ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Instituto de Ciencias Matemáticas

Ingeniería en Estadística Informática

"Modelo de Detección de Fraudes en Aseguramiento de Vehículos utilizando Componentes Principales y Análisis RIDIT"

TESIS DE GRADO

Previa a la obtención del Título de:

INGENIERA EN ESTADÍSTICA INFORMÁTICA

Presentada por:

Heydi Mariana Roa López

GUAYAQUIL - ECUADOR

AÑO

2004

AGRADECIMIENTO

A todas las personas que de una u otra manera colaboraron en la realización de este trabajo y específicamente a la gran ayuda de mi Director el Msc. Fernando Sandoya Sánchez.

DEDICATORIA

A Dios

A mis padres

A mis hermanos

Y a mis familiares y amigos

TRIBUNAL DE GRADUACIÓN

_____	_____
ING. PABLO ÁLVAREZ	**MSC. FERNANDO SANDOYA**
PRESIDENTE DEL TRIBUNAL	**DIRECTOR DE TESIS**
_____	_____
ING. JORGE FERNÁNDEZ	**ING. SORAYA SOLÍS**
VOCAL	**VOCAL**

DECLARACIÓN EXPRESA

“La responsabilidad del contenido de esta tesis de grado, me corresponde exclusivamente; y el patrimonio intelectual de la misma a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL”

(Reglamento de graduación de la ESPOL)

_____

Heydi Mariana Roa López

RESUMEN

El presente trabajo tiene como propósito proporcionar una metodología dirigida a cuantificar la probabilidad de fraude en las declaraciones de siniestros vehiculares, denominada Análisis PRIDIT, específicamente cuando se tienen variables cualitativas involucradas puesto que en el Ecuador no son muy numerosos los estudios dedicados a la detección de fraudes con este tipo de variables.

En el primer capítulo se presenta información acerca de lo que representa el fraude en el mercado asegurador de vehículos, en el segundo capítulo se da a conocer los antecedentes del cobro de siniestros en los seguros de automóviles en nuestro país, Ecuador, desde 1.993 hasta 2.003; así como también la situación actual de este tipo de seguro frente a los demás ramos de seguros del mercado ecuatoriano. El tercer capítulo presenta información del Análisis de Componentes Principales, técnica que es utilizada para el desarrollo de la nueva metodología aquí propuesta. En el cuarto capítulo se realiza la descripción, desarrollo y presentación de resultados de la aplicación del Análisis PRIDIT, para finalmente presentar las respectivas conclusiones y recomendaciones.

INDICE GENERAL

Pág.

II. COBRO DE SINIESTROS EN SEGUROS DE VEHÍCULOS DURANTE LA ÚLTIMA DÉCADA EN EL ECUADOR

III. ANÁLISIS DE COMPONENTES PRINCIPALES

ABREVIATURAS

ICEA	Investigación Cooperativa entre Entidades Aseguradoras
RIDIT	Relative to an Identified Distribution
ACP	Análisis de Componentes Principales
SCREE	Score de los Eigenvalores
G1	Grupo 1 o Grupo Fraudulento
G2	Grupo 2 o Grupo No Fraudulento

SIMBOLOGÍA

I	Número índice
X_t	Monto del siniestro pagado en el período t
X_0	Monto del siniestro pagado en el período base
X	Matriz de datos
Σ	Matriz de Varianzas y Covarianzas
Y_j	j-ésima Componente Principal
a_j	j-ésimo eigenvector normalizado de Σ
λ_j	j-ésimo eigenvalor de Σ
A	Matriz de eigenvectores de Σ
Λ	Matriz de Varianzas y Covarianzas de Y
Y	Matriz de Componentes Principales
Var (Y_j)	Varianza de la j-ésima Componente Principal
tr(Σ)	Traza de la Matriz de Varianza y Covarianza
tr(Λ)	Traza de la Matriz de Varianzas y Covarianzas de Y

SIMBOLOGÍA

$Var(X_i)$	Varianza de la i-ésima variable original
$\sum_{i=1}^{p} \lambda_j$	Sumatoria desde j igual a 1 hasta p de los eigenvalores
n	Número de unidades experimentales
S	Matriz de Varianzas y Covarianzas Muestral
$\hat{\lambda}_j$	Estimador del j-ésimo eigenvalor
$\hat{a}_j$	Estimador del j-ésimo eigenvector
d	Dimensionalidad del espacio en el que están los datos
k	Número de eigenvalores iguales a cero
C	Vector de carga de las Componentes Principales
P	Matriz de Correlación
λ_j^+	j-ésimo eigenvalor de la Matriz de Correlación
a_j^+	j-ésimo eigenvector de la Matriz de Correlación
X_{rj}	Valor de la r-ésima categoría de la j-ésima variable indicadora

SIMBOLOGÍA

Z_{rj}	Valor estandarizado de la r-ésima categoría de la j-ésima variable indicadora de fraude
k_t	Número de categorías disponibles y ordenadas para la variable indicadora t
$\hat{p}_t$	Vector de proporciones de respuesta observada de las demandas
B_{ti}	Puntuación RIDIT de la i-ésima categoría de la variable t
$\sum \hat{p}_{ti} B_{ti}$	Valor esperado de las puntuaciones RIDIT
H_0	Hipótesis Nula
H_1	Hipótesis Alterna
R_{ti}	Puntuación RIDIT original de Bross de la i-ésima categoría de la variable t
n_1	Tamaño de la muestra 1
n_2	Tamaño de la muestra 2

SIMBOLOGÍA

U_{α}	Estadístico de prueba U de Mann-Whitney
μ_{U_1}	Media del estadístico U_1
σ_{U_1}	Varianza del estadístico U_1
$\bar{B_t}$	Media de las puntuaciones RIDIT de la variable t
W_1^*	Estadístico de prueba de Wilcoxon
N	Número de demandas
$\pi_t^{(1)}$	Vector multinomial de la probabilidad de respuesta del Grupo 1
$\pi_t^{(2)}$	Vector multinomial de la probabilidad de respuesta del Grupo 2
θ	Proporción de demandas que pertenecen al grupo fraudulento
(θ-1) A_t	Proporción de demandas que pertenecen al grupo no fraudulento
$\theta\ A_t$	Puntuación esperada para una demanda fraudulenta

SIMBOLOGÍA

A_t	Medida de Poder Discriminatorio de las t variables indicadoras de fraude
I_A	Función indicadora de A
F	Matriz de puntuaciones individuales PRIDIT
f_{it}	Valor RIDIT de la i-ésima categoría de la variable t
$\hat{W}_t^{\infty}$	Ponderación PRIDIT estimada de las t variables indicadoras de fraude
E[F'F]	Valor esperado de la matriz F transpuesta por la matriz F
$\hat{U}$	Matriz diagonal de las varianzas singulares de un Análisis de Factores
q	Tipo de grupo
B_{qt}^*	Puntuación RIDIT de una demanda seleccionada al azar del grupo q

INDICE DE TABLAS

Pág.

INDICE DE CUADROS

Pág.

INDICE DE GRÁFICOS

Pág.

INDICE DE FIGURAS

Pág.

INTRODUCCIÓN

A pesar de que el fraude se da en casi todas las ramas, los fraudes en los seguros se han convertido en una práctica común. El mercado asegurador considera el fraude como un factor ineludible de riesgo y, hoy en día las entidades luchan por desarrollar un foco de acción frente al mismo.

Precisamente, el tema de investigación realizado es un modelo de detección de fraude en las declaraciones de siniestros de vehículos utilizando una nueva técnica denominada PRIDIT, el mismo que es aplicable a las compañías aseguradoras.

El **objetivo general** del presente trabajo de investigación es:

- La aplicación de un modelo de detección de fraude a una cartera real de seguro de automóviles, enfocado en las declaraciones de siniestros de automóviles para clasificar y cuantificar el nivel de fraude de cada una de estas declaraciones, realizadas por los asegurados.

Los **objetivos específicos** son:

- Reducir la incertidumbre e incrementar las oportunidades de clasificar las demandas correcta y eficientemente a cada grupo (fraudulentas / no fraudulentas) sin importar el tipo de variables que intervengan.

- Transformar respuestas categóricas en un conjunto de valores numéricos que estén dentro de un intervalo [-1,1], lo cual refleje la relativa anormalidad de una respuesta en particular.
- Determinar una ponderación de fraude para cada variable involucrada en el análisis.
- Determinar una medida de poder discriminatorio que permita clasificar las demandas en fraudulentas y no fraudulentas.

CAPÍTULO 1

1. EL FRAUDE EN EL MERCADO ASEGURADOR DE AUTOMÓVILES.

Introducción

Un término desgraciadamente muy utilizado en toda la sociedad es el de "fraude". Conocemos por fraude cualquier actividad en la que para derivar un beneficio económico, se crean situaciones ficticias o se exageran daños.

El fraude está considerado como una de las industrias criminales más grandes en la sociedad, y según estudios de varios investigadores aumentan en épocas donde la gente necesita dinero tales como la Navidad, Fin de año, etc.

A pesar de que el fraude se da en casi todas las actividades humanas, los fraudes en los seguros se han convertido en una práctica común.

Precisamente, este es uno de los problemas que enfrentan las compañías aseguradoras al verse defraudadas por algunos de sus clientes.

Como se conoce, un contrato de seguro involucra un acuerdo entre la compañía aseguradora y el asegurado para que la compañía cubra un riesgo, pero el comportamiento del asegurado, quien recibe compensación en caso de un accidente, no siempre es honesto. La existencia de comportamientos fraudulentos en el seguro de automóviles se ha convertido en un asunto de gran preocupación para las compañías de seguros y también para sus usuarios. La influencia de las acciones deshonestas por parte de los asegurados se deja sentir tanto en el número de siniestros declarados como en la cuantía de los mismos. Si consideramos el peso que ello puede tener a la hora de justificar la aparición de resultados técnicos negativos durante los últimos años en el seguro de vehículos y el incremento del valor de las primas por la contratación de estos mismos seguros, queda más que justificada la necesidad de diseñar herramientas que ayuden a las entidades en la detección y lucha contra el fraude.

En el Ecuador no se han hecho estudios serios al respecto, sin embargo en otros países este tipo de estudios están bien adelantados, observemos

de manera general las situaciones que se han dado al respecto en algunos países:

De acuerdo a un estudio[1] realizado en 1.991 sobre el seguro de automóviles en Massachussets, el 72 por ciento de las demandas no tuvieron problema de litigación, pero de las demandas que resultaron en litigación, cerca del 99 por ciento fueron decretadas con problemas de litigación antes de que el veredicto real del jurado sea dictaminado. Consecuentemente, el número de situaciones en las cuales uno puede realmente observar el verdadero valor de la variable dependiente (determinación legal de fraude o legalmente no comprobable como fraude) es apenas del 28 por ciento.

En España, según la ICEA[2], el sector del automóvil es el que más fraudes registró, puesto que de los 46.228 casos detectados en el 2001, el 90 por ciento correspondía a esta rama. En lo que respecta al año en

---

[1] Estudio desarrollado en 1.991 por Weisberg y Derrig referente al seguro de automóviles en Massachusetts.

[2] ICEA (Investigación Cooperativa entre Entidades Aseguradoras y Fondos de Pensión) es una asociación cuyo objetivo es estudiar e investigar materias relacionadas con el seguro. Anualmente organiza un concurso en el que las aseguradoras del mercado español participan aportando los casos de fraude que han detectado, esto les permite compartir información y luchar contra esa práctica.

curso (2004) según datos de la ICEA, más del 75 por ciento de los casos de fraude detectados corresponde a la rama de vehículos.

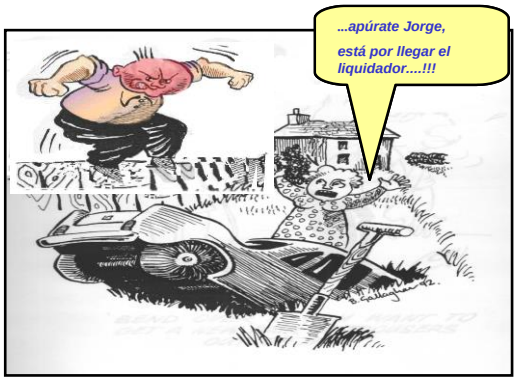
Las compañías aseguradoras del Ecuador no cuentan aún con una unidad investigativa de fraude que se dedique a realizar este tipo de estudios. Tampoco las aseguradoras poseen mecanismos de control para la detección de fraudes, limitadamente se basan en la percepción y experiencia que tienen en el sector para determinar si un reporte de siniestro muestra fraude en su contenido.

Diversos artículos publicados en revistas especializadas del sector asegurador e incluso manuales como la *Guía Anti-Fraude*, editada por el Comité Europeo de Seguros o el libro, *Manual de Investigación de Siniestros y Lucha contra el Fraude en el Seguro de Automóviles*, alertan sobre la existencia del problema, sobre sus diferentes formas de manifestarse y señalan pautas de actuación frente al mismo. En todos ellos se hace referencia a la importancia de diseñar un perfil del asegurado defraudador y a la conveniencia de identificar determinados factores que incrementan la probabilidad de aparición de fraude, ya que en los casos de declaraciones fraudulentas de siniestros de vehículos hay determinadas características que aparecen con mayor frecuencia y que

ayudan de manera importante a detectar la situación delictiva del asegurado.

A continuación se presentan diferentes escenarios donde se puede observar el comportamiento fraudulento del asegurado, razones por las cuales las compañías de seguros deberían estar mayormente interesadas en apoyar al desarrollo de nuevas y mejores técnicas que permitan detectar estos comportamientos dolosos. Ver FIGURA 1.1, FIGURA 1.2 y FIGURA 1.3.

FIGURA 1. 1
FRAUDE: CASO 1



Fuente: www.aach.cl/html/informacion/seminarios/fraude.asp
Elaborado por: HRoa

FIGURA 1. 2

FRAUDE: CASO 2

Fuente: www.aach.cl/html/informacion/seminarios/fraude.asp
Elaborado por: HRoa

FIGURA 1. 3

FRAUDE: CASO 3

Fuente: www.aach.cl/html/informacion/seminarios/fraude.asp
Elaborado por: HRoa

A pesar del creciente índice de fraude que las aseguradoras de vehículos están sufriendo en los últimos años, el diseño de métodos cuantitativos dirigidos al estudio de dichos factores o de las circunstancias comúnmente asociadas a la aparición de comportamientos deshonestos

ha sido hasta hace muy poco tiempo inexistente, no sólo dentro del marco nacional, sino también del internacional.

Incluso en países en los que las compañías aseguradoras tienen de algún modo un mecanismo destinado al control de fraude, cuantificar la influencia del comportamiento deshonesto de los asegurados en la siniestralidad declarada a la compañía no les es nada fácil, más aún, teniendo en cuenta la divergencia existente entre las cifras de sospecha de fraude y las cifras de fraudes realmente detectados, presentadas desde diferentes fuentes del sector. El hecho de que las primeras sean notablemente superiores a las segundas hace pensar que la investigación de siniestros realizada hasta la fecha por las entidades aseguradoras no ha sido realmente suficiente. Ante ello, cabe esperar que un determinado número de siniestros, clasificados de legítimos por la entidad, sean en realidad fraudes no detectados.

La alarma desencadenada dentro del sector asegurador en torno al problema y, la posibilidad de tratar el mismo desde un punto de vista aplicado, ha despertado el interés de los investigadores. Los últimos años han supuesto un giro de 90 grados y si hasta hace poco tiempo la literatura existente sobre el tema era prácticamente inexistente, hoy en día es posible encontrar estudios con modelos cada vez más exhaustivos

y sofisticados, dirigidos a validar los denominados *“indicadores de fraude”.*

No obstante, para poder aplicar cualquier método de detección de fraude es necesario que las compañías aseguradoras cuenten con la información suficiente y necesaria del siniestro, que les permita detectar la acción fraudulenta. Pero precisamente, el proceso de obtención de esta información lleva consigo el uso de recursos humanos y financieros así como también de tiempo, lo que dificulta determinar en un periodo reducido y sin mucho costo para la aseguradora si las declaraciones de los asegurados son honestas o no. Además este no es el único problema en lo que respecta a la obtención de información, tanto investigadores, ajustadores de seguros como los administradores de las demandas de seguros se enfrentan a menudo con situaciones donde hay información incompleta para tomar una decisión concerniente a la validez o posible estado fraudulento de un reporte de siniestro en particular. O muchas veces se cuenta con una cantidad sustancial de información sobre un determinado reporte, pero no se sabe cómo clasificarla de acuerdo a su validez. En cualquier circunstancia, algunas decisiones estratégicas y tácticas tienen que ser tomadas, como por ejemplo, si pagar la demanda del asegurado o enviar el archivo a una unidad investigativa especial o tal vez enviar el caso al departamento legal de la compañía. Para el caso de

las compañías de seguros de nuestro país les es más fácil pagar los siniestros reclamados por los asegurados que entrar en trámites legales, ya que tampoco en el país se cuenta con una unidad investigativa a la que puedan recurrir las aseguradoras para verificar la honestidad de los reportes de siniestros. Por todo esto, es que las aseguradoras deberían estar interesadas en buscar métodos que les permitan detectar y clasificar una demanda como fraudulenta o no.

Las técnicas estadísticas clásicas y modernas (análisis discriminante, regresión logística, análisis de redes neuronales, y otras) requieren que los investigadores de la compañía aseguradora o los ajustadores de las demandas utilicen información existente de un conjunto de demandas previamente clasificadas como fraudulentas o no fraudulentas para desarrollar un programa calificador para nuevas demandas, además se necesita que los datos estén en un intervalo numérico para el análisis estadístico. Por ejemplo, una aplicación típica de detección de fraude usando estadística clásica podría ser: usar datos sobre transacciones de tarjetas de crédito fraudulentas y no fraudulentas para desarrollar un modelo que podría permitir la clasificación de nuevas demandas de acuerdo a su posible probabilidad de ser fraudulentas.

Las técnicas más tradicionales como regresión, análisis discriminante y regresión logística, no son tan útiles cuando el costo de obtener información válida sobre la variable de criterio (fraude versus no fraude) es excesivo o cuando tal información es imposible de obtener desde un punto de vista práctico. Sin embargo están tomando auge nuevas técnicas estadísticas que permiten validar los denominados *indicadores de fraude* o *banderas rojas.*

Técnica estadística para validación de los indicadores de fraude.

