef.Conting.Cramer	0,10		
Coef.Conting.Pearson	0,17		

Nivel de Instrucción alcanzado - Fumigación en el hogar

Estadístico	Valor	gl	p
Chi Cuadrado Pearson	4,64	2	0,0980
Chi Cuadrado MV-G2	5,07	2	0,0792
Coef.Conting.Cramer	0,10		
Coef.Conting.Pearson	0,14		

Fumigación en el hogar – Consumo de Agua de Pozo

Frecuencias absolutas
En columnas:POZO

FUMIH	0,00	1,00	Total
0,00	132	12	144
1,00	85	31	116
Total	217	43	260

Estadístico	Valor	gl	p
Chi Cuadrado Pearson	15,74	1	0,0001
Chi Cuadrado MV-G2	15,93	1	0,0001
Irwin-Fisher bilateral	0,18		0,0001
Coef.Conting.Cramer	0,17		
Coef.Conting.Pearson	0,24		
Coeficiente Phi	0,25		

Cocientes de chance (odds ratio) y riesgos relativos

Estadístico	Estim	LI 95%	LS 95%
Odds Ratio 1/2	4,01	1,97	8,15
Odds Ratio 2/1	0,25	0,12	0,51

Apéndice 4: Pruebas t y ANOVA para determinar cambios en valores de biomarcadores y parámetros antropométricos según variables categóricas.

Perímetro Cefálico corregido - Consumo de agua de pozo

Clasific	Variable	n(no)	n(si)	media(no)	media(si)	p
POZO	Pesocorr	201	42	0,23	0,18	0,6321
POZO	Tallacorr	201	42	-0,65	-0,64	0,4689
POZO	PCcorr	201	42	0,17	0,70	**0,0033
POZO	PPla	203	37	597,93	621,35	0,1686
POZO	AChE	90	18	12,10	12,32	0,4160
POZO	CAT	19	7	0,15	0,14	0,5983

Perímetro Cefálico corregido – Fumigación en el hogar

Clasific	Variable	n(no)	n(si)	media(no)	media(si)	p
FUMIH	Pesocorr	116	97	0,12	0,30	0,0689
FUMIH	Tallacorr	116	97	-0,71	-0,63	0,3078
FUMIH	PCcorr	116	97	0,07	0,47	**0,0067
FUMIH	PPla	111	98	591,89	620,10	0,0672
FUMIH	AChE	49	44	12,58	12,46	0,5605
FUMIH	CAT	14	12	0,14	0,15	0,2170

Acetilcolinesterasa y Catalasa – Época de parto

Variable	n(R)	n(P)	media(R)	media(P)	p
Pesocorr	141	95	0,23	0,24	0,4758
Tallacorr	141	95	-0,65	-0,72	0,6807
PCcorr	141	95	0,32	0,34	0,4519
PPla	134	96	610,45	600,00	0,7166
AChE	49	55	10,07	14,03	**<0,0001
CAT	13	13	0,12	0,17	**0,0001

R: Receso P: Pulverización

Peso corregido – Hábito de Fumar

Variable	N	R^2	R^2 Aj	CV
Pesocorr	245	0,04	0,03	378,27

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	6,84	2	3,42	4,56	0,0114
FUMAR	6,84	2	3,42	4,56	**0,0114
Error	181,66	242	0,75		
Total	188,51	244			

Test:LSD Fisher Alfa=0,05 DMS=0,27406

Error: 0,7507 gl: 242

FUMAR	Medias	n		
FUMA ACTIVO	-0,10	49	A	
NO FUMA	0,27	89		B
FUMA PASIVO	0,34	107		B

Letras distintas indican diferencias significativas(p<= 0,05)

