Licenciatura en Estadística

Facultades de Ciencias y Ciencias Económicas y Administración

Univesidad de la República

EQUIPOS – MORI

Elecciones 2004

¿A quién votan los indecisos?

Teresita Fuster Bardier

Tutora: Prof. Laura Nalbarte

Junio-Diciembre 2004

RESUMEN

El presente trabajo se enmarca dentro del programa de pasantías de la Licenciatura en Estadística en convenio con la empresa Equipos Mori y tiene como objetivo central la modelización de los datos provistos por las encuestas con el fin último de predecir el comportamiento de los indecisos y, por tanto, el resultado de la Elección Nacional de octubre de 2004.

Los datos utilizados en este trabajo provienen de la encuesta realizada por la empresa en agosto de 2004 y consta de un total de 993 individuos (856 decididos y 137 indecisos).

El modelo obtenido se basa en una Regresión Logística Multinomial (usando como base de datos “Decididos”) en la cual la variable de respuesta es la intención de voto. Las variables explicativas son seleccionadas de forma tal que acumulen la mayor cantidad de información posible de los datos originales (realizándose en algunos casos Análisis Factoriales para reducir dimensiones), tengan relevancia para explicar el comportamiento de la variable de respuesta y logren los menores errores de clasificación. Con el modelo obtenido se proyecta la base de Indecisos.

El modelo final tiene un alto poder predictivo: un 10% de error de clasificación global, siendo los valores totales para cada categoría de la variable de respuesta: Encuentro Progresista: 50.76%; Partido Nacional: 34.43% y Partido Colorado y Otras Respuestas 14.81%.

I- INTRODUCCIÓN

En el año 2004 se realizan Elecciones Nacionales en Uruguay. Como todos los años electorales, se caracteriza por la sobre información presentado por las distintas empresas de opinión pública que funcionan en plaza.

Desde el nacimiento del país a su vida independiente, han existido dos partidos, llamados normalmente Partidos Tradicionales, que son el Partido Colorado y el Partido Nacional. Durante el siglo XX fueron creándose otros partidos menores (algunos de ellos con fuertes referencias internacionales, como el Partido Comunista o el Partido Socialista). En 1971 se forma una nueva fuerza política con parte de estos partidos y grupos escindidos de los partidos tradicionales, el Frente Amplio. En 1994 esta coalición se amplía, pasando a denominarse Encuentro Progresista – Frente Amplio.

Hasta noviembre de 1994, la elección nacional se hacía en forma conjunta con las elecciones departamentales (Intendentes y Juntas Departamentales). La elección del presidente de la República era directa: el candidato más votado del partido con mayoría simple de votos era el elegido.

A raíz de la Reforma Constitucional aprobada por Plebiscito el 8 de diciembre de 1996, el sistema electoral uruguayo tuvo un cambio radical en cuanto a la elección de las autoridades de gobierno. Por un lado, las elecciones nacionales se separan de las departamentales. Además, la elección presidencial consta de dos etapas. En la primera (último domingo de octubre) se eligen las Cámaras de Senadores y Diputados[1]. Si uno de los partidos políticos alcanza la mitad más uno de los votos emitidos, también se elige el Presidente de la República.[2] Si esto no sucede, se lleva a cabo la segunda etapa o Balotaje (el último domingo de noviembre) donde se elige el presidente entre los candidatos de los dos partidos que hayan obtenido mayor número de votos en la instancia anterior. En este caso será electo quién obtenga la mayoría simple de votos emitidos.

Durante el último año, los resultados de las distintas encuestas de opinión han dado a uno de los partidos políticos (el Encuentro Progresista – Frente Amplio – Nueva Mayoría) un porcentaje de intención de voto muy cercano al 50%. Este valor es muy importante, ya que están en juego dos cosas: la mayoría parlamentaria (que se logra con el 50% de los votos válidos, o sea, sin tener en cuenta los votos en blanco o anulados) y la posible elección directa, sin Balotaje, del Presidente y Vicepresidente de la República.

[1] Artículo 88 Constitución de la República - 1996
[2] Artículo 151 Constitución de la República - 1996

En estas circunstancias el pronunciamiento de las personas que al momento de la encuesta no tenían decidido su voto es fundamental, ya que pueden definir la elección en primera vuelta o la necesidad de Balotaje.

El presente trabajo se enmarca dentro del programa de pasantías de la Licenciatura en Estadística de las Facultades de Ciencias y Ciencias Económicas y Administración (Universidad de la República – Montevideo – Uruguay) en convenio con la empresa Equipos Mori y tiene como objetivo central la modelización de los datos provistos por las encuestas con el fin último de predecir el comportamiento de los indecisos y, por tanto, el resultado de la Elección Nacional de octubre de 2004.

I.1 OBJETIVOS

El objetivo principal es encontrar modelos que permitan predecir el comportamiento electoral de los ciudadanos indecisos, o sea, aquellos que al momento de efectuar la encuesta no tienen decidido qué partido van a votar o no manifiestan esta intención.

Durante el período de pasantía se trabaja basándose en dos resultados: en uno de ellos se trata de encontrar un modelo con pocas variables, fácil y rápido de implementar, que pueda ser usado durante las últimas semanas de la campaña electoral, momento en el cual se realizan encuestas de opinión semanales o incluso, a diario en forma telefónica.

El otro modelo es el que se presenta en este trabajo. Tiene mayor complejidad e implica diversas técnicas, aunque su implementación no sea tan sencilla.

I.2 ANTECEDENTES

Los antecedentes sobre proyección de indecisos en las distintas empresas consultoras de Opinión Pública de Uruguay indican que las estimaciones se basan en criterios de corte más bien sociológicos e históricos: de acuerdo a variables demográficas, económicas y de definiciones políticas básicas, se intenta predecir el voto de los indecisos.

En algunos casos se han efectuado regresiones logísticas binomiales (realizando tantas regresiones como combinaciones de Partidos Políticos interesen). Estas estimaciones tienen como característica desfavorable que las probabilidades de pertenencia a cada partido no suman uno, lo cual es imprescindible al tratarse de categorías exhaustivas y mutuamente excluyentes. Asimismo, los modelos se han confeccionado tomando como base la totalidad de los individuos, sin discriminar si ya tienen decidido o no su voto (lo que implica que también esté presente la modalidad “Indecisos”).

II - ASPECTOS METODOLÓGICOS

II.1 DATOS

Los datos utilizados en este trabajo provienen de la encuesta realizada por la empresa Equipos Mori entre los días 14 y 23 de agosto de 2004[3] y consta de un total de 993 individuos. La misma se realizó bajo la modalidad de entrevista "cara a cara". La población objetivo son los ciudadanos uruguayos que habitan el territorio de la República. La selección de la muestra sigue un diseño en varias etapas, siendo la primera de ellas una estratificación del país en ciudades, pueblos y zonas rurales, teniendo en cuenta la cantidad de habitantes según el Censo Nacional de Población y Viviendas de 1996. El número de encuestados por estrato es proporcional a la población del mismo. Se han tenido en cuenta las modificaciones más importantes desde el último censo, particularmente la creación de nuevos barrios en las periferias de las ciudades principales. La selección de las ciudades participantes en la muestra se realiza mediante un sorteo en el cual el peso de cada una es proporcional al número de habitantes.

En la encuesta se realizan preguntas referentes al comportamiento electoral presente e histórico de la persona encuestada, además de su caracterización en aspectos socio demográficos. De ese total se seleccionan aquellas preguntas que tienen relación directa con el objetivo del trabajo y que pueden ser relevantes para el mismo. Esta selección se realiza en acuerdo con los técnicos de la empresa.

Estas variables se dividen en bloques usando un criterio temático para la clasificación. Estos bloques son: Variables Básicas, Variables de Coyuntura, Variables de Trayectoria Electoral, Variables de Definiciones Políticas Básicas, Variables sobre popularidad de líderes.[4]

[3] La decisión de cuál de las distintas muestras sería usada en el trabajo quedó a cargo de la empresa.
[4] Ver cuadro con variables en Anexo 1.

II.2 MODELO

En el trabajo se utiliza un modelo de Regresión Logística Multinomial cuya variable de respuesta es la intención de voto para las próximas elecciones y las variables explicativas son definidas conjuntamente con el personal especializado de la Empresa (en especial los Licenciados en Sociología Ignacio Zuasnabar y Agustín Canzani).
El número de variables propuestas inicialmente como explicativas es muy numeroso, por lo que se estudian en los distintos agrupamientos mencionados anteriormente a efectos de poder determinar el conjunto final de variables a ser consideradas en el modelo. En cada bloque se analizan asociaciones y/o correlaciones (dependiendo si el bloque es de variables cualitativas o cuantitativas) así como técnicas de Análisis Factoriales con el objetivo fundamental de reducir dimensiones. A su vez este análisis se complementa con otra técnica exploratoria como Árboles de Clasificación, con el fin de una mejor comprensión de los datos, que permita definir la estrategia de análisis final, o sea, definir el conjunto de variables explicativas.

II.3 SUPUESTOS

El supuesto fundamental de trabajo es que las mismas variables que son importantes para las personas que ya tienen definido su voto, también son relevantes para la definición del voto de los indecisos.

II.4 CONCEPTOS TEÓRICOS

Como se menciona anteriormente, en el presente trabajo se utiliza como modelo el de Regresión Logística Multinomial, el cual es complementado por un conjunto de técnicas exploratorias. A continuación se presentan los conceptos teóricos de las diferentes técnicas utilizadas.

II.4.1 REGRESIÓN LOGÍSTICA MULTINOMIAL

La Regresión Logística Multinomial es una técnica que pretende explicar el comportamiento de una variable cualitativa con tres o más categorías, a partir de variables explicativas que pueden ser cuantitativas y/o cualitativas. Es una generalización del modelo de Regresión Logística Binomial. Los objetivos generales tienen tres características: descriptivo, predictivo y de reclasificación. El análisis es descriptivo pues intenta estudiar las relaciones entre las variables explicativas y de respuesta, de acuerdo a los datos de la

muestra. Es predictivo al utilizarlo para proyectar los resultados obtenidos en nuevos individuos, de los cuales se posee las características expresadas en las variables explicativas. La reclasificación es utilizada cuando la variable de respuesta proviene de técnicas previas de clasificación tales como Análisis de Clusters (y no de la realidad), pudiéndose obtener una mejor clasificación de los individuos en los grupos propuestos.

La función de unión (link) entre la variable de respuesta y las explicativas es la llamada "logística", y sus valores están siempre entre 0 y 1, lo que permite modelar una probabilidad. Se utiliza el logaritmo neperiano de esta función de unión para linealizar el modelo. Se denomina "odds" al cociente de probabilidades de dos categorías de la variable. El logaritmo neperiano de los odds es llamado "transformación logit".

En modelos donde la variable de respuesta presenta J categorías se obtienen J-1 logits.[5] En general se toma una de las categorías como referencia (baseline) para plantear los cocientes respecto a la misma.

Sea: $\pi_j = \Pr ob(Y = j \mid X)$ con j = 0, 1, 2.

Si se toma como categoría de referencia j = 0:

$$g_1(X) = Ln\left(\frac{\pi_1}{\pi_0}\right) = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + ... + \beta_{1K}x_K$$

siendo β_{1k} los parámetros a estimar en el primer logit y K el total de variables explicativas

$$g_2(X) = Ln\left(\frac{\pi_2}{\pi_0}\right) = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + ... + \beta_K x_K$$

siendo β_{2k} los parámetros a estimar en el segundo logit

La relación entre las modalidades 1 y 2 se obtiene al operar con los logits planteados: $Ln\left(\frac{\pi_2}{\pi_1}\right) = g_2(X) - g_1(X)$

Las probabilidades de pertenencia a cada modalidad de la variable de respuesta quedan dadas por:

$$\pi_0 = \frac{1}{1 + \exp(g_1(X)) + \exp(g_2(X))}$$

siendo $X_{n*(K+1)}$ la matriz de diseño

$$\pi_1 = \frac{\exp(g_1(X))}{1 + \exp(g_1(X)) + \exp(g_2(X))}$$

[5] La ilustración del modelo se realiza con tres categorías como una forma de simplificar la notación.

$$\pi_2 = \frac{\exp(g_2(X))}{1+\exp(g_1(X))+\exp(g_2(X))}$$

Las estimaciones de los parámetros β_{jk} se realizan mediante el método de máxima verosimilitud (o sea, hallando el logaritmo de la función de verosimilitud de la distribución de la variable y luego las derivadas respecto a cada uno de los parámetros para encontrar el vector que maximice esta función). En general se utiliza el algoritmo iterativo de Newton Raphson, que necesita como insumo para la convergencia una matriz estrictamente dominante.

La variancia estimada asintótica de los parámetros se obtiene como:

$\hat{Var}(\hat{\beta}) = (X'VX)^{-1}$ siendo V la matriz diagonal de n x n, cuyo elemento general es de la forma $\hat{\pi}_{ij}(1-\hat{\pi}_{ij})$

Los intervalos de confianza para los parámetros se obtienen de la forma habitual:

$L(\beta) = \hat{\beta} - s(\hat{\beta}) * z_{\alpha/2}$ con α = nivel de confianza

$U(\beta) = \hat{\beta} + s(\hat{\beta}) * z_{\alpha/2}$

Los intervalos de confianza para exp(β) se obtienen a partir de los intervalos de confianza de los parámetros, con la salvedad de que no son simétricos en torno a la estimación puntual máximo verosímil.

$L*(e^{\beta}) = \exp(L(\beta_j))$ con j = 1, 2

$U*(e^{\beta}) = \exp(U(\beta_j))$

La prueba de significación del modelo se puede realizar a través de distintos test, entre ellos el de razón de verosimilitudes (RV) y la deviance.

- En el caso de la *razón de verosimilitudes*, la prueba de hipótesis planteada es:

H_0) $\beta_k = 0$ $\forall\ \beta_k$ k=1, 2 , … K

H_a) alguno de los β_k es distinto de cero

El estadístico a usar es:

$RV = -2Ln\left(\frac{L_0}{L_M}\right)$ siendo L_0 la función de verosimilitud de un modelo bajo H_0 cierta y L_M la verosimilitud del modelo bajo H_a. Este estadístico se distribuye aproximadamente χ^2_l con l grados de libertad, siendo l = (N° categorías –1)*(Σ grado de libertad de cada variable).

- En el caso de la *deviance*, la prueba planteada es:

H_0) el modelo M planteado se ajusta a los datos

H_a) el modelo no se ajusta

$Deviance = 2Ln(L_s) - 2Ln(L_M)$ siendo L_s el modelo saturado (o sea, el que pasa por todos los puntos de la muestra).

Este estadístico se distribuye aproximadamente χ^2_l con l = (n-(k+1))*(categorías –1)

Si el modelo es anidado, la Deviance coincide con la Razón de Verosimilitudes.

La prueba de significación de los parámetros en los distintos logits se realiza planteando

H_0) $\beta_{jk} = 0$

H_a) $\beta_{jk} \neq 0$

El estadístico a utilizar es el de Wald, que se distribuye asintóticamente normal y cuya fórmula es: $z = \frac{\hat{\beta}_{jk}}{s(\hat{\beta}_{jk})}$

El poder predictivo del modelo se analiza teniendo en cuenta la clasificación final del modelo, es decir, los errores de clasificación en cada categoría así como el error de clasificación global. El error de clasificación en cada categoría i se define como el porcentaje de individuos que fueron clasificados en la categoría i cuando pertenecen a la categoría j. En este caso se define P(i | j) a la probabilidad de clasificación errónea. Existen varias formas de asignar un individuo a un grupo determinado, entre ellas:

- Minimizar errores de clasificación

  Probabilidad de error: es la probabilidad de asignar un individuo x_i al grupo G_j cuando pertenece al grupo G_h. La notación usual es P(j | h)

  Sea $\pi_j = P(x_i \in G_j)$ y $f_j(x)$ = función de densidad o cuantía de X si $x \in G_j$

  Probabilidad de error total: $\sum P(j \mid h)\pi_h$

  La mejor regla es la que minimiza el error total.

- Maximizar la probabilidad a posteriori

$$P(x_i \in G_j \mid \text{mod } elo) = \frac{P(\text{mod } elo \mid x_i \in G_j)P(x_i \in G_j)}{\sum_k P(x_i \in G_k)P(\text{mod } elo \mid x_i \in G_k)} = \frac{f_j(x)\pi_j}{\sum_k f_k \pi_k}$$

Se asigna el individuo x_i al grupo G_j cuando $\frac{f_j(x)}{f_k(x)} > \frac{\pi_j}{\pi_k} \forall k \neq j$

- Maximizar la verosimilitud (probabilidad a priori iguales)
  Si los π_j son desconocidos, entonces asigno x_i al grupo G_j de forma tal que se maximice la función de verosimilitud de x_i (es el caso particular en que $\pi_j = 1/J \ \forall$ j).
- En algunos casos pueden existir costos de clasificación errónea y pueden ser incorporados en la decisión. Ésta se toma de forma tal de minimizar los costos de clasificación errónea.

Para la exploración y reducción de las variables a incluir en el modelo, se utilizan otras técnicas multivariadas que se describen a continuación.

II.4.2 ANÁLISIS FACTORIAL

El Análisis Factorial se enmarca dentro de los métodos multidimensionales, ya que trata dos o más variables simultáneamente. Permite la confrontación de información numerosa, lo cual lo hace mucho más rico que estudiar cada variable por separado o en combinaciones binarias. Se extraen las tendencias más sobresalientes de datos demasiado numerosos para ser aprehendidos directamente, se jerarquizan y eliminan efectos marginales o puntuales que perturban la percepción global de los hechos. Al utilizar métodos gráficos permite transformar en distancias euclídeas las proximidades entre los datos.

Aunque existen varios métodos comprendidos dentro del Análisis Factorial, todos tienen en común el partir de tablas rectangulares de individuos por variables. Los objetivos comunes son: en cuanto a los individuos, evaluar su semejanza (dos individuos son más semejantes cuanto más próximos sean sus valores en el conjunto de las variables); en cuanto a las variables, se trata de evaluar su relación (la proximidad entre variables estará dada en cuanto más individuos compartan simultáneamente). Otro de los objetivos comunes es tratar de reducir dimensiones del análisis sin perder demasiada información. En la práctica se busca una serie de direcciones llamadas ejes factoriales. Cada dirección hace máxima la inercia respecto al baricentro. Una vez encontrada la primera, se impone a las siguientes ser ortogonales a la ya encontradas. El plano formado por las dos primeras direcciones halladas, hace máxima la inercia proyectada sobre él y así sucesivamente. El

hacer máxima la inercia es equivalente a minimizar la desviación entre la nube y su proyección. La inercia de un elemento M con relación al centro de coordenadas 0 se define como el producto del peso del elemento por el cuadrado de la distancia entre M y O. La inercia de un conjunto de elementos es la suma de las inercias de cada uno de ellos. La noción mecánica de inercia de una nube de puntos respecto a su baricentro se corresponde con la noción estadística de variancia.

En este trabajo se utilizan en particular dos métodos de Análisis Factorial: el Análisis de Componentes Principales (ACP) y el Análisis de Correspondencia Múltiple (ACM).

II.4.2.1 Análisis de Componentes Principales

El Análisis de Componentes Principales (ACP) se aplica a tablas bidimensionales que cruzan individuos (filas) con variables continuas (columnas). Se anota como I al número total de individuos y J al total de variables.

Si el interés se centra en los individuos, se debe mirar la tabla como una yuxtaposición de filas, en donde a cada individuo se le asocia un conjunto de J números (los individuos pertenecen al espacio R^J). Se llama “nube de individuos” a este conjunto.

Si el interés es en las variables, se mira la tabla como yuxtaposición de columnas. A cada variable le corresponde una sucesión de I números (las variables pertenecen a un espacio de R^I) Se denomina “nube de variables” a este conjunto.

