Universidad de Chile
Facultad de Medicina
Escuela de Salud Pública

TESIS PARA OPTAR AL GRADO ACADEMICO DE
MAGÍSTER EN BIOESTADÍSTICA

"HACIA UNA NUEVA PROPUESTA EN LA SELECCION DE LAS FAMILIAS DEL PROGRAMA CHILE SOLIDARIO: Aplicación de Multivariate Adaptive Regression Splines (MARS) y Análisis Discriminante Basado en Distancias (DB)"

Profesor Guía: Sergio Alvarado Orellana
Alumna : Alina Oyarzún Chicuy

1 ANTECEDENTES

1.1 Política Social

Los grandes lineamientos de la política social provienen de la reformas de inicios de los años 80s y tienen como objetivo la superación de la pobreza e igualación de oportunidades. Se introducen conceptos de focalización, descentralización, subsidios a la demanda y participación privada. Estas características se aplican actualmente con distinto énfasis en la gama de políticas y programas sociales del país (Larrañaga, 2005).

Según el mismo autor las políticas sociales se pueden agrupar en tres categorías:

Asistenciales: Que se orientan a grupos de población con carencias importantes y su objetivo es *Alivio de la pobreza*. Ejemplo de este tipo son los programas de subsidios monetarios. La focalización es aquí una característica clave y parte importante del diseño se centra en resolver el problema de identificación de beneficiarios.

Servicios Sociales: Los sectores tradicionales de educación, salud y vivienda canalizan gran parte del gasto social, incluye una variedad de programas tanto focalizados como universales. Sus objetivos son: superación de la pobreza, cobertura de necesidades básicas e Igualdad de oportunidades

Inversión Productiva: Esta dirigido a potenciar las oportunidades productivas de los grupos mas pobres. Por ejemplo, incluir acciones de incremento de

capital humano, iniciativas que aumenten la productividad de las microempresas y demás activos físicos.

1.2 Focalización

El objetivo de la focalización es aumentar la efectividad del gasto social, asignando los escasos recursos a los grupos que presentan las mayores carencias.

La motivación para aplicar un método de focalización proviene de las siguientes características en el contexto de las políticas públicas (Coday *et al.*2004):

Reducir al máximo la pobreza, o en términos más amplios, el aumento del bienestar social.

Un presupuesto limitado para destinarlo a los fines de reducción de la pobreza La disyuntiva entre la cantidad de beneficiarios que cubre la intervención y el nivel de transferencias.

1.3 Mecanismos de Elegibilidad para Programas Sociales

La elección de un mecanismo particular de elegibilidad depende de varios factores: (a) factibilidad presupuestaria y administrativa; (b) factibilidad técnica, dado el grado de informalidad en la economía, y (c) aceptabilidad política. Considerando estos factores (Castañeda, 2005), hace referencia a tres tipos de mecanismos de selección:

1. *Test de Medios Verificado (TMV):* Produce resultados “gold standard” con respecto a la exactitud de la focalización. Verificación extensiva de información puede también promover transparencia y credibilidad. Estados Unidos utiliza este sistema. Sin embargo, TMV puede ser extremadamente costoso de implementar y técnica y administrativamente poco factible de desarrollar en países con alto niveles de trabajo informal.

2. *Test de Medios No verificados (TMN):* Puede ser de menor costo, y una alternativa factible, particularmente en situaciones en las cuales se requieren decisiones rápidas, tales como admisiones a hospitales en sistemas con subsidio de salud para familias de bajos ingresos. La focalización a través de este procedimiento no es tan poderosa como Test de Medios Verificados (TMV) o Test de Medios Proxy (TMP). Sin embargo, preocupaciones como la falta de transparencia, error de mediciones e incentivos adversos para el auto reporte hacen de TMN menos atractivo desde un punto de vista técnico y político cuando la elegibilidad para una gran cantidad de beneficios esta siendo determinada. Brasil utiliza TMN para la elegibilidad de los beneficiarios de sus programas sociales.

3. *Test de Medios Proxy (TMP)*: Es una buena alternativa para la focalización de transferencias monetarias en países en desarrollo con altos niveles de trabajo informal. TMP puede ser más transparente y exacto que TMN. Los costos de entrevistas de estos sistemas son más bajos con respecto a los costos de entrevistas en sistemas TMV, además presentas altos niveles

de transparencias con respectos a los dos sistemas anteriores. Países como Chile, Colombia, Costa Rica y México utilizan este sistema para un amplio grupo de programas sociales.

Los sistemas TMP usualmente involucran tres pasos; (1) determinación de las variables y ponderaciones a utilizar en la predicción; (2) recolección de los datos de los hogares, y (3) determinación de la elegibilidad del hogar calculando puntajes TMP compuesto.

1.4 Problema de Investigación

La necesidad de desarrollar un mecanismo que permita medir objetivamente el nivel de bienestar o nivel socioeconómico de los hogares, surge de la presencia de incentivos adversos asociada al auto-reporte de información de los individuos.

Los incentivos adversos, como subestimar el nivel de ingreso o de gasto, ocurren cuando el individuo sabe que el indicador socioeconómico en el que se le clasifique afecta sus posibilidades de acceder a las transferencias de los programas públicos. Por ello, es que la mayoría de los instrumentos de focalización buscan identificar un conjunto de características, observables y verificables, que permitan determinar objetivamente la calificación de los individuos como potenciales beneficiarios de programas públicos.

Las características observables están formadas en su gran mayoría por variables de tipo binaria o nominal, es decir toman valores discretos. Por lo tanto, para aplicar metodologías paramétricas, se deben transformar dichas variables a continuas. Además, se requiere conocer la función de distribución asociada a cada una de ellas. Este proceso de transformación generalmente es realizado de modo arbitrario. Asignando valores continuos a cada una de las variables categóricas.

En conclusión, existe la necesidad de desarrollar una metodología estadística que permita utilizar datos de tipo binario, nominal y/o continuo para la clasificación de población beneficiaria y no beneficiaria sin necesidad de transformaciones arbitrarias de variables, permitiendo de este modo trabajar con la verdadera relación de las variables predictoras con la variable respuesta y de este modo hacer que la interpretación de los resultados sea mas simple.

1.5 Hipótesis

Un modelo estadístico de clasificación que pertenezca a la familia de los modelos aditivos generalizados (GAM), usando metodología Splines, permitiría realizar mejores predicciones que un modelo clásico. Ya que esta ultima metodología esta ligada a distribuciones de probabilidad y se hace complejo el trabajar con variables predictoras de distinta naturaleza.

2 OBJETIVOS

2.1 Objetivo General

Generar un modelo de discriminación explicito usando la metodología Multivariate Adaptive Regression Splines (MARS) para la selección de beneficiarios del Programa Chile Solidario, y comparar las clasificaciones con Análisis Discriminante Basado en Distancias (DB).

2.2 Objetivos Específicos

1. Aplicar técnicas estadísticas multivariadas para datos mixtos para la selección de beneficiarios del Sistema Chile Solidario.

2. Comparar los resultados de clasificación obtenido con el Análisis Discriminante Basado en Distancia con el proceso de selección realizado por MIDEPLAN.

3. Comparar los resultados de clasificación obtenido mediante MARS con el proceso de selección realizado por MIDEPLAN.

4. Evaluación de las metodologías de clasificación de beneficiarios en base a su correcta selección.

5. Comparar las ventajas y desventajas del análisis discriminante Basado en Distancias y MARS.

6. Generar un modelo de discriminación explicito utilizando MARS.

7. Especificar que variables son importantes en la clasificación de beneficiarios y entregar valores de corte de las variables predictoras que permitan optimizar la toma de decisión.

3 MATERIAL Y METODO

3.1 Material

Este estudio utiliza dos instrumentos desarrollados por MIDEPLAN para la planificación y el accionar de la política social, uno para la selección de beneficiarios a los diversos programas sociales denominado Ficha CAS y otro para evaluación de los programas sociales y caracterización socioeconómica de la población denominada Encuesta de Caracterización Socioeconómica Nacional (CASEN).

3.1.1 La Ficha CAS

La ficha CAS es el principal instrumento de focalización de los programas gubernamentales existentes en el país. Así, todos los subsidios monetarios (SUF, PASIS, etc.) utilizan este instrumento para identificar a las familias que presentan las mayores carencias. La Ficha CAS es también un insumo importante en la asignación de beneficios de los Programas de Vivienda Social, Junta Nacional de Jardines Infantiles, Mejoramiento de Barrios y Sernam, entre otros.

La Ficha CAS es un registro de las principales características socioeconómicas de las familias. Incluye un total de 50 variables agrupadas en 9 secciones. Debe ser aplicadas a las familias que postulan a aquellos beneficios sociales que se asignan sobre la base del índice CAS. (Ver anexo 1).

En su actual versión, el índice CAS es un promedio ponderado sobre la base de 13 variables agrupadas en cuatro factores principales: vivienda, ocupación, educación y patrimonio-ingresos. De esta manera el índice CAS es una medida de la condición socioeconómica de cada familia. En el anexo 2 se presentan las variables utilizadas en el cálculo del índice CAS.

3.1.2 La Encuesta CASEN

La Encuesta de Caracterización Socioeconómica Nacional (CASEN), es una herramienta básica para la formulación del diagnóstico y evaluación del impacto de la política social en los hogares y programas más importantes que componen el gasto social. Además, proporciona información acerca de las condiciones socioeconómicas de los diferentes sectores sociales del país, sus carencias más importantes, la dimensión y características de la pobreza, así como la distribución del ingreso de los hogares.

Se ha aplicado desde el año 1985 con una periodicidad de dos años, excepto la del año 89, que debió realizarse en 1990; y la del 2002 que se realizó el 2003. Las encuestas realizadas hasta la fecha, corresponden a los años 1985, 1987, 1990, 1992, 1994, 1996, 1998, 2000 y 2003. La información que proporciona esta encuesta, constituye un antecedente básico para focalizar el gasto social y sirve de manera sustantiva al proceso de descentralización de la gestión del Estado. Sus resultados se obtienen a nivel regional y los mismos están referidos a nivel de comunas, para un número significativo de ellas.

El formulario de la Encuesta CASEN está organizado en seis módulos que contienen series de preguntas relativas a las siguientes temáticas: residentes del hogar, vivienda, educación, salud, empleo e ingresos del trabajo y otros ingresos.

3.1.3 Homologación ficha CAS y Encuesta CASEN

Para poder calcular la distribución nacional de los índices CAS a partir de la información disponible en la Encuesta CASEN 2000, fue necesario construir una Matriz de Homologación (base de datos), comparando cada una de las variables CAS (que otorgan puntajes) con sus símiles CASEN. Una vez definida la matriz, fue necesario replicar el modelo base de cálculo de Índice CAS obteniéndose las ponderaciones de los factores y sub-factores obtenidas a partir de diversos análisis factoriales y métodos valorativos.

Esta metodología permite analizar las características socioeconómicas de las familias según el índice CAS a nivel nacional, de acuerdo, a la realidad actual de la situación de pobreza. En el anexo 3 se indica el detalle de las variables homologadas.

3.1.4 Programa Chile Solidario

Chile Solidario es un sistema de Protección Social diseñado por el Gobierno de Chile en el año 2002 que combina dos elementos centrales:

asistencia y promoción, desde una perspectiva integradora para abordar la extrema pobreza en que viven hoy 225.000 familias.

Este sistema de protección integral surge a partir de visualizar la extrema pobreza como un problema multidimensional relacionado con las variables de: ingresos monetarios insuficientes; escasa presencia de capital humano; débil capital social; y alta vulnerabilidad de las familias ante sucesos que las afectan como enfermedades, accidentes, cesantía y otros.

El Sistema Chile Solidario comprende cuatro componentes:

- Apoyo psicosocial intensivo y bono de protección a la familia.
- Subsidios monetarios garantizados.
- Acceso preferente a programas de promoción social.
- Beneficios previsionales y de inserción laboral.

3.1.4.1 Selección de beneficiarios del Programa Chile Solidario

La homologación CAS-CASEN permitió determinar un índice CAS de corte, particular para cada una de las regiones del país (ver anexo 4), entendiéndose que aquellas familias en extrema pobreza que obtienen, en la aplicación de la ficha CAS, un índice igual o inferior al índice de corte de la correspondiente región, son **exclusivamente** aquellas elegibles para ser invitadas a integrarse y participar en el Sistema Chile Solidario.

El índice de corte para cada región se estableció entre el percentil de puntaje y el porcentaje de indigencia asociado para cada región del país.

Con estos índices, la cobertura regional de familias a atender en el período 2002 – 2005 es la que indica en el anexo 5.

3.1.5 Programas Estadísticos

Se utilizo el paquete estadístico SPSS versión 12.0 para Windows para el análisis exploratorio de los datos. Para la aplicación del análisis de partición fuzzy C-means y el análisis discriminante Basado en Distancias se utilizo el programa Ginkgo versión 1.4 de Julio 2005 desarrollado por el Departamento de Biología Vegetal y del Departamento de Estadística de la Universidad de Barcelona. Ginkgo realiza métodos de análisis multivariantes a partir de una matriz de distancias o similaridad, permitiendo elegir entre distintos coeficientes de similaridad y disimilaridad adecuados a los datos que se dispone.

Para el análisis de regresión multivariada MARS se utilizo el programa del mismo nombre versión 2.0 del año 2000, desarrollado por Salford Systems. Este programa permite ejecutar el método multivariate adaptive regresion splines, el cual es una técnica flexible e innovativa que construye modelos para variable dependiente y/o variables independientes de tipo continuo, categórica o binaria.