Una de las técnicas que está permitiendo la validación de estos indicadores es el Análisis de Componentes Principales en conjunción con puntuaciones RIDIT's, conocido también como Análisis PRIDIT, el mismo que es una técnica no paramétrica más fácil de entender e implementar. Las suposiciones necesarias sobre la información para el desarrollo del análisis PRIDIT difieren de las requeridas en la regresión, regresión logística, análisis discriminante u otra metodología de clasificación, además estas suposiciones son menos estrictas que las de técnicas estadísticas tradicionales, las mismas que necesitan que la información esté previamente clasificada como demandas honestas o fraudulentas y que los datos se encuentren en un intervalo numérico para poder realizar cualquier tipo de análisis estadístico. En cambio, la metodología PRIDIT

no requiere un delineamiento de una muestra de casos en las cuales la ocurrencia de fraude junto con sus covarianzas, sean conocidas así como tampoco se necesita conocer el delineamiento de una muestra de casos en las cuales se deba tener la ausencia de fraude junto con sus covarianzas.

El contexto de los indicadores de fraude en el seguro de automóviles está a menudo ligado a contestaciones binarias en las declaraciones de las demandas, con una respuesta positiva que indica una creciente probabilidad de fraude. A través de un algoritmo computacional todos los archivos de las demandas pueden ser ordenados de forma decreciente de acuerdo a los niveles de sospecha de fraude y, si se desea, pueden ser asignados a un grupo (fraudulentos/no fraudulentos) en base a su puntuación. Simultáneamente y con igual importancia, se puede obtener una medida del poder discriminatorio de la variable predictora. Esta medida es de hecho una transformación del método de puntuación RIDIT introducida por Bross en 1.958 para estudios epidemiológicos.

Toda esta evaluación es hecha únicamente basándonos en los datos existentes, sin necesidad de contratar costosos ajustadores o investigadores para examinar todas las demandas, calificarlas y ordenarlas de acuerdo a su nivel de fraude. Así, se puede lograr ahorros

significativos como consistencia interna por las empresas si deciden aplicar esta nueva metodología.

Entre las características de las variables que van a permitir la validación de los indicadores de fraude tenemos: características del accidente, características de la demanda, características del asegurado y características del vehículo.

CAPÍTULO 2

2. COBRO DE SINIESTROS EN SEGUROS DE VEHÍCULOS DURANTE LA ÚLTIMA DÉCADA EN EL ECUADOR

2.1. Introducción

En el presente capítulo se pretende dar a conocer los antecedentes del cobro de siniestros en los seguros de automóviles en nuestro país, Ecuador, desde 1.993 hasta 2.003; así como también la situación actual de este tipo de seguro frente a los demás ramos de seguros del mercado ecuatoriano.

Durante la década (1.993 – 2003) en el Ecuador, muchas de las aseguradoras que ofrecen seguros para automóviles se percataron de que el monto de indemnización pagado a sus asegurados por causa de un siniestro específico era muy alto. El mercado asegurador del Ecuador durante esta última década ha visto en el ramo de seguros de vehículos,

el ramo con mayor monto pagado por siniestros en lo que respecta a los demás ramos de seguros.

La Superintendencia de Bancos y Seguros del Ecuador dispone de la información de los montos totales de siniestros pagados de todas las aseguradoras en cada uno de los ramos. En lo que respecta al seguro de vehículos, que es el que es objeto de esta investigación, la Superintendencia de Bancos y Seguros del Ecuador posee sólo información general de los montos pagados por todas las aseguradoras, no dispone de la información de los montos pagados por siniestros desglosada por tipos de siniestros, tales como: robo total del vehículo, daños a terceros, pérdidas parciales, etc. Esto nos hace pensar que en el Ecuador es difícil hacer un estudio como el que se propone por la falta de información fiable y consistente o por la excesiva agregación de la poca información disponible.

2.2. Información Contable de Siniestros Pagados en Seguros de Vehículos de los Archivos de la Superintendencia de Bancos.

En base a la información proporcionada por la Superintendencia de Bancos y Seguros se puede constatar que el monto total pagado por siniestros de vehículos a Diciembre de 1.993 (en miles de dólares)

alcanzó la cifra de $ 28,058 y continúo creciendo notablemente hasta Diciembre de 1.998, fecha en la cual las aseguradoras tuvieron que pagar un monto total de $ 59,907 miles de dólares; a Diciembre de 1.999[(1)] hubo cierta disminución en el monto de pago de siniestros, pero para el período del 2000[(2)], se vuelve a tener una reducción de estos montos pagados pero en pequeña proporción; sin embargo esta disminución es sólo visual porque en términos monetarios los montos pagados en el 2.000 representan un valor mayor debido a la adopción del dólar como moneda oficial. Desde este período 2.000 el monto total de siniestros pagados hasta el 2003 continuó con una tendencia creciente. Ver CUADRO 2.1 y GRÁFICO 2.1

---

(1) Año 1.999, período hasta el cual los montos totales de pagos de siniestros se hacían en sucres, por lo que para presentarlos en dólares se ha realizado la conversión respectiva de acuerdo a la inflación de cada año.

(2) Año 2.000, período en el cual el Ecuador adoptó como moneda oficial el dólar de los Estados Unidos de América, fijando el valor de un dólar en 25.000 sucres.

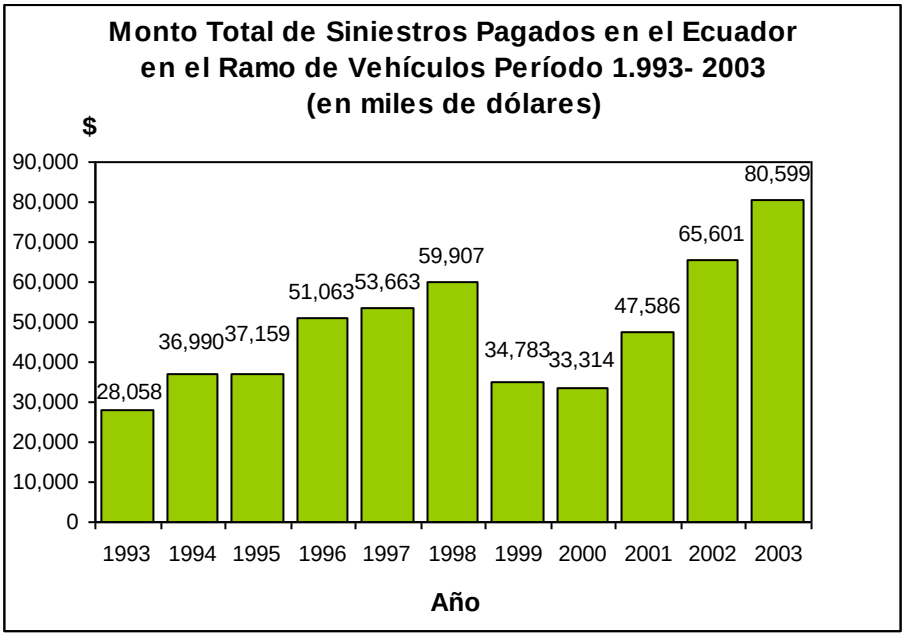
CUADRO 2. 1

Monto Total de Indemnización Liquidada por Siniestros Vehiculares para el Período 1.993 – 2.003

	Indemnización Pagada (en miles de dólares)
Dec-93	28,058
Dec-94	36,990
Dec-95	37,159
Dec-96	51,063
Dec-97	53,663
Dec-98	59,907
Dec-99	34,783
Dec-00	33,314
Dec-01	47,586
Dec-02	65,601
Dec-03	80,599

Fuente: Superintendencia de Bancos y Seguros del Ecuador
Elaborado por: H. Roa

GRÁFICO 2. 1



Elaboración: H. Roa

En el GRÁFICO 2.1 se puede visualizar la tendencia creciente de los pagos de siniestros a los asegurados desde 1.993 hasta 1.998, notándose claramente la reducción de los pagos en el año 1.999 y 2.000 por los temores económicos que en ese período se suscitaron; sin embargo a partir del 2.001 vuelve la tendencia creciente, una vez adoptado el dólar como moneda oficial.

La variación porcentual corriente del monto de siniestros pagados a Diciembre de 1.994 creció en 24.15 % en relación al año anterior 1.993. Entre el período de 1.994 y 1.995 no se registró mayor aumento de

siniestros liquidados, apenas aumentó un 0.45 % de un año a otro. Sin embargo, al siguiente año, en 1.996 se pagó un 27.23 % más que en 1.995. En 1.997 las aseguradoras pagaron 4.85 % más que el año anterior, así mismo a Diciembre de 1.998 se registró un aumento de siniestros liquidados de 10.42 % con respecto a 1.997. Pero entre el período de 1.998 y 1.999 hubo una disminución considerable del pago de indemnizaciones, el monto de siniestros liquidados se redujo en 72.23 %. A Diciembre del 2000 se registró un pago del 4.22 % menos que el registrado en 1.999, esto también se debe a que para este período el país tuvo que dolarizarse. En 2.001, el mercado asegurador vuelve a tener una variación creciente, tal es el caso que pagó 29.99 % con respecto al año anterior. A Diciembre de 2002, se produjo un aumento del 27.46 % más que lo registrado en el 2001. En el 2.003 se registró un 18.61% más en el pago total de siniestros que el realizado en el 2.002. Ver TABLA I y GRAFICO 2.2

TABLA I

Variación Porcentual Corriente

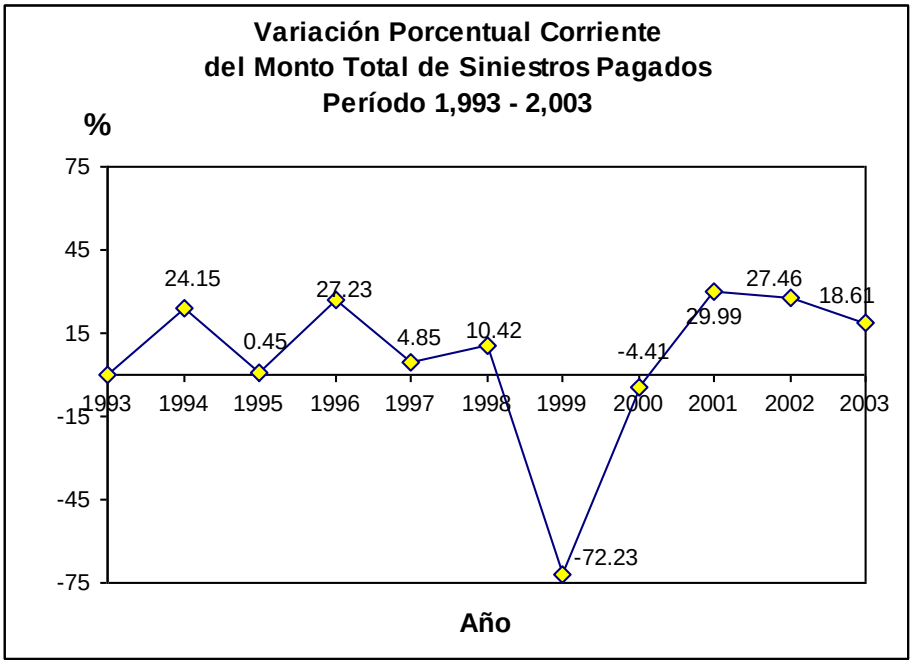
Año a Año para el Período 1.993 – 2.003

	Variación % Corriente
Dec-93	-
Dec-94	24.15
Dec-95	0.45
Dec-96	27.23
Dec-97	4.85
Dec-98	10.42
Dec-99	-72.23
Dec-00	-4.41
Dec-01	29.99
Dec-02	27.46
Dec-03	18.61

Elaborado por: H. Roa

En el GRÁFICO 2.2 se puede apreciar el aumento año a año del total de siniestros pagados por las aseguradoras del país, a excepción del año 1.999, en el cual el mercado asegurador tuve una reducción sustancial de estos montos lo cual se debió a la poca demanda de pólizas de seguros de automóviles en ese año por la inestabilidad económica sufrida en este lapso. Luego, nuevamente a partir del año 2.001 se vuelve a apreciar variaciones crecientes en relación de un año a otro.

GRÁFICO 2. 2



Elaborado por: H. Roa

2.3. Cálculo de Índices de los Montos Totales de Siniestros Pagados

Un índice es una medida estadística que tiene la propiedad de informar de los cambios de valor que experimenta una variable o magnitud en dos situaciones, una de las cuales se toma como referencia. La comparación suele hacerse por cociente. A la situación inicial se le llama período base y a la situación que queremos comparar período actual o corriente.

Para reflejar con precisión la evolución de los montos de siniestros pagados por las aseguradoras se ha procedido a calcular los números

índices con año base 1.993 = 100. La conversión de los datos a índices facilita la estimación de la tendencia en una serie compuesta por números muy grandes con se está manejando (miles de dólares).

Para determinar los índices de nuestra serie de datos se utiliza la siguiente fórmula:

$$I = \frac{X_t}{X_0}$$

Donde:

X_t es el monto del siniestro pagado en el período dado.

X_0 es el monto del siniestro pagado en el período base.

Los índices calculados se los ha determinado tomando como período base el año 1.993 = 100. Ver TABLA II.

TABLA II

Monto Total de Indemnización Pagada e Índices

con Año Base 1.993 = 100

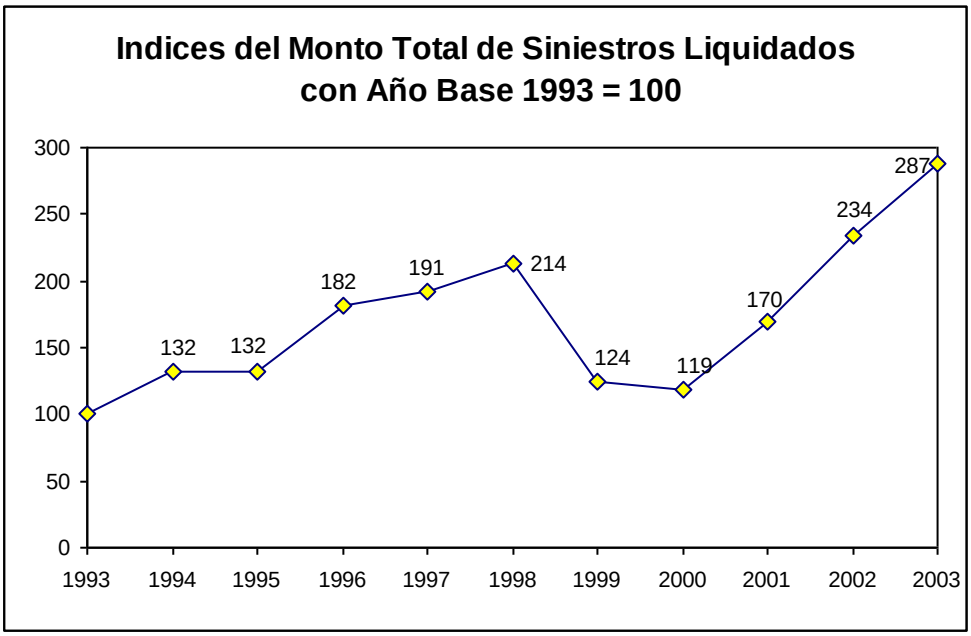
	Indemnización Pagada (en miles de dólares)	**Ajustado a 100 con Año Base 1.993**
Dec-93	28,058	100
Dec-94	36,990	132
Dec-95	37,159	132
Dec-96	51,063	182
Dec-97	53,663	191
Dec-98	59,907	214
Dec-99	34,783	124
Dec-00	33,314	119
Dec-01	47,586	170
Dec-02	65,601	234
Dec-03	80,599	287

Elaborado por: H. Roa

De acuerdo a los índices obtenidos con período base 1.993 = 100, podemos determinar con mejor claridad la evolución del pago de seguros por siniestros de vehículos que las compañías aseguradoras han venido pagando desde 1.993 hasta el 2003. Basados en los datos obtenidos, 2.003 ha sido hasta el momento el año que mayor variación ha reflejado con respecto a 1.993 en el pago de siniestros con 187 puntos más.

Como se puede apreciar en el GRÁFICO 2.3, la tendencia del pago de siniestros vehiculares siempre ha sido creciente con respecto a 1.993, lo que lleva a las compañías aseguradoras a exigir mayor atención a sus departamentos de siniestros. La correcta administración del departamento de siniestros es fundamental para el resultado de una compañía de seguros. El desorden administrativo podría obligar al asegurador a pagar siniestros que no debiera.

GRÁFICO 2. 3



Elaborado por: H. Roa

Actualmente en nuestro país, Ecuador, existen 40 compañías dedicadas al mercado asegurador, de las cuales 27 ofrecen pólizas de seguros para vehículos. El seguro de vehículos dentro de este mercado se encuentra clasificado como un seguro perteneciente al ramo de Seguros Generales. Haciendo referencia al año 2.003, que es el último año con el que se cuenta de información completa, además de ser este año el que mayor variación ha experimentado desde hace una década, podemos realizar una comparación entre los montos pagados por siniestros de vehículos y los montos pagados por siniestros de los otros segmentos de realización en ese mismo ramo de Seguros Generales y darnos cuenta que también aquí la mayor preocupación de las aseguradoras se centra en el seguro de vehículos puesto que es el que mayor gasto ocasiona a las aseguradoras, a quienes les toca desembolsar pagos importantes. Y hasta el primer trimestre del 2.004 se tiene una tendencia igual que la del 2.003. Ver CUADRO 2.4

CUADRO 2. 2
Monto y Porcentaje de Siniestros Pagados por Segmento de Realización en el Ramo de Seguros Generales

	MONTO DE SINIESTROS PAGADOS (miles de dólares)		PORCENTAJE DE SINIESTROS PAGADOS	
SEGMENTO DE REALIZACIÓN	**Dec-03**	**Mar-04**	**% Dic-03**	**% Mar-04**
ENFERMEDAD O ASIST. MED.CONSOLIDADO	15,898	4,295	8.77	9.27
ACCIDENTES PERSONALES CONSOLIDADO	7,401	1,832	4.08	3.95
INCENDIO Y LINEAS ALIADAS	9,587	2,703	5.29	5.83
LUCRO CESANTE CONSOLIDADO (1)	501	108	0.28	0.23
RIESGO CATASTROFICO	2,672	677	1.47	1.46
VEHICULOS	**80,599**	**19,991**	**44.46**	**43.14**
TRANSPORTE	18,761	3,950	10.35	8.52
DINERO Y VALORES	453	787	0.25	1.70
CASCO DE BUQUES	4,515	1,235	2.49	2.67
CASCO AEREO	7,617	1,232	4.20	2.66
ROBO	4,173	1,138	2.30	2.46
AGROPECUARIO	213	87	0.12	0.19
RIESGOS TECNICOS (2)	15,850	5,105	8.74	11.02
RESPONSABILIDAD CIVIL	3,742	1,528	2.06	3.30
FIDELIDAD Y BBB	2,603	783	1.44	1.69
MULTIRIES. O SEG. COMBINADO	20	3	0.01	0.01
FIANZAS (3)	6,342	836	3.50	1.80
RIESGOS DIVERSOS	347	49	0.19	0.11
TOTAL RAMOS DE GENERALES	**181,294**	**46,339**	**100.00**	**100.00**

Fuente: Superintendencia de Bancos y Seguros del Ecuador.
Elaborado por: H. Roa

Notoriamente en el CUADRO 2.2 podemos observar que tanto los montos de siniestros pagados en el segmento de vehículos del ramo de Seguros Generales como el porcentaje del mismo, representa la mayor proporción con respecto a los demás segmentos de este ramo. Para el año 2.003, el pago de siniestros de vehículos representó un 44.46% del total de pagos que se hizo por el ramo en general. Y hasta el primer trimestre de este

año (2.004), este segmento representa el 43.14% de los pagos totales del ramo de Seguros Generales.