Previa a la utilización del ACP es conveniente, en la mayoría de los casos, tipificar los datos. Si x_{ij} es el valor genérico del individuo i en la variable j, los nuevos valores se obtienen al aplicar la fórmula: $z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$. De esta forma, el baricentro de la nube de individuos aparece como el origen de los ejes (importa la distancia entre los mismos) y las variables se encuentran en una hiperesfera de radio 1 (importa el ángulo que forman los vectores, ya que el coseno de este ángulo es igual a la correlación entre las variables). En este caso la inercia proyectada es igual a J, o sea, el número de variables.

En el caso de la nube de individuos, debido al centrado, el criterio de máxima inercia respecto al centro de gravedad permite interpretar a los ejes factoriales como direcciones de máximo alargamiento de la nube. Se habla también de factores de máxima variabilidad, ya que reflejan lo más posible la diversidad de los individuos. En el caso de la nube de variables se trata de que los ángulos formados por los vectores que las representan resulten lo menos deformados posible.

La inercia asociada a cada eje puede interpretarse como la porción de la varianza de los datos captada por cada uno de ellos. Por el hecho de ser los ejes ortogonales, la inercia acumulada por un conjunto de ejes es igual a la suma de las inercias en cada uno de ellos. Una forma de saber cuántos ejes serán utilizados en el análisis es, justamente, mirando el porcentaje de variabilidad acumulada por ellos o en qué momento la incorporación de un nuevo eje no aumenta sustancialmente la inercia acumulada.

Como resultado del análisis se obtiene un conjunto de variables (factores) que son combinaciones lineales de las variables originales.

En la nube de individuos (filas), los factores están dados por la fórmula:

$$F_s = XMu_s$$

donde X es la matriz de datos (de I x J) ya centrada o estandarizada

M es la métrica de las variables (generalmente la matriz identidad de J x J)

u_s es el autovector asociado al s-ésimo valor propio que se obtienen al diagonalizar la matriz $X'DX$ (con D = matriz de pesos de los individuos, generalmente la matriz diagonal con elemento genérico 1/I)

En la nube de variables (columnas) los factores están dados por la fórmula

$$G_s = X'Dv_s$$

donde las matrices son las vistas. Se cumple además que los autovalores en ambos casos son los mismos.

Existen fórmulas de transición que relacionan ambos análisis:

$$F_s = \frac{1}{\sqrt{\lambda_s}} XMG_s \qquad \text{y} \qquad G_s = \frac{1}{\sqrt{\lambda_s}} X'DF_s$$

La interpretación del análisis se realiza básicamente con los siguientes elementos:

- Calidad de representación de un individuo: se mide por la razón

  Inercia de la proyección de i sobre el eje s / Inercia total de i

  Coincide con el cuadrado de ángulo que forman Oi y el eje s
- Calidad de representación de la nube sobre un eje. Se llama también porcentaje de inercia asociado al eje. Se mide por

  Inercia de la proyección de la nube / Inercia total de la nube

Si se desea calcularla respecto a varios ejes, es la suma de las inercias asociadas a cada uno de ellos.

- Contribución de un elemento a la inercia de un eje: es el cociente entre la inercia del elemento en el eje sobre la inercia total de la nube en ese eje
- El estudio de outliers se realiza a través de la norma (distancia del individuo al baricentro) y su calidad de representación.

II.4.2.2 Análisis de Correspondencia Múltiple

El Análisis de Correspondencia Múltiple permite estudiar una población de I individuos descriptos por J variables cualitativas. Una de las aplicaciones más comunes del ACM es el tratamiento del conjunto de respuesta a una encuesta: cada pregunta constituye una variable cuyas modalidades son las respuestas propuestas (entre las cuales el individuo debe elegir una y solo una: o sea, se trata de modalidades exhaustivas y mutuamente excluyentes).

El Análisis de Correspondencia Múltiple es una técnica que al igual que el Análisis de Componentes Principales, tendrá como uno de sus objetivos reducir dimensiones. Al ser los datos cualitativos se definen diferentes distancias. En general se trabaja con nubes de perfiles (fila y columna) y la distancia χ^2.

Existen dos maneras de presentar los datos para su análisis: la Tabla Disyuntiva Completa (en la cual las filas son los individuos y las columnas son las diferentes modalidades de cada una de las variables. En la intersección de la fila i con la columna k solo existen dos valores posibles: 1 si el individuo posee esa modalidad y 0 si no la posee) o la Tabla de Burt (tabla en la cual se cruzan las variables dos a dos. Es una tabla simétrica, en la cual las subtablas diagonales son a su vez diagonales, ya que en la misma variable un individuo puede tener una única modalidad). Los análisis realizados a partir de cada una de ellas son equivalentes.

Con respecto a la inercia en un ACM hay que destacar que, cualquiera sea la estructura de la tabla, el porcentaje de inercia asociado a cada factor (y en particular al primero) es necesariamente débil cuando las variables tienen muchas modalidades. Incluso si un factor está muy ligado a una variable es imposible que todas sus modalidades estén bien representadas. Si el ACM se realiza a partir de una tabla de Burt, la inercia total de la nube tiene relación con la estructura misma de los datos. Si solamente se toman dos variables (Análisis de Correspondencia Simple), la inercia de la nube es proporcional al estadístico χ^2 de independencia. Utilizando el hecho de que las marginales de la tabla de Burt son proporcionales a las marginales de las subtablas que cruzan las variables dos a

dos, se puede mostrar que la inercia total es igual a la suma de los χ^2 de independencia asociadas a cada una de las J^2 subtablas. En esta suma, las tablas que cruzan dos variables distintas intervienen dos veces y las que cruzan una variable con ella misma son diagonales y su χ^2 nunca es nulo. Por eso la inercia en una tabla de Burt nunca es nula, aún cuando todas las variables sean independientes.

En el caso del ACM, no se pueden usar las mismas indicaciones que en el ACP para definir el número de ejes a utilizar, ya que el aporte de cada uno de ellos es muy pequeño. Existen indicadores que permiten decidir la cantidad de ejes a incorporar, tales como el índice de Benzecri o el de Greenacre, que utilizan los valores propios asociados a cada eje pero relativizando su importancia. En ambos casos se utilizan únicamente los valores propios que son mayores a la media de los mismos. Las fórmulas respectivas son:

Benzecri: $\varphi(\lambda)=\left(\frac{J}{J-1}\right)^2 * \left(\lambda-\frac{1}{J}\right)^2$ Greenacre: $\varphi(\lambda)=\left(\frac{J}{J-1}\right)^2 * \left(\sqrt{\lambda}-\frac{1}{J}\right)^2$

II.4.3 ÁRBOLES DE CLASIFICACIÓN Y REGRESIÓN

Es una técnica no paramétrica muy usada debido a su sencillez y aplicabilidad.

Puede verse como la estructura resultante de la partición recursiva del espacio de representación a partir de un conjunto numeroso de prototipos. Esta partición recursiva se traduce en una organización jerárquica del espacio de representación que puede modelarse mediante una estructura de tipo árbol. Cada nodo interior contiene una pregunta sobre un atributo concreto (con un hijo por cada posible respuesta) y cada *nodo hoja* se refiere a una decisión (clasificación). Un Árbol de Decisión genera sub grupos que contienen elementos homogéneos dentro de ellos y heterogéneos entre distintos subgrupos. Pueden definirse como una forma de representar el conocimiento obtenido en el proceso de aprendizaje inductivo.

La clasificación de patrones se realiza basándose en una serie de preguntas sobre los valores de sus atributos, empezado por el nodo raíz y siguiendo el camino determinado por las respuestas a las preguntas de los nodos internos, hasta llegar a un nodo hoja. La etiqueta asignada a esta hoja es la que se asignará al patrón a clasificar.

Los componentes del análisis son: la variable dependiente (que puede ser cuantitativa o cualitativa según de trate de un Árbol de Regresión o Clasificación

respectivamente) y variables independientes (dependiendo de la técnica usada pueden ser categóricas, cuantitativas o ambas). Pueden existir grupos de datos de entrenamiento y de testeo (si el número de datos originales es lo suficientemente grande como para dividirlo) Existen distintos tipos de Árboles de Clasificación: binarios (si la división de un nodo se realiza en dos nodos hijos) o multiway (si un nodo puede dividirse en tres o más nodos hijos).

En el presente trabajo se utiliza la técnica CART (Classification and Regression Tree)[6], que se desarrolla a continuación. Se refiere a un Árbol binario, que puede ser de Regresión o de Clasificación. Las variables explicativas pueden ser tanto cualitativas como cuantitativas. Se obtienen como resultados las variables que más discriminen, los aciertos de clasificación, tanto global como en grupos, determinando así el poder predictivo.

La construcción del Árbol se realiza siguiendo las siguientes reglas, que siguen un esquema recursivo:

1. El avance está basado en la partición de un nodo de acuerdo a alguna regla, normalmente evaluando una condición sobre el valor de alguna variable. Los prototipos que verifican la condición se asignan a uno de los dos nodos hijo (normalmente el izquierdo) y los restantes, al otro. Cuando un nodo se particiona, pasa a ser un nodo *intermedio*.
2. El caso base *o* condición de parada tiene como objetivo detener el proceso de partición de nodos. Cuando se verifica la condición de parada en un nodo, éste es un *nodo hoja*. Los prototipos asociados a un nodo hoja constituyen un agrupamiento homogéneo, por lo que al nodo se le asigna una *etiqueta*. En ocasiones, se simplifica el árbol resultante utilizando alguna regla de poda.

Para hacer la división se debe elegir la variable y el punto de corte que maximicen los cambios en la deviance: deviance – deviance_{Iz} – deviance_{Der}. La deviance en el nodo t se define, en este caso, de la siguiente manera:

$$D(t) = -2\sum_j n_{jt}\, p(j \mid t) Ln(p(j \mid t))$$

siendo: p(j|t) la probabilidad de pertenencia a la clase j dado que pertenece al nodo t.

n_{jt} el número de individuos de la clase j que pertenecen al nodo t.

Otra forma de efectuar la división es usar el índice de Gini, que mide la diversidad de clases en un nodo:

$$IndicedeGini = 1 - \sum_j p(j \mid t)^2$$

[6] Esta técnica fue desarrollada por Breiman, Friedman, Olshen and Stone a partir de 1984

Si se tiene información sobre costos de clasificación errónea, entonces se trata de minimizar ésta.

En cuanto a la etiqueta que se le asigna a un nodo, la forma más fácil es por medio de la clase que sea mayoría en ese nodo. Si el máximo es compartido por dos o más clases se realiza un sorteo y la asignación será arbitraria a cualquiera de ellas.

III - RESULTADOS

III.1 CONSIDERACIONES GENERALES

En este trabajo se busca explicar la opción del elector por alguna de las modalidades de la variable SIMPAT (simpatía política) de acuerdo a diferentes variables. En el modelo final, alguna de las variables son las originales de la encuesta (recodificadas en caso de ser necesario), otras se crean en base a dos o más variables y las restantes resultan de aplicar previamente técnicas de Análisis Factorial (ACP o ACM según corresponda). La selección de las variables a incluir depende principalmente de la significación de las mismas en el modelo.

Para la creación del modelo la base original se divide en dos a partir de la variable que se pretende explicar. En una de estas bases se encuentran todos los individuos que ya tenían decidido su voto en las Elecciones Nacionales al momento de realizarse la encuesta ("Decididos")[7]. Es el punto de partida para la formulación del modelo. La otra base cuenta con los individuos que, a la fecha de la encuesta, aún no sabían qué partido iban a votar ("Indecisos").[8] En estos individuos se usan los parámetros obtenidos en el modelo para efectuar la predicción de cuál es la categoría con mayor probabilidad de pertenencia.

En una primera instancia se estudian las variables de cada bloque por separado, con los objetivos de tener una mejor comprensión de las mismas y reducir, siempre que sea posible, la cantidad de variables a incluir en el análisis. En cada uno de los bloques se realiza un Análisis Factorial.

Los análisis multivariados se realizan en el programa S-plus (versión para Estudiantes) y en R.

III.2 ESTUDIO DE CADA BLOQUE DE VARIABLES

III.2.1 VARIABLES BÁSICAS

Las variables de este bloque que intervienen en el análisis son las siguientes:
Género, Edad, Índice socio económico, Educación y Región Geográfica[9]

Con el objetivo de reducir dimensiones se realiza un Análisis de Correspondencia Múltiple con las variables del bloque. Las modalidades de las variables usadas no están, en general, bien representadas y los ejes no tienen una interpretación clara. Por estos motivos,

[7] Ver Estadística Descriptiva en Anexo 2
[8] ídem

se realiza otro tipo de análisis exploratorio previo (en este caso, un Árbol de Clasificación) para estudiar cuáles de ellas pueden tener mayor poder explicativo para la variable de respuesta. De allí se extraen las variables que finalmente son usadas en el modelo, en el cual se decide finalmente su significación. Los resultados de este análisis se presentan en el Anexo 3.

III.2.2 VARIABLES DE COYUNTURA

Las variables consideradas en este bloque son: Situación económica personal y familiar actual; Situación económica del país actual; Percepción de la situación económica personal y familiar dentro de un año; Percepción de la situación económica del país dentro de un año y Opinión sobre la gestión del Presidente de la República.[10] Con la finalidad de reducir el número de variables a incluir en el modelo final, se realiza un ACM con algunas de las variables del bloque.

III.2.2.1 Análisis de Correspondencia Múltiple

El análisis se realiza usando las variables de situación económica personal y de país actual y las perspectivas a un año (No se usa la variable Gestión del Presidente por las características que posee: solo el 7% de los encuestados pertenecen a la categoría "Aprueba").

Se utilizan en el análisis los dos primeros ejes factoriales. La inercia acumulada por ellos es del 55 % de la inercia total, si se considera el índice ponderado de Benzecri.

Las variables que quedan mejor representadas son situación personal y de país dentro de un año, siendo las modalidades extremas (Mejor y Peor en ambos casos) las que tienen mayor calidad de representación.

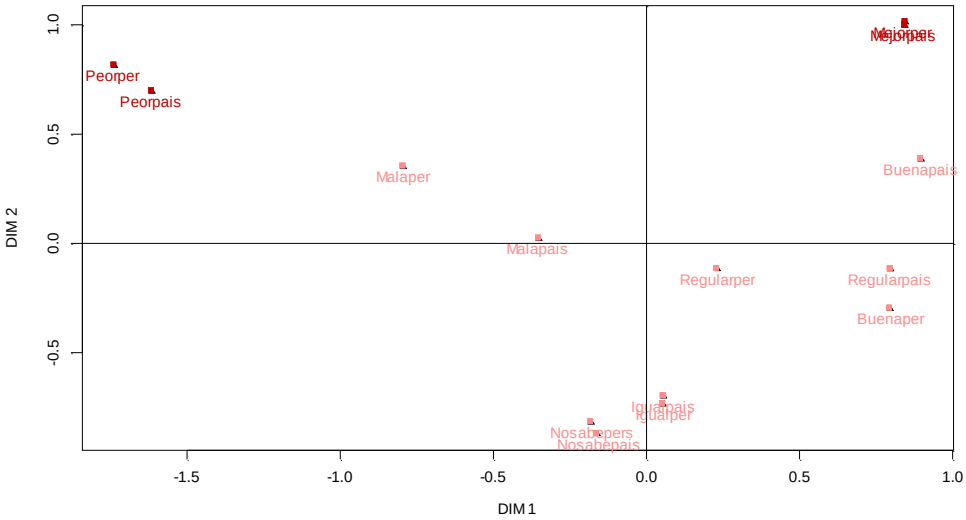
La interpretación de los factores es la siguiente:

- ✓ Primer eje factorial: Separa las categoría Peor o Mala (según la variable) con coordenadas negativas, de Mejor o Buena con coordenadas positivas. Se puede decir que ordena los individuos según su visión de la situación económica general y de los cambios que espera. Se observa una cercanía en general entre las modalidades correspondientes a la situación personal y del país.

[9] Ver definiciones de cada variable en Anexo 1 y Estadísticas Descriptivas en Anexo 2
[10] Ver definiciones de cada variable en Anexo 1 y Estadística Descriptiva en Anexo 2

- ✓ Segundo eje factorial: básicamente separa las categorías extremas de las centrales.
- ✓ El primer plano factorial divide claramente cuatro situaciones. En el primer cuadrante: perspectivas positivas sobre economía personal y general; en el segundo cuadrante: perspectivas negativas de situación económica; en el tercero: no emite opinión sobre sus perspectivas; y en el cuarto cuadrante: situación actual regular y pocas perspectivas de cambio a futuro.

Figura 1 - ACM Variables de Coyuntura



Fuente: Elaboración propia – Base agosto 04 – Equipos Mori

III.2.2.2 Conclusiones

El primer eje factorial (que posee un 33 % de la inercia total) puede tomarse como un buen resumen de las variables del bloque: tienen "puntajes" negativos los individuos que tienen una visión negativa de la situación económica (tanto personal como a nivel país) y que no esperan cambios a corto plazo. En el otro extremo se ubican los que piensan que la situación económica actual es buena y tienen perspectivas favorables en el futuro inmediato. Teniendo en cuenta lo planteado y la necesidad de significación del modelo final, este primer eje factorial es usado como insumo en los análisis posteriores.

III.2.3 VARIABLES DE TRAYECTORIA ELECTORAL

Las variables correspondientes a este bloque son las siguientes: Voto en Elección Nacional 1994; Voto en Elección Nacional 1999; Voto en Balotaje 1999; Voto en Plebiscito por ANCAP.[11] Por razones análogas a las descriptas anteriormente, se realiza un ACM tratando de reducir las dimensiones de los datos.

III.2.3.1 Análisis de Correspondencia Múltiple

Se utilizan todas las variables del bloque.[12]

Si se aplica la ponderación de Benzecri, se obtiene que los tres primeros ejes acumulan un 81 % de la inercia total.

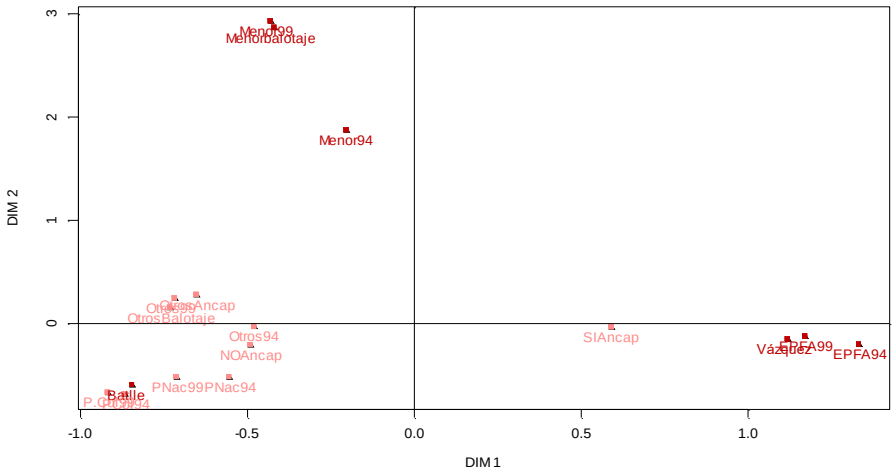
Las modalidades que tienen mayor calidad de representación son: EPFA99 (0.88); Menor99 (0.85); Vázquez (0.86); Batlle (0.63); EPFA94 (0.74) y Menor94[13] (0.68).