3.1.6 Descripción de los Datos

La información utilizada corresponde a una muestra de jefes de familias de las comunas de la V región de Chile, de la base de datos homologada CAS-CASEN del año 2000. En esta región se seleccionaron aleatoriamente dos muestras una de entrenamiento y otra de validación. El tamaño de cada una de las muestra es de 1.000 jefes de familias de los cuales 199 jefes de familias corresponden a beneficiarios del Programa Chile Solidario según la clasificación de MIDEPAN. En el anexo 6 y anexo 7 se muestra la distribución de la muestra de entrenamiento y de validación por comunas de la V región.

Las variables predictoras del modelo son 13 de las cuales 11 son variables cualitativa de tipo categórica y dos variables son cuantitativas de tipo continuas (años de estudio del jefe de familia e ingreso familiar per cápita).

Las variables predictoras del estudio y sus respectivas codificaciones son las siguientes:

I. Vivienda

1. Material muro (p19) (Categórica)

1 Ladrillo, concreto o bloque

2 Albañilería de piedra

3 Tabique forrado

4 Adobe

5 Barro, quincha o pirca

6 Tabique sin forro interior

7 Desecho

2. Material piso (p20) (Categórica)

1 Radier revestido

2 Radier no revestido

3 Madera sobre solera

4 madera, plástico o pastelón sobre tierra

5 Piso de tierra

3. Material del techo (p21) (Categórica)

1 Teja, tejuela o losa

2 Zinc, pizarreño con cielo interior

3 Zinc, pizarreño sin cielo interior

4 Fonolita

5 Paja, coirón, totora o caña

6 Desecho

4. Abastecimiento de agua (p23) (Categórica)

1 Red Pública con llave dentro de la vivienda

2 Red Pública con llave dentro del sitio

3 Red Pública con llave fuera del sitio

4 Sin Red Pública con llave dentro de la vivienda

5 Sin Red Pública con llave dentro del sitio

6 Sin Red Pública, acarreo

5. Sistema de eliminación de excreta (p24) (Categórica)

1 Uso exclusivo alcantarillado

2 Uso exclusivo fosa séptica

3 Uso exclusivo letrina sanitaria

4 Uso exclusivo pozo negro

5 Uso compartido alcantarillado

6 Uso compartido fosa séptica

7 Uso compartido letrina sanitaria

8 Uso compartido pozo negro

9 No tiene sistema

6. Disponibilidad de tina o ducha (p25) (Categórica)

1 Tina o ducha de uso exclusivo con agua caliente

2 Tina o ducha de uso exclusivo con agua fría

3 Tina o ducha de uso compartida con agua caliente

4 Tina o ducha de uso compartida con agua fría

5 No tiene Tina o ducha

7. Hacinamiento (hacin2) (Binaria)

0 Con Hacinamiento

1 Sin Hacinamiento

II. Educación

1. Años de escolaridad (p45jefe) (Continua)

III. Ocupación

1. Categoría ocupacional (catcas1) (Categórica)

0 Familiar no remunerado

3 Trabajador por cuenta propia

4 Trabajador dependiente urbano

5 Asalariado agrícola

6 Pequeño productor agrícola

7 Empleado sector público o equivalente

8 Jubilado o dependiente de terceros

9 Actividad mejor remunerada

10 No tiene actividad

IV. Ingreso/patrimonio

1. Sitio* (p47) *(Categórica)

1 Sitio propio sin deuda

2 Sitio propio sin deudas atrasadas

3 Sitio propio con deuda

4 Arrienda el sitio

5 Usan el sitio pero no creen que serán desalojados

6 Usan el sitio pero creen que serán desalojados

2. Tenencia de Refrigerador (p49) (Binaria)

0 No tiene

1 Si tiene

3. Tenencia de Calefont (p50) (Binaria)

0 No tiene

1 Si tiene

4. Ing. autónomo familiar per cápita (ifpc) (Continua)

Por otra parte, la variable respuesta es cualitativa de tipo binaria, asignándose el valor 1, en caso de ser beneficiario del Programa Chile Solidario es decir si el índice CAS de corte es menor o igual 494. A su vez, si el índice CAS de corte es mayor que 494 es no beneficiario del Programa Chile Solidario y en este caso se le asigno el valor 0. Es decir, el valor 1 se considera como el evento que para este estudio es ser seleccionado como beneficiario del Programa Chile Solidario. El valor 0 no constituye evento, es decir no es seleccionado beneficiario del Programa Chile Solidario.

3.2 METODO

3.2.1 Multivariate Adaptive Regresión Spline (MARS)

Multivariate Adaptive Regression Splines (MARS) es una metodología, desarrollada por Jerome Friedman el año 1991. MARS es una generalización de Regresión Recursiva Particionada (PR) que utiliza funciones *splines* para ajustar un modelo en lugar de otras funciones mas simples. Dado un conjunto de variables predictoras, MARS ajusta un modelo en la forma de una expansión de productos de funciones bases de predictores elegidos durante una estrategia recursiva particionada *forward y backward*. MARS produce modelos continuos con derivadas continuas. MARS es poderoso y flexible para modelar relaciones que son aditivas o involucran interacciones en las variables. El modelo puede ser representado de tal forma que separadamente identifica las contribuciones aditivas y estas asociadas con las diferentes interacciones multivariadas. (Friedman, 1991).

3.2.1.1 Regresión Recursiva Particionada

El origen de la metodología de Partición Recursiva es desarrollado por Morgan y Sonquist el año 1963, la cual aparece inicialmente utilizada en la técnica en AID (Automatic Interaction Detection). Extensiones y contribuciones a esta metodología han sido realizadas por Breiman, Friedman, Olsen y Stone el año 1984.

Se define Partición Recursiva como una aproximación a la función desconocida $f(x)$ en x usando una expansión en un conjunto de funciones

bases, dónde cada función base es un producto de funciones de salto univariante. La Partición Recursiva es considerada el precursor de MARS, el cual usa una expansión en un conjunto de funciones bases que son productos de funciones univariantes spline.

Siguiendo a Alvarado (2002) y Friedman (1991) la Regresión Recursiva Particionada consiste en N muestras de y y $x=(x_1, x_2, ..., x_N)$ definidas como $\{y_i, x_i\}_{i=1}^{N}$. Sea $\{R_j\}_{j=1}^{S}$ un conjunto de subregiones disjuntas de *D* tal que $D = \bigcup_{j=1}^{S} R_j$. Dada las subregiones de $\{R_j\}_{j=1}^{S}$, la partición recursiva estima la función desconocida $f(x)$ en x con

$$\hat{f}(x) = \sum_{m=1}^{M} a_m B_m(\mathrm{x}) \quad (2)$$

La función base B_m toma la forma:

$$B_m(x) = I[x \in \mathrm{R}_m]$$

Dónde $I[.]$ es una función indicadora con valor 1 si el argumento es verdadero y 0 en otro caso. Los $\{a_m\}_1^M$ son los coeficientes de la expansión cuyos valores son conjuntamente ajustados para realizar el mejor ajuste de los datos. Los $\{R_m\}_1^M$ son las subregiones del espacio de las covariables. Sea

$H[\eta\,]$ una función indicadora, la cual es un producto de funciones de salto univariante,

$$H(\eta) = \begin{cases} 1, & \text{si } \eta \geq 0, \\ 0, & \text{otro caso} \end{cases}$$

Esto describe cada subregión R_m. Asi $B_m(\mathrm{x})$ es una función con valores 1 si y solo si x es un miembro de los $R_m th$ subregión de D.

A pesar de que Partición Recursiva es una metodología muy robusta, tiene algunas desventajas tales como (Friedman, 1991):

1. Modelos de partición recursiva tienen subregiones disjuntas y son usualmente discontinuas en subregiones límites.

2. La partición recursiva no tiene habilidad para estimar adecuadamente funciones $f(\mathrm{x})$ que son lineales con más que unos pocos coeficientes no ceros y funciones aditivas.

3. La forma del modelo de partición recursiva (2), en una combinación aditiva de funciones de variables predictoras en regiones disjuntas, lo que hace que la estimación de la verdadera forma de la función desconocida $f(x)$ sea dificultosa para tamaños grandes de *p*.

3.2.1.2 Regresión Spline para un predictor

Para enfrentar los problemas de Partición Recursiva señalados anteriormente es que se desarrolla Regresión Spline.

Por tanto, una manera de estimar la función $f(\mathrm{x})$, sería usar como una aproximación la función spline $\hat{f}_q(x)$ la que se obtiene dividiendo el rango de la variable predictora x en $K+1$ regiones disjuntas separadas por K nodos. Esta aproximación posee la forma de un polinomio de grado q en cada una de las subregiones generadas por la partición, para que la función y sus primeras $q-1$ derivadas sean continuas en cada región. Estas restricciones y el uso de polinomios dentro de cada subregión producen funciones suavizadas y ajustadas.

Usualmente el orden que debe considerarse en la función *spline* debe ser menor o igual que tres, debido a que cada polinomio de grado q se define por $q-1$ parámetros y como se tiene $K+1$ regiones, finalmente se debe estimar $(K+1)\times(q+1)$ parámetros, los que usualmente se estiman por mínimos cuadrados.

La continuidad en cada región, agrega q restricciones para cada nodo generando de esta manera $K\times q$ restricciones, luego se deben estimar $K+q+1$ parámetros.

El ajuste de la regresión spline se realiza escogiendo un conjunto de funciones bases de la forma $\left\{ B_k^{(q)}(x) \right\}_0^{K+q}$, que genera el espacio de las funciones spline de orden q, a su vez aplicando el ajuste de mínimos cuadrados.

Bajo estas consideraciones, la aproximación toma la forma $\hat{f}(x) = \sum_{k=0}^{K+q} a_k B_k^{(q)}(x)$, en donde $\{ a_k \}_0^{K+q}$ son coeficientes sin restricciones y las restricciones de continuidad son incorporadas en las funciones bases $\left\{ B_k^{(q)}(x) \right\}_0^{K+q}$, una función esta compuesta por las funciones:

$$\left\{ x^j \right\}_0^q, \left\{ \left(x - t_k \right)_+^q \right\}_1^K$$

Las ubicaciones de los nodos que definen las $K+1$ regiones y las funciones de dominio truncadas se definen por:

$$\left(x - t_k \right)_+^q = \begin{cases} 0 & , x \leq t_k \\ (x - t_k)^q & , x \geq t_k \end{cases}$$

3.2.1.3 Selección Adaptiva de Nodos

El nodo marca el fin de una región de los datos y el inicio de otra. Es decir, el nodo describe la conducta de los cambios de la función. En el

enfoque clásico spline, los nodos son predeterminados e igualmente espaciados, sin embargo en MARS, los nodos son determinados por un procedimiento automático de búsqueda. La estrategia involucra una minimización del criterio Generalizad Cross-Validation

$$\sum_{i=1}^{N} (y_i - \sum_{k=0}^{K+q} a_q B_k^{(q)}(x))^2$$

Smith (1982) citado por Friedman (1991) sugiere el uso de funciones básales truncadas que al ser reemplazadas en la ecuación anterior da como resultado:

$$\sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{q} b_j x^j - \sum_{k=1}^{K} a_k (x - t_k)_+^q \right\}^2$$

En dónde los coeficientes $\{b_j\}_0^q, \{a_k\}_1^K$ se pueden considerar como los parámetros asociados con una regresión lineal múltiple de la respuesta y sobre las variables predictoras $\{x^j\}_0^q$ y $\{(x - t_k)_+^q\}^K$ respectivamente.

Si se agrega o elimina un nodo, esto involucra agregar o eliminar la respectiva variable $(x - t_k)_+^q$, ya que es seleccionado automáticamente el nodo y su ubicación.

La estrategia de Smith consiste en comenzar con un número grande de nodos $(t_1, ..., t_{K\max})$ dónde $K_{\max} = N - 2$ y considerar las

correspondientes variables $\left\{(x - t_k)_+^q\right\}^{K\max}$ como candidatas a ser seleccionadas por procedimiento stepwise. Una de las características de la regresión spline es que las observaciones anómalas afectan las respuestas localmente y no globalmente.

3.2.1.4 Producto Tensor Spline, Extensión a *p* Predictores

Para el caso multivariado se define la aproximación spline de p variables $x = (x_1, ..., x_p)$, en dónde el espacio dimensional R^p es dividido en un conjunto de regiones disjuntas y en cada una de ellas $\hat{f}_q(x)$ se toma como un polinomio de p variables. La aproximación $\hat{f}_q(x)$ se restringe para una de ellas y todas sus derivadas de orden $q - 1$ sean continuas en todas partes. De esta manera se imponen restricciones sobre los polinomios en las regiones disjuntas y en sus límites. La aproximación se construye seleccionando un conjunto de funciones bases que generen el espacio de todas las funciones spline.