Para tener una visión no sólo desde el punto de vista de lo que representa el pago de siniestros de vehículos en el ramo al que pertenece sino de lo que representa en el mercado asegurador en general se puede hacer una comparación entre el monto total de lo que paga el mercado asegurador por siniestros en todos sus ramos contra lo que se paga a los asegurados por siniestros vehiculares y también descubrir que es éste, el que más gasto representa. Ver CUADRO 2.3

CUADRO 2. 3

Porcentaje por Segmento de Realización

de Siniestros Pagados

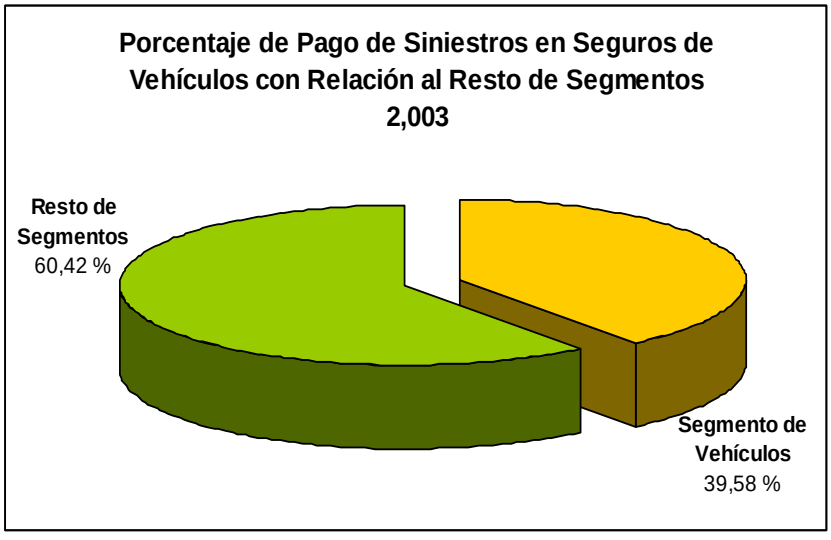
	PORCENTAJE DE SINIESTROS PAGADOS	
	% Dic-03	% Mar-04
VIDA INDIVIDUAL	0.72	0.11
VIDA GRUPO	10.22	11.78
RENTA VITALICIA	0.03	0.03
ENFERMEDAD O ASISTENCIA MEDICA CONSOLIDADO	7.81	8.16
ACCIDENTES PERSONALES CONSOLIDADO	3.63	3.48
INCENDIO Y LINEAS ALIADAS	4.71	5.14
LUCRO CESANTE CONSOLIDADO (1)	0.25	0.21
RIESGO CATASTROFICO	1.31	1.29
VEHICULOS	39.58	38.00
TRANSPORTE	9.21	7.51
DINERO Y VALORES	0.22	1.50
CASCO DE BUQUES	2.22	2.35
CASCO AEREO	3.74	2.34
ROBO	2.05	2.16
AGROPECUARIO	0.10	0.17
RIESGOS TECNICOS (2)	7.78	9.70
RESPONSABILIDAD CIVIL	1.84	2.90
FIDELIDAD Y BBB	1.28	1.49
MULTIRIES. O SEG. COMBINADO	0.01	0.01
FIANZAS (3)	3.11	1.59
RIESGOS DIVERSOS	0.17	0.09
TOTAL RAMOS DE VIDA	**10.97**	**11.92**
TOTAL RAMOS DE GENERALES	**89.03**	**88.08**
TOTAL	**100**	**100**

Fuente: Superintendencia de Bancos y Seguros del Ecuador
Elaborado por: H. Roa

Con información del último año (2.003) y con la información del primer trimestre del año en curso (2.004) que la Superintendencia de Bancos y Seguros del Ecuador tiene en sus registros sobre los siniestros pagados

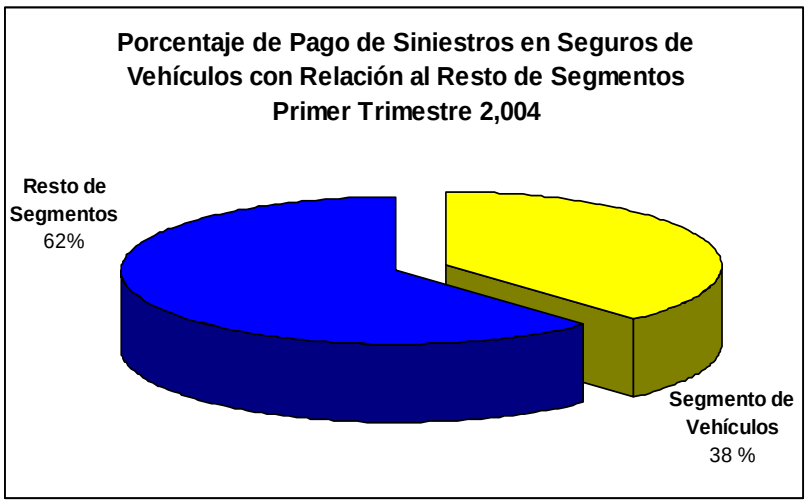
por las compañías aseguradoras, se ha determinado la representación, en porcentaje, de cada segmento de realización y se ha obtenido que el segmento de vehículos es el que mayor porcentaje representa en pagos de siniestros con relación a los demás segmentos. En el 2.003, un 39.58% del pago total de siniestros que realizó el mercado asegurador correspondió a pagos por siniestros de vehículos (Ver GRÁFICO 2.4); mientras que en el último trimestre del 2.004 del monto total pagado por siniestros se tiene que el 38% ha sido hecho para pagos de siniestros de vehículos. Ver GRÁFICO 2.5

GRÁFICO 2. 4



Elaborado por: H. Roa

GRÁFICO 2. 5



Elaborado por: H. Roa

Esta información nos permite vislumbrar que el pago de siniestros de seguros vehiculares es un segmento que tiene no sólo la mayor representación en el ramo al que pertenece, que es el de Seguros Generales, sino que también dentro de todo el mercado asegurador es el que mayor porcentaje tiene en el pago de siniestros.

2.4. Datos de Seguros Equinoccial del Ecuador sobre Pagos de Siniestros en Seguros Vehiculares y Desglose por Causas

Seguros Equinoccial es una de las compañías aseguradoras del país, mayormente involucrada con el seguro de vehículos. Durante la última década, esta compañía ha sido una de las que mayores pagos ha realizado a sus clientes por siniestros vehiculares, es por esto que gracias a la información que poseen en sus bases de datos podemos encontrar los desgloses de los pagos por siniestros pero a su vez divididos por las causas por las que han sido cobrados, entre las cuales están: daño parcial, daño total, robo parcial y robo total. La información con la que se cuenta es con los montos pagados por siniestros vehiculares en cada una de sus sucursales desde el 2.000 hasta el 2.003. (Véase CUADRO 2.4).

CUADRO 2. 4

Montos Pagados por Siniestros Vehiculares
por Seguros Equinoccial
(Período 2.000 – 2.003)

		VALOR PAGADO (en dólares)			
Sucursal	**Causas**	**2000**	**2001**	**2002**	**2003**
CUENCA	DAÑO PARCIAL	51,048	58,949	129,634	87,864
	DAÑO TOTAL	20,954	27,404	42,029	81,016
	ROBO PARCIAL	5,189	3,016	6,668	13,443
	ROBO TOTAL	20,580	9,941	8,730	32,899
Total CUENCA		**97,770**	**99,310**	**187,061**	**215,223**
GUAYAQUIL	DAÑO PARCIAL	109,411	161,996	215,893	539,112
	DAÑO TOTAL	51,080	75,260	119,330	152,875
	ROBO PARCIAL	40,947	23,149	25,525	73,717
	ROBO TOTAL	123,591	67,015	40,452	43,817
Total GUAYAQUIL		**325,030**	**327,420**	**401,200**	**809,522**
QUITO	DAÑO PARCIAL	615,191	971,582	1,531,313	1,997,830
	DAÑO TOTAL	163,257	425,845	1,105,290	1,093,183
	ROBO PARCIAL	180,119	247,309	292,380	422,735
	ROBO TOTAL	258,075	312,623	333,322	244,809
Total QUITO		**1,216,643**	**1,957,359**	**3,262,305**	**3,758,557**
TOTAL		**1,639,443**	**2,384,088**	**3,850,565**	**4,783,301**

Fuente: Base de Datos Seguros Equinoccial del Ecuador
Elaborado por: H. Roa

Así mismo, la compañía de seguros tiene estadísticas de cada una de sus sucursales de los siniestros que los asegurados han reportado para

el cobro de los mismos. Estas estadísticas van desde el 2.000 hasta Diciembre del 2.003. Véase CUADRO 2.5

CUADRO 2. 5
Número de Siniestros Vehiculares Pagados por Seguros Equinoccial
(Período 2.000 – 2.003)

		Número de Siniestros			
Sucursal	**Causas**	**2000**	**2001**	**2002**	**2003**
CUENCA	DAÑO PARCIAL	87	75	135	110
	DAÑO TOTAL	2	2	4	4
	ROBO PARCIAL	25	16	16	29
	ROBO TOTAL	11	4	2	10
Total CUENCA		**125**	**97**	**157**	**153**
GUAYAQUIL	DAÑO PARCIAL	104	103	199	407
	DAÑO TOTAL	4	7	10	16
	ROBO PARCIAL	35	35	40	98
	ROBO TOTAL	16	12	13	17
Total GUAYAQUIL		**159**	**157**	**262**	**538**
QUITO	DAÑO PARCIAL	890	967	1529	2033
	DAÑO TOTAL	7	39	91	95
	ROBO PARCIAL	484	508	550	647
	ROBO TOTAL	64	62	53	45
Total QUITO		**1445**	**1576**	**2223**	**2820**
TOTAL		**1729**	**1830**	**2642**	**3511**

Fuente: Base de Datos Seguros Equinoccial del Ecuador
Elaborado por: H. Roa

Por la información presentada se puede concluir que los siniestros de vehículos son el punto fijo de investigación por parte de los departamentos de siniestros, quienes tendrán de ahora en adelante la tarea de buscar e implementar nuevas técnicas de detección de fraude en las declaraciones de siniestros.

CAPÍTULO 3

3. ANÁLISIS DE COMPONENTES PRINCIPALES

3.1. Introducción

A fin de examinar las relaciones entre un conjunto de *p* variables correlacionadas, es útil transformar el conjunto original de variables en un nuevo conjunto de variables no correlacionadas, llamadas *componentes principales*. Estas nuevas variables son combinaciones lineales de las variables originales y se obtienen en orden decreciente de importancia, de manera que la primera componente principal explique tanta variación en los datos originales como sea posible.

La técnica para encontrar esta transformación es llamada análisis de componentes principales (abreviada ACP). Es una técnica *dirigida por las variables* que resulta adecuada cuando las variables surgen 'sobre un fundamento igual' de manera que, por ejemplo, no se tiene una variable dependiente y varias variables explicativas como sucede en regresión múltiple.

Algunas de las razones para considerar el análisis de componentes principales son:

Cribado de datos

Los análisis de seguimiento sobre las componentes principales son útiles para comprobar las hipótesis que el investigador podría establecer acerca de un conjunto de datos multivariados y para identificar posibles datos atípicos en el conjunto.

Agrupación

El análisis de componentes principales también es útil cuando el investigador desea agrupar las unidades experimentales en subgrupos de tipos semejantes.

Análisis Discriminante

El análisis de componentes principales revela que unas cuantas componentes principales se tiene casi toda la información contenida en las variables originales. Obteniendo los valores para las componentes principales para cada unidad experimental y usando estas nuevas variables como variables de entrada en una función de análisis

discriminante se puede producir una regla de discriminación para clasificar las observaciones.

Regresión

Se sabe que la regresión múltiple puede ser inexacta y conducir a resultados dudosos cuando las variables predictoras están altamente correlacionadas (multicolinealidad). El análisis de componentes principales puede ayudar a determinar si ocurre multicolinealidad entre las variables predictoras.

3.2. Objetivos del Análisis de Componentes Principales

Esta técnica debe usarse principalmente como una técnica exploratoria y debe ayudar a los investigadores a que adquieran cierta percepción respecto a un conjunto de datos. Los objetivos principales del análisis de componentes principales son:

1. Reducir la dimensionalidad del conjunto de datos.
2. Identificar nuevas variables subyacentes.

Con respecto al primer objetivo, lo que en realidad se está intentando hacer es descubrir la verdadera dimensionalidad de los datos. Se puede usar el ACP para determinar la dimensionalidad real de los datos y,

cuando esa dimensionalidad es menor que *p*, las variables originales se pueden reemplazar por un número menor de variables subyacentes, sin que se pierda información.

Con respecto al segundo objetivo, desafortunadamente no se puede garantizar que las nuevas variables sean significativas, de hecho lo común es no esperar que se puedan interpretar las variables componentes principales. Sin embargo, es importante no perder de vista que un ACP es muy útil sin importar si se pueden interpretar estas componentes.

3.3. Obtención de las Componentes Principales

Las nuevas variables componentes principales deben ser tales que:

1. No estén correlacionadas.
2. La primera componente principal explique tanto de la variabilidad de los datos como sea posible.
3. Cada componente subsiguiente tome en cuenta tanto de la variabilidad restante como sea posible.

Supongamos que $\mathbf{X}^T = [X_1, ..., X_p]$ es una variable aleatoria *p*-dimensional con media μ y matriz de covarianzas $\sum$. El problema es encontrar un

nuevo conjunto de variables, digamos Y_1, Y_2, ...,Y_p, las cuales son no correlacionadas y cuyas varianzas son decrecientes de la primera a la última. Cada Y_j será una combinación lineal de las X's, de manera que:

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + ... + a_{pj}X_p$$

$$Y_j = \mathbf{a}_j^T \mathbf{X} \qquad (1)$$

Donde:

$\mathbf{a}_j^T = [a_{1j,...,} a_{pj}]$ es un vector de constantes.

La ecuación (1) contiene un factor de escala arbitrario. Por lo tanto, imponemos la condición $\mathbf{a}_j^T \mathbf{a}_j = \sum_{k=1}^{p} a_{kj}^2 = 1$. Esta normalización asegura que las distancias en el *p*-espacio se preservan.

El primer componente principal, Y_1, se encuentra eligiendo $\mathbf{a}_1$ de manera tal que la varianza de Y_1 se maximiza. En otras palabras, se elige $\mathbf{a}_1$ de manera tal que se maximice la varianza de $\mathbf{a}_1^T \mathbf{X}$ sujeta a la condición $\mathbf{a}_1^T \mathbf{a}_1$ =1.

Se puede demostrar que el valor máximo de la varianza de $\mathbf{a}_1^T\mathbf{X}$ entre todos los vectores $\mathbf{a}_1$ que satisfacen $\mathbf{a}_1^T\mathbf{a}_1$=1 es igual a λ_1, el eigenvalor más grande de $\sum$, y que este máximo ocurre cuando $\mathbf{a}_1$ es un eigenvector de $\sum$ correspondiente al eigenvalor λ_1.

La segunda componente principal, Y_2, se encuentra eligiendo $\mathbf{a}_2$ de tal manera que Y_2 tenga la mayor varianza posible para todas las combinaciones de la forma de la ecuación (1), las cuales no están correlacionadas con Y_1. Es decir, $\mathbf{a}_2$ se elige de modo que la varianza de $\mathbf{a}_2^T\mathbf{X}$ sea un máximo entre todas las combinaciones lineales de **X** que no están correlacionadas con la primera variable componente principal y tenga $\mathbf{a}_2^T\mathbf{a}_2$=1.

Se puede demostrar que el máximo indicado arriba es igual a λ_2, el segundo eigenvalor más grande de $\sum$, y que este máximo ocurre cuando $\mathbf{a}_2$ es un eigenvector de $\sum$ correspondiente al eigenvalor λ_2.

De manera similar, pueden definirse las componentes principales restantes $Y_3,...,Y_p$. La j-ésima componente principal (j = 3,4, ..., p) se expresa por $\mathbf{a}_j^T\mathbf{X}$ en donde $\mathbf{a}_j$ se elige de modo que $\mathbf{a}_j^T\mathbf{a}_j$=1 y de forma que la varianza de $\mathbf{a}_j^T\mathbf{X}$ sea un máximo entre todas esas combinaciones

lineales de **X** que no estén correlacionadas con las *j*-1 componentes principales restantes. Es posible demostrar que este máximo es igual a λ_j, el *j*-ésimo eigenvalor más grande de $\sum$ correspondiente al eigenvalor λ_j y que satisface $\mathbf{a}_j^T \mathbf{a}_j = 1$.

De esta manera $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ denotan los eigenvalores ordenados de $\sum$ y $\mathbf{a}_1, \mathbf{a}_2, .., \mathbf{a}_p$ denotan los eigenvectores normalizados correspondientes. Nótese que los *p* eigenvalores de $\sum$ deben ser todos no negativos debido a que $\sum$ es semidefinida positiva.

3.4. Varianza Explicada

Denotemos por *A* a la matriz *p*×*p* de eigenvectores, donde:

$$A = [\mathbf{a}_1, ..., \mathbf{a}_p]$$

y al vector de *p*×1 de componentes principales por **Y**. Entonces:

$$\mathrm{Y} = A^T \mathbf{X} \qquad (2)$$

La matriz de covarianzas de **Y** se denotará por Λ y está dada por:

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & . & . & . & 0 \\ 0 & \lambda_2 & . & . & . & 0 \\ . & & & & & \\ . & & & & & \\ . & & & & & \\ 0 & & . & . & . & \lambda_p \end{bmatrix}$$

Nótese que la matriz es diagonal debido a que los componentes se han elegido de manera que no estén correlacionados. Es posible notar que los eigenvalores pueden interpretarse como las respectivas varianzas de los distintos componentes. Recordemos que tr($\sum$) = $\sigma_{11} + \sigma_{22} + ... + \sigma_{pp}$. Por lo tanto tr($\sum$), en cierto sentido, mide la variación total en las variables originales. Ahora bien, la suma de las varianzas de los componentes está dada por:

$$\sum_{i=1}^{p} \mathrm{Var}(Y_i) = \sum_{i=1}^{p} \lambda_i = \mathrm{tr}(\Lambda)$$

Sin embargo, puede demostrarse que:

$$\text{tr}(\Lambda) = \text{tr}(\Sigma) = \sum_{i=1}^{p} \text{Var}(X_i)$$

Así pues, tenemos un resultado importante: la suma de las varianzas de las variables originales y las de sus componentes principales son iguales. En otras palabras, la variación total explicada por las variables componentes principales es igual a la cantidad total de la variación medida por las variables originales.