La interpretación de los ejes es la siguiente:

- ✓ Primer eje factorial: separa las modalidades correspondientes al EPFA en todas las variables y Sí Ancap (coordenadas positivas) del resto. Ordena los individuos de Derecha a Izquierda del espectro político nacional.
- ✓ Segundo eje factorial: separa las modalidades Menor (tanto en Balotaje como en ambas Elecciones Nacionales) de las modalidades correspondientes a los distintos partidos. Las categorías correspondientes a Otros son baricéntricas.
- ✓ Tercer eje factorial: no tiene una interpretación clara.
- ✓ La representación de la nube de individuos en los ejes indica que existe una buena representación, ya que prácticamente no quedan puntos de esa nube en el primer cuadrante, lo que indica que la posición del individuo en el primer plano factorial puede traducir su trayectoria de voto en las últimas elecciones nacionales

[11] Ver definiciones de variables en Anexo 1 y Estadística Descriptiva en Anexo 2
[12] Ver salidas en Anexo 3
[13] La modalidad Menor representa a aquellas personas que no votaron en la elección correspondiente al año mencionado por ser menores de 18 años. Son “nuevos votantes”

Figura 2 – ACM Variables de Trayectoria Electoral



Fuente: Elaboración propia – Base agosto 04 – Equipos Mori

III.2.3.2 Conclusiones

Se entiende que el primer eje factorial efectúa una buena clasificación de los individuos según su trayectoria electoral, por lo cual será usado como insumo en los análisis posteriores. Este eje posee una inercia igual al 32 % del total, según el índice ponderado de Benzecri.[14]

III.2.4 VARIABLES DE DEFINICIONES POLÍTICAS BÁSICAS

En este bloque se incluyen las siguientes variables: Interés por la política; Simpatía hacia su partido político; Nivel de cercanía al Partido Colorado; Nivel de cercanía al Partido Nacional; Nivel de cercanía al Encuentro Progresista; Auto identificación ideológica e Intención de voto en la elección nacional. [15]

[14] Una alternativa al uso del primer eje factorial es una variable creada por los técnicos de la empresa que contiene la trayectoria electoral de las dos últimas elecciones nacionales (no incluye Balotaje ni Referéndum de Ancap) en la cual las categoría son (primera modalidad nombrada corresponde a 1994 y segunda a 1999): PC-PC; PN-PN; EP-EP; Tradicional – Tradicional (cambios dentro de los partidos Nacional y Colorado en ambos sentidos); Tradicional – EP (Partido Nacional o Colorado en 1994 y EP en 19999) y Otros (en esta categoría se incluyen todas las combinaciones posibles restantes, que no se justifican estar por separado debido a que las frecuencias son muy chicas).

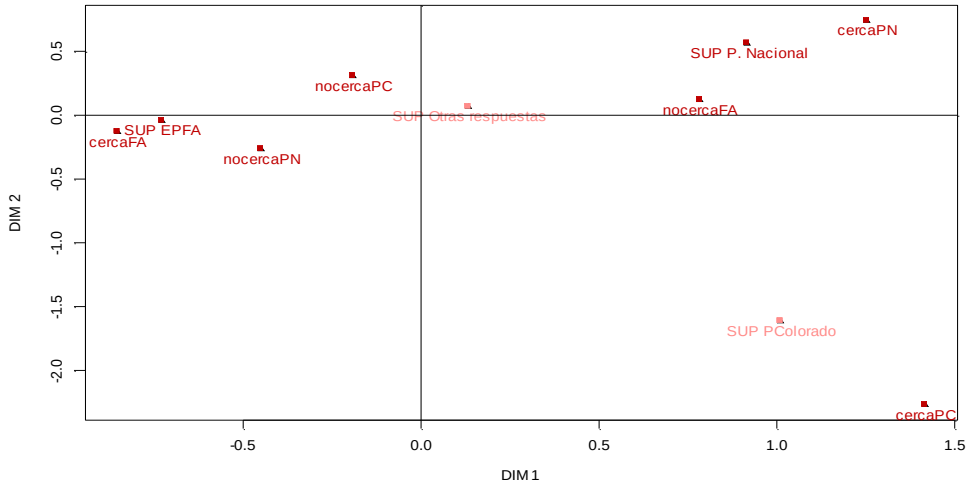
[15] Ver definición de las variables en Anexo 1 y resumen de las mismas en Anexo 2.

Se realiza un análisis factorial con todas las variables, excepto Simpatía Política[16], buscando como en los casos anteriores, simplificar el análisis posterior al disminuir el número de variables.

La inercia explicada por los primeros ejes es del 68 % del total (usando el índice de Benzecri para la ponderación). La interpretación de los ejes no es muy clara. Debido a este inconveniente y dada la importancia que los técnicos asignan a estas variables a la hora de decidir el voto, se decide efectuar otro análisis exploratorio (un Árbol de Clasificación) tratando de hallar cuáles de ellas tienen mayor influencia sobre la variable de respuesta, con el fin de usarlas por separado en el modelo y ver cuáles de ellas son significativas.[17]

Si el ACM se realiza usando únicamente las variables de cercanía a los tres partidos políticos y se proyecta la variable SIMPAT como suplementaria, se observa la importante relación que existe entre la cercanía a determinado partido y la intención de voto. El gráfico de este análisis es el siguiente:

Figura 3 – ACM Variables Definiciones Políticas Básicas



Fuente: creación propia – Base: Encuesta Agosto - Equipos Mori

[16] Ver salidas en Anexo 3.
[17] Ver gráficas y resumen en Anexo 3.

III.2.5 BLOQUE DE VARIABLES SOBRE POPULARIDAD DE LÍDERES

En la encuesta se les pregunta a los individuos su opinión sobre distintos líderes políticos con puntajes que van de 0 a 10. En esta oportunidad se seleccionan aquellos líderes cuyo nivel de conocimiento entre los individuos de la muestra supere el 90%. Para evitar los problemas ocasionados por los datos faltantes, se decide otorgar a las no respuestas un valor central (en este caso, igual a 5). De esta manera quedan como variables los niveles de popularidad de siguientes líderes de los partidos principales:

- Partido Colorado: Jorge Batlle, Julio María Sanguinetti, Guillermo Stirling, Alejandro Atchugarry
- Partido Nacional: Luis Alberto Lacalle, Jorge Larrañaga
- Encuentro Progresista: Tabaré Vázquez, Rodolfo Nin Novoa, José Mujica, Mariano Arana, Danilo Astori, Rafael Michelini[18]

Las correlaciones más altas se dan entre el candidato presidencial de la coalición de izquierda y los demás líderes de este partido: Vázquez-Mujica (0.72); Vázquez-Arana (0.69); Vázquez-Nin (0.62) y Vázquez-Michelini (0.57). También son altas las correlaciones entre Mujica-Nin y Mujica-Arana (0.64 en ambos casos). Dentro del Partido Colorado, las correlaciones más altas (aunque menores a las anteriores) se dan entre Sanguinetti-Stirling (0.62) y entre Batlle-Sanguinetti y Batlle-Stirling (0.55 en los dos casos). Entre los líderes nacionalistas considerados la correlación es de 0.53.
Estos datos harían aparecer a los líderes del Encuentro Progresista como un conjunto más compacto.

III.2.5.1 Análisis Factorial

Se realiza un ACP con todas las variables del bloque.[19] En este análisis los tres primeros ejes factoriales poseen un 69% de la inercia acumulada.

La interpretación de los ejes es la siguiente:

- ✓ <u>Primer eje factorial</u> - Separa los líderes de los partidos tradicionales (con coordenadas positivas) de los líderes del Frente Amplio. Ordena los individuos políticamente de Izquierda a Derecha. Las correlaciones positivas más altas se dan con Lacalle (0.528), Larrañaga (0.505), Sanguinetti (0.503)

[18] Ver Estadística Descriptiva en Anexo 2
[19] Ver salidas en Anexo 3

y Batlle (0.473). Las correlaciones negativas con mayores valores absolutos son con Vázquez (-0.839), Mujica (-0.795), Arana (-0.761) y Nin (-0.7).

- ✓ Segundo eje factorial – Todas las variables tienen coordenadas positivas, aunque son más altas las de los líderes del Partido Colorado: Atchugarri (0.67), Stirling (0.655), Sanguinetti (0.59) y Batlle (0.58). Las coordenadas de los líderes del Partido Nacional son intermedias y las menores son las correlaciones de los líderes del Encuentro Progresista. Puede considerarse un eje del Partido Colorado.
- ✓ Tercer eje factorial – Separa los líderes del partido Colorado (coordenadas positivas) de los del Partido Nacional (negativas). En este eje, los líderes del Frente Amplio quedan baricéntricos. Se puede considerar como el eje de los Partidos Tradicionales. Las coordenadas positivas más altas son: Batlle (0.294), Sanguinetti (0.286) y Stirling (0.203). Los valores negativos más altos se dan en Larrañaga (-0.621) y Lacalle (-0.404).
- ✓ En el primer plano factorial se puede observar una clara división en dos bloques: por un lado los líderes del EPFA y por otro los dirigentes de los Partidos Fundacionales.

- ➢ Estudio de outliers

Para el estudio de los outliers se considera la norma de cada uno de los individuos (distancia al centro de la nube) así como la calidad de representación del mismo en los ejes (medida en el coseno cuadrado respecto a cada uno de los tres ejes factoriales que se consideran para el análisis). Se tienen en cuenta dos posibilidades:

- ✓ Si se considera como criterio el tener una norma mayor a la media más tres veces el desvío, se encuentran 9 individuos pero todos ellos quedan bien representados en el conjunto de los ejes considerados
- ✓ Si el criterio es estar en el tercer cuartil de la distribución de la norma, se encuentran 249 casos. De ellos se seleccionan los que no quedan bien representados en ninguno de los tres ejes (el criterio elegido es pertenecer al primer cuartil de la distribución del coseno cuadrado de cada eje) Existen 12 individuos en estas condiciones. Como regla general, son personas que dan puntajes altos a líderes de distintos partidos (que en primera instancia aparecen como opuestos) o que dan puntajes máximos y mínimos a líderes que tienen altos coeficientes de correlación.

Se decide no excluir ninguno de estas observaciones de la muestra, ya que los técnicos entienden que son respuestas válidas y representan una forma de pensar válida de una parte del electorado (en todos los casos declaran tener poco o nada de interés en la política, lo que explicaría las contradicciones aparentes)

III.2.5.2 Conclusiones

En los modelos se incorporan, en principio, los tres primeros ejes factoriales, ya que tienen una proporción importante de la variabilidad de los datos. Estas nuevas variables son: Lider1 (primer eje factorial, toma valores positivos para los líderes de los partidos Colorado y Nacional y negativos para los del Frente Amplio); Lider2 (segundo eje factorial, eje del Partido Colorado) y Lider3 (tercer eje factorial, toma valores positivos para los líderes del Partido Colorado y negativos para los del Partido Nacional)

III.3 ÁRBOLES DE CLASIFICACIÓN

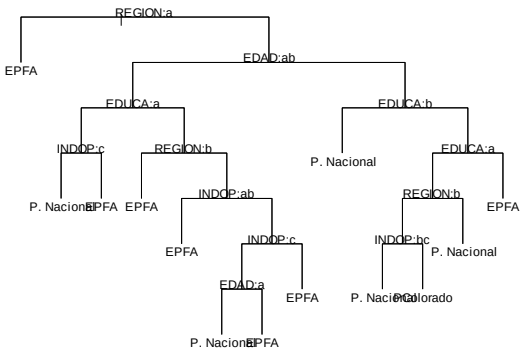
Como última etapa previa a la utilización del modelo, se procede a analizar los bloques de variables que no pudieron ser reducidas en los respectivos Análisis Factoriales. Se busca a través de la técnica de Árboles de Clasificación ver si existe alguna variable que discrimine más que otras para así reducir las dimensiones del bloque.
Se utiliza la técnica CART (Classification and Regression Tree). La base a usar es la de "Decididos".

III.3.1 Variables Básicas

Un primer análisis se efectúa usando SIMPAT como variable a explicar y el grupo de "Variables Básicas" como explicativas: edad, educación, índice socio económico y región política. Se usa como criterio de corte el costo optimal recursivo. La gráfica de los errores de clasificación en función de la complejidad del Árbol detiene prácticamente la caída alrededor de un tamaño de 10 nodos terminales (u hojas). Se obtiene finalmente un Árbol con 13 nodos terminales, partiendo de uno completo de 66 nodos hojas. El error de clasificación no varía sustancialmente: 39.37% en el árbol completo y 39.727% en el árbol podado. La deviance media de los residuos es de 1.94.[20] Se desprende de este análisis que las variables socio demográficas no tienen por sí misma un valor explicativo fuerte al

momento de decidir el voto en una Elección Nacional, ya que los errores de clasificación son muy altos.

Figura 4 – Árbol de Clasificación Variables Básicas



Fuente: elaboración propia – Base: Encuesta Agosto - Equipos Mori

Interpretación: la variable que discrimina en primer lugar es Región, separando en la rama izquierda Montevideo y en la rama derecha el resto del país. La segunda variable en importancia es Edad y luego Educación. La variable Índice socio económico aparece como la menos significativa del grupo, por lo que no se usa en los análisis posteriores. En la primera clasificación y en la rama de la izquierda (Montevideo) aparece un único nodo terminal que tiene como etiqueta "EPFA". Este resultado se corresponde con los datos originales que indican que el 67% de los montevideanos que tienen decidido su voto, lo harían por el EPFA. Otro dato interesante es que en este Árbol no aparece ningún nodo hoja etiquetados como "Otras respuestas" y un único nodo terminal etiquetado como"P. Colorado"

III.3.2 Variables políticas

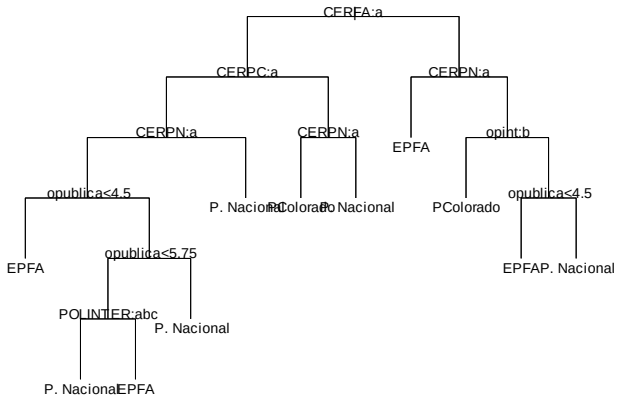
Un segundo análisis se realiza usando como variables explicativas las correspondientes al bloque "Definiciones políticas básicas": Cercanía al P. Colorado (CERCAPC), Cercanía al P. Nacional (CERCPN), Cercanía al Encuentro Progresista (CERCAFA) (estas tres variables se toman como indicatrices: no cerca o cerca al Partido), Identificación Ideológica (OPUBLICA) y Opinión sobre Gestión del Intendente (OPINT). Esta última variable se crea en función de otras dos: la región política y la opinión respecto

[20] Ver salida en Anexo 3

a la gestión del intendente de su departamento. Se sospecha que puede tener gran influencia en la decisión del voto, sobre todo en los ciudadanos del interior del país. Las modalidades correspondientes son: ApMdeo (vive en Montevideo y aprueba la gestión de la intendencia frenteamplista); ApBlanca (vive en uno de los trece departamento cuya intendencia está a cargo del Partido Nacional y aprueba su gestión); ApColorada (vive en uno de los cinco departamentos con Intendente perteneciente al partido Colorado y aprueba su gestión) y Otrosinten (neutro o no aprueba en todos los casos).[21] Las modalidades de las variables de cercanía a los distintos partidos se unen quedando únicamente dos: cerca del partido correspondiente o no cerca.

Se poda el árbol obtenido usando el mismo criterio que en el caso anterior: la mayor caída en la gráfica de los errores de clasificación en función de la complejidad del árbol se da para un tamaño del mismo de alrededor de 10 nodos terminales. El número de nodos terminales es de 11 y la deviance media por nodo es de 0.88. Los errores de clasificación son de 15.7% en el árbol podado y de 14.6 % en el completo.[22] Se puede observar que estas variables tienen un poder explicativo por sí mismas mucho mayor que las variables básicas.

Figura 5 – Árbol de Clasificación Variables Políticas



Fuente: Elaboración propia – Base: Encuesta Agosto - Equipos

Interpretación: la primera variable que discrimina los individuos es CERCAFA: “No cerca” que se dirigen a la izquierda del árbol y “Cerca” hacia la rama derecha. La siguen en importancia las variables CERCAPN y CERCAPC. En este Árbol de Clasificación no

[21] Ver tabla de datos cruzados en Anexo 2
[22] Ver reporte en Anexo 3

aparece como significativa la variable Interés en la política, por lo que no es considerada en los análisis posteriores.

III.3.3 Conclusiones

En los análisis posteriores se utilizan las variables que aparecen con mayor poder para la explicación de la variable de respuesta. Estas son, en el grupo de Variables Básicas: Edad, Educación y Región; en el grupo de Variables de Definiciones Políticas Básicas, las tres variables indicatrices de cercanía a cada uno de los tres grandes partidos (CERCAFA, CERCAPN y CERCAPC), la variable sobre identificación ideológica (OPUBLICA) y la variable de Opinión sobre Gestión del Intendente (OPINT).

III.4 - REGRESIÓN LOGÍSTICA MULTINOMIAL

La Regresión Logística Multinomial se realiza partiendo de la base “Decididos” y luego se pronostica en la base “Indecisos”. La variable de respuesta es Simpatía Política u opción de voto y las variables explicativas son las descriptas con anterioridad y que surgen de los análisis previos. La elección del modelo final depende de indicadores tales como la significación general del modelo, significación de las variables y del poder predictivo del mismo.

Previo a la consideración del modelo final estimado, se detallan a continuación procedimientos que llevaron a la elección del mismo. Se corrieron distintas alternativas en las que la variable de respuesta presenta cuatro modalidades y otros modelos con dicha variable codificada en tres categorías. A su vez, la estrategia inicial consistió en incluir el conjunto de variables explicativas que surgieron de los análisis previos y que eran consideradas de importancia para explicar el fenómeno en estudio. El modelo con el que finalmente se trabajó considera tres modalidades de la variable de respuesta y un conjunto reducido de las variables originales como explicativas.

En todos los modelos que se han analizado se observa que las variables Género e Índice socio económico no son significativas. Esto implicaría que ni el sexo del ciudadano ni su situación económica tienen influencia a la hora de decidir su voto.

Tampoco es significativa la variable creada a partir del eje factorial derivado del Análisis de Correspondencia Múltiple del bloque Variables de Coyuntura. Una explicación posible es que las modalidades “positivas” de todas las variables del bloque tienen frecuencias relativas muy bajas.

En los primeros modelos observados se encuentra que la variable Edad presenta problemas de significación en algunas de sus modalidades en todos o alguno de los logits, por lo que se resuelve volver a codificarla, quedando con tres categorías: de 18 a 29 años (jóvenes), de 30 a 49 años (edades medias) y de 50 años en adelante (mayores).

Con respecto a la utilización de los ejes factoriales se decidió usar en el bloque Líderes Políticos el primer eje (separa Encuentro Progresista de los demás partidos) y el tercero (divide los líderes Nacionalistas de los Colorados y deja baricéntricos los frenteamplistas). El segundo eje no tiene una clara explicación y además no es significativo en ninguno de los modelos analizados. Entre los dos ejes considerados acumulan un 47% de la inercia. Los nombres otorgados a estas nuevas variables son Lider1 y Lider3 respectivamente.

En cuanto a la alternativa de usar el primer eje factorial del bloque Trayectoria Electoral o la variable con 6 modalidades creada a partir de los datos en las dos últimas elecciones, se optó por la primera, ya que es un buen resumen de los datos e incluye información sobre otras dos instancias: el balotaje de 1999 y el plebiscito por Ancap. Además, al ser una variable continua no crea variables indicatrices adicionales, lo que redunda en un modelo más parsimonioso. A esta nueva variable se le denomina Hist1.

La variable de respuesta, Simpatía Política, es una variable con distribución multinomial, con 4 categorías: EPFA, P. Nacional, P. Colorado y Otras respuestas (incluye partidos menores, en blanco y anulado). En las Regresiones se toma la modalidad “EPFA” como referencia debido a que es la que tiene mayor frecuencia relativa y, en cierta manera, la que interesa particularmente en esta oportunidad por los motivos ya explicitados (alcance o no de la mayoría absoluta de los electores).

En los modelos analizados se encuentra que los errores de clasificación son muy grandes para las modalidades P. Colorado y Otras Respuestas. Esto puede deberse, en principio, a que estas categorías tienen frecuencias muy bajas comparadas con las otras (se recuerda que en el caso de la base Decididos, las frecuencias relativas son: EPFA: 54.55%; P. Nacional: 32%; P. Colorado: 9.81% y Otras Respuestas: 3.64%). Esta división se basó en criterios políticos e históricos, dada la importancia del Partido Colorado en la vida institucional. Se presenta un modelo con estas características en el Anexo 4.