Para *p*> 2 se toman regiones disjuntas que definen la aproximación spline como productos tensores de intervalos disjuntos en cada una de las variables delineadas por la ubicación del nodo. Así se ubican K_j nodos en cada una de las variables $(0 \leq j \leq p)$ produciendo $\prod_{j=1}^{p} (K_j + 1)$ regiones. Un

conjunto de funciones bases que generen el espacio de funciones splines sobre todo el conjunto de regiones es el producto tensorial de las correspondientes básales splines unidimensionales asociadas a la ubicación de los nodos a cada variable

$$\hat{f}(x) = \sum_{k_1=0}^{K_1+q} \cdots\cdots\cdots \sum_{k_n=0}^{K_p+q} a_{k_1 \ldots . k_p} \prod_{j=1}^{p} B_{k_j}^{(q)}(x_j)$$

Dónde $\left\{ B_{k_j}^{(q)}(x_j) \right\}_{K_j=0}^{K_j}$ es el conjunto de funciones bases para la aproximación splines de orden q dada la ubicación de los K_j nodos en x_j. El número de coeficientes para ser estimados es de $\prod_{j=1}^{p} (K_j + q + 1)$.

3.2.1.5 Suavizamiento

Una de las ideas centrales de MARS en la generalización de la partición recursiva es la de reemplazar las funciones de salto por una potencia truncada. De esta manera se logra una aproximación en la forma de una expansión de productos tensores de funciones bases splines, la propiedad de continuidad de la aproximación incluida en el producto tensorial esta dominada por la elección del orden q para las funciones splines univariantes.

Una dificultad de la regresión splines de orden mayor se centra en el llamado efecto extremo. La ubicación de x cerca de los bordes del dominio de D, hace que se genere largas contribuciones del error cuadrático medio.

Un enfoque sugerido por Stone y Koo (1985), sugieren modificar las funciones basales splines de modo que cerca de los extremos de los intervalos, estos sean unidos suavemente por una función lineal, estas funciones que parecen una spline de orden $q = 1$ poseen derivadas continuas, la estrategia consiste en reemplazar cada función $b(x/s,t) = (s(x-t))_+$ por funciones cúbicas truncadas de la forma:

$$C(x/S=+1,t_-,t,t_+) = \begin{cases} 0 & ,x \le t_- \\ p_+(x-t_-)^2 + r_+(x-t_-)^3 & ,t_- < x < t_+ \\ (x-t) & ,x \ge t \end{cases}$$

$$C(x/S=+1,t_-,t,t_+) = \begin{cases} -(x-t) & ,x \le t_- \\ p-(x-t_+)^2 + r_-(x-t_+)^3 & ,t_- < x < t_+ \\ 0 & ,x \ge t \end{cases}$$

Dónde $t_- < t < t_+$ y

$$p_+ = \frac{(2t_+ + t_- - 3t)}{(t_+ - t_-)^2} \qquad p_- = \frac{(3t - 2t_- - t_+)}{(t_- - t_+)^2}$$

$$r_+ = \frac{(2t - t_+ - t_-)}{(t_+ - t_-)^3} \qquad r_- = \frac{(t_- + t_+ - 2t)}{(t_- - t_+)^3}$$

De esta manera $C(x/S=+1,t_-,t,t_+)$ será continua y tendrá primeras derivadas continuas. Cada función lineal truncada $b(x/s,t)$ es caracterizada

por una sola ubicación del nodo t, en cambio la función cúbica truncada es caracterizada por tres nodos: un nodo central t y nodos laterales t_+ y t_-.

Para tratar la colinealidad de las variables predictoras MARS, ajusta una secuencia de modelos de manera de ir incrementando su orden de interacción y comparar el puntaje GCV. Además, computacionalmente existe la opción de manejar la colinealidad de las variables predictoras incorporando un orden de penalización que puede ser, moderada, alta y sin penalización.

3.2.1.6 Algoritmo MARS

Si se elige una sub-base del producto tensorial completo de las funciones bases splines de las p-variables, con nodos distintos de los datos marginales, entonces la función base esta representada de la siguiente forma:

$$B_m(x) = \prod_{K=1}^{K_m} \left[S_{k_m} (x_{v(k,m)} - t_{k_m}) \right]^2$$

Dónde K_m es el número de factores en la m-esima función base, S_{k_m} toma solo dos valores siendo $S_{k_m} = \pm 1$ e indica el sentido izquierdo o derecho del truncamiento $v(k,m)$ es la etiqueta de la variable predictora $1 \leq v(k,m) \leq p$, t_{k_m} es la ubicación del nodo en cada una de las variables correspondientes. El exponente q es el orden de la aproximación splines. Estos dos lados de la base potencia truncada son equivalentes a las ecuaciones del producto

tensorial truncado, cuando se incluyen los monomios $\left\{ x_j^K \right\}_{k=1 j=1}^{q=1.....p}$ en cada una de las variables y la constante $B_0(x) = 1$.

3.2.1.7 Selección de Funciones Bases

MARS utiliza la estrategia setpwise (forward/backward) para generar un conjunto de funciones bases, forward es un procedimiento iterativo en la que cada interacción construye una lista expandida de funciones bases que se consideran simultáneamente y luego se decide cuales ingresan en este paso.

Cada iteración agrega dos nuevas funciones bases al modelo actual y este procedimiento continua hasta tener un número grande de funciones bases incluidas (sobreajuste).

El procedimiento forward se inicia usando una función base $B_0(x) = 1$, después de la M-esima iteración hay $2M + 1$ funciones en el modelo $\left\{ B_m(x) = 1 \right\}_0^{2M}$.

En donde cada $B_m(x)$ tiene la forma $B_m(x) = \prod_{K=1}^{K_m} \left[S_{km}(x_v(k,m) - t_{km}) \right]^q$, la $M + 1$ iteración agrega dos nuevas funciones bases:

$$B_{2M+1}(x) = B_{l(M+1)} \left[+ (x_{v(M+1)} - t_{M+1}) \right]_+^q$$

$$B_{2M+2}(x) = B_{v(M+1)} \left[- (x_{v(M+1)} - t_{M+1}) \right]_+^q$$

Donde $B_{l(M+1)}(x)$ es una de las $2M+1$ funciones bases ya elegidas con $0 \le l(M+1) \le 2M$, $v(M+1)$ es una de las variables predictoras y t_{M+1} es la ubicación del nodo en la variable. Los parámetros $l(M+1), v(M+1)$ y t_{M+1} definen las dos nuevas funciones bases que son seleccionadas por las que proporcionan un mejor ajuste del modelo, cuyo procedimiento esta definido por:

$$(l(M+1), v(M+1), t_{M+1}) = \arg\min_{l,v,t} \sum_{i=1}^{N} \left\{ y_i - \sum_{m=0}^{2M} a_m B_m(x) - a_{2M+1} B_l(x) \left[+(x_v - t) \right]_+^q - a_{2M+2} B_l(x) \left[-(x_v - t) \right]_+^q \right\}^2$$

Sin embargo, en el algoritmo MARS se deben ingresar las funciones bases de menor orden de interacción antes de que puedan ingresar las de mayor orden.

3.2.1.8 Selección del Modelo

MARS elige un número grande de funciones bases como posibles candidatas para luego eliminar las funciones bases que están en exceso, la estrategia de eliminación es análogo a una regresión lineal estándar en que el procedimiento de selección backward parte con $M_{\max}$ funciones bases que representan el stock de variables que son factibles de ser seleccionadas para después eliminar las funciones bases que resultan excesivas.

Este tipo de modelo requiere estimar la falta de ajuste sobre un conjunto de datos representativos que no forme parte de la muestra de entrenamiento. El modelo que minimice el criterio de selección stepwise

backward se considera la estimación final de la función. Como MARS es un procedimiento no lineal un criterio basado en muestras reutilizables como validación cruzada (CV) o bootstrapping puede ser justificado, el cual se define como:

$$CV = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}_{M/i}(x_i))^2$$

Donde $\hat{f}_{M/i}$ es la M función base del modelo considerado en el proceso de eliminación forward backward, estimado con la *i*-esima observación removida. Dada la estructura jerárquica del conjunto de modelos considerados en la estrategia stepwise, el criterio anterior puede ser evaluado para todos $0 \leq M \leq M_{\max}$ los modelos con el mismo esfuerzo computacional requerido para la evaluación de uno de ellos.

El criterio de validación cruzada requiere que el modelo completo sea replicado N veces con cada una de las observaciones removidas. Esto es a menudo aproximado por un procedimiento análogo que replica el modelo $F<N$ veces con N/F observaciones diferentes siendo removidas cada vez. Sin embargo Friedman propuso una modificación al criterio de validación cruzada generalizado (GCV) originalmente propuesto por Craven y Wahba (1979), el cual requiere de una sola evaluación del modelo:

$$GCV(M) = \frac{1}{N}\sum_i (y_i - \hat{f}_q(x_i))^2 / \left[1 - \frac{C(M)}{N}\right]^2 \qquad (3)$$

Dónde aquí la dependencia de $\hat{f}_q$ y el criterio sobre el número de funciones bases M es explícitamente indicado. El GCV es el error cuadrado medio del ajuste de los datos (numerador) y el denominador es un termino penalizado que representa el incremento de la varianza asociada con el incremento de la complejidad del modelo (numero de funciones bases M).

Si los valores de los parámetros de las funciones bases asociados con el modelo MARS fueron determinados independientemente de los valores respuestas $(y_1, ..., y_N)$, entonces solo los coeficientes $(a_0, ..., a_M)$ están siendo ajustados para los datos. Por consiguiente la complejidad de la función de costo es:

$$C(M) = traza(B(B'B)^{-1}B') + 1 \qquad (4)$$

Dónde *B* es la matriz de datos $M \times N$ de las M funciones bases. Esto es igual al número de funciones bases linealmente independientes y por lo tanto $C(M)$ es el número de parámetros que esta siendo ajustado. Utilizando las ecuaciones (3) y (4) conducen al criterio GCV propuesto por Craven y Wahba (1979).

Friedman y Silverman (1989) sugieren usar (3) como criterio de validación, pero con un incremento de la función de complejidad de costo

$\dot{C}(M)$ para reflejar los parámetros adicionales que, según los coeficientes de expansión $(a_0,.....,a_M)$ están siendo ajustados para los datos. Tal función de complejidad de costo puede ser expresada como:

$$\dot{C}(M) = C(M) + d \times M$$

La cantidad d representa un costo para cada función base optimizada y es un parámetro del procedimiento y M representa el número de funciones basales.

En el caso de modelos aditivos, Friedman y Silverman (1989) proponen elegir el valor $d = 2$, basado en la esperada disminución del error cuadrado medio por agregar un único nodo para hacer un modelo lineal piecewise.

3.2.1.9 Descomposición ANOVA

El resultado de la aplicación del algoritmo MARS es un modelo de la forma:

$$\hat{f}_q(x) = a_0 + \sum_{m=1}^{M} a_m \prod_{k=1}^{K_m} \left[S_{km}(x_{v(k,m)} - t_{km} \right]_+$$

Donde a_0 es el coeficiente de la función base constante B_1, y la sumatoria comprende todas las funciones bases B_m que permanecieron después de aplicar el procedimiento backward y $S_{k_m} = \pm 1$. Esta representación del modelo no proporciona una visión sobre la naturaleza de

la aproximación, sin embargo ella se puede obtener reestructurando los términos del modelo de tal forma que entregue información acerca de la relación predictiva entre la variable respuesta y las variables predictoras o covariables. De esta manera el modelo puede ser reescrito de la siguiente forma:

$$\hat{f}_q(x) = a_0 + \sum_{K_m=1} f_i(x_i) + \sum_{K_m=2} f_{ij}(x_i, x_j) + \sum_{K_m=3} f_{ijk}(x_i, x_j, x_k) + \ldots\ldots$$

La primera suma es sobre todas las funciones bases que involucran una sola variable, la segunda suma es sobre todas las funciones bases que involucran exactamente dos variables, representado la interacción entre las dos variables. De forma similar, la tercera suma, representa la contribución del efecto de interacción entre tres variables y así sucesivamente.

Sea $V(m) = \{v(k,m)\}_1^{K_m}$ el conjunto de variables asociadas con la m-esima función base B_m. Entonces cada función de la primera suma puede ser expresada como:

$$f_i(x_i) = \sum_{\substack{K_m=1 \\ i \in V(m)}} a_m B_m(x_i)$$

Esta es una suma sobre todas las funciones bases que involucran solo una variable x_i y es una representación spline $q = 1$ de una función

univariante. Cada función bivariante en la segunda suma puede ser expresada como:

$f_{ij}(x_i, x_j) = \sum_{\substack{K=2 \\ (i,j) \in V(m)}} a_m B_m(x_i, x_j)$, la cual es una suma sobre todas las funciones bases que involucran dos variables x_i y x_j . Agregando esto a las correspondientes contribuciones univariantes se tiene:

$$f_{ij}^{*}(x_i, x_j) = f_i(x_i) + f_j(x_j) + f_{ij}(x_i, x_j)$$

Dado $q = 1$, la aproximación del producto tensor spline representando la contribución bivariada conjunta de x_i y x_j para el modelo. Los términos que involucran más variables pueden ser reunidos y representados similarmente.

La interpretación del modelo MARS es facilitada a través de la descomposición de la tabla ANOVA. Esta representación identifica las variables que entran al modelo, si ellas ingresan aditivamente o involucran interacciones con otras variables, el nivel de interacciones y las otras variables que participan.

3.2.2 Análisis Discriminante Basado en Distancias

3.2.2.1 Distancias

Una distancia δ sobre un conjunto (finito o no) Ω es una aplicación que

a cada par de individuos $(\omega_i, \omega_j) \in \Omega \times \Omega$, le hace corresponder un número real $\delta\left(\omega_i, \omega_j\right) = \delta_{ij}$, que cumple con las siguientes propiedades:

1. $\delta_{ij} \geq 0$
2. $\delta_{ii} = 0$
3. $\delta_{ij} = \delta_{ji}$
4. $\delta_{ij} \leq \delta_{ik} + \delta_{kj}$ (desigualdad triangular)

Cuando además, se cumple la propiedad 4 se dice que la distancia es métrica.