Por lo tanto, tenemos que el *i*-ésimo componente principal explica una proporción $\lambda_i / \sum_{j=1}^{p} \lambda_j$ de la variación total en los datos originales. También decimos que los primeros *m* componentes explican una proporción $\sum_{j=1}^{m} \lambda_j / \sum_{j=1}^{p} \lambda_j$ de la variación total.

3.5. Calificaciones de los componentes principales

A fin de utilizar las variables componentes principales en análisis estadísticos subsecuentes, es necesario calcular las calificaciones de tales componentes para cada unidad experimental en el conjunto de

datos. Estas calificaciones proporcionan las ubicaciones de las observaciones en un conjunto de datos con respecto a sus ejes componentes principales.

Sea $\mathbf{x}_r$ el vector de variables medidas para la *r*-ésima unidad experimental. Entonces el valor (calificación) de la *j*-ésima variable componente principal, para la *r*-ésima unidad experimental estará dada por:

$$y_{rj} = \mathbf{a}_j^T \mathbf{x}_r$$

para $j = 1, 2, ..., p$ y $r = 1, 2, ..., n$

(*n* es el número de unidades experimentales)

3.6. Estimación de Componentes Principales

La obtención de los componentes principales de **X** descrita en la sección 3.3 supone que $\sum$ es conocida. Difícilmente ésta será la situación. Por consiguiente, $\sum$ será reemplazada por *S*, la matriz de correlaciones muestral.

La obtención de las componentes principales de **X** utilizando las varianzas y covarianzas muestrales se llevará a cabo de la misma manera que la descrita en la sección 3.3. Las componentes principales serán los eigenvectores de *S*.

Denotaremos los eigenvectores de *S*, en orden descendente, por $\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_p$, y los eigenvectores correspondientes por $\hat{\mathbf{a}}_1, \hat{\mathbf{a}}_2, ..., \hat{\mathbf{a}}_p$. Debido a que *S* es semidefinida positiva, sus eigenvalores serán todos no negativos y representarán las varianzas estimadas de las distintas componentes.

Si nuestra muestra de 'individuos' es una muestra aleatoria de una población más grande, entonces $\{\hat{\lambda}_i\}$ y $\{\hat{\mathbf{a}}_i\}$ pueden considerarse como estimadores de los eigenvalores y eigenvectores de $\sum$, proporcionando estimadores de las componentes principales de **X**. Sin embargo, no se realizan supuestos acerca de la población subyacente, lo cual hace imposible derivar las propiedades muestrales de los estimadores.

La tendencia moderna es visualizar el ACP como una técnica matemática que no hace uso de un modelo estadístico subyacente. Los componentes

principales obtenidos a partir de la matriz de correlaciones muestral *S* son vistos como *los* componentes principales y no como estimadores de las cantidades correspondientes obtenidas a partir de $\sum$. Los 'sombreritos' sobre λ_i y $\mathbf{a}_i$ se omiten con frecuencia. De hecho ni siquiera es necesario considerar a **X** y **Y** como variables aleatorias.

Cuando se trabaja con los datos muestrales, las calificaciones de componentes principales se estiman por:

$$y_{rj} = \hat{\mathbf{a}}_j \mathbf{x}_r$$

para j =1, 2, ..., p y r = 1, 2, ..., n

3.7. Determinación del Número de Componentes Principales

Cuando se lleva a cabo un ACP, es necesario determinar la dimensionalidad real del espacio en el que contiene a los datos; es decir, el número de componentes principales que tienen varianzas mayores que cero.

Existen dos métodos que ayudan a elegir el número de componentes principales que deben usarse cuando se está aplicando el ACP a *S*.

Ambos se basan en los eigenvalores de *S*. Sea *d* la dimensionalidad del espacio en el cual se encuentran en realidad los datos.

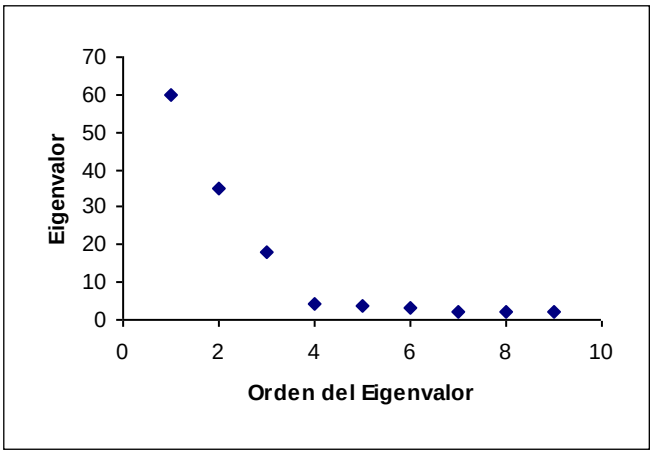
3.7.1. Método 1

Supongamos que se desea tomar en cuenta $\gamma 100\%$ de la variabilidad total en las variables originales. En este método para estimar *d* se considera $V = \sum_{j=1}^{k} \lambda_j / \sum_{j=1}^{p} \lambda_j$ para valores sucesivos de *k* = 1, ..., *p*. Entonces *d* se estima por el menor de los valores de *k* en el que, por primera vez, *V* sobrepasa $\gamma 100\%$. Para datos del tipo de laboratorio puede resultar bastante fácil explicar más de 95% de la variabilidad total con sólo dos o tres componentes principales. Por otro lado, para datos del 'tipo de individuos' es posible que se requieran cinco o seis componentes principales para explicar más de 70 al 75% de la variación total. Desafortunadamente, entre más componentes principales se requieran, menos útil se vuelve cada una de ellas.

3.7.2. Método 2

En este método se utiliza una gráfica SCREE (Score de los eigenvalores). Ésta se construye graficando las parejas $(1, \hat{\lambda}_1)$, $(2, \hat{\lambda}_2)$,..., $(p, \hat{\lambda}_p)$. Cuando los puntos de la gráfica tienden a nivelarse, estos eigenvalores suelen estar suficientemente cercanos a cero como para que puedan ignorarse. Es probable que los más pequeños no estén midiendo otra cosa sino ruido aleatorio. Por lo tanto, en este método se supone que la dimensionalidad del espacio de los datos es la que corresponde al orden del eigenvalor grande más pequeño. En la figura 3.1 se muestra un ejemplo de una gráfica SCREE, la cual sugeriría que la dimensionalidad real del espacio en el que se encuentran los datos es tres y, en consecuencia, el número apropiado de componentes principales que deben usarse también es tres.

GRÁFICO 3. 1
Score de Eigenvalores



Elaborado por: H. Roa

3.8. Resultados Útiles sobre ACP

3.8.1. Eigenvalores iguales a cero

Si algunas de las variables originales son linealmente dependientes, entonces algunos de los eigenvalores de $\sum$ serán iguales a cero. La dimensión del espacio que contiene a las observaciones es igual al rango de $\sum$, y este está dado por (*p* - número de eigenvalores iguales a cero). Si existen *k* eigenvalores iguales a cero, entonces podemos encontrar *k* restricciones lineales independientes en las variables. Estas restricciones en ocasiones son llamadas *relaciones estructurales*. En otras palabras, los eigenvectores que

corresponden a los eigenvalores iguales a cero definen las relaciones estructurales entre las variables originales.

Las ecuaciones de relaciones estructurales proporcionan información acerca de cómo están relacionadas las variables entre sí, pero no son útiles para el cribado de datos multivariados. Sin embargo sí son útiles para examinar las relaciones de multicolinealidad entre un conjunto de variables predictoras en los problemas de regresión.

Si tenemos una muestra proveniente de una población, entonces una dependencia lineal exacta en la población también existirá en la muestra. Si observáramos los eigenvalores de *S*, los estimadores muestrales de los eigenvalores iguales a cero de $\sum$ también deben ser cero.

La ocurrencia de dependencia lineal exacta no es común en la práctica. Un problema práctico más importante es detectar dependencia lineal aproximada. Los componentes principales correspondientes a eigenvalores pequeños pueden tomarse como los estimadores de las relaciones lineales subyacentes.

3.8.2. Vectores de Carga de Componentes

Recordemos que los eigenvectores de $\sum$ que se están usando para definir componentes principales se normalizan para tener longitud 1; esto es, $\mathbf{a}_j^T\mathbf{a}_j$=1 para j = 1,...,*p*. Este hecho puede crear confusión cuando se está tratando de interpretar componentes principales mediante el examen de los elementos de los eigenvectores que definen esas componentes. Es decir, los elementos dentro de un eigenvector son comparables entre sí, pero los elementos en eigenvectores diferentes no son comparables. Esto es en virtud de que los eigenvectores se normalizan para tener longitud 1, lo cual requiere que la suma de los cuadrados de los elementos en cada vector debe ser igual a 1. Por consiguiente, entre más elementos haya en un solo eigenvector que sean en realidad diferentes de cero, más pequeño debe ser cada elemento.

A fin de hacer comparaciones entre eigenvectores se cambia su escala multiplicando los elementos en cada vector por la raíz cuadrada de su eigenvalor correspondiente. Sea

$$\mathbf{a}_j^* = \lambda_j^{1/2}\mathbf{a}_j$$

para j = 1,...,p.

Estos vectores son llamados *vectores de carga de componentes* y son también eigenvectores de $\sum$, pero tienen longitudes iguales a λ_j en lugar de tener longitud 1. Defínase:

$$C = [\mathbf{a}_1^*, \mathbf{a}_2^*, ..., \mathbf{a}_p^*],$$

Es decir, $C = A\Lambda^{1/2}$

Los elementos de C son tales que los coeficientes de los componentes más importantes están escalados para, en general, ser más grandes que aquellos correspondientes a los componentes menos importantes.

Todos los elementos en todos los $\mathbf{a}_j^*$ son comparables entre sí. El i-ésimo elemento de $\mathbf{a}_j^*$ da la covarianza entre la i-ésima variable original y la j-ésima componente principal.

3.9. ACP sobre la Matriz de Correlaciones P

Tal como se indicó al inicio del capítulo, el análisis de componentes principales sólo es apropiado en aquellos casos en los que todas las variables surgen “sobre un fundamento igual”. Esto significa que:

1. Todas las variables deben estar medidas en las mismas unidades o, por lo menos, en unidades comparables.
2. Las variables deben tener varianzas que tengan tamaños aproximadamente semejantes.

Como desarrollo del punto 1, si las variables respuesta no se miden en las mismas unidades, entonces cualquier cambio en la escala de medición en una o mas de las variables tendrá un efecto sobre las componentes principales. Ese cambio de escala podría invertir los papeles de las variables importantes y las no importantes.

Como desarrollo del punto 2, en general las componentes principales se modifican por un cambio de escala de las variables; por lo que no son una característica única de los datos. Si una de las variables tiene una varianza mucho más grande que las demás, dominará la primera componente principal, sin importar la estructura de las covarianzas de las variables y, en este caso, tiene poco objeto la realización de un ACP.

Cuando no parezca que las variables están ocurriendo sobre un fundamento igual, muchos investigadores aplicarán el ACP a la matriz de correlación de las respuestas, en lugar de a la matriz de covarianzas. Esto es equivalente a aplicar el ACP a los datos estandarizados, en lugar de aplicarlo a los valores de los datos en bruto. En este caso, los componentes principales se definen por los eigenvalores y eigenvectores de *P*, la matriz de correlación, en lugar de por aquellos correspondientes a $\sum$, la matriz de covarianzas. Los eigenvalores y eigenvectores de P son distintos a los de $\sum$ y no existe simplificación sencilla para pasar de un conjunto de valores a otro.

Los eigenvalores y eigenvectores de *P* se denotarán por $\lambda_1^+ \geq \lambda_2^+ \geq ... \geq \lambda_p^+$ y $\mathbf{a}_1^+, \mathbf{a}_2^+, ..., \mathbf{a}_p^+$, respectivamente.

Notación. Una notación que usaremos a lo largo del capítulo es colocar un signo de mas (+) en las cantidades que provienen de los valores estandarizados o de las matrices de correlaciones. Cuando el '+' no aparezca, las funciones que se estarán considerando son de los datos en bruto o de matrices de covarianzas.

3.9.1. Datos Estandarizados (valores Z)

Antes de explicar algunos puntos importantes acerca del ACP sobre la matriz de correlaciones, describiremos como se obtiene un conjunto de datos estandarizados.

Al llevar a cabo una estandarización, estamos haciendo que las variables se midan en unidades comparables.

Defínase:

$$Z_{rj} = \frac{x_{rj} - \overline{x_j}}{\sqrt{s_{jj}}}$$

para $r = 1, 2, ..., n$ y $j = 1, 2,..., p$.

Donde:

x_{rj} son los valores de las variables medidas en sus unidades originales.

Las variables Z_{rj} son los valores estandarizados de las variables x_{rj}. Se les conoce como 'valores *Z*'. Estos datos pueden acomodarse en una matriz como sigue:

$$Z = \begin{bmatrix} z_{11} & z_{12} & . & . & . & z_{1p} \\ z_{21} & z_{22} & . & . & . & z_{2p} \\ . & & & & & . \\ . & & & & & . \\ . & & & & & . \\ z_{n1} & z_{n2} & . & . & . & z_{np} \end{bmatrix}.$$

3.9.1.1. Calificaciones de Componentes Principales

Cuando se realiza el ACP sobre una matriz de correlaciones, las calificaciones de componentes principales deben calcularse a partir de los valores de la variable estandarizada (valores *Z*). En este caso, la calificación de la *j*-ésima componente principal para la *r*-ésima unidad experimental se define por:

$$y_{rj}^{+} = \mathbf{a}_{j}^{+T}\mathbf{z}_{r}$$

para $j = 1,2,...,p$ y $r = 1,...,n$

3.9.1.2. Vectores de Correlaciones de Componentes

Cuando se ha realizado el ACP sobre una matriz de correlaciones, las correlaciones entre las variables originales y la *j*-ésima componente principal están dadas por:

$$\mathbf{c}_j^+ = \lambda_j^{+1/2}\mathbf{a}_j^+$$

Estos vectores se llaman *vectores de correlaciones de componentes*. Nótese que c_{ij}^+ es el *i*-ésimo elemento del *j*-ésimo vector de carga de componentes.

3.9.1.3. Matriz de Correlaciones de la Muestra

Si se aplica el ACP a la matriz de correlación *R* de la muestra, los estimadores de λ_j^+ y de $\mathbf{a}_j^+$ se denotan por $\hat{\lambda}_j^+$ y $\hat{\mathbf{a}}_j^+$, respectivamente, en donde los $\hat{\lambda}_j^+$'s y los $\hat{\mathbf{a}}_j^+$'s son los eigenvalores y eigenvectores de *R*.

Si el análisis se lleva a cabo sobre *R*, las calificaciones de componentes principales se calculan a partir de los valores Z, según la fórmula:

$$y_{rj}^+ = \hat{\mathrm{a}}_j^+ \mathbf{z}_r$$

para j = 1,2…,p y r = 1,…,n

3.9.1.4. Determinación del Número de Componentes Principales

Los dos métodos descritos en la sección 3.7 para determinar la dimensionalidad del espacio en que se encuentran los datos, también se pueden utilizar cuando se está utilizando un ACP sobre una matriz de correlación. Adicionalmente, se puede usar un tercer método cuando se analiza la matriz de correlaciones.

En éste, se buscan aquellos eigenvalores que sean mayores que 1 y se estima que la dimensionalidad del espacio muestral es el del número de eigenvalores que sean mayores que 1. La razón para comparar los eigenvalores con 1 es que, cuando se está realizando el análisis sobre datos estandarizados, la varianza de cada variable estandarizada es igual a 1. Se considera que si una componente principal no puede explicar más variación que una variable por sí misma, entonces es probable que no sea importante, por lo que

frecuentemente se ignoran componentes cuyos eigenvalores son menores que 1. En todos los casos, la decisión por lo que toca a cuántas componentes principales considerar es subjetiva.

3.10. Posibles Interpretaciones de las Componentes Principales

Una situación común que puede surgir al realizar un ACP surge cuando todas las variables están correlacionadas de manera positiva. La primera componente principal es entonces una especie de promedio ponderado de las variables y puede considerarse como una medida de *tamaño*. En particular, si se analiza la matriz de correlación, las variables (estandarizadas) tendrán pesos casi iguales.

Para matrices de correlación que contienen tanto elementos positivos como negativos, la posición es menos clara. Al tratar de dar alguna interpretación significativa a un componente en particular, el procedimiento usual parece consistir en examinar el eigenvector correspondiente y tomar las variables para las cuales los coeficientes son relativamente grandes, ya sean positivos o negativos. Una vez que se ha establecido el subconjunto de variables que son importantes para un componente en particular, el investigador tratará de ver que tienen en

común estas variables. Sin embargo, si las componentes son utilizadas para agrupar las variables, con frecuencia es también el caso que estos grupos pueden encontrarse mediante inspección visual directa de la matriz de correlaciones.

De cualquier manera, la conclusión general es que más bien resulta peligroso tratar de encontrar un significado para las componentes.

3.11 El Uso de Componentes en Análisis Subsecuentes

Como se discutió en secciones previas, un beneficio importante de un ACP es que éste constituye un camino para determinar la dimensionalidad efectiva de un conjunto de datos. Si los primeros componentes explican gran parte de la variación en los datos originales, con frecuencia es buena idea utilizar estas primeras calificaciones de componentes en los análisis subsecuentes. No es necesario hacer ninguna hipótesis distribucional para llevar esto a cabo.

Siempre es una buena idea graficar los resultados. Si los dos primeros componentes explican una proporción grande de la variación total,

entonces será de utilidad graficar los valores de las calificaciones de las componentes para cada unidad experimental. En otras palabras, un ACP nos permite graficar los datos en dos dimensiones. En particular, es posible detectar outliers o grupos de individuos.

Asimismo, es posible graficar las correlaciones de cada una de las variables con las dos primeras componentes. Es decir, graficar c_{i1} contra c_{i2} para cada i (i =1,...,p.) y buscar grupos de variables.