Se recuerda que el objetivo principal del trabajo es confeccionar un modelo que permita predecir el resultado de las Elecciones Nacionales del 31 de octubre de 2004, proyectando qué votarán las personas que estaban indecisas al momento de realizarse la

encuesta. Por las particularidades de esta elección, lo fundamental es predecir si el EPFA obtendría o no la mayoría absoluta de votos. En los análisis realizados se encuentra que las modalidades menores (P. Colorado y Otras Respuestas) tienen errores grandes de clasificación. No es posible eliminar esta última categoría, ya que la medición del 50% del electorado más 1 que necesita obtener uno de los partidos para que no haya Balotaje se realiza sobre el total del electorado. Para evitar este problema y tratando de reducir los errores de clasificación, se recodifica la variable de respuesta, colapsando las categorías P. Colorado y Otras Respuestas. Se tienen entonces tres categorías: EPFA, P. Nacional y PC y Otros.

Tomando en consideración lo expuesto anteriormente el modelo final se ajusta tomando como variable de respuesta Simpatía Política u Opción de Voto (SIMPAT), dividida ahora en tres categorías y como variables explicativas las siguientes:

- Opinión sobre gestión del intendente (OPINT): con cuatro modalidades: Aprueba Montevideo, Aprueba Intendencia Blanca, Aprueba Intendencia Colorada y Otras respuestas.
- Cercanía al Encuentro Progresista (CERCAFA): variable dicotómica con valores Cerca y No cerca.
- Cercanía al Partido Nacional (CERCAPN): con dos modalidades Cerca y No cerca
- Cercanía al Partido Colorado (CERCAPC) con las mismas características que las anteriores
- Variables resultantes del Análisis Factorial aplicado al bloque de Popularidad de líderes: primer eje factorial (LIDER1) y tercer eje (LIDER3)
- Primer eje factorial del bloque Trayectoria electoral (HIST1)
- Identificación ideológica (OPUBLICA): variable cuantitativa con valores de 1 a 10).

Una última consideración respecto a las variables: dada la falta de significación en los modelos de Regresión Logística vistos en las primeras etapas del estudio de la variable Región geográfica, se estimó conveniente la inclusión de una nueva variable llamada OPINT. Esta variable es creada a partir de dos variables: la región política (teniendo el cuenta el partido que tiene a cargo la intendencia del departamento) y la aprobación o no de la gestión del intendente de ese departamento.[23] Por la información provista por los técnicos, se sospecha que puede tener valor explicativo para el caso en estudio, sobre todo

[23] Ver estadística descriptiva y tablas cruzadas con Región Política en Anexo 2

en las personas del interior del país. Esta sospecha es confirmada por el árbol de clasificación realizado, en la cual la variable aparece como significativa.

En un primer análisis, el modelo es significativo en su conjunto, sin embargo no se puede decir lo mismo en cuanto a la significación de las variables. Si se analiza la significación global de las variables (no en cada uno de los dos logits) Edad y Educación aparecen como no significativas, por lo que se decide no incluirlas en el modelo final. El cuadro de significación del modelo general así como de las variables se presenta en el anexo 5. Si se considera la variable Edad, por ejemplo, la prueba de significación es la siguiente:

H_0) $\beta_{edad} = 0$

H_a) $\beta_{edad} \neq 0$

El estadístico a usar es la razón de verosimilitudes, que por ser un modelo anidado coincide con la Deviance.

$$Deviance = -2(\ln(L_M) - \ln(L_A)) = 414.14 - 407.47 = 6.67$$

Donde LM es la verosimilitud del modelo con Edad y Educación y LA la verosimilitud del modelo sin Edad. Este estadístico se distribuye $\chi 2$ con 4 grados de libertad. El p-valor es de 0.84, por lo que no se rechaza la hipótesis nula.

III.4.1 MODELO FINAL

Los **resultados** de la regresión logística son los siguientes:

- La modalidad de referencia es *EPFA*, quedando los logits formados de la siguiente manera:

  $$Ln\left(\frac{\pi_j}{\pi_{EP}}\right) = X\beta$$

  donde j = "P. Nacional", "PC y Otros"
  X es la matriz de diseño, cuyas dimensiones son 856 x11
  β es el vector de coeficientes de la regresión, de dimensiones 11 x 1
  π_j = Prob (SIMPAT = j / X)

En cuanto a las variables cualitativas, se introducen en el modelo tantas variables indicatrices como número de modalidades tenga la variable menos uno. En este caso, se obtiene:

Tabla 2 – Variables indicatrices usadas por el modelo

Variable	Indicatriz	Interpretación (igual a 1 si)	Modalidad de referencia
OPINT	OPINT1	Ap. Montevideo	Otras respuestas (no aprueba o neutro, cualquier intendencia)
	OPINT2	Ap. Int. Blanca	
	OPINT3	Ap. Int. Colorada	
CERCAFA	CERCAFA	Cercano al EPFA	No cerca al EPFA
CERCAPN	CERCAPN	Cercano al P.Nac.	No cerca al P.Nac.
CERCAPC	CERCAPC	Cercano al P.Col.	No cerca al P.Col.

III.4.1.1 Significación del modelo y bondad de ajuste

Acorde a los resultados del estadístico de Razón de Verosimilitudes, la prueba de hipótesis de la significación del modelo es rechazada (no todos los parámetros son iguales a cero).

Modelo sólo con intersección: -2 ln(verosimilitud) = 1651.887
Modelo final: -2 ln(verosimilitud) =419.94
χ^2_{28} = 1244.409 cuyo p-valor es 0.000

La bondad de ajuste del modelo se estudia mediante la Deviance:
Deviance: 419.9453

El p-valor es 1, por lo que la hipótesis de que el modelo se ajusta a los datos no puede ser rechazada.

Se pueden considerar, al igual que en los Modelos Lineales Generalizados, una medida de bondad de ajuste semejante al R^2. Se hallan varios indicadores denominados pseudo R^2 para este tipo de modelos, entre ellos el de Mac Fadden, que se detalla a continuación. Uno de los inconvenientes de este tipo de estadísticos es que los valores hallados empíricamente no son muy altos.

$$pseudoR^2 = 1 - \frac{L_M}{L_0}$$

En el caso planteado, el valor del estadístico es 0,746, lo que podría estar indicando un buen ajuste del modelo planteado.

III.4.1.2 Significación de los parámetros

Para analizar la significación se usa el estadístico de Wald, cuya fórmula es

$$z = \frac{\hat{\beta}_{jk}}{s(\hat{\beta}_{jk})}$$ y que se distribuye asintóticamente normal.

Es importante remarcar que si bien las variables son significativas para explicar la intención de voto, su comportamiento es diferenciado en los distintos logits. No es significativa la variable Cerca P. Colorado en el primero de los logits planteados (P. Nacional versus E. Progresista), pero sí lo es en el segundo. Lo contrario sucede con la variable Cerca P. Nacional: es significativa en el segundo logit (P: Colorado y otros sobre E. Progresista) y no en el primero. La variable Opint es significativa en general, pero la modalidad "Opint2" no lo es en ninguno de los logits. Una opción es recodificar la variable, pero se desecha esta posibilidad dada la importancia de distinguir entre las diferentes regiones político – administrativas.

En el modelo planteado, las ecuaciones estimadas resultantes son las siguientes:[24]

$$Ln\left(\frac{\hat{\pi}_{PCyO}}{\hat{\pi}_{EP}}\right) = -4.05 + 0.65 * CercaPC + 0.59 * CercaPN - 0.81 * CercaFa + 0.31 * Opublica$$
$$- 0.51 * Op\,\text{int}\,1 - 0.19 * Op\,\text{int}\,2 - 0.22 * Op\,\text{int}\,3 - 2.03 * Hist1 + 1.42 * Lider1 + 0.61 * Lider3$$

$$Ln\left(\frac{\hat{\pi}_{PN}}{\hat{\pi}_{EP}}\right) = -3.98 - 0.08 * CercaPC + 1.40 * CercaPN - 1.30 * CercaFA - 0.08 * Op\,\text{int}\,1$$
$$- 0.23 * Op\,\text{int}\,2 - 0.51 * Op\,\text{int}\,3 + 0.44 * Opublica - 1.69 * Hist1 + 1.40 * Lider1 - 1.13 * Lider3$$

III.4.1.3 Interpretación de los parámetros

Recordando la expresión $\frac{\hat{\pi}_j}{\hat{\pi}_{EP}} = \exp(X\hat{\beta}_j)$ la interpretación de los parámetros puede realizarse en base a los odds o a los cocientes de odds.

Valores negativos en el exponente indican términos menores que 1 (lo que hace disminuir el cociente considerado) y valores positivos indican términos mayores que 1 (aumento del cociente).

Si la variable es continua, el parámetro indica el impacto en el cociente de los odds cuando la variable considerada aumenta en 1 unidad, dejando constantes las demás.

$$\left(\frac{P(Y = j / X_k = a+1) / P(Y = EP / X_k = a+1)}{P(Y = j / X_k = a) / P(Y = EP / X_k = a)}\right) = \exp(\beta_{jk})$$

[24] Ver salida en Anexo 5

Si la variable es dicotómica, indica la variación que se produce en el cociente si está o no presente esta característica, eliminando el efecto lineal de las restantes.

$$\left(\frac{P(Y = j / X_k = 1) / P(Y = EP / X_k = 1)}{P(Y = j / X_k = 0) / P(Y = EP / X_k = 0)} \right) = \exp(\beta_{jk})$$

El intervalo de confianza para $\exp(\beta)$ en cada logit se halla a partir del intervalo de confianza del parámetro respectivo:

$[\exp(L(\beta)), \exp(U(\beta))]$ siendo $[L(\beta), U(\beta)]$ el intervalo de confianza de cada parámetro.

Si el intervalo de confianza contiene el valor 1 significa que las probabilidades de pertenencia a cualquiera de las dos categorías son iguales, por lo que la variable por sí misma no proporciona información suficiente para distinguirlas.

Por lo expuesto anteriormente, valores de $e^{\beta} > 1$ refieren a factores de "riesgo relativo" para la probabilidad que se está modelizando. Estos se asocian a coeficientes de $\beta>0$, por lo que producen un incremento de la probabilidad. Cuanto mayor sea el valor de e^{β}, mayor será el impacto en el cociente de los odds.

Volviendo al caso en estudio, las variables que tienen mayor impacto para la explicación de la variable de respuesta son el eje factorial derivado del bloque de Trayectoria Electoral, los ejes derivados del bloque de Líderes y las variables de cercanía a los tres Partidos Políticos principales.

<u>Logit P. Nacional versus Encuentro Progresista</u>

- ✓ El mayor impacto está dado por la variable HIST1. Esta variable es un eje factorial extraído de un ACM usando las variables del bloque Trayectoria Electoral (elecciones y plebiscitos desde el año 1994). La variable toma valores positivos si esta votación refiere al EPFA y negativos si fue a los Partidos Tradicionales. El coeficiente negativo está indicando que si la variable toma valores positivos, el exponente es menor que cero por lo que el logit es menor que 1 y aumenta la probabilidad de voto al EPFA comparada con la probabilidad de voto al P.Nacional. Si la variable toma valores negativos entonces el exponente es mayor que cero, lo que hace aumentar la probabilidad de pertenencia a la categoría P. Nacional con relación a la categoría de referencia. El modelo está indicando que la trayectoria

electoral es un dato importante a tener en cuenta a la hora de decidir el voto. Por cada unidad que aumenta la variable, quedando las demás constantes, la probabilidad de pertenencia a la categoría EPFA es 5.4 veces mayor que la probabilidad de pertenencia a la categoría P. Nacional (ya que $\exp(\beta_{PN,hist1}) = 5.4$).

- ✓ La variable LIDER1 (eje factorial que toma valores positivos para puntajes altos a líderes de los Partidos Tradicionales y negativos para los líderes del EPFA) tiene el mayor coeficiente positivo. Por cada unidad que aumenta la variable, dejando las otras variables constantes, aumenta 4.1 veces la probabilidad de votar al P. Nacional con respecto a la probabilidad de hacerlo al EPFA.
- ✓ Otra variable importante es CercaPN. Si el individuo posee esta modalidad, el coeficiente positivo indica que la probabilidad de votar al P. Nacional en lugar de al EPFA es 4.1 veces mayor, frente a los que no tienen esta característica y dejando el resto constante.
- ✓ Algo similar ocurre con la variable CercaFA: el coeficiente negativo indica que si la persona tiene esta característica la probabilidad de voto al EPFA es 3.7 veces mayor que la probabilidad de voto al P. Nacional (quedando lo demás constante).
- ✓ En cuanto a la variable Opinión Pública, un incremento en su valor indica un corrimiento hacia la derecha del espectro político nacional. El signo positivo de su coeficiente indica que al crecer en magnitud su valor, aumenta la probabilidad de votar al P. Nacional en contra de la probabilidad de votar al EPFA.
- ✓ El signo negativo del término independiente indicaría que, sin tener en cuenta ninguna de las variables, es más probable que la persona vote al EPFA que al P. Nacional. Esto se condice con los datos obtenidos de la muestra, ya que esta modalidad es la que posee mayor frecuencia.

<u>Logit “PC y Otros” versus “Encuentro Progresista”.</u>

- ✓ Los parámetros son similares a los del logit ya explicado. También el mayor valor absoluto lo tiene la variable HIST1 y con coeficiente negativo, por lo que la interpretación es la misma a la ya realizada. Por una unidad que aumenta la variable, la probabilidad de pertenencia a la categoría EPFA es 7.6 veces mayor que la probabilidad de pertenencia a la categoría “P. Colorado y Otros”.
- ✓ La variable CercaPC es significativa en este logit y tiene coeficiente positivo. El tener esta característica aumenta la probabilidad de pertenencia a la categoría “P.

Colorado y Otros" casi 2 veces frente a la probabilidad de pertenencia a la categoría EPFA (manteniendo las demás características constantes).

- ✓ La variable Lider1 tiene coeficiente positivo de forma tal que si aumenta en una unidad, la probabilidad de voto al "P. Colorado y Otros" es 4 veces mayor que la probabilidad de voto al EPFA.
- ✓ La variable Lider3 es un eje factorial derivado del bloque Popularidad de Líderes. Este eje separa los líderes del P. Colorado (coordenadas positivas) de los ejes del P. Nacional. En este logit su coeficiente es positivo: si aumenta una unidad el valor de la variable, la probabilidad de pertenencia a la categoría "P.C y Otros" es casi el doble (1.85) que la probabilidad de pertenencia a la categoría EPFA. Si se halla la razón de probabilidades entre "P.C. y Otros" y "P. Nacional" se observa que por cada unidad que aumenta la variable, la probabilidad de pertenencia a la primera modalidad mencionada es 5.7 veces mayor que a la segunda.
- ✓ También es significativa la modalidad ApMontevideo de la variable OPINT y su coeficiente es negativo. Esto significa que si el encuestado vive en Montevideo y aprueba la gestión del Intendente, es mayor la probabilidad de que vote al EPFA que al P. Colorado u otras opciones, comparado con los individuos de la categoría Otrosint que es la de referencia. También en este caso se supone que las otras características se mantienen constantes.

III.4.1.4 Errores de clasificación producidos por el modelo

Los errores y aciertos de clasificación se observan al cruzar los datos observados en la muestra con los obtenidos al aplicar el modelo. Los mismos se presentan en la siguiente tabla

Tabla 2 – Datos predichos vs. observados

		Predice				
Observados	Categoría	EPFA	P.N.	PCyOtros	Total	% Aciertos
	EPFA	448	12	7	467	95.93
	P.Nacional	12	242	20	274	88.32
	PC y Otros	8	28	79	115	68.70
	Total	468	282	106	856	89,84

Fuente: Elaboración propia – Base: Encuesta Agosto - Equipos Mori

El error de clasificación total es de 10.16%, si bien el mismo no es homogéneo en las distintas modalidades. El mayor error se produce en la categoría "PC y Otros", que es la

que tiene menor frecuencia relativa (31,3%) y el menor se da en la modalidad EPFA (4.07%).

Teniendo en cuenta el poder predictivo y el modelo obtenido, las estimaciones de las distintas modalidades son:

Modalidad	Estimación puntual (como % total Decididos)
EPFA	54.67
P.Nacional	32.94
P.C y Otros	12.38

El intervalo de confianza al 95% para la modalidad EPFA es (52.4 , 60.1). La construcción de los intervalos de confianza se puede encontrar en Anexo 6.

III.4.2 PROYECCIÓN DE INDECISOS

La proyección de los indecisos se basa en aplicar el modelo encontrado para los "Decididos" a la base de "Indecisos".

Los valores que se obtienen una vez realizada la predicción son los siguientes

Tabla 3 – Proyección de indecisos

Categoría	Proyecta indecisos	% indecisos
EPFA	37	27
P.N.	68	50
Otros	32	23
Total	137	100

Fuente: Elaboración propia – Base: Encuesta Agosto - Equipos Mori

Los individuos son asignados a las distintas categorías según la probabilidad estimada por el modelo. Los indecisos son clasificados en un 50% en la categoría P.Nacional, mientras que el resto se reparte casi en forma equitativa en el resto de las opciones.

Si en lugar de mirar la categoría a la que se asigna el individuo se le observan las probabilidades de pertenencia a cada categoría, pueden encontrarse aquellas asignaciones que son, en cierto sentido, dudosas (o sea, que las diferencias entre las probabilidades de pertenencia a cada modalidad no sean lo suficientemente grandes).

Al estudiar los individuos indecisos según estas probabilidades de pertenencia a las distintas categorías se observa lo siguiente:

- ✓ Categoría EPFA – posee 37 individuos. De ellos 3 individuos están en duda, por tener poca diferencia la probabilidad de pertenencia a esta categoría con la correspondiente a la categoría P. Nacional (2 individuos) o a la categoría P. Colorado y Otros (1 individuo). Todos ellos son de centro, con poco y nada interés en la política, no se sienten cercano a ninguno de los partidos. Son votantes de los partidos tradicionales (Hist1 negativo).
- ✓ Categoría P. Nacional – posee 68 individuos. De ellos 9 tienen probabilidades cercanas a la categoría P. Colorado y Otros y 5 con EPFA. En general son individuos que tienen poco y nada interés en la política, de centro y no se sienten cercano a ningún partido. Son votantes de los partidos tradicionales.
- ✓ Categoría P. Colorado y Otros – posee 32 individuos. De ellos, 3 tienen probabilidad parecida a la de la categoría P. Nacional, en un caso es cercana a la del EPFA y en dos individuos las probabilidades de pertenencia a las tres modalidades son muy cercanas.
- ✓ Se observan 23 individuos (16.7% del total de Indecisos) cuya clasificación por el modelo genera dudas.

III.4.3 RESULTADOS FINALES

Al considerar los resultados obtenidos por el modelo para las bases Decididos e Indecisos, se obtienen los siguientes valores:

Tabla 4 – Proyecciones del modelo

Modalidad	Decididos		Indecisos	
EPFA	468	54,67%	37	27.01%
P. Nacional	282	32.94%	68	49.64%
P.Colorado y Otros	106	12.39%	32	23.35%
Total	856	100%	137	100%

Fuente: Elaboración propia – Base: Encuesta Agosto - Equipos Mori

.

Las proyecciones del modelo para la totalidad de los individuos se observan en la siguiente tabla.

Tabla 5 – Proyecciones totales

Modalidad	Número individuos	Porcentaje
EPFA	505	50.85%
P. Nacional	350	35.25%
P.Colorado y Otros	138	13.90%
Total	993	100%

Fuente: Elaboración propia – Base: Encuesta Agosto - Equipos Mori

Al considerar al conjunto de individuos de la muestra (decididos y no decididos) se produce un ajuste en la estimación de las distintas modalidades. El impacto mayor lo recoge el EPFA en la medida que la incorporación de la proyeccion del comportamiento de los indecisos reduce la intención de voto hacia esa colectividad en aproximadamente el 4%.