Si Ω es un conjunto finito, que indicaremos como $\Omega = \{1, 2, ..., n\}$, las distancias δ_{ij} se expresan mediante la matriz simétrica Δ , llamada matriz de distancias sobre Ω :

$$\Delta = \begin{pmatrix} \delta_{11} & \delta_{12} \cdots & \delta_{1n} \\ \delta_{21} & \delta_{22} \cdots & \delta_{2n} \\ \cdots & \cdots & \cdots \\ \delta_{n1} & \delta_{n2} \cdots & \delta_{nn} \end{pmatrix} \; ; \delta_{ii} = 0, \; \delta_{ij} = \delta_{ji}$$

Se denomina preordenación de Ω asociada a Δ , a la ordenación de menor a mayor de los $m = n \times (n-1)/2$ pares de distancias no nulas:

$$\delta_{i_1 j_1} \leq \delta_{i_2 j_2} \leq ..... \leq \delta_{i_m j_m} ,$$

Es decir, la ordenación de los pares (*i,j*) de Ω , de acuerdo a su proximidad.

Una matriz de distancia Δ puede ser transformada como:

Transformación aditiva: que consiste en sumar una constante fuera de la diagonal de Δ , es decir:

$$\delta_{ij}^{*} = \begin{cases} 0 & , i = j \\ \delta_{ij} + c, & i \neq j \end{cases}$$

Transformación *q*-aditiva: que afecta el cuadrado de la distancia se define como:

$$\overset{\cdot}{\delta}_{ij}^{2} \begin{cases} 0, & i = j \\ \delta_{ij}^{2} + c, & i \neq j \end{cases}$$

Ambas transformaciones son útiles para lograr que la nueva distancia cumpla propiedades que la distancia original no posee, pero conservando la preordenación, es decir, las relaciones de proximidad.

3.2.2.1.1 Distancia para variables cuantitativas

Supongamos ahora que cada individuo de Ω puede ser representado por un punto $x = (x_1, x_2, ..., x_p) \in$ ¡ ^{p}. Algunas distancias especialmente

interesantes entre dos puntos $x, y \in ¡^p$, son:

a. Distancia Euclídea

$$d_E(x,y) = \sqrt{\sum_{h=1}^{p} (x_i - y_i)^2}$$

Esta distancia supone implícitamente que las variables son incorrelacionadas y no es invariante a los cambios de escala.

b. Distancia "ciudad"

$$d_1(x,y) = \sum_{i=1}^{p} |x_i - y_i|$$

c. Distancia "valor absoluto"

$$d_A(x,y) = \sqrt{\sum_{i=1}^{p} |x_i - y_i|}$$

d. Distancia de Mahalanobis

$$d^2_M(x,y) = (x-y)' \sum^{-1} (x-y)$$

Esta distancia (al cuadrado) puede tener las siguientes versiones:

1. Distancia de una observación x_i al vector de medias $\bar{x}$ de x:

$$(x_i - \bar{x})' S^{-1} (x_i - \bar{x})$$

2. Distancia entre dos poblaciones representadas por dos matrices de datos $X_{n_1 \times p}, Y_{n_2 \times p}$:

$$(\bar{x} - \bar{y})' S^{-1} (\bar{x} - \bar{y})$$

Donde $\bar{x}, \bar{y}$ son los vectores de medias y

$$S = (n_1 S_1 + n_2 S_2) / (n_1 + n_2)$$

Es la media ponderada de las correspondientes matrices de covarianzas.

Las distancias d_E y d_P son casos particulares de d_M cuando la matriz de covarianza es la I_p y la diag(S), respectivamente, es decir:

$$d_E(i,j)^2 = (x_i - x_j)'(x_i - x_j)$$

$$d_P(i,j)^2 = (x_i - x_j)' \left[diag(S) \right]^{-1} (x_i - x_j)$$

3.2.2.2 Similaridad

Una similaridad *s* en un conjunto de Ω, es una aplicación que asigna a cada par $(\omega_i, \omega_j) \in \Omega \times \Omega$ un número real $s_{ij} = s(i,j)$, que cumple:

1. $0 \leq s_{ij} \leq s_{ii} = 1$

2. $s_{ij} = s_{ji}$

Cuando Ω es un conjunto finito, entonces la matriz

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

Se denomina matriz de similaridades sobre Ω .

Es inmediato pasar de similaridad a distancia y recíprocamente. Las dos transformaciones básicas son:

$$\delta_{ij} = 1 - s_{ij}$$
$$\delta_{ij} = \sqrt{1 - s_{ij}}$$

En general una matriz de similaridades puede tener en su diagonal elementos $s_{ii} \neq 1$. La transformación que nos permite pasar de similaridad a distancia es:

$$\delta_{ij} = \sqrt{s_{ii} + s_{jj} - 2s_{ij}}$$

3.2.2.2.1 Similaridad con variables binarias

Supongamos que tenemos *p* variables binarias $X_1, X_2, ..., X_p$, donde cada X_i toma los valores 0 ó 1. Para cada par de individuos (i, j), se tienen los siguientes coeficientes de similaridad:

a. Sokal-Michener

$$s_{ij} = \frac{a + d}{p}$$

b. Jaccard

$$s_{ij} = \frac{a}{a + b + c}$$

Siendo a, b, c, d las frecuencias de (1,1), (1,0), (0,1) y (0,0), respectivamente y, $p = a + b + c + d$.

3.2.2.2.2 Similaridad con variables mixtas

Si las variables son mixtas, continuas, binarias o cualitativas, es entonces adecuado utilizar la distancia de Gower $d_{ij}^2 = 1 - s_{ij}$, siendo

$$s_{ij} = \left(\sum_{h=1}^{p_1} (1 - |x_{ih} - x_{jh}| G_h) + a + \alpha \right) / (p_1 + (p_2 - d) + p_3)$$

Dónde:

p_1 = es el número de variables cuantitativas.

a = es el número de coincidencias para las p_2 variables binarias.

d = es el número de no coincidencias para las p_2 variables binarias.

α = es el número de coincidencias para las p_3 variables cuantitativas.

G_h = es el rango de la h-ésima variable cuantitativa.

3.2.2.3 Teorema de Caracterización

Sea $A = (a_{ij})$ la matriz con $a_{ij} = -\delta_{ij}^2 / 2$ y ***B=HAH***.

La matriz de distancias Δ es Euclídea en dimensión *m* si y solo si ***B***≥0 (***B*** es semidefinida positiva) y el rango (***B***)=*m* , *m* ≤ *n*-1.

Si Δ es Euclídea, es posible obtener la siguiente descomposición espectral:

$$B = U\Lambda U' = XX'$$

Donde:

$X = U\Lambda^{1/2}$, contiene los *m* vectores propios de ***B***.

Λ , es una matriz diagonal que contiene los valores propios ordenados $\lambda_1 > .... > \lambda_m > 0$.

B= proporciona las coordenadas Euclídeas del conjunto $\Omega = \{1, 2, ..., n\}$.

x_i de *x*=contiene las coordenadas principales de i.

3.2.2.4 Propiedades de las Coordenadas Principales

a) Las filas $x_1,.....x_n$ de *X* verifican $\delta_{ij}^2 = (x_i - x_j)'(x_i - x_j)$, es decir, sus distancias Euclídeas se igualan a los términos (i,j) en Δ.

b) Las columnas $\mathrm{X}_1,.....\mathrm{X}_m$ de X, entendidas como variables, tienen media 0.

c) Cada columna X_j de X tiene varianza igual a λ_j / n.

d) Las columnas de X son ortogonales (incorrelacionadas).

e) Las columnas de X pueden ser interpretadas como componentes principales.

f) La representación de 1,2,…,n utilizando las filas de X es óptima.

La cantidad

$$\sum \delta_{ij}^2(k) = 2n(\lambda_1 + ...+ \lambda_k)$$

Es máxima en dimensión reducida k.

$\delta_{ij}(k)$ es la distancia utilizando las $k \prec m$ primeras coordenadas y

$\lambda_1 > ..... > \lambda_k$ son los k primeros valores propios de B, ordenados de mayor a menor.

3.2.2.5 Métodos de Particiones

3.2.2.5.1 Partición K-means

K-means es una algoritmo de clasificación que particiona un conjunto de n objetos en c grupos o clusters. El criterio para definir los grupos es que estos deben tener un mínimo de dispersión. La función a ser minimizada es la suma al cuadrado del error total (TESS):

$$TESS_c = \sum_{k=1}^{c} E_{(k)}^2 = \sum_{k=1}^{c} \sum_{i=1}^{n} I\left[\omega_i \subset \Omega_k\right] e_{ik}^2$$

Dónde $E_{(k)}^2$ es la suma al cuadrado del error (ESS) para el grupo Ω_k, y $I\left[\omega_i \subset \Omega_k\right] = 1$ si el objeto ω_i pertenece a Ω_k, y $I\left[\omega_i \subset \Omega_k\right] = 0$ otro caso. e_{ik}^2 es la distancia al cuadrado del objeto ω_i con respecto al centroide:

$$e_{ik}^2 = d^2(\omega_i, \Omega_k) = \sum_{j=1}^{p} (x_{ij} - \bar{x}_{(k)j})^2 = (\mathbf{x}_i - \bar{\mathbf{x}}_{(k)})'(\mathbf{x} - \bar{\mathbf{x}}_{(k)})$$

El número de posibles particiones de objetos dentro del grupo es limitado y uno debería explorar todos ellos y mantener la partición con el menor TESS.

K-means puede ser realizado a partir de una matriz descriptora de objetos (X) o una matriz simétrica de disimilaridad (D).

3.2.2.5.2 Particiones fuzzy sets, fuzzy logic y fuzzy

La teoría de fuzzy set (FST) es una extensión de la teoría clasica *set* desarrollada por Zadeh (1965) como un modo de dar con conceptos vagos, tales como “Jordi es alto”. La teoría set clásica (hard o crisp) considera un objeto como miembro de un conjunto dado (conjunto de la gente alta) o no. Se expresa la pertenencia a un grupo usando una variable indicadora (I), la cual tomara valores 1 si el objeto pertenece al grupo y 0 en otro caso. En set fuzzy, la variable indicadora binaria es ampliada para una variable continua (u) llamada miembro, la cual puede tomar valores intermedios en el intervalo $[0,1]$.

Una partición fuzzy es una partición de objetos dentro del grupo o clases permitiendo miembros intermedios. En una partición fuzzy, cada objeto se divide los miembros de 1 entre los diferentes cluster. La matriz $\mathrm{U}_{(n\times c)} = [u_{ik}]$ es una partición fuzzy no degenerada si satisface:

$$0 \le u_{ik} \le 1 \quad \text{para todo } i = 1,.....,n \text{ y } k = 1,.....,c; \quad (1)$$

$$\sum_{i=1}^{c} u_{ik} = 1 \quad \text{para todo } i = 1,.....,n; \quad (2)$$

$$\sum_{i=1}^{n} u_{ik} \ge 0 \quad \text{para todo } i = 1,.....,c; \quad (3)$$

Particiones – c hard o crip implican las mismas condiciones excepto por (1), la cual es reemplazada por:

$$u_{ij} \in \{0,1\} \text{ para } i = 1,.....,n \text{ y } j = 1,.....,c$$

3.2.2.5.3 Partición fuzzy C-means

Fuzzy C-means es una extensión de K-means usando el concepto de fussy logic. Una de las maneras de introducir fuzzy logic en K-means es:

$$FTESS_{c,m} = \sum_{k=1}^{c} J_{k,m}^2 = \sum_{k=1}^{c} \sum_{i=1}^{n} u_{ik}^m e_{ik}^2$$

Dónde u_{ik} es el miembro fuzzy de un objeto ω_i para el conjunto fuzzy Ω_k, $m \in (1,\infty)$ es un exponente fuzzines el cual determina la incidencia de los valores fuzzy sobre los cálculos. Aquí e_{ik}^2 tiene el mismo significado como en K-means, aunque las coordenadas del centroide son ahora calculadas como:

$$\bar{x}_{kj} = \frac{\sum_{i=1}^{n} u_{ik}^m x_{ij}}{\sum_{i=1}^{n} u_{ik}^m} \qquad (1)$$

A partir de la minimización de la ecuación funcional se obtiene la siguiente expresión para calcular los miembros:

$$u_{ik} = \frac{1}{\sum_{i=1}^{c} \left[\frac{e_{ik}}{e_{il}} \right]^{2/(m-1)}} \qquad (2)$$

El algoritmo Fuzzy C-means trabaja en esencia de la misma forma que K-means. Se debe especificar las condiciones de inicio y el algoritmo minimizara la función proveyendo que los centroides sean calculados como en (1) y los objetos serán reasignados usando la ecuación (2). Sin embargo fuzzy C-means necesita una más compleja regla de finalización. La cual se realiza comparando la partición fuzzy de dos sucesivas iteraciones. Esta finalizara cuando U^t y U^{t-1} difieran menos que un valor especificado.

3.2.2.6 Análisis Discriminante

Supongamos que Ω_1, Ω_2 son dos poblaciones, $X_1, X_2, ..., X_p$ son p variables observables y $\mathbf{x} = (x_1, ..., x_p)$ contiene las observaciones de las variables sobre un individuo ω que se quiere asignar a una de las poblaciones.