Finalmente, ya se ha hecho notar que la regresión múltiple puede ser inexacta si las variables independientes están muy correlacionadas entre sí. Una alternativa en estos casos es llevar a cabo la regresión, no sobre las variables originales, sino sobre aquellas componentes que se encuentran más altamente correlacionadas con la variable dependiente.

CAPÍTULO 4

4. USO DE COMPONENTES PRINCIPALES Y PUNTUACIONES RIDIT EN LA DETECCIÓN DE FRAUDE EN EL COBRO DE SEGUROS DE VEHÍCULOS.

4.1. ¿Qué son los RIDIT'S y el Análisis PRIDIT?

Los RIDIT’s (Relative to an Identified Distribution) introducidos por primera vez por Bross en 1.958, son un sistema de puntuación simple relativo a una distribución identificada. Este sistema transforma un conjunto de respuestas categóricas en un conjunto de valores numéricos que están dentro de un intervalo [-1,1], lo cual refleja la relativa anormalidad de una respuesta en particular.

Así, el análisis PRIDIT constituye una nueva técnica no-paramétrica más simple y fácil de entender e implementar, además puede satisfacer necesidades gerenciales debido a que las aplicaciones de esta metodología pueden extenderse a clasificaciones mucho más finas que la prueba binaria de detección fraude/no fraude, puesto que la metodología PRIDIT provee un valor adicional en su capacidad de probar la consistencia de su modelo de puntuaciones con los patrones de las variables de entrada. Específicamente, los pesos y las puntuaciones obtenidos de la metodología PRIDIT son representativos de las variables de entrada y pueden ser probados a través de la correlación con otros modelos de puntuaciones ya determinados. Además las correlaciones del PRIDIT con variables exógenas podrían ser usadas para “hacer un perfil” por ejemplo, de un marketing objetivo si se está hablando en un contexto

de marketing, o de “indicadores de fraude” en el caso de un contexto de seguros.

El análisis PRIDIT difiere de otras pruebas estadísticas no-paramétricas, como la Chi Cuadrado, en la cual se asume un orden natural de los datos, ya que esta técnica estadística puede emplearse incluso con variables categóricas que pueden estar en escalas subjetivas (severo, moderado, menor) o pueden tomar una forma numérica donde el sistema de medición confía fuertemente en el método experimental o en la habilidad técnica del científico involucradas en la medición de la variable en cuestión.

4.2. Tipo de Datos

La mayoría de los métodos estadísticos utilizados para este tipo de estudios requieren que los datos sean de tipo intervalo y en algunos casos estos métodos requieren que los datos sean continuos y estén normalmente distribuidos. Pero la metodología PRIDIT no pone ninguna restricción en lo que respecta a los datos, así esta metodología trabaja ya sea con variables ordinales discretas o con variables categóricas, pero para el caso de seguros de automóviles por lo general la mayoría de las variables son dicotómicas.

Aquí cabe señalar que no todas las compañías aseguradoras tienen el mismo formato para la declaración de un siniestro ni tampoco utilizan las mismas variables, sin embargo las variables utilizadas en esta investigación son las variables que se encuentran en la mayoría de los formularios de las compañías de seguro no sólo a nivel nacional sino a nivel mundial, por lo que se tomaron estas variables como variables de un formulario estándar. Ver CUADRO 4.1.

CUADRO 4. 1
Variables Utilizadas en la Investigación

Características del Asegurado y de la Demanda	
COBERT	El asegurado tiene cobertura de daños a terceros.
FRAQCIA	El asegurado tiene franquicia en la póliza.
ACCESOR	El asegurado tiene cobertura de accesorios.
Características del Vehículo	
VEHUSO	El vehículo es de uso privado del asegurado.
Características del Accidente	
ACULPA	El asegurado acepta la culpa del siniestro.

ZNURB	El siniestro ha ocurrido en una zona no urbana.
ACCNOCHE	El siniestro ocurrió en la noche.
ACCFINS	El siniestro ocurrió en el fin de semana.
TESTIGOS	Existen testigos.
REPPOLIC	Existe un reporte policial.
ZONA1	El siniestro ocurrió en una zona de elevada siniestralidad.
ZONA3	El siniestro ocurrió en una zona de baja siniestralidad.
REPSOSP	Existe presencia de relatos sospechosos.
PARENTEZ	Existe coincidencia de apellidos entre las partes.
RETRASO	El siniestro no fue reportado a la compañía aseguradora dentro del período establecido.

Elaborado por: H. Roa

4.2.1 Descripción y Codificación de la Variables

A continuación se presenta la descripción y codificación de las variables utilizadas a lo largo de esta investigación.

- **COBERT**

La cobertura de daños a terceros en una póliza de seguro es un compromiso aceptado por el asegurador en virtud del cual se hace cargo de las consecuencias económicas de un siniestro, por las cuales el asegurado es responsable frente a un tercero.

Es sinónimo de garantías. La variable COBERT identifica si el asegurado cuenta con cobertura de daños a terceros, de ser así se codifica con 1 caso contrario 0.

- **FRAQCIA**

  La franquicia es la cantidad que el asegurado asume de su propio riesgo, soportando la parte proporcional correspondiente de los daños en caso de que ocurra un siniestro. La franquicia puede ser un porcentaje sobre la suma asegurada o una cantidad fija asumida por el asegurado. La variable FRAQCIA se codifica 1 si el asegurado tiene franquicia en la póliza caso contrario 0.

- **ACCESOR**

  La cobertura de accesorios en una póliza de seguro es el compromiso aceptado por el asegurador en virtud del cual se hace cargo de daños o pérdidas de los accesorios del vehículo establecidos dentro de la póliza. Es sinónimo de garantías. La variable ACCESOR identifica si el asegurado cuenta con cobertura de accesorios, de ser así se codifica con 1 caso contrario 0.

- **VEHUSO**

La variable VEHUSO identifica si el vehículo es de uso privado del asegurado, de ser así VEHUSO se codifica con 1 caso contrario 0.

- **ACULPA**

  La variable ACULPA identifica si el asegurado acepta que el siniestro fue su responsabilidad, de ser así ACULPA se codifica con 1 caso contrario 0.

- **ZNURB**

  La variable ZNURB identifica si el siniestro del asegurado ocurrió en una zona no urbana, de ser así ZNURB se codifica con 1 caso contrario 0.

- **ACCNOCHE**

  La variable ACCNOCHE identifica si el siniestro ocurrido fue en horas de la noche, de ser el caso ACCNOCHE se codifica con 1 caso contrario 0.

- **ACCFINS**

  La variable ACCFINS identifica si el siniestro ocurrió en los días del fin de semana, de ser el caso ACCFINS se codifica con 1 caso contrario 0.

- **TESTIGOS**

  La variable TESTIGOS identifica si existen testigos del siniestro que puedan corroborar las causas del mismo, de ser así TESTIGOS se codifica con 1 caso contrario 0.

- **REPPOLIC**

  La variable REPPOLIC identifica si existe un reporte policial del siniestro, de ser así REPPOLIC se codifica con 1 caso contrario 0.

- **ZONA1**

  La variable ZONA1 identifica si el siniestro aconteció en una zona de alta siniestralidad, de ser así ZONA1 se codifica con 1 caso contrario 0.

- **ZONA3**

  La variable ZONA3 identifica si el siniestro aconteció en una zona de baja siniestralidad, de ser así ZONA3 se codifica con 1 caso contrario 0.

- **REPSOSP**

  La variable REPSOSP identifica si en la declaración del siniestro existe presencia de relatos sospechosos, de ser así REPSOSP se codifica con 1 caso contrario 0.

- **PARENTEZ**

  La variable PARENTEZ identifica si existe coincidencia de apellidos entre las partes involucradas en el siniestro, de ser así PARENTEZ se codifica con 1 caso contrario 0.

- **RETRASO**

  La variable RETRASO identifica si el siniestro no fue reportado a la compañía aseguradora dentro del período establecido, de ser así RETRASO se codifica con 1 caso contrario 0.

4.3. Validez de las Variables Instrumentales

El proceso de recolección de las variables involucradas en este tipo de estudios es muy costoso en dinero y tiempo, por lo que el investigador del fraude está especialmente interesado en determinar el valor relativo de cada variable predictora para discriminar entre grupos de fraude y no fraude. Este valor relativo debería ser análogo al coeficiente que se obtiene en la regresión, o al coeficiente del análisis discriminante, en el cual la variable dependiente de hecho es una variable observable dicotómica. Esta situación es deseada porque el análisis de muchas de las variables consume demasiado dinero y tiempo en lo que respecta a su obtención y deberían entonces ser evaluadas de acuerdo al valor que tiene cada variable para determinar el fraude en las demandas presentadas por los asegurados que han sufrido siniestros. Por otra parte, agregar variables de valores ambiguos puede disminuir notablemente la capacidad del modelo para distinguir los casos de fraude.

4.4. Apreciación Global de PRIDIT en el Caso de Datos Ordinales Discretos

El análisis PRIDIT es, esencialmente, una técnica que recoge el contenido de una demanda en un conjunto de variables (independientes) predictoras tanto individualmente como en un conjunto entero. A través de un algoritmo computacional, todas las demandas pueden ser ordenadas en forma decreciente según el nivel o grado de sospecha de fraude y, si se desea, puede ser asignada a un grupo (sea éste al grupo de fraudulentas o al de no fraudulentas) en base a esta puntuación.

4.5. Asignación de Puntuaciones Numéricas a las Categorías de Variables Ordinales Cualitativas

La tarea básica a la que se enfrenta el investigador en el uso del Análisis PRIDIT, es desarrollar una clasificación relativa de cada una de las demandas de los asegurados siniestrados de acuerdo a una variable fundamental y latente que en este caso es la probabilidad de fraude, cuando la variable de criterio que es la determinación actual de fraude o no fraude no es observada ni siquiera en una submuestra.

Así, han sido desarrolladas varias técnicas de puntuaciones para datos categóricos por investigadores como Brockett (1.981), Brockett y Levine (1977), tanto de manera empírica como teórica. Sin embargo estos investigadores han basado sus estudios en métodos estadísticos clásicos, en los cuales se asume que los datos de entrada son de tipo intervalo y con una distribución específica, y de aquí entonces se supone que el método de puntuación crea datos con “igual intervalo”, entre las categorías. En un estudio empírico realizado por Golden y Brockett en 1.987, se encontró que otro método de puntuaciones produce mejores resultados que muchos de los análisis estándares estadísticos, y sobretodo, que constituye un método de puntuación de mejor desempeño para variables categóricas, en muchos aspectos. El objetivo de este método de puntuación es cuantificar el nivel de sospecha de fraude producida por la representación categórica de cada variable indicadora de fraude en una declaración de siniestro. Además, se desea simultáneamente obtener una puntuación global de fraude para cada una de las demandas de siniestro.

Sea k_t el número de categorías disponibles y ordenadas para la variable indicadora t, y sea $\hat{p}_t = (\hat{p}_{t1},..., \hat{p}_{tk_t})$ las proporciones de respuesta observadas en el conjunto entero de demandas. Se asume que las categorías están ordenadas en relación a la probabilidad de sospecha de

fraude, en forma decreciente; por lo que una respuesta categórica más alta indica menos sospecha de fraude. Para la opción categórica i de la variable t, se asigna el valor numérico o puntuación siguiente:

$$B_{ti} = \sum_{j<i} \hat{p}_{tj} - \sum_{j>i} \hat{p}_{tj} \qquad i = 1, 2, \ldots, k_t \qquad \textbf{(A)}$$

A esta puntuación se la denomina, puntuación RIDIT para respuestas categóricas. (1)

Así, este procedimiento transforma cualquier conjunto de respuestas categóricas en un conjunto de valores numéricos dentro de un intervalo [-1, 1], lo cual refleja la relativa "anormalidad" de una respuesta particular. Por ejemplo, para una respuesta binaria de una variable indicadora de fraude (se tiene que si = 1 y no = 2 en una escala natural entera) un **"si"** podría ser mas indicativo de la existencia de fraude potencial que un **"no".** Asumiendo, por ejemplo, que un 10 por ciento de las demandas de siniestro indican un "si" en la variable indicadora de fraude, y un 90 por ciento tienen un "no" como respuesta, la puntuación RIDIT calculada para

(1) Esta puntuación es de hecho una transformación del método de puntuación RIDIT introducido al principio por Bross(1.958) para estudios epidemiológicos, pero la versión en la Ecuación (A) es más conveniente para nuestro análisis.

el “SI” sería: $\mathbf{B_{t1}}$**(“si”) = -0.9** y la puntuación RIDIT para el “NO” sería: $\mathbf{B_{t2}}$ **(“no”) = 0.1.**

Este mecanismo de puntuación por lo tanto produce un valor numérico asignable a cada categoría donde, en lugar de usar 1 y 2 que son las codificaciones en valor entero de SI y NO, se usa -0.9 y 0.1. Al igual que en la puntuación entera natural, en la puntuación RIDIT calculada en la Ecuación (A), se advierte que ésta aumenta a medida que la probabilidad de fraude disminuye, pero a diferencia de la puntuación entera natural, la puntuación RIDIT refleja la magnitud o grado para el cual una respuesta en particular es anormal, y en qué dirección. Otra respuesta binaria para una variable indicadora de fraude con 50 por ciento de las demandas indicando un “si” en esta variable y con 50 por ciento indicando un “no” habría producido las mismas puntuaciones enteras como en la anterior, pero produciría en RIDIT puntuaciones de -0.5 y 0.5, respectivamente, indicando al investigador que un “si” en la primera variable indicadora es más anormal que un “si” en la segunda variable indicadora (-0.9 versus -0.5).

Efectivamente, este método produce una puntuación para la variable indicadora, la cual es positiva cuando la mayoría de las demandas resultan en una categoría clasificada “más baja” (esto es, la mayoría de

las demandas son más probables a ser fraudulentas que la demanda de ese momento) y negativa si la mayoría de las demandas resultan en una respuesta “más alta” en la clasificación de la categoría (son más probables a ser no fraudulentas que la demanda de ese momento) para esa variable. Dentro de este estudio se cuenta con 15 variables que tienen datos categóricos por lo que es necesario y aconsejable aplicar el método PRIDIT para obtener una mejor puntuación numérica. Ver TABLA III.

TABLA III

CALCULO DE PUNTUACIONES RIDIT

Variables	% Si	B_{t1}("si")	B_{t2}("no")
El asegurado tiene cobertura de daños a terceros	96%	-0.04	0.96
El asegurado tiene franquicia en la póliza	90%	-0.10	0.90
El asegurado tiene cobertura de accesorios	93%	-0.07	0.93
El vehículo es de uso privado del asegurado	55%	-0.45	0.55
El asegurado acepta su culpa	6%	-0.94	0.06
Accidente ocurrido en zona no urbana	10%	-0.90	0.10
Accidente ocurrido en la noche	57%	-0.43	0.57
Accidente ocurrido durante un fin de semana	38%	-0.62	0.38
Existen testigos	12%	-0.88	0.12
Existe un reporte policial	95%	-0.05	0.95
Siniestro ocurrido en zona de elevada siniestralidad	81%	-0.19	0.81
Siniestro ocurrido en zona de baja siniestralidad	21%	-0.79	0.21
Existe presencia de relatos sospechosos	3%	-0.97	0.03
Existe coincidencia de apellidos entre las partes	1%	-0.99	0.01
El siniestro no fue reportado a la compañía aseguradora dentro del período establecido.	14%	-0.86	0.14

Elaborado por: H. Roa

- **Ejemplo de cómo se calculan las puntuaciones RIDIT: VEHUSO**

TABLA IV

VEHUSO: El vehículo es de uso particular						
				Proporción		
Valor	Código	Número	Proporción	Debajo	Encima	RIDIT
Si	1	55	0.550	0.000	0.450	-0.450
No	2	45	0.450	0.550	0.000	0.550

Elaborado por: H. Roa

Como lo demuestran los resultados de la TABLA III, consistente con el orden del rango categórico natural de las variables, el método PRIDIT es monótonamente creciente. Esto quiere decir que puntuaciones numéricas más altas corresponden a las opciones más altas de la clasificación categórica y, por ende una probabilidad más alta de no ser fraudulenta; además cada puntuación está en general ***"centrada"*** por lo que el valor esperado es $\sum_i \hat{p}_{ti} B_{ti} = 0$ para cada variable indicadora de fraude t.

Ya que todas las variables tienen la misma escala [-1,1], las variables predictoras con varias categorías se convierten en variables fácilmente comparables, puesto que una respuesta alta en una pregunta con diez opciones de clasificación categórica no influye tanto en la sumatoria total de la puntuación de sospecha de fraude que una alta clasificación en una variable con cinco opciones de categorías. De hecho, aquí se realizaron las pruebas de bondad de ajuste de Kolgomorov-Smirnov para probar que

las puntuaciones RIDIT's de las variables se ajustan a una uniforme en [-1,1], como se ve en la TABLA V.

PRUEBA DE HIPOTESIS

H_o: Las puntuaciones RIDIT provienen de una distribución Uniforme en

[-1,1]

vs

H_1: No es verdad H_o

TABLA V

PRUEBA DE KOLMOGOROV-SMIRNOV
DE LAS PUNTUACIONES RIDIT

		RIDIT's
Parámetros uniformes[a,b]	Mínimo	-1.00
	Máximo	1.00
Diferencia más extrema	Absoluta	.130
Z de Kolmogorov-Smirnov		.730
Sig. asintót. (bilateral)		.691

a. La distribución de contraste es la Uniforme [-1,1].

b. Se han calculado a partir de los datos.

Elaborado por: H. Roa

Como el valor p de la prueba es mayor que 0.05, existe suficiente evidencia estadística para no rechazar H_o por lo que se puede concluir

que las puntuaciones RIDIT provienen de una distribución uniforme en [-1,1].

La metodología PRIDIT se ha aplicado sólo para los casos en que las variables predictoras son categóricas puesto que es preferible cuantificar el nivel de sospecha de fraude para estas variables que tener simplemente sus valores enteros, Las puntuaciones RIDIT calculadas y mostradas en la TABLA III, son puntuaciones de sospecha de fraude únicamente para cada variable, pero el propósito de esta investigación no es simplemente determinar una puntuación que indique la existencia de fraude o no en determinada variable sino que se logre la puntuación de fraude de una demanda como un todo.

4.6. Desarrollo de una Medida de Poder Discriminatorio de una Variable

Antes de empezar con la discusión del poder discriminatorio para esta metodología, se comenzará con algunos aspectos de fondo. Aún sin asumir una estructura probabilística específica para los grupos fraude/no fraude, varias consecuencias interesantes de este sistema de puntuación resaltan por su importancia. Primero, la puntuación RIDIT para valores a respuestas categóricas (B_{ti}) está linealmente relacionada al método de puntuación RIDIT introducido por Bross (1958), comúnmente utilizado en

epidemiología en la actualidad. De hecho, si R_{ti} es la puntuación RIDIT original de Bross para la categoría *i* de la variable *t*, entonces $B_{ti} = 2R_{ti} - 1$. De este modo, las justificaciones heurísticas dadas para el uso de la puntuación RIDIT se trasladan directamente a esta situación.