La proyeccion final estaría indicando una intención de voto del 50,85 % de los votantes para el Encuentro Progresista FA, un 35,25% para el Partido Nacional y un 13.9% para el Partido Colorado y otras respuestas.

Estos resultados estarían indicando que no habría segunda vuelta, ya que el Presidente de la República sería electo el 31 de octubre de 2004.

IV CONCLUSIONES

El objetivo principal del trabajo es formular un modelo en el cual se pueda predecir el resultado de la Elección Nacional de octubre de 2004 mediante la proyección de voto de las personas que no tenían decidido su voto al momento de realizarse la encuesta.

Se logra un modelo de Regresión Logística Multinomial cuya variable de respuesta es la intención de voto de los encuestados. Las variables explicativas son seleccionadas de forma tal que acumulen la mayor cantidad de información posible de los datos originales (para lo que se realizan en algunos casos Análisis Factoriales para reducir dimensiones), tengan relevancia para explicar el comportamiento de la variable de respuesta y logren los menores errores de clasificación.

Las variables Género, Índice socio económico, Edad y Educación no son significativas. Esto implicaría que ni el sexo del ciudadano ni su situación económica así como su edad no tienen influencia relevante a la hora de decidir su voto.

La variable de mayor impacto es la Trayectoria Electoral, seguida de la opinión que tiene el ciudadano respecto a los líderes políticos y su cercanía a cada uno de los partidos principales.

El modelo finalmente elegido tiene un alto poder predictivo: un 10% de error de clasificación global (teniendo la categoría Encuentro Progresista un error de clasificación de un 4%).

Los valores finalmente obtenidos, teniendo en cuenta a los ciudadanos decididos como a los no decididos, para cada categoría de la variable de respuesta son: Encuentro Progresista – Frente Amplio: 50.76%; Partido Nacional: 34.43% y Partido Colorado y Otras Respuestas 14.81%.

Con los datos obtenidos se puede concluir que el Encuentro Progresista obtendría la mayoría absoluta de ambas Cámaras del Parlamento y la Presidencia de la República en primera vuelta, o sea, sin necesidad de Balotaje.

BIBLIOGRAFÍA

- Agresti, Alan. Categorical Data Analyssis.
- Agresti, Alan. Introduction to Categorical Data Analyssis.
- Blanco, Jorge. Introducción al Análisis Multivariado. Oficina de Apuntes CECEA. Montevideo. 2001
- Breiman, Friedman, Olsen, Stone. Classification and Regression Trees. 1984
- Cortijo Bon, Francisco José. Técnicas Supervisadas II: Aproximación No Paramétrica. Octubre 2001. Capítulo 7: Árboles de clasificación.
- Escoffier Brigitte, Pages Jerome. Análisis Factoriales Simples y Múltiples. Universidad del País Vasco. Bilbao. 1992
- Hosmer David W, Lemeshow Stanley. Applied Logistic Regression. Second Edition. 2000
- Nalbarte Laura, Álvarez Ramón, Castrillejo Andrés. Notas del Curso Análisis Multivariado II. Licenciatura en Estadística de las Facultades de Ciencias y Ciencia Económicas. Montevideo. 2004
- Netter, Kutner, Nachistsheim, Wasserman. Applied Linear Statistical Models.
- Puerta Goicochea, Aita. Imputación basada en Árboles de Clasificación. EUSTAT. Junio 2002

ANEXO 1: DESCRIPCIÓN DE LAS VARIABLES QUE INTERVIENEN EN EL ESTUDIO

A1.1 BLOQUES DE VARIABLES

Bloque	Variables
Variables básicas	Sexo
	Edad
	Zona geográfica
	Educación
	Nivel socio económico
Trayectoria electoral	Voto 1994
	Voto octubre 1999
	Voto balotaje 1999
	Voto plebiscito Ancap 2003
Variables de coyuntura	Situación económica actual del país
	Situación económica personal y familiar actual
	Visión de la situación del país en un año
	Visión de la situación personal en un año
	Gestión del presidente
Definición política	Interés en la política
	Simpatía por su partido político
	Cercanía P Colorado
	Cercanía P Nacional
	Cercanía EPFA
	Identificación ideológica
	Intención de voto actual
Popularidad de líderes	J.Batlle – J.M.Sanguinetti – G.Stirling – A.Atchugarry
	J. Larrañaga – LA Lacalle
	T. Vázquez – R. Nin Novoa – J. Mujica – D. Astori –M. Arana – R. Michelini

A1.2 BLOQUE: VARIABLES BÁSICAS

Las variables de este bloque que intervienen en el análisis son las siguientes:

- Género (VB_SEXO)
- Edad: (EDAD) dividida en principio en 6 tramos: de 18 a 22 años (nuevos votantes); de 23 a 29 años; de 30 a 39 años; de 40 a 49 años; de 50 a 59 años; mayores de 60 años
- Índice socio económico: (INDOP) variable creada a partir de datos extraídos de las respuestas dadas por los individuos encuestados, tales como educación, ocupación, salario, posesión de ciertos bienes, etc. Se divide en 4 categorías: alto y medio alto; medio; medio bajo y bajo.
- Educación: (EDUCAOP) dividida en tres categorías: primaria, secundaria y terciaria

- Región geográfica: (REGION) en principio el país se divide en tres grandes regiones: Montevideo, Canelones y Resto del país. Otra división factible de usar (dadas las características del estudio y su importancia histórica) es la división del país de acuerdo con el partido político que gobierne el departamento (REPOL). Tenemos así las siguientes modalidades: Montevideo (intendencia a cargo del EPFA); Intendencias Coloradas (Canelones, Artigas, Salto, Río Negro, Rivera) ; Intendencias Blancas (Paysandú, Soriano, Colonia, San José, Tacuarembó, Durazno, Flores, Florida, Maldonado, Lavalleja, Rocha, Treinta y Tres, Cerro Largo). Una última alternativa (CIUDAD) es dividir al país en estratos semejantes a los que se usan para extraer la muestra, o sea, de acuerdo al tamaño de la ciudad o pueblo. En este caso, las modalidades serían las siguientes: Montevideo, Ciudades Grandes (más de 20.000 habitantes); Ciudades Medianas (entre 10 y 20 mil habitantes); Ciudades Chicas (entre 500 y 10.000 habitantes) y Zonas Rurales (poblaciones menores a 500 habitantes y zonas rurales propiamente dichas).

A1.3 BLOQUE: VARIABLES DE COYUNTURA

Las variables que se incluyen en este bloque son:

- Situación económica personal y familiar actual. (SECFAMAC) La variable original tiene las siguientes categorías: Muy Buena, Buena, Regular, Mala, Muy Mala, No sabe o no contesta. Debido a que las categorías Muy Buena y No sabe tienen muy pocas observaciones, se determina recodificar la variable, quedando entonces las siguientes modalidades: Buena (une las dos primeras); Regular (incluye No sabe, no contesta) y Mala (une las categorías Mala y Muy Mala).
- Situación económica del país actual (SECPAISA). Las modalidades son las mismas que en el caso anterior.
- Percepción de la situación económica personal y familiar dentro de un año (SECRPRO1) Las modalidades para esta variable son: Mejor, Igual, Peor y No sabe.
- Percepción de la situación económica del país dentro de un año (SECPPRO1). Aquí también las modalidades son Mejor, Igual, Peor y No sabe.
- Opinión sobre la gestión del Presidente de la República (OPEVPRES). Las modalidades en este caso son: Aprueba; Ni uno ni otro; Desaprueba; No sabe, no contesta.

A1.4 BLOQUE: TRAYECTORIA ELECTORAL

Este bloque incluye las siguientes variables:

- Voto en Elección Nacional de 1994 (VOTOP2) cuyas modalidades son: Partido Colorado; Partido Nacional; Encuentro Progresista – Frente Amplio; Otros (incluye no votó, en blanco, anulado y partidos menores) ; Menor.
- Voto en Elección Nacional de 1999 (VOTOE99A) Las modalidades son las mismas que en el caso anterior.
- Voto en Balotaje noviembre 1999 (VOTOBAL) cuyas modalidades son: Tabaré Vázquez; Jorge Batlle; Otros (incluye no respuestas, votos en blanco y anulados, no votó por distintas circunstancias); Menor.
- Voto en Plebiscito por ANCAP en diciembre 2003 (REF1203) Las modalidades en este caso son: Sí; No; Otros.

A1.5 BLOQUE: DEFINICIONES POLÍTICAS BÁSICAS

En este bloque se incluyen las siguientes variables:

- Interés por la política (POLINTER), cuyas modalidades son: Mucho, Bastante, Poco y Nada.
- Simpatía hacia su partido político (POLIDENT), con las modalidades Simpatía Fuerte, Simpatía a Secas e Indefinida.
- Nivel de cercanía al Partido Colorado (CERCAPC), cuyas modalidades originales son: Muy Cerca; Cerca; Ni cerca ni lejos; Lejos; Muy Lejos y No sabe / no contesta. Esta variable se recodifica, quedando las modalidades siguientes: Cercano (une las dos primeras); Neutro (Ni cerca ni lejos más No sabe) y Lejano (une Lejos y Muy Lejos).
- Nivel de cercanía al Partido Nacional (CERCAPN) con las mismas características que la variable anterior.
- Nivel de cercanía al Encuentro Progresista – Frente Amplio (CERCAFA) con iguales modalidades que las anteriores.
- Auto identificación ideológica (OPUBLICA). Identificación del propio entrevistado sobre su posición en el especto político nacional, con puntajes del 1 (izquierda) a 10 (derecha). Se categoriza de la siguiente manera: Izquierda, Centro izquierda, Centro, Centro Derecha, Derecha y No sabe en la variable AUTID.
- Intención de voto en las Elecciones Nacionales del 31 de octubre de 2004 (SIMPAT) Es la variable que se usará como respuesta en la Regresión Logística.

Tiene 5 modalidades: Partido Colorado, Partido Nacional, Encuentro Progresista – Frente Amplio – Nueva Mayoría, Otras respuestas (incluye partidos menores, en blanco y anulado) e Indecisos.

ANEXO 2:ESTADÍSTICA DESCRIPTIVA DE LAS VARIABLES USADAS EN EL ESTUDIO

A2.1 BLOQUE: VARIABLES BÁSICAS

```
       VB.SEXO                   INDOP          EDAD             EDUCAOP
 Masculino :447   Alto y medio alto:236   menor22:102        Primaria: 94
 Femenino  :546              Medio:243   de23a29:138      Secundaria:702
                        Medio bajo:205   de30a39:160       Terciaria:197
                               Bajo:309   de40a49:195
                                          de50a59:159
                                          mayor60:239

        REPOL           REGION          CIUDAD
      Mdeo:425   Montevideo:425   Montevideo:425
  Intblanca:240    Canelones:186      Grandes:253
 Intcolorada:328    Restopais:382     Medianas: 70
                                        Chicas:134
                                       Rurales:111
```

A2.2 BLOQUE: VARIABLES DE COYUNTURA

```
  SECFAMAC         SECFPRO1           SECPAISA         SECPPRO1
   Buenaper:145    Mejorper :298     Buenapais : 40   Mejorpais :301
 Regularper:544     Igualper:409   Regularpais :260   Igualpais :366
    Malaper:304     Peorper :143      Malapais :693    Peorpais :153
                 Nosabepers:143                      Nosabepais:173

               OPEVPRES
           Aprueba : 70
     Ni uno, ni otro :147
          Desaprueba :764
No sabe, no contesta : 12
```

A2.3 BLOQUE: TRAYECTORIA ELECTORAL

```
     VOTOE99A             VOTOBAL            REF1203        VOTOP2
  P.Col99:201         Vázquez  :399         NOAncap:245    PCol94:193
   PNac99:195        Batlle    :368         SIAncap :488    PNac94:197
   EPFA99:384     OtrosBalotaje:134       OtrosAncap:260    EPFA94:287
Menor99  : 88   Menorbalotaje  : 92                       Menor94:161
Otros99  :125                                            Otros94:155
```

A2.4 BLOQUE: DEFINICIONES POLÍTICAS BÁSICAS

```
     POLINTER                 POLIDENT           CERCAPC         CERCAPN
   Mucho : 86   Simpat\355a fuerte :259        CercPCol:107     CercPNac:237
Bastante :194   Simpat\355a a secas:361      NeutroPCol:234   NeutroPNac:241
    Poco :374          Indefinidas :373         LejPCol:652      LejPNac:515
    Nada :339
```

```
     CERCAFA                  AUTID                    SIMPAT              OPUBLICA
   CercFA:423          Izquierda: 67        P. Colorado: 84        Min:  1.000000
 NeutroFA:232   Centro-izquierda:305        P. Nacional:274     1st Qu.:  4.000000
    LejFA:338             Centro:260               EPFA:467     Mean:  5.143001
                Centro-derecha:166   Otras respuestas: 31     Median:  5.000000
                       Derecha: 69         Indecisos:137     3rd Qu.:  6.000000
                    Nosabeiden:126                           Max: 10.000000
                                                          Std Dev.:  2.072816
```

A2.5 BLOQUE: POPULARIDAD DE LÍDERES

```
             BATLLE   LACALLE    MUJICA LARRANAGA  NINNOVOA  STIRLING
     Min:  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
 1st Qu.:  0.000000  0.000000  0.000000  2.000000  1.000000  0.000000
    Mean:  1.922457  2.677744  4.638469  4.551863  4.126888  2.830816
  Median:  0.000000  2.000000  5.000000  5.000000  5.000000  2.000000
 3rd Qu.:  4.000000  5.000000  8.000000  7.000000  6.000000  5.000000
     Max: 10.000000 10.000000 10.000000 10.000000 10.000000 10.000000
Std Dev.:  2.574501  3.037547  3.569756  3.204328  2.953267  2.823000

           MICHELINI SANGUINETTI   VAZQUEZ    ARANA   ASTORI  ATCHUGARRI
     Min:  0.000000   0.000000  0.000000  0.000000  0.000000   0.000000
 1st Qu.:  2.000000   0.000000  1.000000  3.000000  3.000000   0.000000
    Mean:  3.899295   2.271903  5.049345  5.117825  4.928499   3.832830
  Median:  5.000000   0.000000  5.000000  5.000000  5.000000   5.000000
 3rd Qu.:  5.000000   5.000000  8.000000  8.000000  7.000000   6.000000
     Max: 10.000000  10.000000 10.000000 10.000000 10.000000  10.000000
Std Dev.:  2.657523   2.880795  3.699469  3.253288  2.966534   2.975412
```

Matriz de correlaciones

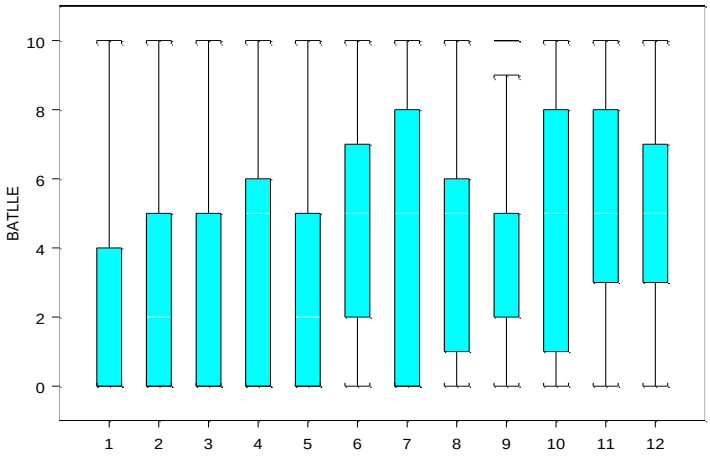
```
             SIMJBATL      SIMLAL      MUJICA    SIMLARRA    SIMNOVOA
SIMJBATL  1.000000000  0.48226046 -0.19039958   0.30848393 -0.09496091
  SIMLAL  0.482260456  1.00000000 -0.24512402   0.52556905 -0.13567888
  MUJICA -0.190399579 -0.24512402  1.00000000  -0.26172902  0.64242562
SIMLARRA  0.308483929  0.52556905 -0.26172902   1.00000000 -0.12021634
SIMNOVOA -0.094960910 -0.13567888  0.64242562  -0.12021634  1.00000000
SIMSTIRL  0.556332059  0.42672135 -0.24125085   0.44126965 -0.06827775
SIMRAFAM -0.037387950 -0.09368718  0.50142811  -0.08426371  0.60659369
  SIMJMS  0.555224130  0.44905120 -0.22343723   0.36397698 -0.11022435
SIMTVAZQ -0.247796132 -0.29667957  0.72307761  -0.32170143  0.62601126
SIMARANA -0.132384400 -0.21455724  0.63610740  -0.20931481  0.59460432
SIMDASTO  0.009436661 -0.05424383  0.50816643  -0.05003526  0.50328722
 AATCHUG  0.405600911  0.32853233 -0.06131190   0.38355244  0.01159398

             SIMSTIRL      SIMRAFAM      SIMJMS     SIMTVAZQ
SIMJBATL   0.556332059  -0.037387950  0.55522413 -0.247796132
  SIMLAL   0.426721348  -0.093687175  0.44905120 -0.296679568
  MUJICA  -0.241250852   0.501428110 -0.22343723  0.723077613
SIMLARRA   0.441269651  -0.084263714  0.36397698 -0.321701432
SIMNOVOA  -0.068277749   0.606593693 -0.11022435  0.626011260
SIMSTIRL   1.000000000  -0.008185572  0.61911699 -0.251225526
SIMRAFAM  -0.008185572   1.000000000 -0.03131327  0.568652081
  SIMJMS   0.619116986  -0.031313274  1.00000000 -0.243595009
SIMTVAZQ  -0.251225526   0.568652081 -0.24359501  1.000000000
SIMARANA  -0.142165255   0.558009598 -0.14723040  0.690687562
SIMDASTO   0.044295698   0.447135238 -0.03924395  0.521043919
 AATCHUG   0.536570477   0.043636462  0.43562810 -0.078924663
```

	SIMARANA	SIMDASTO	AATCHUG
SIMJBATL	-0.13238440	0.009436661	0.40560091
SIMLAL	-0.21455724	-0.054243831	0.32853233
MUJICA	0.63610740	0.508166434	-0.06131190
SIMLARRA	-0.20931481	-0.050035261	0.38355244
SIMNOVOA	0.59460432	0.503287219	0.01159398
SIMSTIRL	-0.14216525	0.044295698	0.53657048
SIMRAFAM	0.55800960	0.447135238	0.04363646
SIMJMS	-0.14723040	-0.039243952	0.43562810
SIMTVAZQ	0.69068756	0.521043919	-0.07892466
SIMARANA	1.00000000	0.633747909	0.03671558
SIMDASTO	0.63374791	1.000000000	0.22123349
AATCHUG	0.03671558	0.221233486	1.00000000

En la siguiente gráfica se puede ver el box plot correspondientes a estas variables, siendo la codificación la siguiente:

1- Batlle 2- Stirling 3- Sanguinetti 4- Atchugarry
5- Lacalle 6- Larrañaga
7- Mujica 8-Nin 9- Michelini 10- Vázquez 11- Arana 12- Astori



A2.6 ESTADÍSTICA DESCRIPTIVA BASE "DECIDIDOS"

```
              SIMPAT                    INDOP         EDAD            EDUCA
           EPFA:467   Alto y medio alto:220  de18a29:202     Primaria:322
     P. Nacional:274              Medio:203  de30a49:315   Secundaria:388
       PColorado: 84         Medio bajo:171  mayor50:339    Terciaria:146
 Otras respuestas: 31               Bajo:262
```

```
        REGION          CERPC           CERPN          CERFA         POLINTER
 Montevideo:374   nocercaPC:754   nocercaPN:630  nocercaFA:446      Mucho : 84
  Canelones:166     cercaPC:102     cercaPN:226    cercaFA:410   Bastante :181
  Restopais:316                                                      Poco :335
                                                                     Nada :256
```

```
              LIDER1       LIDER3       HIST1        opublica
    Min: -14.4900316  -9.65382113 -0.97170436       1.000000
   Mean:  -0.4160581  -0.09639145  0.08084951       5.098131
 Median:  -0.5260593   0.11419995 -0.23667031       5.000000
    Max:  15.4426237  10.74814124  1.25091948      10.000000
Std Dev.:   6.9954321   2.79659596  0.86113015       2.195358
```