Apéndice 5: Programa en R para simular datos mixtos (5 variables cualitativas-5 cuantitativas)

```
###################### Ingreso el número de categorías ####################

categ<-c(3,2,2,2,2)

######## Ingreso el vector de probabilidades correspondientes a cada cruce es decir 3*2*2*2*2 = 48 cruces, teniendo en cuenta que la suma debe dar 1 #########

mult<-c(9, 13, 8, 6, 7, 4, 6, 3, 1, 0, 1, 1, 2, 3, 2, 7, 15, 8,10, 5, 9, 6, 8, 10, 1, 1, 3,
0, 4, 3, 1, 1, 6, 8, 3, 2, 1, 3, 4, 3, 1, 1,1, 0, 1, 2, 1, 1)/196

#### Ingreso el vector de medias para cada posible resultado de la multinomial. Si son 5 variables normales multivariables entonces queda una matriz de 48 x 5 ####

matriz.medias <- read.delim("clipboard", na.string="", dec=",", header=FALSE)
matriz.medias<-as.matrix(matriz.medias)

############ Todos los posibles cruces quedan en la matriz cruce ############

cruce<-read.delim("clipboard", na.string="", dec=",", header=FALSE)
cruce<-as.matrix(cruce)

###### Ingreso la matriz de varianza-covarianza de las variables normales ######

readcov<-read.delim("clipboard", na.string="", dec=",", header=FALSE)
varcov<-cov(readcov, use="pairwise.complete.obs")
n<-50

#########################################################################
#################### SIMULACIÓN DE DATOS MIXTOS ####################
#########################################################################
library(mvtnorm)
simixtos <- function (n, categ, mult, cruce, matriz.medias, varcov)
{p=round(mult*10000)
indices=c()
for (i in (1:length(p))) {indices=c(indices,rep(i,p[i]))}
k=length(indices)