4.6.1. Relación entre el Análisis PRIDIT y la Prueba No Paramétrica de Wilcoxon

- **Prueba No Paramétrica de Wilcoxon**

  La prueba no paramétrica de Wilcoxon es un procedimiento no paramétrico que se utiliza con dos muestras relacionadas para contrastar la hipótesis de que las dos variables tienen la misma distribución. No hace supuestos sobre las formas de las distribuciones de las dos variables.

La prueba de Wilcoxon consiste en seguir los siguientes pasos:

1. Las dos muestras se combinan en un conjunto ordenado de forma ascendente, en el que cada valor muestral se identifica según el grupo muestral original (G1 o G2).
2. Los valores ordenados se clasifican entonces de menor a mayor, asignando el rango 1 al menor valor muestral observado. En caso de valores iguales, se les asigna el rango medio.
3. Luego se calcula la suma de rangos del grupo 1, W_1, y la suma de rangos del grupo 2, W_2.
4. Luego los estadísticos correspondientes son:

$$U_1 = W_1 - \frac{n_1(n_1+1)}{2}$$

$$U_2 = W_2 - \frac{n_2(n_2+1)}{2}$$

5. El estadístico de prueba U es:

$$U = \min\{U_1, U_2)$$

6. Se rechaza la hipótesis nula si el estadístico U es menor que U_α.

Para muestras de gran tamaño, el estadístico de prueba puede fundamentarse en:

$$Z = \frac{U_1 - \mu_{U_1}}{\sigma_{U_1}}$$

donde:

$$\mu_{U_1} = \frac{n_1 n_2}{2} \qquad \text{y} \qquad \sigma_{U_1} = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

La relación entre la Prueba no Paramétrica de Wilcoxon y las puntuaciones RIDIT, también se trasladan a esta metodología. En particular, uno puede demostrar, asumiendo dos grupos tales como fraude y no fraude, que la media de la puntuación RIDIT, $\overline{B}_t^{(1)}$ para un miembro del grupo denominado fraudulento de la variable *t* es $2W_1^* / N_1 - 2$, donde W_1^* es el estadístico de la Prueba no Paramétrica de Wilcoxon para comparar el grupo denominado fraudulento contra las repuestas de los no fraudulentos para la variable *t*.

En consecuencia, este método de puntuación RIDIT produce una medida del *"poder discriminatorio de la variable"*, asumiendo que los miembros del grupo son conocidos. De hecho, este método de puntuación y los

métodos de puntuación relacionados a rangos son a menudo mejores que las puntuaciones enteras para el análisis discriminante lineal de datos no normales.

Se enfatiza sin embargo, que en la situación aquí considerada de detección de perjuicio por fraude en seguros de automóviles, no se conoce el número de miembros del grupo y por ende la Prueba de Wilcoxon no es aplicable. A pesar de eso, el recurso intuitivo de la conexión y la correspondencia con metodologías estadísticas conocidas bajo suposiciones adicionales del número de miembros del grupo conocido da gran seguridad y confianza en los resultados, en la situación de esta investigación donde el número de miembros del grupo es desconocido.

Ahora se procede al problema de determinar el poder discriminatorio de la variable y la discriminación entre dos grupos cuando *no se conoce* la variable de criterio (es decir, cuando la muestra no está clasificada con respecto a la característica del grupo de probabilidad de fraude versus la de no fraude). Se asume que las *N* demandas tienen variables categóricas que incluyen valores ordinales, en las cuales están representados dos grupos (el de las demandas fraudulentas y el de las demandas no fraudulentas); sin embargo, no se tiene conocimiento a

priori de qué demanda individual pertenece a tal o cual grupo. También se asume que las variables fueron construidas de tal manera que exista una relación dominante entre los dos grupos de cada variable (por ejemplo, que las demandas fraudulentas tiendan a caer con mayor probabilidad en clasificaciones categóricas más bajas que las demandas no fraudulentas). A partir de esto, se debe llamar entonces al grupo con más alta tendencia a responder hacia las clasificaciones categóricas más bajas, para una variable indicadora de fraude en particular, como el grupo fraudulento o Grupo 1; mientras que al grupo con una más alta tendencia a contestar hacia las clasificaciones categóricas más altas se lo denominará grupo no fraudulento o Grupo 2 (aunque cabe recalcar que no se tiene conocimiento de qué demandas pertenecen a cada grupo respectivamente).

El vector de las proporciones de respuestas observadas para las k_t categorías de la variable t usando as N demandas, p_t, puede ser modelado como una combinación de estos dos grupos:

$$p_t = \theta\pi_t^{(1)} + (1-\theta)\pi_t^{(2)} \qquad \text{(B)}$$

donde para $q = 1$ y 2, $\pi_t^{(q)} = \left(\pi_{t1}^{(q)},...,\pi_{k_t}^{(q)}\right)$ es el vector multinomial de la probabilidad de respuesta para el grupo q. Por lo que $\pi_{tj}^{(1)}$ es la

proporción de las demandas fraudulentas, o grupo 1, que caen en la categoría j de la variable indicadora de fraude t; $\pi_{tj}^{(2)}$ es la proporción de las demandas no fraudulentas, o grupo 2, que caen en la categoría j de la variable indicadora de fraude t; y $\theta = N_1 / N$ es la proporción de demandas que pertenecen al grupo fraudulento. Se demostrará que la puntuación esperada (sospecha) para una demanda perteneciente al grupo no fraudulenta es $(\theta - 1)A_t$ y la puntuación esperada de la demanda para el grupo fraudulento es θA_t, donde

$$A_t = \sum_{i=1}^{k_{t-1}} \sum_{j>1} \left\{ \pi_{ti}^{(1)} \pi_{ti}^{(2)} - \pi_{tj}^{(2)} \pi_{tj}^{(1)} \right\} \qquad \text{(C)}$$

es una medida del poder discriminatorio de las t variables indicadoras de fraude. Por supuesto, en varias aplicaciones (incluyendo la de esta investigación), $N_1, \theta, \pi_t^{(1)}$ y $\pi_t^{(2)}$ son todos parámetros desconocidos, por lo que se tiene que encontrar una manera alternativa para estimar A_t. Consecuentemente se demostrará que la primera componente principal de la matriz construida de las puntuaciones RIDIT puede ser usada para estimar A_t y así completar el análisis.

Aunque $N_1, \pi_t^{(1)}$ y $\pi_t^{(2)}$ son parámetros desconocidos (y por ende A_t y θ no son directamente calculables), por su importancia, la cantidad A_t

merece ser discutida con más profundidad. Como fue mencionado, el valor de A_t indica el poder discriminatorio de la variable indicadora de fraude t en tanto que $|A_t| \leq 1$ con $A_t = 1$ si y sólo si la variable se diferencia absolutamente entre el grupo de fraude y el de no fraude para la categoría j, $\sum_{i=1}^{j-1} \pi_{ti}^{(1)} = 1$ y $\sum_{i=i}^{k_t} \pi_{ti}^{(2)} = 1$. Si la variable todavía se diferencia absolutamente, pero en dirección contraria, $A_t = -1$. Si la variable no se discrimina del todo en el caso que $\pi_{ti}^{(1)} = \pi_{ti}^{(2)}$ para $i = 1,2,...,k_t$ entonces $A_t = 0$

Ya que, para cada variable, se tiene un ordenamiento estocástico de la función de distribución acumulada de los dos grupos (fraudulentos/no fraudulentos), la medida de A_t indica el poder discriminatorio ordinal y mide la diferencia esperada del entre-grupo de la variable t. Nótese también que el término dentro de las llaves de las sumatorias que definen A_t es la medida de asociación del producto cruzado estándar para tablas de contingencia 2x2. Entonces, A_t es una nueva medida de asociación para tablas de contingencia 2xk_t con categorías de respuesta ordenadas, lo cual generaliza la clásica medida 2x2 para la situación de esta investigación con variables categóricas clasificadas, ordenadas.

Cuando la verdadera clasificación de cada demanda (fraudulenta/no fraudulenta) es conocida, no se puede calcular directamente la medida de la tabla de contingencia A_t mencionada anteriormente. Sin embargo, la relación entre rangos cercanos, la prueba estadística de Wilcoxon y las puntuaciones RIDIT proporcionan una conexión entre el método de puntuación que se ha desarrollado y una medida de la variable discriminatoria cuando el número de miembros del grupo no son conocidos, esta conexión proporciona credibilidad para los resultados en la situación frecuente en la que el número de miembros del grupo no es conocido. Esta conexión puede ser desarrollada de la siguiente manera:

Sea Y_{tj} la notación de la posición del rango de la respuesta j sobre la variable de respuesta categórica t correspondiente a las N demandas que contienen una categorización para la variable t, y sea I_A la función indicadora del conjunto A. El rango de una puntuación de una demanda en particular puede ser representado así:

$$Y_{tj} = \frac{N}{2} \sum_{i=1}^{k_t} B_{ti} I\left[categoría\ i\ dada\right] + \frac{\mathrm{N}}{2} \qquad \text{(D)}$$

Ahora, si las N_q demandas se originan del grupo q, $q = 1, 2,$ entonces (usando la relación entre la media de los rangos relativos de dos grupos

en una muestra combinada versus el estadístico de la suma de rangos Wilcoxon) la puntuación media para una demanda del grupo 1 en t es $\overline{B}_t^{(1)} = 2W_1^* / N_1 - 2$, donde W_1^* es el estadístico de la suma de rangos Wilcoxon para comparar las clasificaciones de los grupos 1 y 2 para t, permitiendo el vínculo.

En consecuencia, si se conoce el número de miembros del grupo (como ocurre, por ejemplo, en el análisis discriminante, redes neuronales supervisadas, o regresión logística) entonces la técnica de puntuación RIDIT, B_{ti}, inmediatamente produce una medida $\overline{B}^{(1)}$ del "poder discriminatorio de la variable", es decir, el estadístico Wilcoxon. Este alentador resultado motiva a seguir adelante en la situación más dificultosa donde el número de miembros del grupo no es conocido. Para esto último, se demuestra cómo usar resultados del análisis de componentes principales y el análisis factorial para obtener una estimación consistente de A_t, y, por ende, obtener una medida de los valores de las variables individuales para demandas discriminadas fraudulentas(2).

(2) Nótese que en el análisis PRIDIT, las tareas de determinar el poder discriminatorio de la variable, puntuaciones de variables, y obtener puntuaciones globales de sospecha de fraude no se las realiza en forma separada como en la mayoría de otros análisis de fraude. Más bien, este análisis demuestra que el análisis de componentes principales usando

Una importante suposición que se hace en el desarrollo de este modelo es que se distribuye con un modelo de probabilidad de primer orden, esto es, si para un conjunto de demandas en particular de cualquier grupo (fraude/no fraude), se desea conocer qué proporción de demandas simultáneamente se clasifican dentro de la categoría i de la variable t y dentro de la categoría j de la variable s, sólo se necesita conocer las probabilidades marginales de la categoría i de la variable t y la categoría j de la variable s. Esta suposición del modelo de primer orden estadísticamente se introduce en la propiedad de falta condicional de correlación de las clasificaciones, dando el número de miembros del grupo (fraude/no fraude). Hay que recordar que una vez que se asume atributos de distribuciones condicionales, esta suposición de un modelo de primer orden no implica falta de correlación entre las clasificaciones de las variables para el total de las demandas de la muestra. De hecho, en la práctica hay mucha dependencia entre las variables (lo cual es porque estas variables son útiles para la detección de fraude). No obstante, es suficiente que esta dependencia baje significativamente en los subgrupos individuales. Nótese también que la suposición de falta

---

puntuaciones RIDIT desarrollada aquí, proporciona una consistencia íntima entre estos aspectos del análisis de sospecha de fraude, lo cual está también relacionado para medidas de validez externa.

condicional de correlación es común para características latentes de los modelos, donde esto es llamado “independencia local”. La característica latente en este análisis es el nivel de sospecha de la demanda.

Retomemos ahora la situación en la que las clasificaciones del grupo son desconocidas. Como se tenía anteriormente $p_t = \theta\pi_t^{(1)} + (1-\theta)\pi_t^{(2)}$ para el grupo fraudulento $q = 1$ y para el grupo no fraudulento $q = 2$ donde $\theta = N_1 / N$ es la proporción de demandas que pertenecen al grupo 1. El siguiente Lema 1 relaciona los valores $\theta, \pi_t^{(1)} \; y \; \pi_t^{(2)}$ teóricos (y no observables) al sistema de puntuación anterior y determina una medida del poder discriminatorio de la variable predictora A_t.

Lema 1: *Asumir que una demanda produce una puntuación B_{ti} sobre la variable predictora indicadora de fraude t. Luego la puntuación esperada para una demanda del grupo 1 con bajo nivel de sospecha es $(\theta - 1)A_t$, y la puntuación esperada para una demanda del grupo 2 con un nivel alto de sospecha es θA_t donde A_t está dado por (C).* Ver demostración del Lema 1 en el Apéndice.

4.7. Evaluación del Poder Discriminatorio de las Variables Indicadoras de Fraude y Obtención de una Puntuación Global de Sospecha de Fraude para una Demanda Entera

Sea $F=(f_{it})$ la notación de la matriz de las puntuaciones individuales PRIDIT para cada una de las $t=1,2,...,m$ variables, para cada una de las $i=1,2,...,N$ demandas. Esto es, $f_{it}=B_{tk}$ si la demanda i contiene k niveles de respuestas categóricas para la variable t. Se obtiene una puntuación de sospecha global para cada demanda sumando simplemente las puntuaciones de las variables individuales respectivas[3].

Sea $W^{(0)}=(1,1,...,1)'$, la notación del primer vector transpuesto. Luego el vector de la suma global de las puntuaciones de sospecha de fraude obtenidas para cada demanda denotada en la matriz es: $S^{(0)}=FW^{(0)}$. Ahora, se tiene una medida de consistencia de la variable indicadora t con las puntuaciones globales de sospecha de fraude para las demandas. Esta medida es similar en naturaleza a la medida de fiabilidad

[3] Este valor inicial de la puntuación de sospecha corresponde al método sencillo de conteo de número de indicadores de fraude. Este simple método de conteo para variables indicadoras es común en muchas áreas y puede ser mejorado por análisis más profundos.

Cronback α usada en análisis de cuestionarios para evaluar la consistencia de las preguntas individuales con la puntuación general de todo el cuestionario. De este modo, $W^{(1)} = F'S^{(0)} / \left\| F'S^{(0)} \right\|$ puede verse como un sistema de *"pesos"* para las variables individuales en el conjunto de demandas, donde los componentes de $W^{(1)}$ dado el producto normalizado de la variable indicadora t con los scores globales de cada demanda y medida la consistencia de la variable individual sea ponderada a la puntuación global para toda la demanda. Un valor más alto para $W^{(1)}$ indica una más alta consistencia de la variable t con el nivel global determinado para toda la demanda.

Tiene sentido entonces, ahora que se conoce qué variables son más consistentes (o "mejores jueces") del nivel global de sospecha para un conjunto de demandas, dar un peso o ponderación más alta en este modelo a éstas "mejores" variables y por ende calcular una puntuación "ponderada" para cada una de las demandas existentes, dando una ponderación más alta a las variables "mejores jueces" o también denominadas variables del fraude. Usando las componentes de $W^{(1)}$ como pesos de la variable produce un vector ponderado de las puntuaciones globales del nivel de sospecha de fraude $S^{(1)} = FW^{(1)}$ para el conjunto de demandas. Utilizando esta mejor valoración del nivel global

de sospecha de fraude para cada demanda, se puede obtener alguna mejor medida de la puntuación global de sospecha de fraude para la demanda "correlacionando" las puntuaciones individuales con esta nueva y mejor puntuación global. Esto a su vez, puede correlacionarse de nuevo con las puntuaciones de las variables individuales para conseguir pesos o ponderaciones aún mejores, $W^{(2)} = F'S^{(1)} / \left\| F'S^{(1)} \right\|$, y así sucesivamente se puede seguir realizando este proceso para obtener mejores ponderaciones.

Desde un punto de vista teórico, el Teorema 1 (Ver Apéndice) garantiza que este proceso matemático converge, y demuestra además que esta fijación de ponderaciones para la variable indicadora de fraude t es de hecho proporcional a la medida de Poder Discriminatorio A_t desarrollado en el Apéndice A. Específicamente, la fijación del peso de la variable $\hat{W}^{(\infty)}$ es la primera componente principal de $F'F$, que es una estimación consistente de la componente principal $\hat{W}^{(\infty)}$ del $E\left[F'F\right]$, el *t*-ésimo componente es explícitamente igual a:

$$W_t^{(\infty)} = \frac{A_t}{\left(\mu_1 - U_{tt}\right)\sqrt{\sum_{s=1}^{m} A_s^2 / \left(\mu_1 - U_{ss}\right)^2}} \qquad \text{(E)}$$

donde μ_1 es el eigenvalor más grande de $E[F'F]$, y para cada s, U_{ss}, es la varianza en un modelo de análisis de factores de puntuaciones RIDIT para variables indicadoras de fraude $E[F'F]$. De aquí,

$$A_t = \sum_{i=1}^{k_{t-1}} \sum_{j>1} \left\{ \pi_{ti}^{(1)} \pi_{ti}^{(2)} - \pi_{tj}^{(2)} \pi_{tj}^{(1)} \right\}$$

donde $\pi_{tj}^{(1)}$ es la proporción de fraude o demandas del grupo 1 que adquiere la categoría j sobre la variable indicadora de fraude t.

El Teorema 1 (demostrado en el Apéndice A) es interesante por varias razones. Puesto que la fijación de pesos es proporcional a $A_t / (\mu_1 - U_{tt})$, y porque se puede estimar esta fijación de pesos por medio de $\hat{W}^{(\infty)}$, que es la primera componente principal de $F'F$, μ_1 por $\hat{\mu}_1$, que no es más que el eigenvalor más grande de $F'F$, y la varianza por medio de un análisis de factores produciendo la matriz diagonal de varianzas singulares $\hat{U}$. Entonces se puede estimar también el vector de los valores del poder discriminatorio de las preguntas por $(\hat{\mu}_1 I - \hat{U})\hat{W}^{(\infty)}$. La medida de asociación para la variable t puede hallarse usando componentes principales y análisis de factores sin tener la necesidad de conocer qué demandas pertenecen a determinado grupo y en qué proporción. Es necesario enfatizar que usar Análisis de Componentes

Principales para ponderar variables no es nuevo. De todos modos, en general no hay garantía de que el resultado tenga alguna interpretación estadística significativa. En esta investigación, se demuestra entonces que por usar este sistema particular de puntuaciones se tiene una interpretación útil en términos del poder discriminatorio A_t de la variable y también una conexión interesante entre el análisis dado por una tabla de contingencia, el análisis de componentes principales y el análisis factorial. No hay garantía que esto ocurra con otros sistemas de puntuaciones. Una vez detallada la metodología se continúa con la aplicación de la misma. Primero se calculará la "Ponderación PRIDIT" para cada una de las variables indicadoras de fraude (Ver TABLA VI) y luego una puntuación global de sospecha de fraude para una demanda completa (Ver TABLA VII).