Tablas de datos cruzados entre las variables Repol y Opint (base completa)

```
      *** Crosstabulations ***
Call:
crosstabs(formula =  ~ REPOL + OPINT, data = base.agosto, na.action = na.fail,
      drop.unused.levels = T)
993 cases in table
+-----------+
|N          |
|N/RowTotal|
|N/ColTotal|
|N/Total    |
+-----------+
REPOL  |OPINT
       |Otrsntn|ApMdeo |ApBlanc|ApColrd|RowTotl|
-------+-------+-------+-------+-------+-------+
Mdeo   |206    |219    |  0    |  0    |425    |
       |0.48   |0.52   |0.00   |0.00   |0.43   |
       |0.34   |1.00   |0.00   |0.00   |       |
       |0.21   |0.22   |0.00   |0.00   |       |
-------+-------+-------+-------+-------+-------+
Intblnc|132    |  0    |108    |  0    |240    |
       |0.55   |0.00   |0.45   |0.00   |0.24   |
       |0.22   |0.00   |1.00   |0.00   |       |
       |0.13   |0.00   |0.11   |0.00   |       |
-------+-------+-------+-------+-------+-------+
Intclrd|268    |  0    |  0    | 60    |328    |
       |0.82   |0.00   |0.00   |0.18   |0.33   |
       |0.44   |0.00   |0.00   |1.00   |       |
       |0.27   |0.00   |0.00   |0.06   |       |
-------+-------+-------+-------+-------+-------+
ColTotl|606    |219    |108    |60     |993    |
       |0.61   |0.22   |0.11   |0.06   |       |
-------+-------+-------+-------+-------+-------+
Test for independence of all factors
      Chi^2 = 788.5779 d.f.= 6 (p=0)
      Yates' correction not used
```

A2.6 ESTADÍSTICA DESCRIPTIVA BASE “INDECISOS”

```
$$$"Factor Summaries":
                INDOP          EDAD           EDUCA          OPINT
 Alto y medio alto:16   de18a29:38     Primaria:60  otroscasos:93
             Medio:40   de30a49:40   Secundaria:65      apMdeo:26
        Medio bajo:34   mayor50:59    Terciaria:12    apBlanca: 8
              Bajo:47                                apColorada:10

        CIUDAD           REPOL          REGION          CERPC
Montevideo:51          Mdeo:51   Montevideo:51   nocercaPC:132
   Grandes:       Intblanca:45    Canelones:20     cercaPC:  5
Medianas:14     Intcolorada:41    Restopais:66
   Chicas:15
  Rurales:15

        CERPN           CERFA        POLINTER
 nocercaPN:126   nocercaFA:124      Mucho : 2
   cercaPN: 11     cercaFA: 13   Bastante :13
                                     Poco :39
                                     Nada :83

              LIDER1     LIDER3      HIST1   opublica
     Min:  -8.635653 -5.3469854 -0.9717044  3.0000000
    Mean:   2.599604  0.6022707 -0.5051619  5.4233577
  Median:   2.430229  0.9477909 -0.6818515  5.5000000
     Max:  13.224394  5.0922640  1.2509195 10.0000000
Std Dev.:   4.079365  1.9819194  0.4794624  0.9737181
```

ANEXO 3: ANÁLISIS FACTORIALES

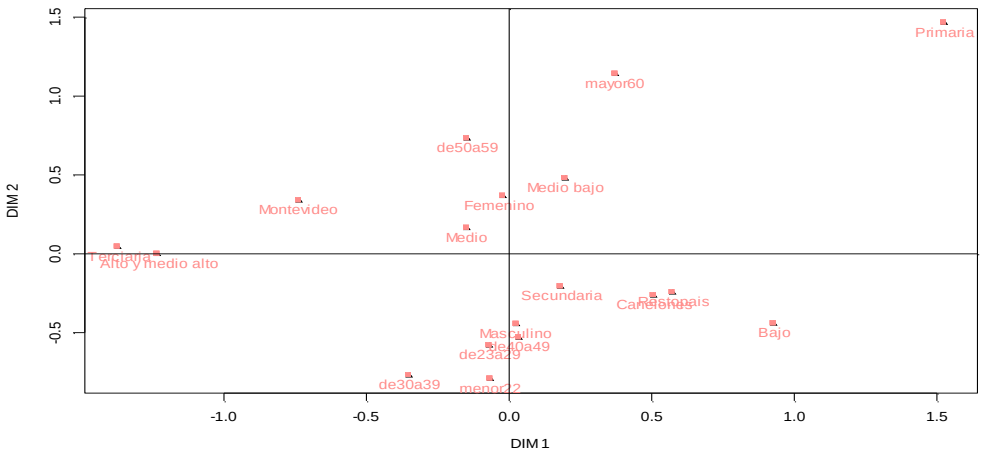
A3.1 BLOQUE: VARIABLES BÁSICAS

El análisis factorial se efectúa con las variables: género, edad, educación, índice socio económico y región del país

Las modalidades que tienen mejor calidad de representación (alrededor de 0.5 en todos los casos) son : niveles socio económico alto y bajo; edad mayor a 60 años; región Montevideo; educación Primaria y Terciaria.

La interpretación de los ejes factoriales puede efectuarse de la siguiente manera:

- ✓ Primer eje factorial - Las variables que identifican el eje son: nivel socio económico: de alto y medio alto (coordenadas negativas) a bajo (coordenadas positivas); región: separa Montevideo (coordenadas negativas) del resto del país; educación: de terciaria (negativa) a primaria (positiva). Las modalidades de la variable género son baricéntricas y las de la variable edad no tienen una interpretación clara. Por lo tanto este eje se puede interpretar como una variable que separa sectores de niveles socio económico alto y medio alto y con educación terciaria de sectores de bajo nivel socio económico y con educación primaria.
- ✓ Segundo eje factorial – Las variables con mayor correlación con el eje son: género (coordenada positiva para femenino); edad (coordenadas positivas para las categorías de edades mayores a 50 años y negativas para las restantes); región (también separa Montevideo del resto del país). Este eje estaría determinado principalmente por la edad.
- ✓ Tercer eje factorial – No tiene una interpretación muy clara



La salida del Análisis es la siguiente:

	Singular Values	Principal Inertias	Chi Squares	Percents	Acumulado
1	0.58729013406	3.449097e-001	1.859600e+003	1.326576e-001	0.1326576
2	0.50129388233	2.512956e-001	1.354874e+003	9.665214e-002	0.2293097
3	0.48761980826	2.377731e-001	1.281967e+003	9.145118e-002	0.3207609
4	0.46884127166	2.198121e-001	1.185130e+003	8.454313e-002	0.4053040
5	0.45792994707	2.096998e-001	1.130609e+003	8.065378e-002	0.4859578
6	0.44931493268	2.018839e-001	1.088469e+003	7.764766e-002	0.5636055
7	0.44725161164	2.000340e-001	1.078495e+003	7.693616e-002	0.6405416
8	0.43864334107	1.924080e-001	1.037379e+003	7.400307e-002	0.7145447
9	0.43279702651	1.873133e-001	1.009910e+003	7.204356e-002	0.7865883
10	0.41895400808	1.755225e-001	9.463393e+002	6.750864e-002	0.8540969
11	0.39741239445	1.579366e-001	8.515242e+002	6.074485e-002	0.9148417
12	0.35767836655	1.279338e-001	6.897624e+002	4.920531e-002	0.9640471
13	0.30574113954	9.347764e-002	5.039900e+002	3.595294e-002	1.0000000
14	NA	2.817789e-009	1.519227e-005	1.083765e-009	1.0000000
15	0.00004899927	2.400928e-009	1.294474e-005	9.234340e-010	1.0000000

	Quality	Mass	Inertia	DIM 1	DIM 2
Masculino	0.16340580	0.09003021	0.02199396	0.02626639	-0.445989259
Femenino	0.16340580	0.10996979	0.01800604	-0.02150380	0.365123075
Alto y medio alto	0.47394337	0.04753273	0.03049345	-1.23297631	-0.001745245
Medio	0.01561284	0.04894260	0.03021148	-0.14814091	0.161993973
Medio bajo	0.06922850	0.04128902	0.03174220	0.19715177	0.476695704
Bajo	0.47691745	0.06223565	0.02755287	0.92739332	-0.442314812
menor22	0.07255030	0.02054381	0.03589124	-0.06621137	-0.793324822
de23a29	0.05532191	0.02779456	0.03444109	-0.06903707	-0.581368393
de30a39	0.13801105	0.03222558	0.03355488	-0.35105865	-0.771542527
de40a49	0.06951609	0.03927492	0.03214502	0.03503185	-0.532216105
de50a59	0.10587281	0.03202417	0.03359517	-0.14756652	0.730449838
mayor60	0.45529947	0.04813696	0.03037261	0.37272772	1.139060211
Montevideo	0.49096492	0.08559919	0.02288016	-0.73663824	0.336933640
Canelones	0.07538878	0.03746224	0.03250755	0.50702731	-0.264600327
Restopais	0.24288659	0.07693857	0.02461229	0.57268108	-0.246023916
Primaria	0.46713203	0.01893253	0.03621349	1.52560189	1.462911518
Secundaria	0.18326825	0.14138973	0.01172205	0.18077961	-0.208059862
Terciaria	0.46643490	0.03967774	0.03206445	-1.37215158	0.043372287

	DIM 3	CONTR 1	CONTR 2	CONTR 3
Masculino	0.09268855	0.0001800875	7.126106e-002	0.003252953
Femenino	-0.07588238	0.0001474342	5.834010e-002	0.002663132
Alto y medio alto	0.74322113	0.2095061620	5.761301e-007	0.110424685
Medio	-1.08225151	0.0031140935	5.110930e-003	0.241090858
Medio bajo	-0.40450563	0.0046529733	3.733639e-002	0.028413260
Bajo	0.55181419	0.1551892904	4.845263e-002	0.079700725
menor22	-0.23794300	0.0002611203	5.145152e-002	0.004891749
de23a29	-0.90557618	0.0003840777	3.738332e-002	0.095862143
de30a39	0.42699412	0.0115147543	7.633710e-002	0.024710518
de40a49	0.30520085	0.0001397449	4.426970e-002	0.015385944
de50a59	0.50712613	0.0020218463	6.799451e-002	0.034637551
mayor60	-0.24780996	0.0193890538	2.485348e-001	0.012432358
Montevideo	-0.22314079	0.1346705974	3.866995e-002	0.017925239
Canelones	0.14121421	0.0279222872	1.043734e-002	0.003141867
Restopais	0.17949999	0.0731584284	1.853164e-002	0.010425821
Primaria	1.35899078	0.1277572705	1.612352e-001	0.147054711
Secundaria	-0.37924395	0.0133971162	2.435621e-002	0.085524884
Terciaria	0.70296506	0.2165936619	2.970208e-004	0.082461604

	COS 1	COS 2	COS 3
Masculino	0.0005648271	1.628410e-001	0.007033428
Femenino	0.0005648271	1.628410e-001	0.007033428
Alto y medio alto	0.4739424232	9.495739e-007	0.172207564
Medio	0.0071104167	8.502423e-003	0.379490938
Medio bajo	0.0101118124	5.911669e-002	0.042567367
Bajo	0.3885351448	8.838231e-002	0.137558712
menor22	0.0005018658	7.204844e-002	0.006481393
de23a29	0.0007692679	5.455264e-002	0.132361887
de30a39	0.0236719664	1.143391e-001	0.035020212
de40a49	0.0002998872	6.921620e-002	0.022761622
de50a59	0.0041515163	1.017213e-001	0.049030131
mayor60	0.0440362098	4.112633e-001	0.019465434
Montevideo	0.4060215775	8.494334e-002	0.037256199
Canelones	0.0592518779	1.613690e-002	0.004596171
Restopais	0.2050443564	3.784224e-002	0.020144245
Primaria	0.2433607842	2.237712e-001	0.193108407
Secundaria	0.0788393435	1.044289e-001	0.346961621
Terciaria	0.4659693374	4.655623e-004	0.122298362

A3.2 BLOQUE: VARIABLES DE COYUNTURA

La salida del Análisis de Correspondencia Múltiple es la siguiente:

Singular Values	Principal Inertias	Chi Squares	Percents	Acumulado
1	0.6851785	0.46946957	2260.5015	0.18778783	0.1877878
2	0.6079659	0.36962249	1779.7366	0.14784899	0.3356368
3	0.6026616	0.36320105	1748.8173	0.14528042	0.4809172
4	0.5249301	0.27555157	1326.7841	0.11022063	0.5911379
5	0.5017940	0.25179720	1212.4065	0.10071888	0.6918567
6	0.4769616	0.22749235	1095.3783	0.09099694	0.7828537
7	0.4235565	0.17940011	863.8136	0.07176004	0.8546137
8	0.3713434	0.13789588	663.9703	0.05515835	0.9097721
9	0.3674364	0.13500952	650.0724	0.05400381	0.9637759
10	0.3009323	0.09056027	436.0488	0.03622411	1.0000000

	Quality	Mass	Inertia	DIM 1	DIM 2
Buenaper	0.12347333	0.03650554	0.08539778	0.79563361	-0.29845139
Regularper	0.08152313	0.13695871	0.04521652	0.23159748	-0.11682962
Malaper	0.33260200	0.07653575	0.06938570	-0.79393389	0.35141700
Mejorper	0.74626418	0.07502518	0.06998993	0.84650498	1.01186839
Igualper	0.38069734	0.10297080	0.05881168	0.05365219	-0.73532907
Peorper	0.61869113	0.03600201	0.08559919	-1.73698828	0.81265393
Nosabepers	0.11809572	0.03600201	0.08559919	-0.18050982	-0.81815875
Buenapais	0.03999169	0.01007049	0.09597180	0.89690172	0.38518743
Regularpais	0.23029163	0.06545821	0.07381672	0.79712940	-0.11760095
Malapais	0.28543602	0.17447130	0.03021148	-0.35083653	0.02188853
Mejorpais	0.74281316	0.07578046	0.06968781	0.84413013	0.99758420
Igualpais	0.28738378	0.09214502	0.06314199	0.05640549	-0.69938531
Peorpais	0.56259252	0.03851964	0.08459215	-1.61392117	0.69570231
Nosabepais	0.16562343	0.04355488	0.08257805	-0.16068000	-0.87133108

	DIM 3	CONTR 1	CONTR 2	CONTR 3	COS 1
Buenaper	0.29672463	0.0492240751	0.0087972634	0.00884950233	0.108242645
Regularper	-0.09087974	0.0156476768	0.0050575157	0.00311441662	0.064986060
Malaper	0.02109706	0.1027603510	0.0255712224	0.00009379103	0.278114122
Mejorper	-0.13239415	0.1145140085	0.2078244792	0.00362074124	0.307249013
Igualper	0.67152092	0.0006313664	0.1506326649	0.12784568127	0.002015976
Peorper	0.23564139	0.2313732417	0.0643249847	0.00550405601	0.507587467
Nosabepers	-1.88038685	0.0024987398	0.0651993964	0.35048877046	0.005481744
Buenapais	-0.55835877	0.0172557175	0.0040423748	0.00864431008	0.033764226
Regularpais	0.18712633	0.0885960413	0.0024492176	0.00631083458	0.225386047
Malapais	-0.03797763	0.0457431591	0.0002261511	0.00069283928	0.284329279
Mejorpais	-0.13931248	0.1150187434	0.2040318649	0.00404939550	0.309941127
Igualpais	0.80521543	0.0006244636	0.1219400784	0.16449356789	0.001857190
Peorpais	0.17478469	0.2137171536	0.0504394916	0.00323997653	0.474435065
Nosabepais	-1.61571010	0.0023952623	0.0894632951	0.31305211719	0.005446982

	COS 2	COS 3
Buenaper	0.015230683	0.0150549506
Regularper	0.016537068	0.0100066039
Malaper	0.054487878	0.0001963805
Mejorper	0.439015164	0.0075156929
Igualper	0.378681368	0.3158126739
Peorper	0.111103666	0.0093415783
Nosabepers	0.112613971	0.5948555560
Buenapais	0.006227465	0.0130856041
Regularpais	0.004905587	0.0124205024
Malapais	0.001106739	0.0033317130
Mejorpais	0.432872034	0.0084419046
Igualpais	0.285526587	0.3784754551
Peorpais	0.088157454	0.0055644074
Nosabepais	0.160176449	0.5507558680

Al aplicar los índices de Benzecri y Greenacre, se obtienen los siguientes valores:

Benzecri			Greenacre		
Inercia	Porcentaje	Acumulado	Inercia	Porcentaje	Acumulado
0,33667614	0,32602489	0,32602489	0,59342391	0,25160473	0,25160473
0,22780371	0,2205968	0,54662169	0,49885344	0,21150797	0,4631127
0,22110259	0,21410768	0,76072937	0,49245367	0,20879454	0,67190723
0,13437611	0,13012492	0,89085429	0,40030187	0,16972326	0,8416305
0,1127115	0,10914571	1	0,37352338	0,1583695	1
1,03267004	1		**2,35855627**	1	

El promedio de los autovalores es de 0.486, por lo que se utilizan los 5 primeros valores propios.