b=c()
for (i in (1:n))            {   b[i]<-indices[round(runif(1,1,k))]  }
L<-matrix(0,n,dim(matriz.medias)[2]+length(categ))
for (i in (1:n))
{    L[i,]<-c(cruce[b[i],],rmvnorm(1,mean=matriz.medias[b[i],], sigma=varcov))     }
L }
```

Apéndice 6: Programa en R para calcular las distancias entre individuos de las configuraciones resultantes en ambas estrategias (5 variables cualitativas-5 cuantitativas)

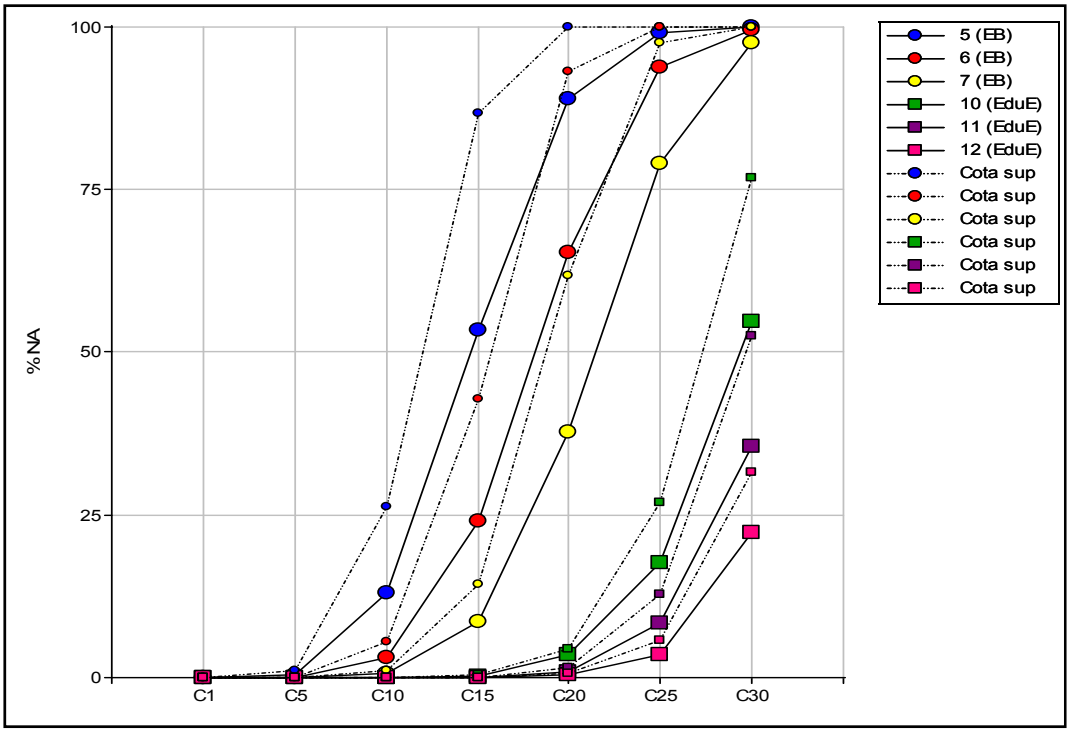
```
#########################################################################
###################### ESTRATEGIA DE UNA ETAPA ##########################
#########################################################################

library(cluster)
library(FactoMineR)
ACoP<-function(X)
        {   if(any(is.na(daisy(X, metric = c("gower"))))==TRUE)
                {   NA   }
            else
                {   P<-dist(cmdscale(daisy(X, metric = c("gower"))))   }
        }

#########################################################################
###################### ESTRATEGIA BIETÁPICA #############################
#########################################################################

mifuncion<-function(X)
              {    if(any(is.na(daisy(X[,-(6:10)],metric = c("gower"))))==TRUE)
        {  NA   }
                 else
         {        P<-as.data.frame(cbind(PCA(X[,-(1:5)], graph=FALSE)$ind$coor,
                  cmdscale(daisy(X[,-(6:10)],  metric = c("gower")),k=5)))
                  t.L <- GPA(P, group=c(5,5),graph = FALSE)
                  dist(t.L$consensus[,(1:2)])   }
                }
```