TABLA VI
PONDERACIONES PRIDIT PARA
LAS VARIABLES INDICADORAS DE FRAUDE

Matriz de componentes[a]

	Componente 1
	W
COBERT	.468
FRAQCIA	-.460
ACCESOR	.354
VEHUSO	.507
ACULPA	-.560
ZNURB	.397
ACCNOCHE	-.105
ACCFINS	-.252
TESTIGOS	.177
REPPOLIC	-.341
ZONA1	.468
ZONA3	-.328
REPSOSP	.317
PARENTEZ	-.104
RETRASO	-.362

a. 1 componentes extraídos

Elaborado por: H. Roa

La TABLA VI muestra que las variables indicadoras de fraude **VEHUSO** y **ACULPA** llevan la mayor ponderación PRIDIT de 100 demandas analizadas, seguidas de **COBERT**, **ZONA1** y **FRAQCIA**, es decir que éstas son las variables que tienen mayor pesos a la hora de determinar un nivel de sospecha de fraude.

TABLA VII

PUNTUACIÓN GLOBAL

PARA UNA DEMANDA COMPLETA

	A_t
COBERT	0.808
FRAQCIA	-0.797
ACCESOR	-1
VEHUSO	0.856
ACULPA	-0.913
ZNURB	0.709
ACCNOCHE	-0.203
ACCFINS	-0.474
TESTIGOS	0.339
REPPOLIC	-0.624
ZONA1	0.868
ZONA3	-0.603
REPSOSP	0.585
PARENTEZ	-0.201
RETRASO	-0.657

Elaborado por: H. Roa

Los valores A_t de la TABLA VII son una medida del poder discriminatorio de las t variables indicadoras de fraude para una demanda entera.

4.8. Clasificación de las Demandas por medio de las Ponderaciones PRIDIT

Con respecto a la clasificación, se consideran dos casos para la proporción de demandas del grupo 1, θ conocido y θ desconocido. Cuando θ es conocido, se ordenan las N demandas por medio de sus puntuaciones unidimensionales $S = \sum_{t=1}^{m} W_t^{(\infty)} X_t$ y luego se clasifican las primeras $N\theta$ demandas en el nivel alto del grupo 1 de sospecha de fraude. Aquí X_t es la puntuación calculada de la demanda obtenida para la variable t, $X_t = \sum_{i=1}^{k_t} B_{ti} I[\text{categoría } i \text{ dada}]$ donde I_A es el indicador del conjunto A.

Si θ es desconocido (como es este caso), se separa los dos grupos de acuerdo a las puntuaciones globales positivas o negativas y se clasifican las demandas dentro del grupo de bajo nivel de sospecha de fraude si la puntuación global del nivel de sospecha de fraude es positivo, es decir en el grupo de las demandas no fraudulentas y las de valor negativo dentro del grupo de demandas fraudulentas. Ver ANEXO 1.

- **Ejemplo de cómo se clasifican las Demandas según las ponderaciones PRIDIT**

TABLA VIII

CLASIFICACIÓN DE LAS DEMANDAS

Demanda	cobertur	fraqcia	accesor	usovehi	aculpa	znurb	accnoche	accfins	testigos	reppolic	zona1	zona3	repsosp	parentez	retraso	Score	Clase
1	-0.04	-0.1	-0.07	0.55	0.06	0.1	-0.43	0.38	-0.88	-0.05	-0.19	-0.79	0.03	0.01	0.14	0.10	2
2	-0.04	-0.1	-0.07	-0.45	0.06	0.1	0.57	0.38	0.12	0.95	0.81	0.21	0.03	0.01	-0.86	-0.54	1
3	-0.04	-0.1	-0.07	-0.45	0.06	0.1	0.57	-0.62	-0.88	-0.05	-0.19	-0.79	0.03	0.01	0.14	0.39	2
4	-0.04	-0.1	-0.07	0.55	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	0.09	2
5	-0.04	-0.1	-0.07	0.55	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	-0.97	0.01	0.14	0.45	2
6	-0.04	0.9	-0.07	-0.45	0.06	0.1	-0.43	0.38	-0.88	-0.05	-0.19	0.21	0.03	0.01	0.14	0.06	2
7	-0.04	-0.1	-0.07	0.55	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.29	1
8	-0.04	-0.1	-0.07	-0.45	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	0.33	2
9	-0.04	-0.1	0.93	0.55	-0.94	0.1	-0.43	-0.62	0.12	-0.05	0.81	-0.79	0.03	0.01	0.14	-0.54	1
10	-0.04	-0.1	-0.07	-0.45	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	0.48	2

Elaborado por: HRoa

Como se demuestra en la TABLA VIII, cada una de las demandas tiene una ponderación o como se la ha llamado en este caso, Score, que permite clasificar cada demanda dentro del grupo que corresponde. Un score positivo significa que la demanda pertenece al Grupo 2, es decir es una demanda no fraudulenta mientras que un score negativo significa que la demanda debe ser discriminada al grupo 1, que es el grupo de las demandas con cierto grado de fraude.

4.9. Prueba de Consistencia para Variaciones de Tiempo

Ahora bien, es importante para cualquier sistema de detección de fraude determinar qué ocurre cuando se producen cambios a través del tiempo. Los grupos a ser detectados pueden cambiar sus comportamientos como resultado del valor de impedimento de fraude de anteriores detecciones. Los patrones indicadores pueden cambiar debido al uso de ajustadores de demandas con entrenamiento más o menos alto. A menos que la relación de los patrones indicadores y las puntuaciones de sospecha de fraude, ya determinada, sea estacionaria (es decir, invariable con el tiempo), los modelos de puntuaciones estimados con datos a priori pueden sufrir deterioro en su precisión. Esta es otra razón por la que no se usan las metodologías estándares para “pruebas de grupos”, puesto que el comportamiento del defraudador está constantemente evolucionando a lo largo del tiempo.

CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

De los resultados obtenidos en la presente investigación, se dan las siguientes conclusiones:

1. La utilización de la técnica PRIDIT en la detección de fraudes dentro del campo de seguros de vehículos, es mucho más eficiente que las técnicas estadísticas tradicionales cuando dentro de las variables involucradas existen variables con respuestas del tipo categóricas, puesto que la técnica PRIDIT transforma el conjunto de respuestas categóricas en un conjunto de valores numéricos dentro de un intervalo [-1,1], lo cual refleja la relativa anormalidad de una respuesta en particular.

2. La técnica PRIDIT provee además una medida que permite determinar qué variables son más consistentes, dando ponderaciones más altas a las variables indicadoras de fraude. En esta investigación

las variables con mayor ponderación son: ACULPA, VEHUSO, COBERT, ZONA1, FRAQCIA. Estas variables son las que indican mayormente fraude.

3. Otra de las ventajas de la utilización de la metodología PRIDIT que se comprobó en esta investigación, es que provee una medida del poder discriminatorio de las variables indicadoras de fraude. Las variables con mayor poder discriminatorio en esta investigación fueron: ACCESOR, ACULPA, ZONA1, VEHUSO, COBERT, FRAQCIA.

4. La medida cuantitativa del poder discriminatorio que resulta de la técnica PRIDIT provee además la capacidad de determinar correlaciones con otras medidas cuantitativas tales como: edad del conductor, número de accidentes anteriores del asegurado, número de años que el asegurado tiene licencia, etcétera. Este procedimiento puede dejarse para un estudio posterior puesto que no es el objetivo de la actual investigación. Además que no se cuenta con datos reales que permitan determinar la verdadera correlación entre las variables.

RECOMENDACIONES

Las recomendaciones aquí expuestas servirán para mejorar y facilitar la realización de futuras investigaciones referentes al seguro de vehículos, puesto que para el desarrollo de la presente investigación se presentaron varias limitaciones que retrasaron la misma:

1. Exhortar a las compañías aseguradoras interesadas en desarrollar estudios para detectar fraudes que se lleve registros de las declaraciones de siniestros en bases de datos, puesto que estos datos son la base de cualquier metodología de detección de fraude.

2. Exhortar a las compañías aseguradoras a llevar registros en bases de datos de las declaraciones de siniestros que no cumplen con las estipulaciones de la póliza, para así lograr un perfil del asegurado defraudador.

3. Siendo Ecuador un país considerado como uno de los más corruptos es recomendable que las empresas, a todo nivel, se interesen en aplicar este tipo de técnicas capaces de detectar fraude especialmente cuando cuentan con información del tipo cualitativa puesto que, de esta manera no sólo se detectará el fraude sino que también se irá creando un historial y perfil del ente defraudador que permitirá la disminución de los intentos de fraude en las diferentes actividades de las empresas.

APÉNDICE

APÉNDICE A

Demostración del Lema 1: Seleccione una demanda al azar del grupo de fraude o grupo 1, y sea $I_i = 1$ si el i-ésimo nivel de clasificación categórica fue asignado a la demanda sobre la variable particular t. Entonces el valor esperado deseado es:

$$\sum_{i=1}^{k_t} \pi_{ti}^{(1)} \; E\left[B_{ti} \mid I_i = 1\right] = \sum_{i=1}^{k_t} \pi_{ti}^{(1)} \left[\sum_{j<1} \frac{(N_1 - 1)\pi_{tj}^{(1)} + N_2 \pi_{tj}^{(2)}}{N} - \sum_{j>1} \frac{(N_1 - 1)\pi_{tj}^{(1)} + N_2 \pi_{tj}^{(2)}}{N} \right]$$

(A1)

Usando las identidades $N_1 = N - N_2 = N\theta$ y $\sum_{i=1}^{k_t-1} \sum_{j<1} x_i y_j = \sum_{i=1}^{k_t-1} \sum_{j>1} x_j y_i$, después de cálculos elementales, el valor esperado es:

$$-\frac{N_2}{N} \sum_{i=1}^{k_t-1} \sum_{j>1} \left\{ \pi_{ti}^{(1)} \pi_{tj}^{(2)} - \pi_{ti}^{(2)} \pi_{tj}^{(1)} \right\} = (\theta - 1) A_t \qquad \text{(A2)}$$

Sea B_{qt}^{*} la notación de la puntuación de una demanda seleccionada al azar del grupo q de nivel de sospecha de fraude sobre la variable t y sea $G = E(F'F)$.

Lema 2: *Sea* $G = M + U$ *donde* $M = N_1 N_2 (A_i A_j)/N$ *y* U *es la matriz diagonal con* $U_{tt} = N\sigma_{1t}^2 + N_2\sigma_{2t}^2$ *donde* $\sigma_{qt}^t = \text{var}\,ianza(B_{qt}^{*})$.*(Note que* ***G*** *tiene la estructura analítica de un factor con* U_{tt} *siendo la componente de singularidad de la varianza para la variable* t).

Demostración del Lema 2: Se tiene $G = (g_{st})$ con

$$g_{st} = E\left[\sum_{i=1}^{N} f_{is} f_{it}\right] = E\left[\sum_{i=1}^{N_1} f_{is} f_{it} + \sum_{i=N_1+1}^{N} f_{is} f_{it}\right] \qquad \text{(A3)}$$

donde se arreglaron las primeras N_1 demandas para que pertenezcan al grupo 1. Teniendo un modelo de primer orden que produce:

$$g_{st} = N_1 E\left[f_{is} f_{it} \mid i - \acute{e}sima\ demanda\ en\ el\ grupo\ 1\ de\ fraude\right] + N_2 E\left[f_{is} f_{it} \mid i - \acute{e}sima\ demanda\ en\ el\ grupo\ 2\ de\ no\ fraude\right]$$

si $s \neq t$, se debe usar la independencia condicional y el Lema 1 para obtener $\frac{N_1 N_2 A_t A_s}{N}$. Si $s = t$, sumar y restar $\frac{N_1 N_2 A_t^{\,2}}{N}$ para obtener:

$$g_{tt} = \frac{N_1 N_2 A_t^{\,2}}{N} + \left[N_1 E\left(B_{2t}^{*2}\right) - \frac{N_1 N_2^2 A_t^{\,2}}{N} \right] + N_2 E\left(B_{2t}^{*2}\right) - \frac{N_2 N_1^2 A_t^{\,2}}{N} \qquad \text{(A4)}$$

De esto se tienen que $G = M + U$ como se afirmó.

Lema 3: L*a función característica polinomial de* $G = E(F'F)$ *es:*

$$\begin{aligned} f(\mu) &= \left\{ \prod_{t=1}^{m} (U_{tt} - \mu) + \sum \frac{N_1 N_2}{N} A_t^2 \coprod_{s,t} (U_{ss} - \varpi) \right\} \\ &= \prod_{t=1}^{m} (U_{tt} - \mu) \left\{ 1 + \sum \frac{N_1 N_2}{N} A_t^2 / (U_{tt} - \mu) \right\} \end{aligned} \qquad \text{(A5)}$$

El eigenvalor más grande μ_1 de G es positivo y más grande que el máximo de U_{tt}. El segundo eigenvalor está entre el segundo valor más grande de U_{tt}.

Demostración del Lema 3: La ecuación para $f(\mu)$ viene de realizar operaciones elementales sobre G.

Para obtener la ubicación de los eigenvalores de G, considerar primero el caso en el cual todos los U_{tt} son distintos y colocarlos en orden $U_1 < U_2 < ... < U_m$. De la fórmula para $f(\mu)$ el signo algebraico de $f(\mu)$ se nota más claramente si $\mu \leq U_1$ y $f(U_k)$ tiene el signo $(-1)^{k+1}$. Entonces $f(U_t)$ tiene signos alternados, f tiene al menos una raíz entre U_t y U_{t+1}, $t = 1, ..., m-1$. Se ha considerado que para $(m-1)$ de los m eigenvalores, y para el resto de eigenvalores, éstos tienen que ser más grandes que U_m. Si todos los U_{tt} no son distintos, es decir, k de ellos son iguales a U_{jj}, entonces $f(\mu)$ tiene una raíz de multiplicidad k en U_{jj} y de nuevo se ha considerado para $(m-1)$ raíces a la izquierda del máximo U_{tt}

.

Teorema 1: L*as secuencias de los pesos de las variables predictoras* $\{W^{(n)}\}$ *y la sumatoria global de las puntuaciones de sospecha de las demandas* $\{S^{(n)}\}$ *convergen. De modo que, la fijación de peso de la variable predictora* $\hat{W}^{(\infty)}$ *es la primera componente principal de* $F'F$ *, la cual es una estimación consistente de la primera componente principal* $W^{(\infty)}$ *del* $E(F'F)$*), la t-ésima componente principal se tiene explícitamente:*

$$W_t^{(\infty)} = \frac{A_t}{(\mu_1 - U_{tt})\sqrt{\sum_{s=1}^{m} A_s^2 / (\mu_1 - U_{ss})^2}} \qquad \text{(A6)}$$

donde μ_1 *es el eigenvalor mas grande de* $E[F'F]$ *y* $U_{tt} = N_1\sigma_{1t}^2 + N_2\sigma_{2t}^2$ *es la "componente de singularidad de la varianza" en un modelo analítico de un solo factor.*

Demostración del Teorema 1: Primero se demostrará que el $\lim_{(n)} W^{(n)}$ y $\lim_{(n)} S^{(n)}$ existen.

Nótese que $W^{(n)} = F'S^{(n-1)} / \left\| F'S^{(n-1)} \right\| = (F'F)^n W^{(0)} / \left\| (F'F)^{(n)} W^{(0)} \right\|$. Por álgebra lineal elemental, se tiene que el $\lim W^{(n)}$ existe y es la proyección de $W^{(0)}$ sobre el eigenespacio generado por el eigenvalor más grande de $F'F$. Para calcular la primera componente principal de $E(F'F))$, se usa el Lema 3 y la matriz $E(F'F) - \mu I)$. Nótese que si V_1 es la primera componente principal, entonces tiene que satisfacer las ecuaciones

$$V_1 = \frac{(\mu_1 - U_{22})A_i}{(\mu_1 - U_{ii})A_2} V_{12} \text{ para } i \neq 2 \qquad \text{(A7)}$$

Esto, junto con el hecho de que $f(\mu_i)=0$ implica $\sum_{i=1}^{m}\frac{N_1N_2A_i^2}{(U_{ii}-\mu_1)}+1=0$ demostrando que con $W_t^{(\infty)}$ dado como en el teorema, se puede tener $E(F'F)W_t^{(\infty)}$, demostrando el teorema.