A3.3 BLOQUE: TRAYECTORIA ELECTORAL

Singular Values	Principal Inertias	Chi Squares	Percents	Acumulado
1	0.8430998	0.71081726	5327.9561	0.218713002	0.2187130
2	0.8024576	0.64393818	4826.6616	0.198134823	0.4168478
3	0.7417795	0.55023684	4124.3199	0.169303642	0.5861515
4	0.6497036	0.42211477	3163.9763	0.129881467	0.7160329
5	0.4713254	0.22214759	1665.1152	0.068353106	0.7843860
6	0.4216213	0.17776449	1332.4401	0.054696767	0.8390828
7	0.3881725	0.15067792	1129.4117	0.046362438	0.8854452
8	0.3130791	0.09801853	734.7014	0.030159548	0.9156048
9	0.2926064	0.08561851	641.7566	0.026344156	0.9419490
10	0.2823276	0.07970889	597.4608	0.024525812	0.9664748
11	0.2588971	0.06702769	502.4084	0.020623904	0.9870987
12	0.1660705	0.02757941	206.7224	0.008485971	0.9955846
13	0.1197912	0.01434992	107.5604	0.004415361	1.0000000

	Quality	Mass	Inertia
P.Col99	0.32822375	0.05060423	0.06135254
PNac99	0.19039814	0.04909366	0.06181734
EPFA99	0.87660269	0.09667674	0.04717639
Menor99	0.84828553	0.02215509	0.07010613
Otros99	0.08207569	0.03147029	0.06723991
Vázquez	0.85963149	0.10045317	0.04601441
Batlle	0.63135484	0.09264854	0.04841583
OtrosBalotaje	0.08667910	0.03373615	0.06654272
Menorbalotaje	0.85285004	0.02316213	0.06979627
NOAncap	0.09316234	0.06168177	0.05794407
SIAncap	0.34070155	0.12286002	0.03911999
OtrosAncap	0.17675032	0.06545821	0.05678209
PCol94	0.29472657	0.04859013	0.06197227
PNac94	0.14357725	0.04959718	0.06166241
EPFA94	0.74077318	0.07225579	0.05469053
Menor94	0.68301666	0.04053374	0.06445116
Otros94	0.04259099	0.03902316	0.06491595

	DIM 1	DIM 2	DIM 3
P.Col99	-0.9165348	-0.67324843	-0.837887694
PNac99	-0.7107296	-0.52347992	-0.158695676
EPFA99	1.1719429	-0.12956452	0.017372983
Menor99	-0.4288075	2.92232264	-0.804683507
Otros99	-0.7158019	0.23991921	2.108016049
Vázquez	1.1204262	-0.15619645	0.030039112
Batlle	-0.8442243	-0.59963248	-0.530564529
OtrosBalotaje	-0.7304716	0.14853658	1.911367032
Menorbalotaje	-0.4183946	2.85960038	-0.791967842
NOAncap	-0.4894153	-0.21190325	-0.522471606
SIAncap	0.5924800	-0.03921405	-0.113769554
OtrosAncap	-0.6508595	0.27327982	0.705865715
PCol94	-0.8653563	-0.68762123	-0.760657984
PNac94	-0.5532660	-0.52348469	-0.219816461
EPFA94	1.3337508	-0.20822888	0.001047082
Menor94	-0.2023425	1.86780191	-0.413820691
Otros94	-0.4787236	-0.03301316	1.654422274

	CONTR 1	CONTR 2	CONTR 3
P.Col99	0.059803527	0.03561995320	6.456673e-002
PNac99	0.034888009	0.02089205924	2.247015e-003
EPFA99	0.186800017	0.00252028689	5.302994e-005
Menor99	0.005731129	0.29382286147	2.607198e-002
Otros99	0.022684463	0.00281310959	2.541552e-001
Vázquez	0.177407586	0.00380593856	1.647359e-004
Batlle	0.092895857	0.05173264805	4.739858e-002
OtrosBalotaje	0.025324696	0.00115589451	2.239928e-001
Menorbalotaje	0.005704173	0.29413391607	2.640244e-002
NOAncap	0.020785189	0.00430118270	3.060079e-002
SIAncap	0.060673627	0.00029339307	2.890101e-003
OtrosAncap	0.039010422	0.00759163040	5.927323e-002
PCol94	0.051189401	0.03567815998	5.109487e-002
PNac94	0.021358318	0.02110672141	4.355397e-003
EPFA94	0.180827327	0.00486530864	1.439741e-007
Menor94	0.002334709	0.21960088932	1.261512e-002
Otros94	0.012581549	0.00006604691	1.941179e-001

	COS 1	COS 2	COS 3
P.Col99	0.213190977	0.1150327696	1.781732e-001
PNac99	0.123435619	0.0669625171	6.154063e-003
EPFA99	0.866017805	0.0105848840	1.903105e-004
Menor99	0.017879643	0.8304058841	6.296284e-002
Otros99	0.073786342	0.0082893474	6.399383e-001
Vázquez	0.843243387	0.0163881051	6.061228e-004
Batlle	0.419646428	0.2117084074	1.657464e-001
OtrosBalotaje	0.083237355	0.0034417431	5.699015e-001
Menorbalotaje	0.017874549	0.8349754900	6.404395e-002
NOAncap	0.078454812	0.0147075287	8.941078e-002
SIAncap	0.339215571	0.0014859763	1.250779e-002
OtrosAncap	0.150260170	0.0264901547	1.767313e-001
PCol94	0.180658028	0.1140685378	1.395874e-001
PNac94	0.075756725	0.0678205212	1.195841e-002
EPFA94	0.723146960	0.0176262168	4.456959e-007
Menor94	0.007922762	0.6750938962	3.313805e-002
Otros94	0.042389404	0.0002015867	5.062679e-001

Valores de la inercia al aplicar los índices de Benzecri y Greenacre

Benzecri			Greenacre		
Inercia	Porcentaje	Acumulado	Inercia	Porcentaje	Acumulado
0,62536422	0,31760233	0,31760233	0,79377328	0,25705925	0,25705925
0,54259449	0,27556626	0,59316859	0,74143575	0,24011002	0,49716927
0,42995036	0,21835793	0,81152652	0,66426025	0,21511715	0,71228642
0,28402305	0,14424616	0,95577268	0,54965801	0,17800382	0,89029024
0,08708432	0,04422732	1	0,33877278	0,10970976	1
1,96901644	1		**3,08790007**	1	

El promedio de los valores propios es de 0.442, por lo que se utilizan los 5 primeros.

A3.4 BLOQUE: DEFINICIONES POLÍTICAS BÁSICAS

	Singular Values	Inertias	ChiSquares	Percents	Acumulado
1	0.7088853	0.50251837	4352.5867	0.18844439	0.1884444
2	0.5758489	0.33160197	2872.1862	0.12435074	0.3127951
3	0.4892613	0.23937659	2073.3717	0.08976622	0.4025613
4	0.4596857	0.21131094	1830.2797	0.07924160	0.4818030
5	0.4208969	0.17715418	1534.4293	0.06643282	0.5482358
6	0.4139115	0.17132276	1483.9202	0.06424603	0.6124818
7	0.3933992	0.15476291	1340.4863	0.05803609	0.6705179
8	0.3862728	0.14920665	1292.3605	0.05595250	0.7264704
9	0.3768800	0.14203851	1230.2733	0.05326444	0.7797348

	Quality	Mass	Inertia	DIM 1	DIM 2
Mucho	0.075792645	0.01443437	0.05708711	-0.81897466	-0.35864818
Bastante	0.086295828	0.03256126	0.05028953	-0.57967078	0.13926971
Poco	0.001771709	0.06277274	0.03896022	-0.01311412	0.05253894
Nada	0.160196280	0.05689829	0.04116314	0.55396057	-0.04667889
Simpatía fuerte	0.269018955	0.04347096	0.04619839	-0.83038021	-0.26992999
Simpatía a secas	0.030372605	0.06059080	0.03977845	-0.20735574	-0.10087958
Indefinidas	0.412357545	0.06260490	0.03902316	0.77727586	0.28506541
CercPCol	0.325261560	0.01795905	0.05576536	0.31591491	-1.61043009
NeutroPCol	0.573146344	0.03927492	0.04777190	1.24065887	0.56552380
LejPCol	0.479690520	0.10943270	0.02146274	-0.49711207	0.06132431
CercPNac	0.458585917	0.03977845	0.04758308	0.32631180	-1.16462511
NeutroPNac	0.580657309	0.04044982	0.04733132	1.12956613	0.73206824
LejPNac	0.536665057	0.08643840	0.03008560	-0.67875987	0.19337418
CercFA	0.735112231	0.07099698	0.03587613	-0.91259426	0.39717563
NeutroFA	0.595799999	0.03893924	0.04789778	1.16653370	0.77040628
LejFA	0.603204868	0.05673045	0.04122608	0.34139513	-1.02585666
Izquierda	0.107417864	0.01124538	0.05828298	-1.13395949	0.44581043
Centro-izquierda	0.464339175	0.05119168	0.04330312	-0.89292191	0.50011784
Centro	0.177859521	0.04363880	0.04613545	0.61873471	0.34437537
Centro-derecha	0.304037452	0.02786170	0.05205186	0.55893772	-1.09648586
Derecha	0.213868960	0.01158107	0.05815710	0.15687041	-1.68504481
Nosabeiden	0.070693138	0.02114804	0.05456949	0.66537875	0.20906276

	DIM 3	CONTR 1	CONTR 2	CONTR 3
Mucho	1.71258881	0.01926579525	0.0055990986	0.1768570921
Bastante	0.29845045	0.02177269080	0.0019045723	0.0121161330
Poco	-0.69132401	0.00002148313	0.0005225364	0.1253293288
Nada	0.15744294	0.03474601195	0.0003738721	0.0058920158
Simpatía fuerte	0.99490083	0.05964874388	0.0095517829	0.1797534245
Simpatía a secas	-0.85350664	0.00518426136	0.0018594998	0.1843906251
Indefinidas	0.13521872	0.07526737204	0.0153419394	0.0047818980
CercPCol	0.24453529	0.00356674120	0.1404591145	0.0044862627
NeutroPCol	0.25253600	0.12030070987	0.0378791325	0.0104635794
LejPCol	-0.13076488	0.05381505388	0.0012410673	0.0078171355
CercPNac	-0.24642808	0.00842871695	0.1627058008	0.0100912867
NeutroPNac	0.33606046	0.10270413440	0.0653736258	0.0190840130
LejPNac	-0.04385848	0.07924778070	0.0097473509	0.0006945959
CercFA	-0.04027501	0.11766394201	0.0337744242	0.0004810936
NeutroFA	0.19097556	0.10544600458	0.0696963460	0.0059328230
LejFA	-0.08068047	0.01315766709	0.1800414312	0.0015426641
Izquierda	1.64438028	0.02877514009	0.0067399651	0.1270273240
Centro-izquierda	-0.40499736	0.08122212641	0.0386124121	0.0350769689
Centro	-0.11946953	0.03324527025	0.0156070170	0.0026019892
Centro-derecha	-0.38742134	0.01732138010	0.1010174846	0.0174700039
Derecha	1.14234580	0.00056712491	0.0991640815	0.0631338243
Nosabeiden	0.23732443	0.01863184917	0.0027874448	0.0049759185

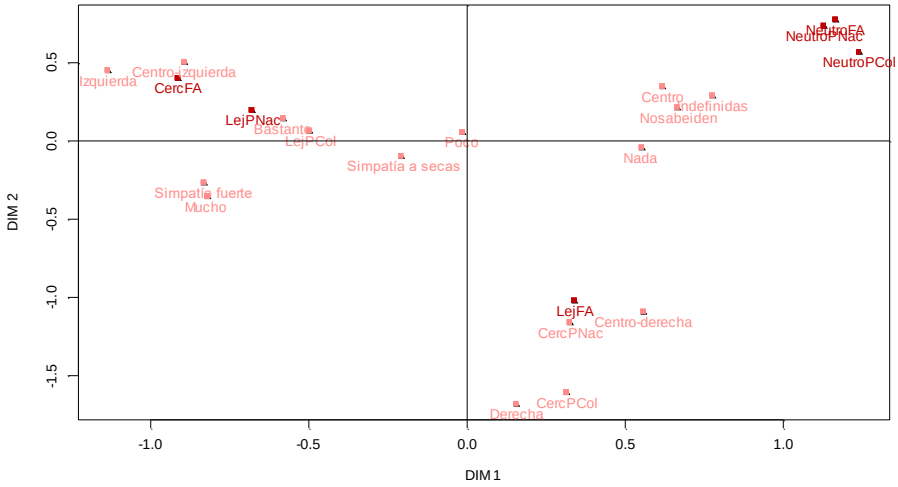
	COS 1	COS 2	COS 3
Mucho	0.0635963355	0.012196309	0.278097681
Bastante	0.0815863989	0.004709429	0.021627157
Poco	0.0001039105	0.001667798	0.288764791
Nada	0.1590668404	0.001129439	0.012848970
Simpatía fuerte	0.2433087239	0.025710232	0.349271611
Simpatía a secas	0.0245596542	0.005812951	0.416105956
Indefinidas	0.3634691052	0.048888440	0.010999952
CercPCol	0.0120528656	0.313208695	0.007221595
NeutroPCol	0.4745465852	0.098599759	0.019661682
LejPCol	0.4725000283	0.007190491	0.032694555
CercPNac	0.0333804443	0.425205473	0.019037369
NeutroPNac	0.4089050974	0.171752211	0.036193788
LejPNac	0.4963770008	0.040288056	0.002072461
CercFA	0.6180462518	0.117065980	0.001203752
NeutroFA	0.4148565108	0.180943488	0.011118825
LejFA	0.0601436848	0.543061183	0.003359017
Izquierda	0.0930376856	0.014380178	0.195644812
Centro-izquierda	0.3534584392	0.110880736	0.072713625
Centro	0.1357932965	0.042066224	0.005062717
Centro-derecha	0.0627089344	0.241328518	0.030127956
Derecha	0.0018376347	0.212031325	0.097447859
Nosabeiden	0.0643412214	0.006351917	0.008185333

Índices de Benzecri y Greenacre para el cálculo de la inercia

Benzecri			Greenacre		
Inercia	Porcentaje	Acumulado	Inercia	Porcentaje	Acumulado
1,76400628	0,35484691	0,35484691	2,73607194	0,22532518	0,22532518
1,0045806	0,2020811	0,556928	2,10406589	0,17327725	0,39860243
0,62440378	0,12560486	0,68253286	1,70329015	0,14027195	0,53887438
0,51516092	0,1036296	0,78616246	1,56877836	0,12919443	0,66806881
0,38779807	0,07800933	0,86417179	1,39451672	0,11484337	0,78291218
0,36678005	0,07378135	0,93795314	1,36341658	0,11228217	0,89519435
0,30844585	0,06204686	1	1,272631	0,10480565	1
4,97117555			**12,1427706**		

El promedio de los autovalores es de 0.389, lo que determina el uso de los 7 primeros para la ponderación.

El gráfico del ACM de este bloque es el siguiente:



Los ejes no tienen una interpretación clara, pero pueden encontrarse grupos de modalidades en el primer plano factorial: el primer cuadrante agrupa las modalidades que pueden identificarse como “apolíticas” : neutro con respecto a los tres partidos, de centro y nada de interés en la política. El segundo cuadrante estaría identificado como de “izquierda”: cercano al EPFA, lejano a los partidos tradicionales, e identificados como de izquierda o centro izquierda. El tercer cuadrante agrupa las modalidades “simpatía fuerte hacia su partido” y mucho interés en la política. El cuarto cuadrante puede ser identificado como “tradicional”: agrupa las modalidades cercano al Partido Colorado y al Nacional, lejano al EPFA, identificación como de derecha y centro derecha.

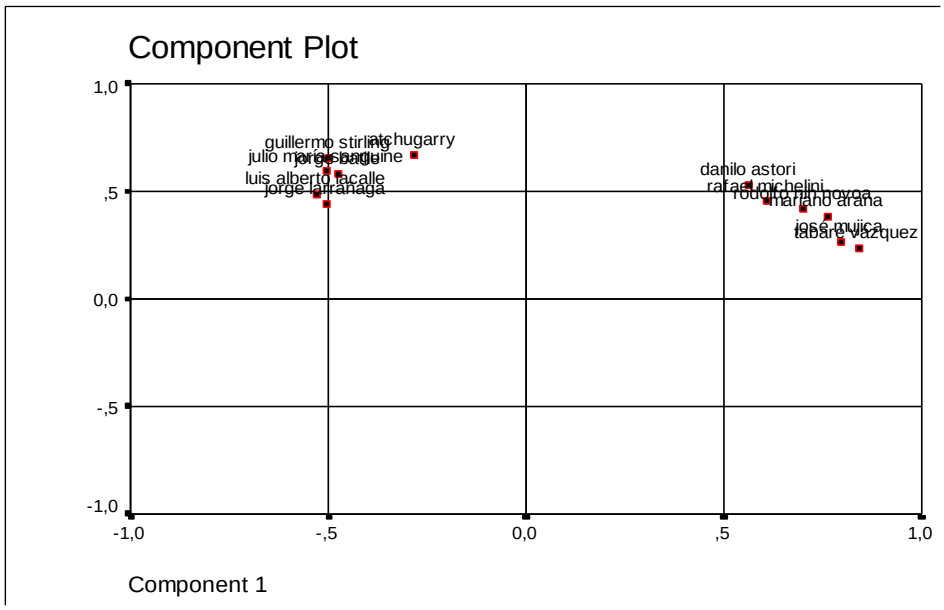
A3.4 BLOQUE: POPULARIDAD DE LÍDERES

```
[1] "Importancia de los Factores"
          Valores Proporcion Acumulado
 COMP 1 4.4367194 0.36972662 0.3697266
 COMP 2 2.9314394 0.24428662 0.6140132
 COMP 3 0.7936319 0.06613600 0.6801492
 COMP 4 0.7212855 0.06010713 0.7402564
 COMP 5 0.5828937 0.04857447 0.7888308
 COMP 6 0.4594997 0.03829164 0.8271225
 COMP 7 0.4338872 0.03615727 0.8632797
 COMP 8 0.4107488 0.03422907 0.8975088
 COMP 9 0.3639272 0.03032727 0.9278361
COMP 10 0.3222909 0.02685758 0.9546937
COMP 11 0.2966850 0.02472375 0.9794174
COMP 12 0.2469912 0.02058260 1.0000000
```

```
[1] "Matriz de Saturaciones"
```

	Comp..1	Comp..2	mp..3
BATLLE	4.733952e-001	5.814317e-001	2.944521e-001
LACALLE	5.284026e-001	4.831256e-001	-4.036623e-001
MUJICA	-7.957090e-001	2.621153e-001	-2.863303e-002
LARRANAGA	5.054575e-001	4.422953e-001	-6.207132e-001
NINNOVOA	-7.000504e-001	4.198999e-001	-1.215206e-001
STIRLING	4.988755e-001	6.548163e-001	2.031976e-001
MICHELINI	-6.094857e-001	4.525121e-001	-7.763206e-002
SANGUINETTI	5.031463e-001	5.922444e-001	2.865148e-001
VAZQUEZ	-8.397260e-001	2.380869e-001	2.888643e-002
ARANA	-7.610370e-001	3.835882e-001	2.922935e-002
ASTORI	-5.625878e-001	5.308169e-001	-4.255344e-003
ATCHUGARRY	2.834933e-001	6.703908e-001	1.095494e-001

Representación gráfica de las variables en el primer plano factorial



A3.5 ÁRBOLES DE CLASIFICACIÓN

A3.5.1 Variables básicas

Se presenta únicamente la primera parte de la salida

```
      *** Tree Model ***

Classification tree:
snip.tree(tree = tree(formula = SIMPAT ~ INDOP + EDAD + EDUCA + REGION, data
       = decididos, na.action = na.exclude, mincut = 5, minsize = 10, mindev
       = 0.01), nodes = c(2, 14, 24, 26, 31, 61, 111, 120, 25, 54, 220))
Number of terminal nodes:  13
Residual mean deviance:  1.939 = 1635 / 843
Misclassification error rate: 0.3972 = 340 / 856
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

  1) root 856 1786.00 EPFA ( 0.5456 0.3201 0.09813 0.03621 )
    2) REGION:Montevideo 374  693.00 EPFA ( 0.6684 0.2166 0.06952 0.04545 ) *
    3) REGION:Canelones,Restopais 482 1044.00 EPFA ( 0.4502 0.4004 0.12030 0.02905)
      6) EDAD:de18a29,de30a49 303  607.70 EPFA ( 0.5413 0.3531 0.07261 0.03300 )
       12) EDUCA:Primaria 113  211.60 EPFA ( 0.5133 0.3894 0.09735 0.00000 )
         24) INDOP:Medio bajo 17   34.97 P. Nacional ( 0.3529 0.4706 0.17650
0.00000 ) *
```

A3.5.2 Variables de definición política

```
      *** Tree Model ***

Classification tree:
snip.tree(tree = tree(formula = SIMPAT ~ CERPC + CERPN + CERFA + POLINTER +
       opublica + opint, data = decididos, na.action = na.exclude, mincut = 5,
       minsize = 10, mindev = 0.01), nodes = c(6, 11, 31, 16, 10, 68, 9, 35,
       69))
Number of terminal nodes:  11
Residual mean deviance:  0.8808 = 744.3 / 845
Misclassification error rate: 0.1577 = 135 / 856
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

 1) root 856 1786.000 EPFA ( 0.54560 0.32010 0.098130 0.036210 )
   2) CERFA:nocercaFA 446  987.500 P. Nacional ( 0.18390 0.57620 0.179400 0.060540
)
     4) CERPC:nocercaPC 362  713.000 P. Nacional ( 0.21820 0.64360 0.066300
0.071820 )
       8) CERPN:nocercaPN 197  483.000 P. Nacional ( 0.35030 0.42130 0.101500
0.126900 )
        16) opublica<4.5 34   40.320 EPFA ( 0.79410 0.17650 0.000000 0.029410 ) *
```

ANEXO 4: MODELO CON VARIABLE DE RESPUESTA EN CUATRO CATEGORÍAS

Se presenta a continuación el modelo de regresión logística multinomial usando la variable de respuesta con 4 categorías:

- Es un *modelo de regresión logística multinomial* en el cual la variable de respuesta es SIMPAT, que consta de 4 categorías (EPFA, P. Nacional, P. Colorado, Otras respuestas) y cuyas variables explicativas son: *Edad* (en tres tramos), *Educación* (con tres modalidades), *Opint* (con cuatro modalidades), *Cerca FA*, *Cerca PN*, *Cerca PC* (como variables indicatices) , *Lider1*, *Lider3* e *Hist1* (como variables continuas ya que son ejes factoriales) y *Opinión Pública* (como variable continua con valores de 1 a 10)

- La <u>modalidad de referencia</u> es *EPFA*, quedando los tres logit formados de la siguiente manera:

  $Ln(\frac{\pi_j}{\pi_{EP}}) = X\beta_j$ donde j = P. Nacional, P .Colorado y Otras Respuestas

  X es la matriz de diseño de dimensiones 856 x 15

  β_j son los vectores de parámetros de la regresión en cada caso, de dimensiones 15 x 1

  π_j = Prob (SIMPAT = j / X)

- <u>Significación del modelo</u> Se realiza mediante la deviance de los residuos, quedando el modelo como válido.
- <u>Interpretación de los parámetros</u>:
  - ✓ En este caso, todas las variables usadas son significativas (por lo menos en alguna de sus modalidades) Para ver la significación se usa el Estadístico de Wald, cuya fórmula es

    $z = \frac{\hat{\beta}_{jk}}{s(\hat{\beta}_{jk})}$ y que se distribuye asintóticamente normal.