Una *regla discriminante* es un criterio que permite asignar ω , y que se plante en términos de una *función discriminante* $D(x_1, ...x_p)$.

Entonces la regla de clasificación es:

Si $D(x_1, ...x_p) \geq 0 \Rightarrow \omega \in \Omega_1$, en otro caso

$D(x_1, ...x_p) < 0 \Rightarrow \omega \in \Omega_2$

Las cuatro reglas más importantes del análisis discriminante son:

1. Regla de Máxima Verosimilitud (MV)

Sea $f_i(x)$ la densidad de **x** en $\Omega_i, i = 1,2.$ La función discriminante es:

$$V(\mathrm{x}) = \log f_1(\mathrm{x}) - \log f_2(\mathrm{x})$$

2. Regla de Bayes (B)

Supongamos que son conocidas las probabilidades a priori:

$$q_1 = P(\Omega_1),\ \ q_2 = P(\Omega_2),\ q_1 + q_2 = 1$$

Entonces la función discriminante es:

$$B(\mathrm{x}) = \log f_1(\mathrm{x}) - \log f_2(\mathrm{x}) + \log(q_1 / q_2)\ \mu$$

3. Regla de Matusita (M)

Dada una distancia $d_i = d(\omega, \Omega_i),\ i = 1,2,$ de ω a cada población. Esta distancia es del tipo $d_i = d(\mathrm{x}, \mu_i),\ \ i = 1,2$, donde ***x*** es el vector de observaciones y μ_i es un vector representante de Ω_i, por ejemplo el vector de medias.

La regla M está basada en la función discriminante

$$M(\mathrm{x}) = d_2^2(\mathrm{x}) - d_1^2(\mathrm{x})$$

4. Regla de Fisher

Si $\Omega_i : (\mu_i, \Sigma)$, es decir se puede identificar Ω_i haciendo uso de un vectorde medias μ_i y una matriz de covarianzas Σ, entoces:

$$L(x)=\left[x-\frac{1}{2}(\mu_1+\mu_2)\right]' \Sigma^{-1}(\mu_1-\mu_2)$$

Las reglas de clasificación son:

1. La regla MV asigna ω a una población Ω_i, tal que la verosimilitud $f_i(x)$ es mayor.

2. La regla de Bayes considera el valor mas grande de la probabilidad "a posteriori " $P(\Omega_i / x)$. MV es un caso particular de *B* si $q_1 = q_2 = 1/2$.

3. La regla M simplemente asigna ω a la población que tiene más próxima.

4. El discriminador lineal de Máxima Verosimilitud, es un caso particular de la regla M cuando $\Omega_i : (\mu_i, \Sigma)$, $i = 1,2$, y $d_j^2 = (x-\mu_j)'\Sigma^{-1}(x-\mu_j)$ es la distancia de Mahalanobis.

Todas las reglas discriminante coinciden en el caso particular de que las poblaciones sean normales multivariantes $N_p(\mu_i, \Sigma_i)$, es decir, la función de densidad en Ω_i es:

$$f_i(\mathrm{x})=\frac{\left|\Sigma_i^{-1}\right|^{1/2}}{(2\pi)^{p/2}}\exp\left\{-\frac{1}{2}(\mathrm{x}-\mu_i)'\Sigma_i^{-1}(\mathrm{x}-\mu_i)\right\}$$

Con $\Sigma_1 = \Sigma_2$.

3.2.2.7 Análisis Discriminante Basado en Distancias

En el análisis discriminante lineal y cuadrático la matriz de datos **(*X*)** es siempre una matriz de datos rectangular. Además, es necesario asumir que la distribución de probabilidad de las variables en cada grupo, se distribuye de manera normal multivariada (con igual o distinta varianza en cada grupo). En cambio, el análisis discriminante Basado en Distancias (DB), creado por el profesor Carles Cuadras de la Universidad de Barcelona (1997), no se requieren supuestos sobre la distribución de los descriptores. Se requiere una matriz simétrica de distancia, permitiendo de esta forma la aplicación de muchos índices de distancias para medir la relación de objetos en ***X*.**

3.2.2.7.1 Fundamentos teóricos del Análisis Discriminante Basado en Distancias

Sea ***X*** un vector aleatorio p -dimensional, definido sobre un espacio de probabilidad $(\Pi, \Lambda, \mathrm{P})$ y que toma valores en $S \subset R^p$ con función de densidad f para una adecuada medida de λ . Sea $d(.,.)$ una función de disimilaridad definida sobre un par de elementos en Π , de modo que al cuadrado es integrable en S. Entonces, la variabilidad geométrica de ***X*** con respecto a $d(.,.)$ es definida como:

$$V_d(\mathrm{X}) = \frac{1}{2}\int_{S\times S} d^2(x_1, x_2) f(x_1) f(x_2) \lambda(dx_1) \lambda(dx_2)$$

$V_\delta(\mathrm{X})$, es el valor esperado de todas las ínter distancias (al cuadrado),

es decir, como una varianza generalizada.

Dada $\boldsymbol{x_0} \in R^p$, se define la proximidad de $\boldsymbol{x_0}$ a la población Π con respecto a $d(.,.)$ como:

$$\phi_d^2(\mathrm{x}_0,\Pi) = \int_S d^2(x_0,x) f(x) \lambda\,(d(x) - V_d(x)$$

Una representación de $d(.,.)$ existe si hay una función $\Psi : R^p \to L$ (donde $L,\langle .,. \rangle$ simboliza un espacio de Hilbert o Euclidiano llenado con un producto escalar $\langle .,. \rangle$, tal como $\|\mathrm{u}\|^2 = \langle \mathrm{u},\mathrm{u} \rangle$ es la norma natural para todo $\mathrm{u} \in L$), y asumiendo que $E(\Psi(\mathrm{X}))$ y $E(\|\Psi(\mathrm{X})\|^2)$ son finitas, entonces la variabilidad geométrica de $\boldsymbol{X}$ y la proximidad de $\boldsymbol{x_0}$ a la población Π son:

$$V_d(\mathrm{X}) = \mathrm{E}(\|\Psi(\mathrm{X}) - \mathrm{E}(\Psi(\mathrm{X}))\|^2)$$

$$\phi_d^2(\mathrm{x}_0) = \|\Psi((\mathrm{x}_0) - E(\Psi(\mathrm{X}))\|^2$$

3.2.2.7.2 La Regla Discriminante

El enfoque de análisis discriminante basado en distancias es calcular la distancia desde un objeto a un grupo a partir de la matriz $D_{(n \times n)} = [d(i,j)]$ de ínter distancias de objetos. Este enfoque es posible para particiones crisp y

fuzzy.

Sea $x(k)_1,.....,x(k)_{n_k}$, y sea n_k el vector de objetos pertenecientes al grupo Ω_k . La distancia al cuadrado de un objeto ω_i al grupo Ω_k es:

$$d^2(x_i,\Omega_k) = \frac{1}{n}\sum_{h=1}^{n_k} d^2(x_i,x(k)_h) - \hat{V}_d(k) \quad (1)$$

Dónde $\hat{V}_d(k)$ es la variabilidad geométrica de:

$$\hat{V}_d(k) = \frac{1}{2n_k^2}\sum_{h,l}^{n_k} d^2(x(k)_h,x(k)_l) \quad (2)$$

Las ecuaciones (1) y (2) se calculan para cada objeto del vector $x_i(i = 1,2,.....,n)$ y cada grupo $\Omega_k\ (k = 1,2,.....,c)$ hasta obtener la matriz cuadrada de distancias desde un objeto a todos los grupos centros.

La regla de clasificación DB es:

Asignar ω_i a al grupo Ω_k si

$$d^2(x_i,\Omega_k) = \min\left\{ d^2(x_i,\Omega_1), d^2(x_i,\Omega_2),.....,d^2(x_i,\Omega_n)\right\}$$

Las ecuaciones (1) y (2) pueden ser generalizada para el caso de particiones fuzzy.

Dada una matriz de partición fuzzy $U_{(n\times c)}$ y m exponentes fuzzines, la

distancia desde un objeto al grupo fuzzy Ω_k y la variabilidad geométrica al grupo fuzzy son:

$$d^2(\mathrm{x}_i,\Omega_k) = \frac{1}{\sum_{h=1}^{n} u_{hk}^m} \sum_{h=1}^{n} u_{hk}^m d^2(\mathrm{x}_i,\mathrm{x}_h) - \hat{V}_{fd}(\Omega_k)$$

$$\hat{V}_{fd}(\Omega_k) = \frac{1}{2(\sum_{i=0}^{n} u_{ik}^m)^2} \sum_{h,l}^{n} u_{hk}^m \times u_{lk}^m d^2(\mathrm{x}_h,\mathrm{x}_l)$$

Las ecuaciones (1) y (2) pueden ser generalizadas en el caso de tener partición fuzzy.

4 RESULTADOS

4.1 Análisis Descriptivo

El factor vivienda esta formado por los subfactores de protección ambiental, hacinamiento y saneamiento y confort. El subfactor de protección ambiental esta formado por las variables tipo de material del muro, piso y techo de la vivienda de los jefes de familia.

En la tabla 4.1 se muestra el tipo de material del muro de la vivienda, el cual nos indica que casi la mitad (48,1 %) de las viviendas de los jefes de familias no beneficiarios del Programa Chile Solidario poseen tipo de material del muro de ladrillo. En tanto un 43,7 % de los jefes beneficiarios del Programa Chile Solidario el material del muro predominante es de desecho. Por otra parte se visualiza que la categoría tabique forrado y adobe no presenta diferencias entre los jefes de familia no beneficiarios y beneficiarios del Programa Chile Solidario (39,5 % v/s 36,7%).

Tabla 4.1 Tipo de Material del Muro de la Vivienda

Material muro	No beneficiario		Beneficiario		Total	
	N	%	N	%	N	%
Ladrillo, concreto o bloque	385	48.1	9	4.5	394	39.4
Albañilería de piedra	4	0.5	1	0.5	5	0.5
Tabique forrado	316	39.5	73	36.7	389	38.9
Adobe	44	5.5	13	6.5	57	5.7
Barro, quincha o pirca	0	0.0	2	1.0	2	0.2
Tabique sin forro interior	2	0.2	11	5.5	13	1.3
Desecho	50	6.2	87	43.7	137	13.7
No responde	0	0.0	3	1.5	3	0.3
Total	801	100.0	199	100.0	1000	100.0

En relación a la variable material del piso de la vivienda se puede señalar que casi un 60 % de los jefes de familia no beneficiarias poseen viviendas cuyo material del piso es de radier revestido. Sin embargo, para los beneficiarios la distribución porcentual es de solo un 5 %. Por otra parte, un 35,2 % de las viviendas de los jefes de familia beneficiarios poseen material precario es decir, madera, plástico o pastelón sobre tierra y piso de tierra cantidad inferior para los jefes de familia no beneficiarios (5,1 %), tal como se señala en la tabla 4.2.

Tabla 4.2 Tipo de Material del Piso de la Vivienda

Material piso	No beneficiario		Beneficiario		Total	
	N	%	N	%	N	%
Radier revestido	458	57.2	10	5.0	468	46.8
Radier no revestido	118	14.7	43	21.6	161	16.1
Madera sobre solera	184	23.0	76	38.2	260	26.0
Madera, plastico o pastelon sobre tierra	34	4.2	28	14.1	62	6.2
Piso de tierra	7	0.9	42	21.1	49	4.9
Total	801	100.0	199	100.0	1000	100.0

El tipo de material del techo en las viviendas de los jefes de familia presenta diferencias en cada una de sus categorías tal como se visualiza en la tabla 4.3. Para los jefes de familia no beneficiarios del Programa Chile Solidario la categoría de techo de zinc, pizarreño con cielo interior alcanza un 80,9 % de las viviendas. Mientras los jefes de familias beneficiario muestran 37,2 % de las viviendas con este tipo de material. Por otra parte aproximadamente un 6,5 % de las familias beneficiarias presenta condiciones precarias (fonolita, paja, coirón, totora, caña o desecho) respecto al tipo de

material del techo. En tanto que para la situación de las viviendas de los jefes de familias no beneficiarias no se presentan condiciones precarias.

Tabla 4.3 Tipo de Material del Techo de la Vivienda

Material del techo	No beneficiario		Beneficiario		Total	
	N	%	N	%	N	%
Teja, tejuela o losa	80	10.0	2	1.0	82	8.2
Zinc, pizarreño con cielo interior	648	80.9	74	37.2	722	72.2
Zinc, pizarreño sin cielo interior	73	9.1	110	55.3	183	18.3
Fonolita	0	0.0	6	3.0	6	0.6
Paja, coiron, totora o caña	0	0.0	3	1.5	3	0.3
Desecho	0	0.0	4	2.0	4	0.4
Total	801	100.0	199	100.0	1000	100.0

El subfactor de hacinamiento esta formado por la variable hacinamiento medido este como la razón entre el número de dormitorios de una vivienda y el número de personas que habitan en dicha vivienda. En la tabla 4.4 se observa que no existen diferencias entre los dos grupos, siendo de un 3,6 % para las familias no beneficiarias y de un 3,5 % para las familias beneficiarias.

Tabla 4.4 Hacinamiento de las Familias

Hacinamiento	No beneficiario		Beneficiario		Total	
	N	%	N	%	N	%
Con Hacinamiento	29	3.6	7	3.5	36	3.6
Sin Hacinamiento	772	96.4	192	96.5	964	96.4
Total	801	100.0	199	100.0	1000	100.0

El subfactor confort y saneamiento esta formado por las variables tipo del abastecimiento de agua, sistema de eliminación de excretas y disponibilidad de tina o ducha.