ANEXOS

ANEXO 1

DATOS ORIGINALES

cobertur	fraqcia	accesor	usovehi	culpa	znurb	accnoche	accfins	testigos	reppolic	zona1	zona3	repsosp	parentez	retraso
1	1	1	0	0	0	1	0	1	1	1	1	0	0	0
1	1	1	1	0	0	0	0	0	0	0	0	0	0	1
1	1	1	1	0	0	0	1	1	1	1	1	0	0	0
1	1	1	0	0	0	1	0	0	1	1	0	0	0	0
1	1	1	0	0	0	1	0	0	1	1	0	1	0	0
1	0	1	1	0	0	1	0	1	1	1	0	0	0	0
1	1	1	0	0	0	1	0	0	1	1	0	0	0	0
1	1	1	1	0	0	1	0	0	1	1	0	0	0	0
1	1	0	0	1	0	1	1	0	1	0	1	0	0	0
1	1	1	1	0	0	1	0	0	1	1	0	0	0	0
1	1	0	0	0	0	1	1	0	1	0	0	0	0	0
1	1	1	0	0	0	0	0	0	1	1	1	0	0	0
1	1	0	0	0	0	1	0	0	1	1	0	0	0	0
1	0	1	1	0	1	1	0	0	1	1	0	0	0	0
1	1	1	0	0	1	0	1	0	1	1	0	0	0	0
1	1	1	1	0	0	1	0	0	1	1	0	0	0	0
1	1	1	0	0	0	0	1	0	1	1	0	0	0	0
1	1	1	1	0	0	1	0	0	1	1	1	0	0	1
1	1	1	0	0	0	0	0	0	1	0	0	0	0	0
1	1	1	1	0	0	0	1	0	1	0	0	0	0	0
1	1	1	1	0	0	0	0	0	1	0	0	0	0	0
1	1	1	0	0	0	0	0	0	1	1	0	0	0	1
1	1	1	1	0	0	1	0	0	1	1	0	0	0	0
0	1	1	0	0	0	0	1	0	1	0	0	0	0	1
1	1	1	1	0	0	0	0	0	1	0	1	0	0	0
1	0	1	0	0	1	1	0	0	1	1	0	0	0	0
1	1	1	1	0	0	0	0	0	1	1	0	0	0	0
1	1	1	0	1	0	1	0	0	1	1	0	0	0	0
1	1	1	0	0	0	0	1	0	1	1	0	0	0	0
1	1	1	1	0	0	1	1	0	1	1	0	0	0	0
1	1	1	0	0	0	1	0	0	1	1	1	0	0	0
1	1	0	1	0	0	0	1	0	1	1	0	0	0	0

cobertur	fraqcia	accesor	usovehi	culpa	znurb	accnoche	accfins	testigos	reppolic	zona1	zona3	repsosp	parentez	retraso
1	1	1	0	0	0	1	0	0	1	0	1	0	0	0
1	0	1	0	0	0	1	0	0	1	1	0	0	0	1
1	0	1	1	0	0	0	1	0	0	1	0	0	0	0
1	1	1	1	0	0	0	0	0	1	1	1	0	0	0
1	1	1	0	0	0	1	0	0	1	1	0	0	0	0
1	1	1	0	0	1	1	0	1	1	1	0	0	0	0
0	1	1	0	1	0	1	1	0	1	1	1	0	0	1
1	1	1	1	0	1	0	0	0	1	1	0	0	0	0
1	1	1	0	0	0	1	0	0	1	1	0	0	0	0
1	1	1	0	0	0	1	0	0	1	1	0	0	0	0
1	1	1	1	0	0	1	0	0	1	0	0	0	0	0
1	1	1	1	0	0	1	1	1	1	1	0	0	0	0
1	1	1	0	1	0	1	0	0	1	1	0	0	0	0
1	1	1	1	0	0	0	0	0	1	1	0	1	0	0
1	1	1	1	0	0	0	1	1	1	1	1	0	0	0
1	0	1	1	0	0	0	1	1	1	0	0	0	0	0
1	1	1	1	0	1	0	1	0	1	1	0	0	0	0
1	1	1	0	0	0	0	1	0	1	0	0	0	0	0
1	1	1	1	0	0	1	0	1	1	0	0	0	0	0
1	1	1	0	0	0	1	1	0	1	1	0	0	0	1
0	1	1	0	0	0	0	1	0	1	1	1	0	0	0
1	1	1	1	0	0	1	1	0	1	1	0	0	0	0
1	1	1	1	0	0	1	1	0	1	1	1	0	0	0
1	1	1	1	0	0	0	1	0	1	1	0	0	0	1
1	1	1	0	0	0	1	0	0	1	0	0	0	0	0
1	1	1	1	0	0	1	0	0	1	1	0	0	0	0
1	1	1	0	0	0	0	1	0	1	1	1	0	0	0
1	1	1	1	0	0	0	0	0	1	1	1	0	0	0
1	1	1	0	0	0	1	0	0	1	1	0	0	0	0
1	1	1	1	0	0	0	1	0	1	1	0	0	0	0
1	1	1	1	0	0	1	0	0	1	1	0	0	0	0
1	1	1	1	0	0	1	0	0	1	1	0	0	0	0
1	1	0	0	1	0	1	0	0	1	0	0	0	0	0
1	1	1	1	0	1	1	0	0	1	1	0	0	0	0
1	1	1	1	0	1	1	1	0	0	1	0	0	0	0
1	0	1	1	0	0	0	0	0	1	1	0	0	0	0
1	1	1	1	0	0	1	1	0	1	1	0	0	0	1
1	1	1	1	0	0	0	1	0	1	1	0	0	0	1
1	1	1	1	0	0	0	0	0	1	1	0	0	0	0
1	1	1	1	0	0	0	0	0	1	1	0	0	0	0

cobertur	fraqcia	accesor	usovehi	culpa	znurb	accnoche	accfins	testigos	reppolic	zona1	zona3	repsosp	parentez	retraso
1	1	1	1	0	0	1	0	0	1	1	0	0	0	0
1	1	1	1	0	0	1	0	0	1	1	0	0	0	0
1	1	1	1	0	0	0	0	0	1	1	0	0	0	0
1	0	1	1	0	0	1	0	0	1	1	1	0	0	0
1	1	0	1	0	0	0	0	0	1	1	0	0	0	0
1	1	1	1	0	0	1	0	0	1	1	1	0	1	1
1	1	1	1	0	0	1	1	0	1	0	0	0	0	0
1	1	1	0	0	0	1	1	0	1	0	1	0	0	0
1	0	1	1	0	0	1	0	0	0	1	0	1	0	0
1	1	1	0	0	0	0	0	0	1	1	0	0	0	1
1	1	1	0	0	0	1	1	0	1	1	1	0	0	1
1	1	1	0	0	0	1	0	0	1	1	0	0	0	0
1	1	1	1	1	0	1	1	0	1	1	0	0	0	1
1	1	0	0	0	0	1	0	0	1	1	0	0	0	0
1	1	1	1	0	0	0	0	0	1	1	0	0	0	0
1	1	1	0	0	0	0	0	0	0	1	1	0	0	0
1	1	1	0	0	0	1	0	1	1	1	0	0	0	0
0	1	1	0	0	0	0	0	0	1	1	0	0	0	0
1	1	1	1	0	0	0	1	0	1	0	0	0	0	0
1	1	1	1	0	0	1	1	0	1	1	0	0	0	0

ANEXO 2

DATOS TRANSFORMADOS CON PUNTUACIONES RIDIT Y CLASES

cobertur	fraqcia	accesor	usovehi	aculpa	znurb	accnoche	accfins	testigos	reppolic	zona1	zona3	repsosp	parentez	retraso	Score	clase
-0.04	-0.1	-0.07	0.55	0.06	0.1	-0.43	0.38	-0.88	-0.05	-0.19	-0.79	0.03	0.01	0.14	0.23	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	0.57	0.38	0.12	0.95	0.81	0.21	0.03	0.01	-0.86	-0.05	1
-0.04	-0.1	-0.07	-0.45	0.06	0.1	0.57	-0.62	-0.88	-0.05	-0.19	-0.79	0.03	0.01	0.14	-0.13	1
-0.04	-0.1	-0.07	0.55	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	0.08	2
-0.04	-0.1	-0.07	0.55	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	-0.97	0.01	0.14	-0.24	1
-0.04	0.9	-0.07	-0.45	0.06	0.1	-0.43	0.38	-0.88	-0.05	-0.19	0.21	0.03	0.01	0.14	-1.07	1
-0.04	-0.1	-0.07	0.55	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	0.08	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.43	1
-0.04	-0.1	0.93	0.55	-0.94	0.1	-0.43	-0.62	0.12	-0.05	0.81	-0.79	0.03	0.01	0.14	2.04	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.43	1
-0.04	-0.1	0.93	0.55	0.06	0.1	-0.43	-0.62	0.12	-0.05	0.81	0.21	0.03	0.01	0.14	1.15	2
-0.04	-0.1	-0.07	0.55	0.06	0.1	0.57	0.38	0.12	-0.05	-0.19	-0.79	0.03	0.01	0.14	0.30	2
-0.04	-0.1	0.93	0.55	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	0.43	2
-0.04	0.9	-0.07	-0.45	0.06	-0.9	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-1.29	1
-0.04	-0.1	-0.07	0.55	0.06	-0.9	0.57	-0.62	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.17	1
-0.04	-0.1	-0.07	-0.45	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.43	1
-0.04	-0.1	-0.07	0.55	0.06	0.1	0.57	-0.62	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	0.22	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	-0.79	0.03	0.01	-0.86	0.26	2
-0.04	-0.1	-0.07	0.55	0.06	0.1	0.57	0.38	0.12	-0.05	0.81	0.21	0.03	0.01	0.14	0.44	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	0.57	-0.62	0.12	-0.05	0.81	0.21	0.03	0.01	0.14	0.18	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	0.57	0.38	0.12	-0.05	0.81	0.21	0.03	0.01	0.14	-0.07	1
-0.04	-0.1	-0.07	0.55	0.06	0.1	0.57	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	-0.86	0.33	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.43	1
0.96	-0.1	-0.07	0.55	0.06	0.1	0.57	-0.62	0.12	-0.05	0.81	0.21	0.03	0.01	-0.86	1.52	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	0.57	0.38	0.12	-0.05	0.81	-0.79	0.03	0.01	0.14	0.26	2
-0.04	0.9	-0.07	0.55	0.06	-0.9	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.78	1
-0.04	-0.1	-0.07	-0.45	0.06	0.1	0.57	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.54	1
-0.04	-0.1	-0.07	0.55	-0.94	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	0.64	2
-0.04	-0.1	-0.07	0.55	0.06	0.1	0.57	-0.62	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	0.22	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	-0.43	-0.62	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.18	1
-0.04	-0.1	-0.07	0.55	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	-0.79	0.03	0.01	0.14	0.40	2
-0.04	-0.1	0.93	-0.45	0.06	0.1	0.57	-0.62	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	0.07	2
-0.04	-0.1	-0.07	0.55	0.06	0.1	-0.43	0.38	0.12	-0.05	0.81	-0.79	0.03	0.01	0.14	0.87	2
-0.04	0.9	-0.07	0.55	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	-0.86	-0.02	1

cobertur	fraqcia	accesor	usovehi	aculpa	znurb	accnoche	accfins	testigos	reppolic	zona1	zona3	repsosp	parentez	retraso	Score	clase
-0.04	0.9	-0.07	-0.45	0.06	0.1	0.57	-0.62	0.12	0.95	-0.19	0.21	0.03	0.01	0.14	-1.09	1
-0.04	-0.1	-0.07	-0.45	0.06	0.1	0.57	0.38	0.12	-0.05	-0.19	-0.79	0.03	0.01	0.14	-0.21	1
-0.04	-0.1	-0.07	0.55	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	0.08	2
-0.04	-0.1	-0.07	0.55	0.06	-0.9	-0.43	0.38	-0.88	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.50	1
0.96	-0.1	-0.07	0.55	-0.94	0.1	-0.43	-0.62	0.12	-0.05	-0.19	-0.79	0.03	0.01	-0.86	2.05	2
-0.04	-0.1	-0.07	-0.45	0.06	-0.9	0.57	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.93	1
-0.04	-0.1	-0.07	0.55	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	0.08	2
-0.04	-0.1	-0.07	0.55	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	0.08	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	-0.43	0.38	0.12	-0.05	0.81	0.21	0.03	0.01	0.14	0.04	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	-0.43	-0.62	-0.88	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.36	1
-0.04	-0.1	-0.07	0.55	-0.94	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	0.64	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	0.57	0.38	0.12	-0.05	-0.19	0.21	-0.97	0.01	0.14	-0.85	1
-0.04	-0.1	-0.07	-0.45	0.06	0.1	0.57	-0.62	-0.88	-0.05	-0.19	-0.79	0.03	0.01	0.14	-0.13	1
-0.04	0.9	-0.07	-0.45	0.06	0.1	0.57	-0.62	-0.88	-0.05	0.81	0.21	0.03	0.01	0.14	-0.45	1
-0.04	-0.1	-0.07	-0.45	0.06	-0.9	0.57	-0.62	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.68	1
-0.04	-0.1	-0.07	0.55	0.06	0.1	0.57	-0.62	0.12	-0.05	0.81	0.21	0.03	0.01	0.14	0.69	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	-0.43	0.38	-0.88	-0.05	0.81	0.21	0.03	0.01	0.14	-0.14	1
-0.04	-0.1	-0.07	0.55	0.06	0.1	-0.43	-0.62	0.12	-0.05	-0.19	0.21	0.03	0.01	-0.86	0.69	2
0.96	-0.1	-0.07	0.55	0.06	0.1	0.57	-0.62	0.12	-0.05	-0.19	-0.79	0.03	0.01	0.14	1.02	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	-0.43	-0.62	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.18	1
-0.04	-0.1	-0.07	-0.45	0.06	0.1	-0.43	-0.62	0.12	-0.05	-0.19	-0.79	0.03	0.01	0.14	0.15	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	0.57	-0.62	0.12	-0.05	-0.19	0.21	0.03	0.01	-0.86	0.08	2
-0.04	-0.1	-0.07	0.55	0.06	0.1	-0.43	0.38	0.12	-0.05	0.81	0.21	0.03	0.01	0.14	0.54	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.43	1
-0.04	-0.1	-0.07	0.55	0.06	0.1	0.57	-0.62	0.12	-0.05	-0.19	-0.79	0.03	0.01	0.14	0.55	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	0.57	0.38	0.12	-0.05	-0.19	-0.79	0.03	0.01	0.14	-0.21	1
-0.04	-0.1	-0.07	0.55	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	0.08	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	0.57	-0.62	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.28	1
-0.04	-0.1	-0.07	-0.45	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.43	1
-0.04	-0.1	-0.07	-0.45	0.06	-0.9	0.57	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.93	1
-0.04	-0.1	0.93	0.55	-0.94	0.1	-0.43	0.38	0.12	-0.05	0.81	0.21	0.03	0.01	0.14	1.46	2
-0.04	-0.1	-0.07	-0.45	0.06	-0.9	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.83	1
-0.04	-0.1	-0.07	-0.45	0.06	-0.9	-0.43	-0.62	0.12	0.95	-0.19	0.21	0.03	0.01	0.14	-0.92	1
-0.04	0.9	-0.07	-0.45	0.06	0.1	0.57	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-1.00	1
-0.04	-0.1	-0.07	-0.45	0.06	0.1	-0.43	-0.62	0.12	-0.05	-0.19	0.21	0.03	0.01	-0.86	0.18	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	0.57	-0.62	0.12	-0.05	-0.19	0.21	0.03	0.01	-0.86	0.08	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	0.57	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.54	1
-0.04	-0.1	-0.07	-0.45	0.06	0.1	0.57	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.54	1
-0.04	0.9	-0.07	0.55	0.06	0.1	-0.43	-0.62	-0.88	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.31	1
-0.04	-0.1	-0.07	0.55	0.06	0.1	-0.43	-0.62	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	0.33	2

cobertur	fraqcia	accesor	usovehi	aculpa	znurb	accnoche	accfins	testigos	reppolic	zona1	zona3	repsosp	parentez	retraso	Score	clase
-0.04	-0.1	-0.07	0.55	0.06	0.1	0.57	0.38	0.12	-0.05	0.81	0.21	0.03	0.01	0.14	0.44	2
-0.04	-0.1	-0.07	0.55	0.06	0.1	0.57	-0.62	-0.88	-0.05	-0.19	0.21	0.03	0.01	0.14	0.05	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	0.57	0.38	-0.88	-0.05	-0.19	-0.79	0.03	0.01	0.14	-0.39	1
-0.04	-0.1	-0.07	-0.45	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.43	1
-0.04	-0.1	-0.07	0.55	0.06	0.1	-0.43	-0.62	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	0.33	2
-0.04	-0.1	-0.07	-0.45	0.06	-0.9	0.57	-0.62	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.68	1
-0.04	-0.1	-0.07	-0.45	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.43	1
-0.04	-0.1	-0.07	-0.45	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.43	1
-0.04	-0.1	-0.07	-0.45	0.06	0.1	0.57	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.54	1
-0.04	0.9	-0.07	-0.45	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	-0.79	0.03	0.01	0.14	-0.56	1
-0.04	-0.1	0.93	-0.45	0.06	0.1	0.57	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.18	1
-0.04	-0.1	-0.07	-0.45	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	-0.79	0.03	-0.99	-0.86	0.36	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	-0.43	-0.62	0.12	-0.05	0.81	0.21	0.03	0.01	0.14	0.29	2
-0.04	-0.1	-0.07	0.55	0.06	0.1	-0.43	-0.62	0.12	-0.05	0.81	-0.79	0.03	0.01	0.14	1.12	2
-0.04	0.9	-0.07	-0.45	0.06	0.1	-0.43	0.38	0.12	0.95	-0.19	0.21	-0.97	0.01	0.14	-1.55	1
-0.04	-0.1	-0.07	0.55	0.06	0.1	0.57	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	-0.86	0.33	2
-0.04	-0.1	-0.07	0.55	0.06	0.1	-0.43	-0.62	0.12	-0.05	-0.19	-0.79	0.03	0.01	-0.86	1.02	2
-0.04	-0.1	-0.07	0.55	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	0.08	2
-0.04	-0.1	-0.07	-0.45	-0.94	0.1	-0.43	-0.62	0.12	-0.05	-0.19	0.21	0.03	0.01	-0.86	0.74	2
-0.04	-0.1	0.93	0.55	0.06	0.1	-0.43	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	0.43	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	0.57	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.54	1
-0.04	-0.1	-0.07	0.55	0.06	0.1	0.57	0.38	0.12	0.95	-0.19	-0.79	0.03	0.01	0.14	-0.04	1
-0.04	-0.1	-0.07	0.55	0.06	0.1	-0.43	0.38	-0.88	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.10	1
0.96	-0.1	-0.07	0.55	0.06	0.1	0.57	0.38	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	0.44	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	0.57	-0.62	0.12	-0.05	0.81	0.21	0.03	0.01	0.14	0.18	2
-0.04	-0.1	-0.07	-0.45	0.06	0.1	-0.43	-0.62	0.12	-0.05	-0.19	0.21	0.03	0.01	0.14	-0.18	1

BIBLIOGRAFÍA

1. **ROA L. HEYDI**, “Modelo de Detección de Fraudes en los Seguros de Vehículos utilizando Componentes Principales y Análisis RIDIT” (Tesis, Instituto de Ciencias Matemáticas, Escuela Superior Politécnica del Litoral, 2004)

2. **FREUN J, WALPOLE R.** Estadística Matemática con aplicaciones, (Prentice Hall Hispanoamericana Cuarta Edición. México, 1990)

3. **JOHNSON, D**, (Métodos Multivariados aplicados al análisis de datos, (Internacional Thompson Editores, México DF, México, 2000)

4. **FERRAN A.** SPSS para Windows: Análisis Estadístico, McGraw-Hill, Madrid, España.,2001)

5. Tutorial paquete estadístico SPSS 10.0 para Windows versión en español, 2004

6. **The Journal of Risk and Insurance**, Fraud Classification Using Principal Component Analysis of RIDIT’s, Vol.69, No. 3, 2002.

7. **2004**, http://www.superban.gov.ec/pages/seguros_privados.htm