Apéndice 7: Programa en R para sacar datos MCAR y hallar correlaciones promedio entre las distancias de las configuraciones originales y con faltantes (5 variables cualitativas-5 cuantitativas)

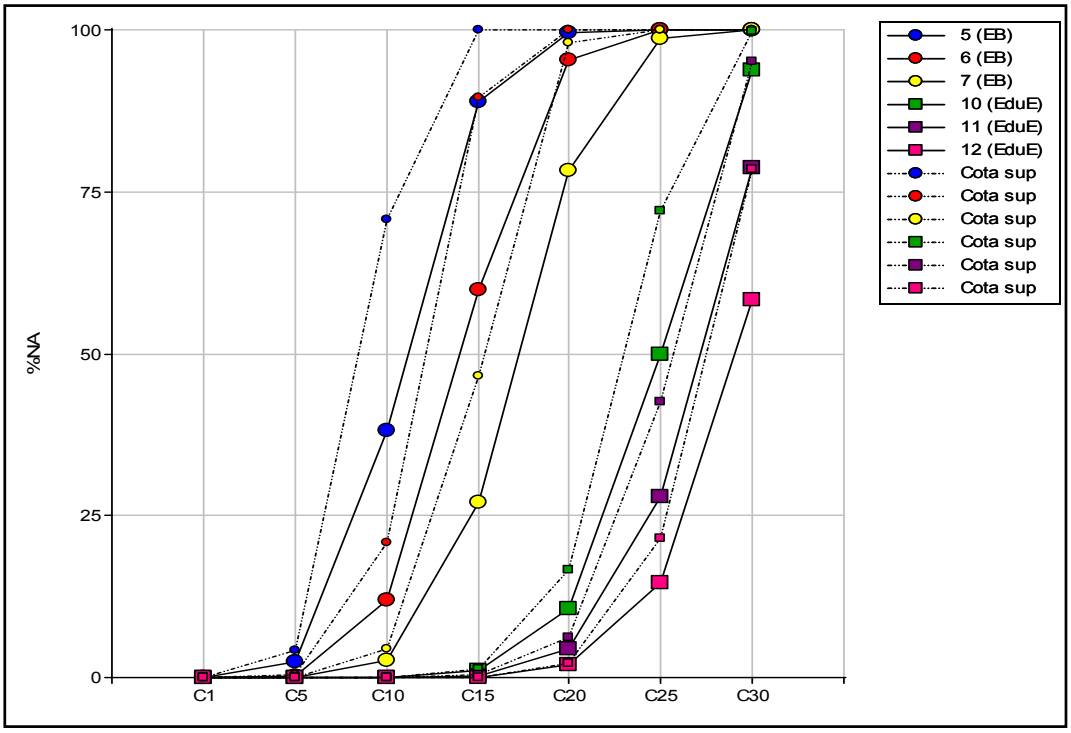
```
#########################################################################
######################## PORCENTAJE DE FALTANTES ########################
#########################################################################

porc.falt<-function(X) {   for (j in 1:1000)    { if ( b>=falt )     {stop}
                                                  else
                           {   fil<-floor(runif(1,1,n+1))
                               colu<-floor(runif(1,1,11))
                                    if(is.na(X[fil,colu])==FALSE)
                                   {    X[fil,colu]<-NA
                                   b<-b+1}
                           }
                   }
        Falt<-X
        }
#########################################################################
############ SIMULACIÓN DE 1000 MATRICES DE 50 INDIVIDUOS ###############
#########################################################################

m<-1000
n<-50
L<-as.list(rep(NA,m))
for (i in 1:m)
                   {    L[[i]]<-simixtos(n, categ, mult, cruce, matriz.medias, varcov)   }
for (i in 1:m)   {
colnames(L[[i]])<-c("V1","V2","V3","V4","V5","V6","V7","V8","V9","V10")
L[[i]]<-as.data.frame(L[[i]])
L[[i]]$V1<-as.factor(L[[i]]$V1)
L[[i]]$V2<-as.factor(L[[i]]$V2)
L[[i]]$V3<-as.factor(L[[i]]$V3)
L[[i]]$V4<-as.factor(L[[i]]$V4)
L[[i]]$V5<-as.factor(L[[i]]$V5)
                   }
```

```
###################### Hallo el conjunto de matrices L ######################
ACoP.L<-lapply(L, ACoP)
Proc.L<-lapply(L, mifuncion)

########################## Saco el 1% de los datos ########################
falt<-0.01*n*10
b<-0
H<-lapply(L,porc.falt)
ACoP.n.1<-lapply(H, ACoP)
Proc.n.1<-lapply(H, mifuncion)

corrACoP.n.1<-rep(NA,m)
corrProc.n.1<-rep(NA,m)

for (i in 1:m)   {
                   if ((any(is.na(ACoP.n.1[[i]])))==TRUE){NA}
                   else{corrACoP.n.1[i]<-cor(ACoP.n.1[[i]],ACoP.L[[i]])}
                  }

for (i in 1:m)  {
                   if ((any(is.na(Proc.n.1[[i]])))==TRUE){NA}
                   else    {    corrProc.n.1[i]<-cor(Proc.n.1[[i]],Proc.L[[i]])    }
                  }

fACoP.50.1<-sum(is.na(ACoP.n.1))
fProc.50.1<-sum(is.na(Proc.n.1)) #para contar matrices NA#
```