En la tabla 4.5 se muestra la distribución porcentual para la variable tipo de abastecimiento de agua.

Tabla 4.5 Tipo de Abastecimiento de Agua de las Viviendas

Abastecimiento de agua	No beneficiario		Beneficiario		Total	
	N	%	N	%	N	%
Red Pública con llave dentro de la vivenda	687	85.8	40	20.1	727	72.7
Red Pública con llave dentro del sitio	49	6.1	61	30.7	110	11.0
Red Pública con llave fuera del sitio	3	0.4	17	8.5	20	2.0
Sin Red Pública con llave dentro de la vivenda	39	4.9	13	6.5	52	5.2
Sin Red Pública con llave dentro del sitio	6	0.7	19	9.5	25	2.5
Sin Red Pública, acarreo	17	2.1	49	24.6	66	6.6
Total	801	100.0	199	100.0	1000	100.0

Se aprecia que un 85,8 % de las viviendas de los jefes de familias no beneficiarios el abastecimiento de agua es de la red pública con llave dentro de la vivienda en cambio, esta proporción es solo de un 20,1 % para las viviendas de los beneficiarios. Por otra parte, solo un 2,1 % de las viviendas de los jefes de familias no beneficiario acarrea el agua, esta cantidad se eleva a un 24,6 % para las viviendas de los beneficiarios.

En la tabla 4.6 se muestra que un 63,4 % de las viviendas de los jefes de familia no beneficiario el sistema de eliminación de excretas es a través de alcantarillado. Sin embargo, para las viviendas de los jefes de familias beneficiarios la cifra es de solo un 11, 1%. Por otra parte, casi un tercio de las viviendas de los jefes de familia beneficiarios no cuenta con un sistema

de eliminación de excretas, cifra significativamente inferior para el caso de las viviendas de los no beneficiarios.

Tabla 4.6 Sistema de Eliminación de Excretas de las Viviendas

Sistema de eliminación de excreta	No beneficiario		Beneficiario		Total	
	N	%	N	%	N	%
Uso exclusivo alcantarillado	508	63.4	22	11.1	530	53.0
Uso exclusivo fosa séptica	155	19.4	9	4.5	164	16.4
Uso compartido letrina sanitaria	44	5.5	26	13.1	70	7.0
Uso compartido pozo negro	78	9.7	80	40.2	158	15.8
No tiene sistema	16	2.0	62	31.2	78	7.8
Total	801	100.0	199	100.0	1000	100.0

En relación a la disponibilidad de tina o ducha se puede visualizar diferencias significativas entre las diversas categorías. De este modo se puede observar en la tabla 4.7 que aproximadamente un 60 % de las viviendas de los jefes de familia no beneficiario posee tina o ducha con agua caliente y de uso exclusivo en tanto, un 46,2 % de las viviendas de los jefes de familias beneficiario no posee tina o ducha.

Tabla 4.7 Disponibilidad de Tina o Ducha de las Viviendas

Disponibilidad de tina o ducha	No beneficiario		Beneficiario		Total	
	N	%	N	%	N	%
Tina o ducha con uso exclusivo agua caliente	472	58.9	0	0.0	472	47.2
Tina o ducha uso exclusivo agua fria	167	20.8	25	12.6	192	19.2
Tina o ducha compartida agua caliente	7	0.9	1	0.5	8	0.8
Tina o ducha compartida agua fria	93	11.6	92	46.2	185	18.5
No tiene tina o ducha	62	7.7	81	40.7	143	14.3
Total	801	100.0	199	100.0	1000	100.0

El factor educación esta formado por la variable años de escolaridad promedio. La escolaridad promedio de los jefes de familia no beneficiarios del Programa Chile Solidario es más de dos veces superior a los beneficiarios, es decir para los jefes de familia no beneficiarios sus años de escolaridad promedio son de 8,9 años en tanto, para los beneficiarios esta alcanza en promedio solo 4,2 años.

De otro modo el factor ocupación lo constituye la variable categoría ocupacional del jefe de familia. Los jefes de familia no beneficiarios se declaran en su gran mayoría como trabajadores dependientes urbanos o como jubilados sin embargo, los beneficiarios en su gran mayoría se declaran sin actividad. Ver tabla 4.8.

Tabla 4.8 Categoría Ocupacional de los Jefes de Familia

Categoría ocupacional	No beneficiario		Beneficiario		Total	
	N	%	N	%	N	%
Familiar no remunerado	5	0.6	0	0.0	5	0.5
Trabajador por cuenta propia	80	10.0	20	10.1	100	10.0
Trabajador dependiente urbano	346	43.2	33	16.6	379	37.9
Pequeño productor agrícola	23	2.9	9	4.5	32	3.2
Empleado sector público o equivalente	29	3.6	0	0.0	29	2.9
Jubilado o dependiente de terceros	125	15.6	24	12.1	149	14.9
Actividad major remunerada	77	9.6	0	0.0	77	7.7
No tiene actividad	116	14.5	113	56.8	229	22.9
Total	801	100.0	199	100.0	1000	100.0

El factor de ingreso/patrimonio esta formado por los subfactores de ingreso, sitio y equipamiento. El subfactor de ingresos esta formado por la variable ingreso familiar per capita. El ingreso familiar per cápita de los jefes de familia no beneficiarios es casi cinco veces mayor que el ingreso de los beneficiarios ($ 109.063 v/s $ 23.311)

El subfactor sitio esta formado por la variable propiedad del sitio. En la tabla 4.9 se puede visualizar que un 55,7 % de los jefes de familia no beneficiarios posee sitio propio sin deuda en tanto, un 52,3 % de los jefes de familia beneficiarios hacen uso del sitio pero creen que no serán desalojados.

Tabla 4.9 Tipo de Propiedad del Sitio

Sitio	No beneficiario		Beneficiario		Total	
	N	%	N	%	N	%
Sitio propio sin deuda	446	55.7	70	35.2	516	51.6
Sitio propio sin deudas atrasadas	67	8.4	2	1.0	69	6.9
Sitio propio con deuda	14	1.7	2	1.0	16	1.6
Arrienda el sitio	72	9.0	5	2.5	77	7.7
Usan el sitio pero no creen que serán desalojados	189	23.6	104	52.3	293	29.3
Usan el sitio pero creen que serán desalojados	13	1.6	16	8.0	29	2.9
Total	801	100.0	199	100.0	1000	100.0

Finalmente, las variables tenencia de refrigerador y calefont forman el subfactor de equipamiento. En relación a la tenencia de refrigerador solo un 25,2 % de los jefes de familias no beneficiarios no posee refrigerador sin embargo, la situación es muy diferente para el caso de los beneficiarios dónde 79,9 % no posee refrigerador tal como se señala en la tabla 4.10.

Tabla 4.10 Tenencia de Refrigerador

Refrigerador	No beneficiario		Beneficiario		Total	
	N	%	N	%	N	%
No tiene	202	25.2	159	79.9	361	36.1
Si tiene	599	74.8	40	20.1	639	63.9
Total	801	100.0	199	100.0	1000	100.0

La tenencia de calefont para el caso de los no beneficiarios es de un 62,4 % en tanto solo un 1,5 % de los beneficiarios declara poseer calefont tal como se muestra en la tabla 4.11.

Tabla 4.11 Tenencia de Calefont

Califont	No beneficiario		Beneficiario		Total	
	N	%	N	%	N	%
No tiene	301	37.6	196	98.5	497	49.7
Si tiene	500	62.4	3	1.5	503	50.3
Total	801	100.0	199	100.0	1000	100.0

4.2 Aplicación de MARS

Para la determinación de los modelos se uso la base de datos de entrenamiento. Se ajustaron 4 modelos MARS, estos modelos tienen como respuesta la variable de tipo binaria definida con el valor 1 si es beneficiario y 0 en caso contrario, es decir para los no beneficiarios. Las variables predictoras fueron señaladas anteriormente. La idea central es modelar:

$$P(y = 1 \mid x_1, ..., x_p)$$

Los modelos ajustados son los siguientes:

MODELO I: 15 funciones bases, ninguna observación entre nodos y sin interacción.

MODELO II: 15 funciones bases, ninguna observación entre nodos y con dos interacciones.

MODELO III: 15 funciones bases, ninguna observación entre nodos y con tres interacciones.

MODELO IV: 15 funciones bases, ninguna observación entre nodos y con cuatro interacciones.

El modelo I es denominado modelo aditivo y los restantes modelos (modelos II, III y IV) se denominan modelos con interacciones o multiplicativos.

En la tabla 4.12 se presentan el número de funciones bases determinadas por MARS para cada uno de los modelos analizados, el número de interacciones en cada uno de los modelos generados, las variables que involucran cada una de las funciones bases y el modelo ajustado. Finalmente, en las últimas dos columnas se presentan dos indicadores de bondad de ajuste del modelo. Estos son los valores para el R^2 ajustado y GCV para cada uno de los modelos generados.

Tabla 4.12 Descripción de los Modelos

MODELO I

Número funciones basales	Número Interacción	Funciones bases	R^2 Ajustado	GCV
8	0	$f_1(abastecimiento\ agua) = (abastecimiento\ agua = 1)$ o $(abastecimiento\ de\ agua = 4)$ $f_4(IFPC) = \max(0, 50.000 - IFPC)$ $f_5(material\ muro) = (material\ muro = 5)$ o $(material\ muro = 6)$ o $(material\ muro = 7)$ o $(material\ muro = 8)$ $f_8(escolaridad\ jefe) = \max(0, 7 - escolaridad\ jefe)$ $f_9(categoria\ ocupacional) = (categoria\ ocupacional = 10)$ $f_{11}(abastecimiento\ agua) = (abastecimiento\ agua = 3)$ o $(abastecimiento\ agua = 4)$ o $(abastecimiento\ agua = 5)$ o $(abastecimiento\ agua = 6)$ $f_{13}(material\ techo) = (material\ techo = 1)$ o $(material\ techo = 2)$ $f_{15}(e\lim inacion\ excretas) = (e\lim inacion\ excretas = 9)$ Y = 0,276 - 0,219 * $f_1(abastecimiento\ agua)$ + 0,0000042* $f_4(IFPC)$ + 0,144 * $f_5(material\ muro)$ + 0,033 * $f_8(escolaridad\ jefe)$ + 0,187 * $f_9(categoria\ ocupacional)$ + 0,140 * $f_{11}(abastecimiento\ agua)$ - 0,144 * $f_{13}(material\ techo)$ + 0,181 * $f_{15}(e\lim inacion\ excretas)$	0,624	0,062

MODELO II

Número funciones basales	Número Interacción	Funciones bases	R2 Ajustado	GCV
8	2	$f_1(abastecimiento\ agua) = (abastecimiento\ agua = 1)$ o $(abastecimiento\ agua = 4)$ $f_2(abastecimiento\ agua) = (abastecimiento\ agua = 2)$ o $(abastecimiento\ agua = 3)$ o $(abastecimiento\ agua = 5)$ o $(abastecimiento\ agua = 6)$ $f_4(IFPC) = \max(0, 50.000 - IFPC)$ $f_5(categoria\ ocupacional) = (categoria\ ocupacional = 0)$ o $(categoria\ ocupacional = 4)$ o $(categoria\ ocupacional$ o $(categoria\ ocupacional = 9) * f_2(abastecimiento\ agua)$ $f_7(disponibilidad\ tina) = (disponibilidad\ tina = 1) *$ $f_4(IFPC)$ $f_9(material\ techo) = (material\ techo = 3)$ o $(material\ techo = 4)$ o $(material\ techo = 5)$ o $(material\ techo = 6)$ $f_{12}(año\ escolaridad) = \max(0, 7 - año\ escolaridad)$ $f_{13}(refrigerador) = (refrigerador = 0) * f_{12}(año\ escolari$ $f_{15}(sitio) = (sitio = 1)$ o $(sitio = 2) * f_{12}(año\ escolaridad)$ Y = 0,372 - 0,395 * $f_1(abastecimiento\ agua)$ + 0,000008 * $f_4(IFPC)$ - 0,341 * $f_5(categoria\ ocupacional)$ - 0,000008 * $f_7(disponibilidad\ tina)$ + 0,190 * $f_9(material\ techo)$ + 0,031 * $f_{12}(año\ escolaridad)$ + 0,046 * $f_{13}(refrigerador)$ - 0,033 * $f_{15}(sitio)$	0,672	0,055