  - ✓ Las variables con mayor influencia para la explicación de la variable de respuesta dependen del logit considerado.
    Para cada individuo h y para cada variable k en el logit j, la fórmula en la que se basa la interpretación es la siguiente:

    $\frac{\hat{\pi}}{\hat{\pi}_{EP}} = \exp(X_h^{'}\hat{\beta}_{jk})$

  - ✓ En el logit P. Nacional / EPFA, la mayor influencia la tiene la variable Hist1 (eje factorial que resume la trayectoria electoral) con coeficiente negativo. Si se recuerda que las coordenadas positivas de este eje representan las modalidades de voto al Frente Amplio, se concluye que valores positivos de la variable dan

valores menores que 1 en el cociente de probabilidades (por ser el exponente negativo), lo que redunda en una mayor probabilidad de pertenencia a la categoría EPFA comparada con la probabilidad de pertenencia a la categoría P. Nacional
Las variables que le siguen en influencia son CercaPN (con coeficiente positivo) y CercaFA con coeficiente negativo. Ambas son variables indicatrices. La interpretación es la siguiente: si el individuo posee la modalidad CercaPN, la probabilidad de pertenencia a la categoría P. Nacional aumenta (comparado con la modalidad de referencia). Disminuye si el individuo considerado posee la modalidad CercaFA

- ✓ En el logit P. Colorado / EPFA, las variables con mayor influencia son semejantes al caso anterior. Predomina Hist1 (también con coeficiente negativo y con la misma interpretación) y le siguen CercaPC (con coeficiente positivo) y CercaFA (con coeficiente negativo)

- ✓ Por último, en el logit Otras Respuestas / EPFA la variable con mayor influencia es Opint. La modalidad "Aprueba Montevideo" es la que tiene coeficiente con mayor valor absoluto, pero negativo. Esto significa que si el individuo considerado vive en Montevideo y aprueba la gestión del intendente frenteamplista crece la probabilidad de que vote al EPFA en contra de que vote a otras opciones como partidos menores, en blanco o anulado, comparada con el mismo cociente pero con "Otras respuestas" en el caso de la gestión del intendente. Le sigue en importancia la modalidad "Aprueba Colorada", con coeficiente positivo. La interpretación es que si el individuo vive en un departamento con Intendencia a cargo del Partido Colorado y aprueba su gestión, aumenta la probabilidad de pertenencia a la categoría Otras respuestas versus la probabilidad de que vote al EPFA

- ✓ Los coeficientes de intersección ($\beta_{j\,0}$) son significativos y negativos en todos los logits. Esto puede interpretarse como que, sin considerar ninguna variable explicativa, es más probable la pertenencia del individuo a la categoría EPFA que a cualquiera de las otras. Se entiende que este resultado es debido a la propia distribución de la variable SIMPAT en la base Decididos, en donde el porcentaje de los encuestados que dicen que votarán al Encuentro Progresista es del 54.5%

- ➢ <u>Poder predictivo del modelo</u>

Se realiza una tabla de datos cruzados entre la variable de respuesta (SIMPAT) y la predicción que efectúa el modelo para cada uno de los individuos en base a los coeficientes estimados. Se obtiene para cada categoría cuántos de los individuos son clasificados correctamente (aciertos) y cuántos son erróneamente asignados a otra modalidad. También se puede observar si las asignaciones erróneas se distribuyen entre las demás categorías o si, por el contrario, todas se vuelcan a una misma (lo que estaría indicando una tendencia que debe ser corregida) En este caso, la tabla muestra lo siguiente:

		Predice					
	Categoría	EPFA	P.N.	P.C.	Otros	Total	Porcentaje Aciertos
Observados	EPFA	449	13	3	2	467	96,15
	P.N.	11	252	9	2	274	91,97
	P.C.	2	16	63	3	84	75
	Otros	7	13	4	7	31	22,58
	Total	469	294	79	14	856	90,1

Fuente: Elaboración propia – Base: Encuesta Agosto - Equipos Mori

Se observa un menor porcentaje de error en las categorías EPFA y P. Nacional, que son las que tienen mayores frecuencias relativas. El mayor porcentaje se da en la categoría Otras respuestas (cuya frecuencia relativa es de 0.36). El porcentaje total de error es de 8.9%.

Al usar la propiedad predictiva del modelo, se obtienen las siguientes proyecciones para el grupo de Indecisos:

Categoría	Proyecta indecisos	% indecisos
EPFA	47	34,31
P.N.	68	49,64
P.C.	14	10,22
Otros	8	5,84
Total	137	100

Si se resume la información en una única tabla, se obtienen los valores totales proyectados (se usan los datos obtenidos en la encuesta para los Decididos y los que se obtienen al aplicar el modelo a los Indecisos)

Modalidad	Decididos		Indecisos	
EPFA	468	54.79%	47	34.31%
P.Nacional	294	34.34%	68	49.64%
P.Col y Otros	79	9.23%	14	10.22%
Total	856	100%	137	100%

Categoría	Número individuos	Total proyectado
EPFA	516	51,76
P.N.	362	34,44
P.C.	93	9,87
Otros	22	3,93
Total	993	100,00

ANEXO 5: MODELO FINAL

A5.1. SIGNIFICACION DEL MODELO GENERAL

Modelo	-2 Log verosimilitud	Chi-Square	GL	Sig.
Solo con costante	1651.887			
Final	407.477	1244.409	28	.000

A5.2. SIGNIFICACIÓN DE LAS VARIABLES EN GENERAL

El estadístico χ^2 es la diferencia en –2ln (verosimilitud) entre el modelo final y el modelo reducido. Éste se forma omitiendo la variable que se desea testear del modelo final. La hipótesis nula es que todos los parámetros de esa variable son cero.

Test razón de verosimilitud

Variables	-2 Log verosimilitud-reducido	Chi-Square	df	Sig.
Intercept	407.477	.000	0	.
HIST1	437.027	29.550	2	.000
LIDER1	470.965	63.487	2	.000
LIDER3	488.479	81.002	2	.000
OPUBICA	420.949	13.472	2	.001
OPINT	420.658	13.181	6	.040
CERCAPC	421.518	14.040	2	.001
CERCAPN	450.428	42.951	2	.000
CERCAFA	434.369	26.891	2	.000
EDAD	414.143	6.666	4	.155
EDUCAOP	413.398	5.921	4	.205

A5.3. SIGNIFICACIÓN DE LAS VARIABLES Y MODELO GENERAL EN MODELO FINAL

Modelo	-2 Log verosimilitud	Chi-Square	GL	Sig.
Solo con costante	1651.887			
Final	419.945	1231.941	20	.000

El estadístico χ^2 es la diferencia en –2ln (verosimilitud) entre el modelo final y el modelo reducido. Éste se forma omitiendo la variable que se desea testear del modelo final. La hipótesis nula es que todos los parámetros de esa variable son cero.

En el modelo finalmente propuesto, se rechaza en todas las variables esta hipótesis nula. El cuadro con los estadísticos correspondientes es el siguiente:

Variable	-2ln(L_M)	χ^2	Grados libertad	p-valor
Intercept	419.945	0	0	0
Lider1	485.526	65.581	2	0.000
Lider3	506.127	86.181	2	0.000
Hist1	452.964	33.019	2	0.000
CercaFA	446.694	26.746	2	0.000
CercaPN	463.755	43.810	2	0.001
CercaPC	434.133	14.187	2	0.000
Opint	434.284	14.339	6	0.026
Opublica	430.475	10.530	2	0.005

A5.4 SIGNIFICACIÓN DE LAS VARIABLES EN CADA LOGIT

```
      *** Multiple Logistic Model ***

Re-fitting to get Hessian

Call:
multinom(formula = INTENCIO ~ CERCAPC + CERCAPN + CERCAFA + OPUBICA + OPINT +
      HIST1 + LIDER1 + LIDER3, data = decididos1, na.action = na.omit, Hess
       = F, trace = F)

Coefficients:
      (Intercept)     CERCAPC   CERCAPN    CERCAFA   OPUBICA      OPINT1
 PNac   -3.979294 -0.08238002 1.4037617 -1.3003415 0.4462914 -0.08285803
Otros   -4.054162  0.65652470 0.5916428 -0.8105496 0.3073694 -0.50668387
         OPINT2     OPINT3     HIST1   LIDER1     LIDER3
 PNac -0.2272377 -0.5198056 -1.690837 1.404414 -1.1292696
Otros -0.1978109 -0.2201888 -2.034587 1.423423  0.6178291

Std. Errors:
      (Intercept)   CERCAPC   CERCAPN   CERCAFA   OPUBICA    OPINT1    OPINT2
 PNac   0.9291572 0.3581973 0.2781490 0.2763025 0.1417044 0.2786524 0.2165817
Otros   0.9983225 0.3626338 0.3207273 0.3217992 0.1539147 0.3352956 0.2582078
        OPINT3     HIST1    LIDER1    LIDER3
 PNac 0.1836827 0.3798696 0.2238130 0.2715103
Otros 0.1902906 0.5104150 0.2460627 0.3101601

Residual Deviance: 419.9453
AIC: 463.9453
```

A5.5 ESTADÍSTICO DE WALD

	PNac	Otros
(Intercept)	-4,28269	-4,06097
CERCAPC	-0,22999	1,810434
CERCAPN	5,046798	1,844691
CERCAFA	-4,70622	-2,51881
OPUBICA	3,149453	1,997011
OPINT1	-0,29735	-1,51116
OPINT2	-1,0492	-0,76609
OPINT3	-2,82991	-1,15712
HIST1	-4,4511	-3,98614
LIDER1	6,274944	5,784798
LIDER3	-4,15921	1,991968

A5.6 INTERVALOS DE CONFIANZA PARA EXP(β)

	Est. puntual	Intervalo	
Pnac/EPFA		Lím. Inferior	Lím. Superior
(Intercept)	0,019	0,003	0,116
CERCAPC	0,921	0,456	1,858
CERCAPN	4,070	2,360	7,021
CERCAFA	0,272	0,159	0,468
OPUBICA	1,563	1,184	2,063
OPINT1	0,920	0,533	1,589
OPINT2	0,797	0,521	1,218
OPINT3	0,595	0,415	0,852
HIST1	0,184	0,088	0,388
LIDER1	4,073	2,627	6,316
LIDER3	0,323	0,190	0,550
PCy Otros			
(Intercept)	0,017	0,002	0,123
CERCAPC	1,928	0,947	3,925
CERCAPN	1,807	0,964	3,388
CERCAFA	0,445	0,237	0,835
OPUBICA	1,360	1,006	1,839
OPINT1	0,602	0,312	1,162
OPINT2	0,821	0,495	1,361
OPINT3	0,802	0,553	1,165
HIST1	0,131	0,048	0,356
LIDER1	4,151	2,563	6,724
LIDER3	1,855	1,010	3,407

ANEXO 6: INTERVALOS DE CONFIANZA PARA LA REGRESIÓN LOGÍSTICA BINOMIAL y MUTINOMIAL

A6.1 INTERVALOS DE CONFIANZA PARA LA REGRESIÓN LOGÍSTICA BINOMIAL

A6.1.1 Inferencia sobre respuesta media

Frecuentemente es requerida una estimación de la probabilidad π para uno o más conjuntos diferentes de valores de las variables predictivas

A.6.1.2 Estimación puntual

El vector de niveles de las variables X para los cuales π será estimado se anota como Xh

$$X_h = \begin{pmatrix} 1 \\ X_{h1} \\ X_{h2} \\ \cdots \\ X_{h,p-1} \end{pmatrix}_{p*1}$$

y la respuesta media de interés para πh:

$$\pi_h = [1 + \exp(-\beta' X_h)]^{-1} = \frac{1}{1 + \exp(-\beta' X_h)} = \frac{\exp(\beta' X_h)}{1 + \exp(\beta' X_h)}$$

El estimador puntual de πh es:

$$\hat{\pi}_h = [1 + \exp(-b' X_h)]^{-1}$$ con b: vector de coeficientes estimados en la regresión

A6.1.3 Estimación de intervalos

Obtendremos un intervalo de confianza para πh en dos etapas: primero calculamos los límites de confianza para el logit de respuesta media π'h y luego usamos la relación

$$E(Y) = [1 + \exp(-\beta' X)]^{-1}$$ para obtener los límites de confianza para πh

Para aclarar, consideramos X = Xh:

$$E(Y_h) = [1 + \exp(-\beta' X_h)]^{-1}$$ y rescribimos la expresión usando el hecho de que E(Yh) = πh y que: $\pi'h = \beta'Xh$

$$\pi_h = [1 + \exp(-\pi'_h)]^{-1}$$

Esta relación se utiliza para convertir los límites de confianza para π'h en límites de confianza para πh

El estimador puntual del logit de respuesta media:

$$\pi'_h = \beta' X_h \quad \text{es} \quad \hat{\pi}'_h = b' X_h$$

Ahora: b'Xh = X'hb (porque es un escalar), entonces se tiene que la variancia estimada aproximada de $\hat{\pi}'_h = b' X_h = X'_h b$ es:

$$s^2\{\hat{\pi}'_h\} = s^2\{X'_h b\} = X'_h s^2(b) X_h$$

donde s2(b) es la matriz de variancias y covariancias aproximadas estimadas de los coeficientes de regresión cuando n es grande.

El intervalo de confianza 1-α aproximado (para muestras grandes) el logit de respuesta media πh se obtiene de la manera habitual:

$$L = \hat{\pi}'_h - z(1-\alpha/2)s(\hat{\pi}'_h)$$
$$U = \hat{\pi}'_h + z(1-\alpha/2)s(\hat{\pi}'_h)$$

Aquí, L y U son respectivamente los límites inferior y superior respectivamente para π'h

Se usa la relación monótona entre πh y π'h para convertir L y U para π' en límites del intervalo de confianza 1-α, L* y U* se tiene:

$$L^* = [1+\exp(-L)]^{-1}$$
$$U^* = [1+\exp(-U)]^{-1}$$

A6.1.4 Intervalos de confianza simultánea para varias medias de respuesta

Si deseamos estimar varios π_h correspondientes a distintos vectores X_h con familia de coeficientes de confianza 1-α , pueden ser usados los intervalos de confianza simultáneos de Bonferroni

El procedimiento para g intervalos de confianza es el mismo que para un solo intervalo, excepto que z(1 - α/2) es reemplazado por z(1 - α/2g).

A6.2 INTERVALOS DE CONFIANZA PARA LA REGRESIÓN LOGÍSTICA MULTINOMIAL

A.6.2.1 Intervalos de confianza para las probabilidades de pertenencia

Efectuando un razonamiento análogo al expuesto para la Regresión Logística Dicotómica, se pueden estimar los intervalos de confianza para la probabilidad de pertenencia a una categoría en la Regresión Logística Multinomial.

Sea Y la variable de respuesta con J categorías y sea j = 1 la categoría de referencia en la regresión logística. En este caso concreto la variable de respuesta presenta 3 categorías denominadas EPFA, P.Nac y P.Col y Otros (siendo la primera la de referencia) los odds se forman de la siguiente manera:

$$\frac{P\{Y = PNac / X_h\}}{P\{Y = EPFA / X_h\}} = \exp(X'_h \beta_{PN}) \qquad \frac{P\{Y = PCol / X_h\}}{P\{Y = EPFA / X_h\}} = \exp(X'_h \beta_{PCyO})$$

Al igual que en el caso de la regresión logística dicotómica, sea:

$\pi'_{jh} = \beta'_j X_h$ con j = PNac, PCol y Otros y $\hat{\pi}'_{jh} = \hat{\beta}'_j X_h$ su valor puntual estimado y la variancia estimada : $s^2\{\pi'_{jh}\} = X'_h s^2\{\hat{\beta}_j\} X_h$

Sean los intervalos de confianza simultáneos de Bonferroni para los dos $\pi_{j'}$, de la forma habitual:

$$L_j = \hat{\pi}'_{jh} - z(1-\alpha/4)s(\hat{\pi}'_{jh}) \qquad U_j = \hat{\pi}'_{jh} + z(1-\alpha/4)s(\hat{\pi}'_{jh})$$

Por lo tanto, con una probabilidad de (1 - α) se cumplirán las siguientes desigualdades:

$$\exp(L_{PN,h}) \le \frac{\pi_{PN,h}}{\pi_{EPFA,h}} \le \exp(U_{PN,h}) \qquad \exp(L_{PCyO,h}) \le \frac{\pi_{PCyO,h}}{\pi_{EPFA,h}} \le \exp(U_{PCyO,h})$$

Al sumar miembro a miembro las desigualdades, se obtiene (con una probabilidad de (1-α):

$$\exp(L_{PN,h}) + \exp(L_{PCyO,h}) \le \frac{\pi_{PN,h} + \pi_{PCyO,h}}{\pi_{EPFA,h}} \le \exp(U_{PN,h}) + \exp(U_{PCyO,h})$$

Por ser Y una variable que con distribución multinomial, se cumple que:

$$\pi_{EPFA,h} + \pi_{Pn,h} + \pi_{PCyO,h} = 1$$

$$\exp(L_{PN,h}) + \exp(L_{PCyO,h}) \le \frac{1-\pi_{EPFA,h}}{\pi_{EPFA,h}} \le \exp(U_{PN,h}) + \exp(U_{PCyO,h})$$

$$1 + \exp(L_{PN,h}) + \exp(L_{PCyO,h}) \le \frac{\pi_{EPFA,h}}{\pi_{EPFA,h}} \le 1 + \exp(U_{PN,h}) + \exp(U_{PCyO,h})$$

$$\exp(L_{PN,h}) + \exp(L_{PCyO,h}) \le \frac{1}{\pi_{EPFA,h}} - 1 \le \exp(U_{PN,h}) + \exp(U_{PCyO,h})$$

$$L^*_{EPFA,h} = \frac{1}{1+\exp(U_{PN,h}) + \exp(U_{PCyO,h})} \le \pi_{EPFA,h} \le \frac{1}{1+\exp(L_{PN,h}) + \exp(L_{PCyO,h})} = U^*_{EPFA,h}$$

Usando los odds, se desprende que:

$$L^*_{PN,h} = \frac{\exp(L_{PN,h})}{1+\exp(U_{PN,h}) + \exp(U_{PCyO,h})} \le \pi_{PN,h} \le \frac{\exp(U_{PN,h})}{1+\exp(L_{PN,h}) + \exp(L_{PCyO,h})} = U^*_{EPFA,h}$$

$$L^*_{PCyO,h} = \frac{\exp(L_{PCyO,h})}{1+\exp(U_{PN,h}) + \exp(U_{PCyO,h})} \le \pi_{PCyO,h} \le \frac{\exp(U_{PCyO,h})}{1+\exp(L_{PN,h}) + \exp(L_{PCyO,h})} = U^*_{EPFA,h}$$

Nota: el análisis se realiza en base a una única observación. Si se requieren más observaciones, el único cambio es usar los intervalos simultáneos de Bonferroni con $z_{1-\alpha/4n}$ siendo n el número de observaciones a estimar.