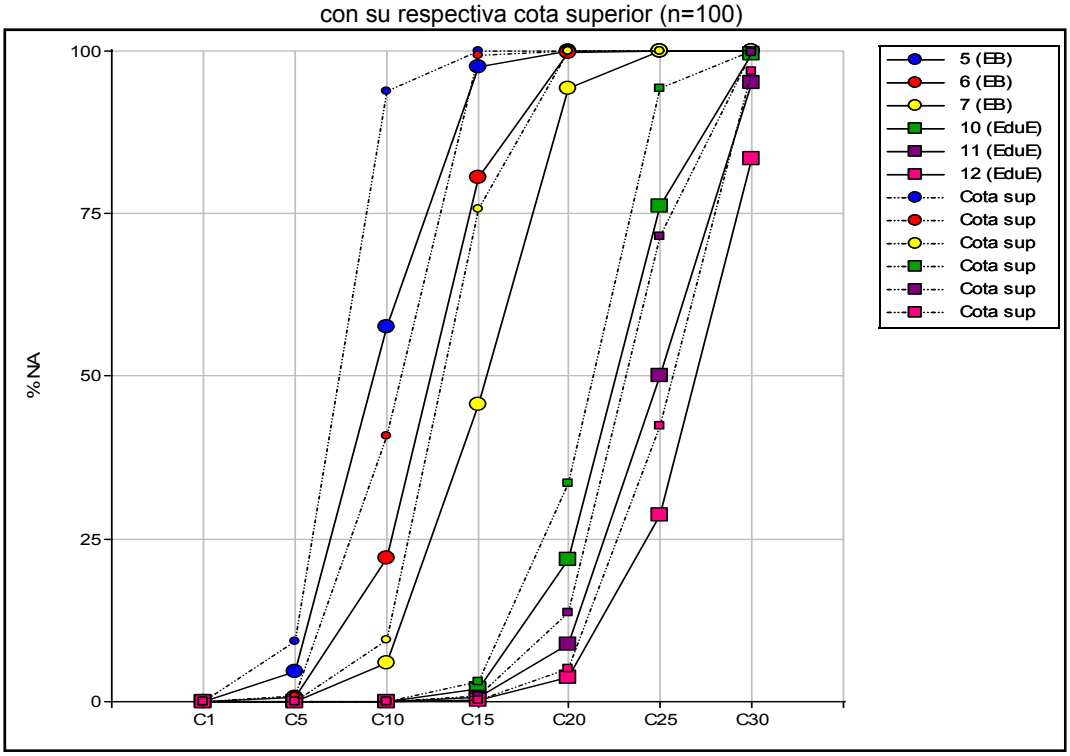
Apéndice 8: Gráficos de la proporción de matrices NA con su respectiva cota superior



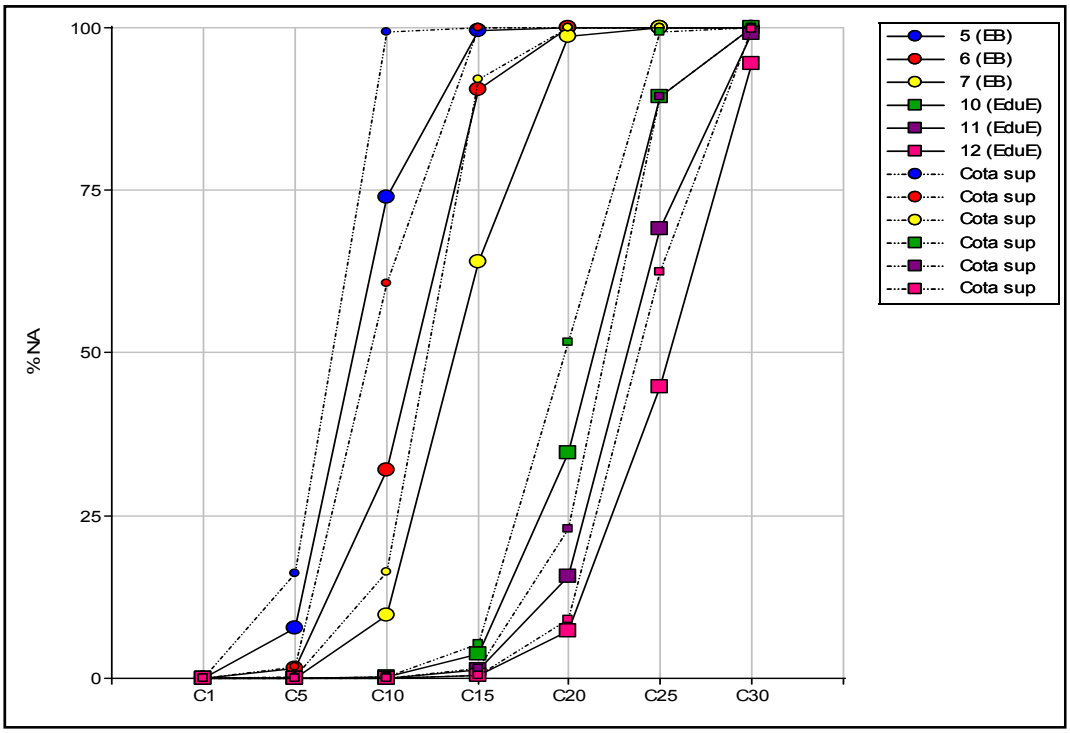
Porcentaje de Matrices NA, según porcentaje de faltantes y número de variables, con su respectiva cota superior (n=50)