MODELO III

Número funciones basales	Número Interaccion	Funciones bases	R2 Ajustado	GCV
8	3	$f_1(abastecimiento\ agua) = (abastecimiento\ agua = 1)$ o $(abastecimiento\ agua = 4)$ $f_2(abastecimiento\ agua) = (abastecimiento\ agua = 2)$ o $(abastecimiento\ agua = 3)$ o $(abastecimiento\ agua = 5)$ o $(abastecimiento\ agua = 6)$ $f_4(IFPC) = \max(0, 50.000 - IFPC)$ $f_5(categoria\ ocupacional) = (categoria\ ocupacional = 0$ o $(categoria\ ocupacional = 4)$ o $(categoria\ ocupacional = 7)$ o $(categoria\ ocupacional =$ $* f_2(abastecimiento\ agua)$ $f_7(disponibilidad\ tina) = (disponibilidad\ tina = 1) *$ $f_4(IFPC)$ $f_8(disponibilidad\ tina) = (disponibilidad\ tina = 2)$ o $(disponibilidad\ tina = 3)$ o $(disponibilidad\ tina = 4)$ o $(disponibilidad\ tina = 5) * f_4(IFPC)$ $f_9(material\ techo) = (material\ techo = 3)$ o $(material\ techo = 4)$ o $(material\ techo = 5)$ o o $(material\ techo = 6)$ $f_{11}(año\ escolaridad) = \max(0 - año\ escolaridad - 0,00000$ $* f_8(disponibilidad\ tina)$ $f_{12}(refrigerador) = (refrigerador = 0) *$ $f_8(disponibilidad\ tina)$ $f_{14}(sitio) = (sitio = 1)$ o $(sitio = 2)$ o $(sitio = 3) * f_8(disponibildad\ tina)$ $Y = 0,454 - 0,451 * f_1(abastecimiento\ agua) + 0,000014 * f_4(IFPC) - 0,373 *$ $f_5(categoria\ ocupacional) - 0,000014 *$ $f_7(disponibilidad\ tina) + 0,190 * f_9(material\ techo)$ $- 0,00000104 * f_{11}(año\ escolaridad) + 0,000006 *$ $f_{12}(refrigerador) - 0,000005 * f_{14}(sitio)$	0,674	0,055

MODELO IV

Número funciones basales	Número Interacción	Funciones bases	R2 Ajustado	GCV
8	4	$f_1(abastecimiento\ agua) = (abastecimiento\ agua = 1)$ o $(abastecimiento\ agua = 4)$ $f_2(abastecimiento\ agua) = (abastecimiento\ agua = 2)$ o $(abastecimiento\ agua = 3)$ o $(abastecimiento\ agua = 5)$ o $(abastecimiento\ agua = 6)$ $f_4(IFPC) = \max(0, 50.000 - IFPC)$ $f_5(categoria\ ocupacional) = (categoria\ ocupacional = 0$ $(categoria\ ocupacional = 4)$ o $(categoria\ ocupacional =$ o $(categoria\ ocupacional = 9) * f_2(abastecimiento\ agua)$ $f_7(disponibilidad\ tina) = (disponibilidad\ tina = 1) *$ $f_4(IFPC)$ $f_8(disponibilidad\ tina) = (disponibilidad\ tina = 2)$ o $(disponibilidad\ tina = 3)$ o $(disponibilidad\ tina = 4)$ o $(disponibilidad\ tina = 5) * f_4(IFPC)$ $f_9(material\ techo) = (material\ techo = 3)$ o $(material\ techo = 4)$ o $(material\ techo = 5)$ o $(material\ techo = 6)$ $f_{11}(año\ escolaridad) = \max(0, año\ escolaridad - 0,0000$ $* f_8(disponibilidad\ tina)$ $f_{12}(refrigerador) = (refrigerador = 0) *$ $f_8(disponibilidad\ tina)$ $f_{13}(refrigerador) = (refrigerador = 1) *$ $f_8(disponibilidad$ tina) $f_{14}(categoria\ ocupacional) = (categoria\ ocupacional = 1$ $f_{13}(refrigerador)$ Y = 0,427 - 0,421 * $f_1(abastecimiento\ agua)$ + 0,000006 * $f_4(IFPC)$ - 0,316 * $f_5(categoria\ ocupacional)$ -0 ,0000065 * $f_7(disponibilidad\ tina)$ + 0,205 * $f_9(material\ techo)$ - 0,00000096* $f_{11}(año\ escolaridad)$ + ,0000104 * $f_{12}(refrigerador)$ + 0,000009 * $f_{14}(categoria\ ocupacional)$	0,674	0,055

El menor GCV lo presentan los modelos II, III y IV, en tanto el mayor R^2 ajustado lo presentan los modelos III y IV sin embargo la diferencia de estos modelos es mínima comparada con el modelo II (0,674 v/s 0,72).

En relación a los modelos se puede señalar que en todos ellos se muestran funciones bases plausibles sin embargo, para el modelo I, las categorías identificados en la función base 11, que se refiere a la variable tipo de abastecimiento de agua no logra diferenciar condiciones optimas de aquellas mas precarias. De igual forma sucede en los modelos III y IV en relación a la función base 8, que selecciona categorías de la variable disponibilidad de tina o ducha, que no diferencian las condiciones óptimas de precarias. Los problemas señalados anteriormente no se ven reflejados en el modelo II.

En la tabla 4.13 se presenta una jeraquización de las variables para cada uno de los modelos. Podemos visualizar que los modelos II, III y IV consideran el mismo orden de importancia para las variables de ingreso familiar per capita y tipo de abastecimiento de agua (primer y segundo lugar de importancia respectivamente) la diferencia radica en el orden de importancia para la variable escolaridad la cual es considerada en tercer lugar para el modelo II, en el sexto lugar para el modelo III y en el séptimo lugar para el modelo IV. Por otra parte el modelo II y III considera las variables de tenencia de refrigerador y sitio que no están consideradas en el modelo I en tanto, el modelo IV considera solo el refrigerador.

Por otra parte se debe señalar que el modelo I presenta las variables de eliminación de excretas y tipo de material en menor orden de importancia que las restantes variables consideradas, las cuales no son consideradas en los modelos II, III y IV.

Tabla 4.13 Importancia de las Variables

Modelo I

VARIABLE	COSTO DE OMISIÓN	IMPORTANCIA	
ABASTECIMIENTO AGUA	0.071	100.000	
ESCOLARIDAD	0.067	73.116	
CATEGORIA OCUPACIONAL	0.066	69.544	
INGRESO FAMILIAR	0.066	61.939	
MATERIAL TECHO	0.064	42.587	
ELIMINACIÓN EXCRETA	0.064	41.123	
MATERIAL MURO	0.063	39.284	

Modelo II

VARIABLE	COSTO DE OMISIÓN	IMPORTANCIA	
INGRESO FAMILIAR	0.067	100.000	
ABASTECIMIENTO AGUA	0.067	99.187	
ESCOLARIDAD	0.063	82.566	
CATEGORIA OCUPACIONAL	0.060	66.816	
TINA O DUCHA	0.059	58.680	
MATERIAL TECHO	0.059	55.566	
REFRIGERADOR	0.058	49.219	
SITIO	0.056	34.238	

Modelo III

VARIABLE	COSTO DE OMISIÓN	IMPORTANCIA	
INGRESO FAMILIAR	0.078	100.000	
ABASTECIMIENTO AGUA	0.072	85.599	
TINA O DUCHA	0.069	77.054	
CATEGORIA OCUPACIONAL	0.062	53.932	
ESCOLARIDAD JEFE	0.060	47.287	
MATERIAL TECHO	0.059	40.254	
REFRIGERADOR	0.057	30.494	
SITIO	0.057	28.973	

Modelo IV

VARIABLE	COSTO DE OMISIÓN	IMPORTANCIA	
INGRESO FAMILIAR	0.078	100.000	
ABASTECIMIENTO AGUA	0.069	77.287	
TINA O DUCHA	0.069	77.117	
CATEGORIA OCUPACIONAL	0.063	59.567	
REFRIGERADOR	0.059	43.993	
MATERIAL TECHO	0.059	43.697	
ESCOLARIDAD JEFE	0.059	43.409	

Las variables consideradas en los diferentes modelos corresponden a aquellas que han sido definidas por diversos estudios como predictoras de la pobreza.

Para seleccionar el mejor modelo debe evaluarse la bondad de este considerando el mayor valor del R^2 ajustado y el menor GCV y la plausibilidad de las funciones bases. También es importante evaluar las predicciones de los modelos en cuanto a la sensibilidad y a la especificidad.

4.3 Evaluación de la bondad del ajuste de los modelos

Cada uno de modelos ajustados señalados anteriormente, fue utilizado para predecir la probabilidad de ser clasificado como beneficiario o no beneficiario del Programa Chile Solidario, utilizando para ello la muestra de validación.

La tabla 4.14 muestra el porcentaje de observaciones cuyos valores fueron correctamente predichos por cada modelo.

Tabla 4.14 Porcentaje de Observaciones Clasificadas Correctamente en los modelos

Modelo	Porcentaje de correctas
I	82,2
II	90,1
III	89,9
IV	84,3

Se observa que el modelo II predice correctamente un 90,1 % del total jefes de familias, en tanto el modelo I solo predice correctamente un 82,3 % del total de jefes de familias.

Para efectos analíticos se consideraran los resultados del análisis MARS y del Análisis Discriminante Basado en Distancias como *Predicciones* y a la clasificación realizada por MIDEPLAN la denominaremos *Actual Clasificación.* Este tipo de análisis permite clasificar a los jefes de familias en cuatro grupos:

1. El primer grupo esta formado por jefes de familias que fueron seleccionados como no beneficiarios y además fueron predichos como no beneficiarios. Es decir son los jefes de familias valorados como 0 en la actual clasificación y predichos como 0 por el modelo.

2. El segundo grupo esta formado por jefes de familias que fueron seleccionados como no beneficiarios, es decir valorados como 0 y fueron predichos como beneficiarios por el modelo, asignándose el valor 1.

3. El tercer grupo lo integran los jefes de familias clasificados como beneficiarios por la actual clasificación pero predichos como 0 por el modelo, es decir no beneficiarios.

4. Finalmente, se considera a los jefes de familias clasificados como beneficiarios por la actual clasificación y predichos como beneficiarios. Este grupo lo constituye aquellos jefes de familias a las cuales se le asigno el valor 1 tato en la actual clasificación como en las predicciones del modelo.

Las tablas números 4.15, 4.16, 4.17 y 4.18 muestran el número de jefes de familias según la actual clasificación y las predicciones. Por otra parte, las tablas 4.19, 4.20, 4.21 y 4.22 muestran la precisión de los modelos en la clasificación de los jefes de familias beneficiarios o no beneficiarios del Programa Chile Solidario.

Tabla 4.15 Modelo I, Número de Jefes de Familias según Clasificación

Actual Clasificación	Predicho 0	Predicho 1	Total Casos
0	625	176	801
1	2	197	199

Tabla 4.16 Modelo I, Porcentaje de Jefes de Familias según Clasificación

Actual Clasificación	Predicho 0	Predicho 1	Total Casos
0	78,03	21,97	100
1	1,00	99,00	100

Tabla 4.17 Modelo II, Número de Jefes de Familias según Clasificación

Actual Clasificación	Predicho 0	Predicho 1	Total Casos
0	711	90	801
1	9	190	199

Tabla 4.18 Modelo II, Porcentaje de Jefes de Familias según Clasificación

Actual Clasificación	Predicho 0	Predicho 1	Total Casos
0	88,76	11,23	100
1	4,52	95,48	100

Tabla 4.19 Modelo III, Número de Jefes de Familias según Clasificación

Actual Clasificación	Predicho 0	Predicho 1	Total Casos
0	710	91	801
1	10	189	199

Tabla 4.20 Modelo III, Porcentaje de Jefes de Familias según Clasificación

Actual Clasificación	Predicho 0	Predicho 1	Total Casos
0	88,64	11,36	100
1	5,03	94,97	100

Tabla 4.21 Modelo IV, Número de Jefes de Familias según Clasificación

Actual Clasificación	Predicho 0	Predicho 1	Total Casos
0	649	152	801
1	5	194	199

Tabla 4.22 Modelo IV, Porcentaje de Jefes de Familias según Clasificación

Actual Clasificación	Predicho 0	Predicho 1	Total Casos
0	81,02	18,98	100
1	2,51	97,49	100

En la tabla 4.23 se presenta la sensibilidad y especificidad de cada uno de los modelos. La sensibilidad se define como el porcentaje de eventos, en este caso ser beneficiario del Programa Chile Solidario, que fueron predichos para ser eventos. En tanto la especificidad es definida como el

porcentaje de no eventos, en este caso no ser beneficiario del Programa Chile Solidario, que fueron predichos para no ser eventos.

Tabla 4.23 Sensibilidad y Especificidad para los Modelos

Modelo	Sensibilidad	Especificidad
I	98,99	78,03
II	95,48	88,76
III	94,97	88,64
IV	97,49	81,02

Los resultados indican el modelo I presenta la mayor sensibilidad, es decir fue mas exacto en predecir el evento de ser beneficiario del Programa Chile Solidario. Por otra parte el modelo II presenta mayor especificidad en comparación con los demás modelos, es decir mas exacto en predecir el no evento (no ser beneficiario del Programa Chile Solidario).

4.4 Aplicación de Análisis Discriminante Basado en Distancias

El Análisis Discriminante Basado en Distancias a diferencia de MARS no genera un modelo matemático explicito, por tanto la aplicación se realizo utilizando la muestra denominada de entrenamiento la misma muestra que se utilizo para generar los modelos MARS.

En este análisis se utilizo el índice de similaridad de Gower y se realizo la transformación $D = \sqrt{1 - S}$ para construir una matriz de distancia necesaria para realizar el análisis de partición Fuzzy C-means y el análisis discriminante.

Los resultados de este análisis se presentan a continuación.