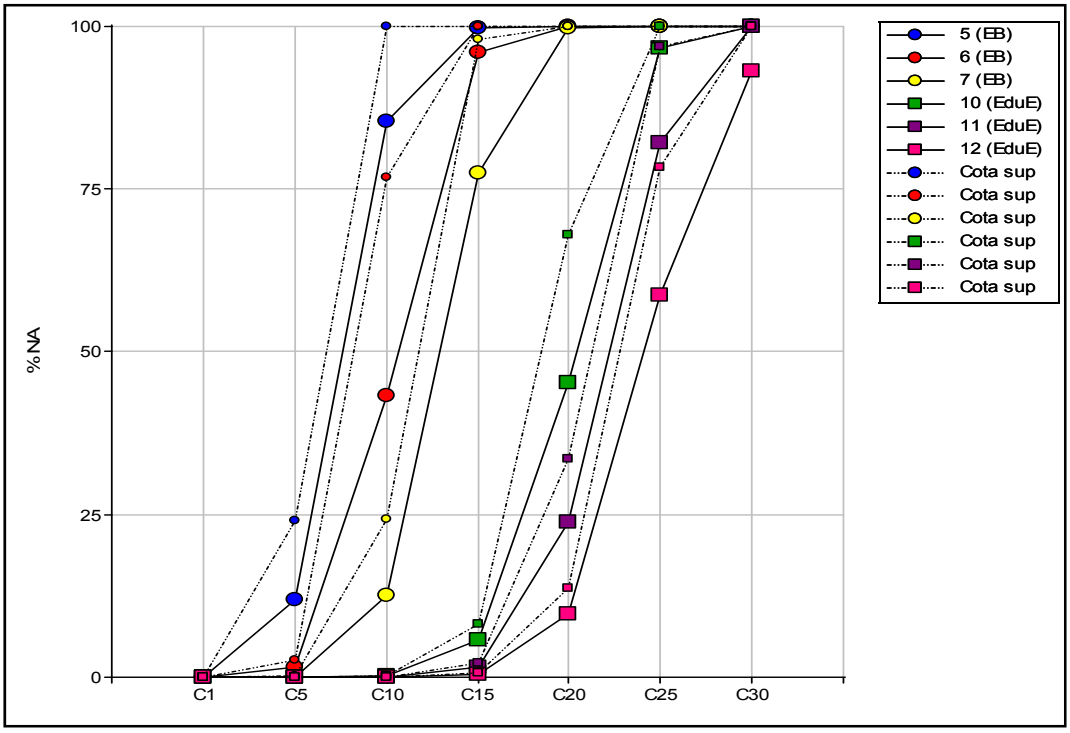
Porcentaje de Matrices NA, según porcentaje de faltantes y número de variables, con su respectiva cota superior (n=100)



Porcentaje de Matrices NA, según porcentaje de faltantes y número de variables, con su respectiva cota superior (n=150)



Porcentaje de Matrices NA, según porcentaje de faltantes y número de variables, con su respectiva cota superior (n=200)



Porcentaje de Matrices NA, según porcentaje de faltantes y número de variables, con su respectiva cota superior (n=250)