	Frequencia	**ESS**	Minima **Distancia**	Maxima **Distancia**	**Distancia**	**Distancia Cercana**
g-1	435.641	111.122	0.000	0.572	2.00	0.408
g-2	564.359	88.038	0.003	0.483	1.00	0.405

Coeficiente Dunn: 0,85681
Dunn Normalizado: 0,71361
Entropía Partición: 0,34775
Eficiencia Partición: 0,34845

Al igual que MARS, el análisis discriminante basado en distancia permite clasificar a los jefes de familias en cuatro grupos. El primer grupo esta formado por el número de no beneficiarios según la actual y nueva clasificación, luego esta el grupo formado por no beneficiarios según la actual clasificación y los beneficiarios según la nueva clasificación. El tercer grupo los constituyen los beneficiarios según la actual clasificación y los no beneficiarios según la nueva clasificación. Finalmente, se tiene el número de jefes de familias beneficiarios según la actual y nueva clasificación.

La tabla 4.24 indica el número de jefes de familias clasificados como no beneficiarios en la actual clasificación y predichos como no beneficiarios por el análisis discriminante basado en distancias es de 545. En tanto el número de jefes de familias beneficiarios según la actual clasificación y predichos por análisis discriminante basado en distancias es de 199.

Tabla 4.24 Número de Jefes de Familias según Clasificación

ACTUAL CLASIFICACION	PREDICHO 0	PREDICHO 1	TOTAL CASOS
0	545	256	801
1	0	199	199

Por otra parte la tabla 4.25 muestra que la especificidad fue de un 68 %, es decir un 32 % de los no beneficiarios es predicho como beneficiario. En tanto la sensibilidad es de un 100 % es decir, el total de hogares clasificados como beneficiarios por la actual clasificación fueron predichos como beneficiarios por el análisis discriminante basado en distancias.

Tabla 4.25 Porcentaje de Jefes de Familias según Clasificación

ACTUAL CLASIFICACION	PREDICHO 0	PREDICHO 1	TOTAL CASOS
0	68,0	32,0	100
1	0	100,0	100

4.5 Resultados de MARS

La tabla 4.26 resume los principales indicadores para seleccionar el mejor modelo. La mayor sensibilidad la tiene el modelo I, la mayor especificidad el modelo II, el mayor porcentaje de correctas el modelo II, el menor GCV los modelos II, III y IV y el mayor R^2 ajustado los modelos II, III y IV. En resumen el modelo II presenta los mejores resultados.

Tabla 4.26 Indicadores de Bondad del Ajuste de los Modelos

Modelo	Sensibilidad	Especificidad	Porcentaje de Correcta Clasificación	GCV	R^2 Ajustado
I	98,99	78,03	82,2	0,062	0,624
II	95,48	88,76	90,1	0,055	0,672
III	94,97	88,64	89,9	0,055	0,674
IV	97,49	81,02	84,3	0,055	0,674

4.6 Análisis del modelo II

La tabla 4.27 muestra la ANOVA del modelo II. En las dos primeras columnas se indica el número de la función base y su respectiva desviación estándar. La tercera columna muestra el costo de omisión o GCV de las funciones bases, es decir el costo de perdida del ajuste del modelo si la función base es eliminada de el modelo. Las próximas dos columnas listan el número de funciones bases agregadas en el modelo y el número de parámetros o de grados de libertad de la función base. Las columnas finales listan el nombre de la variable que entra al modelo, una variable para el efecto principal y dos variables para este caso en que se consideran dos interacciones.

Tabla 4.27 ANOVA

FUNCION °	DESVIACIÓN ESTÁNDAR	COSTO DE OMISION	N°. DE FUNCIONES BASES	N°. PARAMETROS EFECTIVIS	VARIABLES	VARIABLES
1	0,164	0,067	1	3,400	P23	
2	0,135	0,067	1	3,400	IFPC	
3	0,075	0,059	1	3,400	P21	
4	0,069	0,056	1	3,400	P45JEFE	
5	0,094	0,060	1	3,400	P23	CATCAS1
6	0,077	0,059	1	3,400	P25	IFPC

7	0,079	0,058	1	3,400	P45JEFE	P49
8	0,063	0,056	1	3,400	P45JEFE	P47

La tabla 4.28 muestra un ranking de importancia de las variables consideradas en el modelo ajustado. El modelo considera importante solo 8 variables de las 13 ingresadas inicialmente, las cuales son ordenadas desde las más a la menos importante con su respectivo costo de omisión.

Tabla 4.28 Importancia de las Variables

Variable	Costo de Omision	Importancia	
INGRESO FAMILIAR	0,067	100,000	||||||||||||||||||||||||||||||||||||||||||||||||||
ABASTECIMIENTO AGUA	0,067	99,187	|||||||||||||||||||||||||||||||||||||||||||||||||
ESCOLARIDAD	0,063	82,566	|||||||||||||||||||||||||||||||||||||||||
CATEGORIA OCUPACIONAL	0,060	66,816	|||||||||||||||||||||||||||||||||
DISPONIBILIDAD TINA	0,059	58,680	|||||||||||||||||||||||||||||
MATERIAL TECHO	0,059	55,566	||||||||||||||||||||||||||||
REFRIGERADOR	0,058	49,219	|||||||||||||||||||||||||
SITIO	0,056	34,238	|||||||||||||||||

El modelo II ajustado considera las siguientes funciones bases:

$$f_1(abastecimiento\ agua) = (abastecimiento\ agua = 1) \text{ o } (abastecimiento\ agua = 4)$$

Esta función base considera la variable abastecimiento de agua potable cuya categorías sean que la llave se encuentra dentro de la vivienda independientemente de que esta provenga o no de la red pública.

$$f_2(abastecimiento\ agua) = (abastecimiento\ agua = 2) \text{ o } (abastecimiento\ agua = 3) \text{ o } (abastecimiento\ agua = 5) \text{ o } (abastecimiento\ agua = 6)$$

Por otra parte, la función base 2 considera al igual que la función base 1 la variable abastecimiento de agua pero para las categorías cuando la llave se encuentra fuera de la vivienda.

$$f_4(IFPC) = \max(0, 50.000 - IFPC)$$

Esta función base considera el ingreso familiar per cápita dónde el punto de corte queda definido en $50.000.

$$f_5(categoria\ ocupacional) = (categoria\ ocupacional = 0)$$
$$\text{o } (categoria\ ocupacional = 4) \text{ o } (categoria\ ocupacional = 7)$$
$$\text{o } (categoria\ ocupacional = 9) * f_2(abastecimiento\ agua)$$

La función base 5 considera la interacción entre las variables categoría ocupacional y abastecimiento de agua, siendo los puntos de corte (nodos) las ocupaciones optimas para la variable categoría ocupacional y considera como puntos de corte de las condiciones más precarias en cuanto al abastecimiento de agua.

$$f_7(disponibilidad\ tina) = (disponibilidad\ tina = 1) * f_4(IFPC)$$

Esta función considera la interacción entre las variables disponibilidad de tina o ducha y el ingreso familiar per capita siendo los puntos de corte tina o ducha de uso exclusivo con agua caliente y $50.000 respectivamente.

$$f_9(material\ techo) = (material\ techo = 3)\ \text{o}\ (material\ techo = 4)\ \text{o}\ (material\ techo = 5)\ \text{o}\ (material\ techo = 6)$$

La función basal 9 considera la variable tipo de material del techo de la vivienda, considerando como punto de corte las categorías más precarias.

$$f_{12}(año\ escolaridad) = \max(0, 7 - año\ escolaridad)$$

Esta función considera la variable años de escolaridad promedio y define como punto de corte 7 años de escolaridad.

$$f_{13}(refrigerador) = (refrigerador = 0) * f_{12}(año\ escolaridad)$$

La función base 13 considera las variables tenencia de refrigerador y años de escolaridad promedio siendo los puntos de corte no tener refrigerador y 7 años de escolaridad promedio respectivamente.

$$f_{15}(sitio) = (sitio = 1)\ \text{o}\ (sitio = 2) * f_{12}(año\ escolaridad)$$

Por último, la función base 15 considera la tenencia de sitio y años de escolaridad promedio siendo los puntos de corte tener sitio propio y escolaridad promedio de 7 años.

$$\textbf{CHISOL = 0,372 - 0,395 *}\ f_1(abastecimiento\ agua)\ \textbf{+ ,000008 *}\ f_4(IFPC)\ \textbf{- 0,341 *}\ f_5(categoria\ ocupacional)\ \textbf{- ,000008 *}\ f_7(disponibilidad\ tina)\ \textbf{+ 0,190}$$

$* \quad f_9(material\ techo) \quad \mathbf{+ \ 0{,}031} \quad * \quad f_{12}(año\ escolaridad) \quad \mathbf{+ \ 0{,}046} \quad * \quad f_{13}(refrigerador) \quad \mathbf{- \ 0{,}033} \ * \ f_{15}(sitio)$

El modelo explicito es el siguiente:

$$P(CHISOL = 1) = 0{,}372 - 0{,}395 * ((abastecimiento\ agua = 1) \text{ o } (abastecimiento\ agua = 4)) + 0{,}000008 * \max(0, 50.000 - IFPC) - 0{,}341 * ((categoria\ ocupacional = 0) \text{ o } (categoria\ ocupacional = 4) \text{ o } (categoria\ ocupacional = 7) \text{ o } (categoria\ ocupacional = 9)) * f_2(abastecimiento\ de\ agua) - 0{,}000008 * (disponibilidad\ tina = 1) * f_4(IFPC) + 0{,}19 * ((material\ techo = 3) \text{ o } (material\ techo = 4) \text{ o } (material\ techo = 5) \text{ o } (material\ techo = 6)) + 0{,}031 * \max(0, 7 - año\ escolaridad) + 0{,}046 * (refrigerador = 0) * f_{12}(año\ escolaridad) - 0{,}033 * ((sitio = 1) \text{ o } (sitio = 2)) * f_{12}(año\ escolaridad)$$

Los valores predichos del modelo estarán en el rango 0-1 sin embargo, en este caso la modelación por MARS genera valores ya sea negativos o mayores que uno, para tal caso y evaluar que la clasificación sea efectiva se sigue el siguiente criterio. Dado que la muestra de entrenamiento tenia un 20% de jefes de familias beneficiarios del Programa Chile Solidario, se considero en la construcción del modelo un valor de corte de clasificación de 0,20 (prevalencia de jefes de familia beneficiarios del Programa Chile Solidario en la muestra de entrenamiento); entonces para evaluar la pertenencia del jefe de familia al Programa Chile Solidario, al valor predicho máximo se le resta el valor predicho mínimo, es decir generamos el rango predicho y establecemos el siguiente criterio:

$$Chisol = \left\{ \begin{array}{ll} 1\ (Beneficiario) & si:\ Valor_{pred} > Rango_{pred} \times 0{,}20 \\ 0\ (Nobeneficiario) & si: Valor_{pred} < Rango_{pred} \times 0{,}20 \end{array} \right\}$$

A modo de ejemplo, en el siguiente cuadro se muestran 5 casos. Para esta situación se tiene que el $Rango_{predicho}$ = 1,193-(-0,005)=1,198 luego $Rango_{predicho}$*0,2=0,2396, este valor se comprara con cada uno de los valores predichos y se asigna con el valor 1 o 0 según la regla señalada anteriormente.

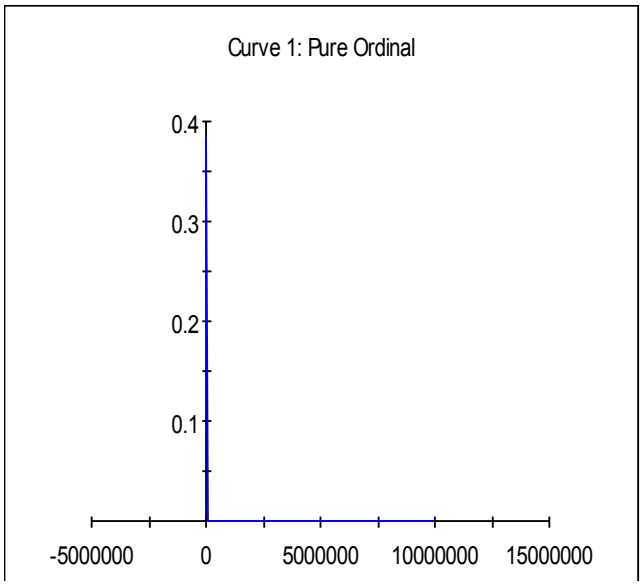
Individuo	**1**	**2**	**3**	**4**	**5**
Chisol observado	0	0	0	1	1
Valor predicho	0,012	0,008	-0,005	0,956	1,193
Chisol predicho	0	0	0	1	1

Las funciones bases y la regresión realizada por MARS ayudan a entender el impacto de la variables predictoras sobre la variable dependiente. A partir del modelo señalado anteriormente, podemos concluir que:

1. La función basal 1 esta formada por una variable categórica, basada en el sistema abastecimiento de agua potable. Si el abastecimiento de agua es 1 o 4, tomara el valor 1 y será 0 para los otros casos. En el modelo el coeficiente de la f_1 es de - 0,395. Así, el valor constante de la regresión decrece en 0,395 cuando la vivienda de un jefe de familia es abastecida de agua potable con llave dentro de la vivienda, provenga ésta de la red pública o no.

2. La función basal 4 (0,000008 * MAX(0, 50.000 - IFPC)) nos indica el punto de corte en que se produce la contribución de la variable IFPC para predecir la probabilidad de ser seleccionado como beneficiario del Programa Chile Solidario. La mayor contribución se dará para los casos en que el IFPC sea 0 y la menor será para los ingresos cercanos a $49.999. Sin embargo, para todos aquellos casos en que el ingreso sea superior a $50.000 la contribución de esta variable al modelo será cero.

Figura 4.1 Contribución de la Variable IFPC

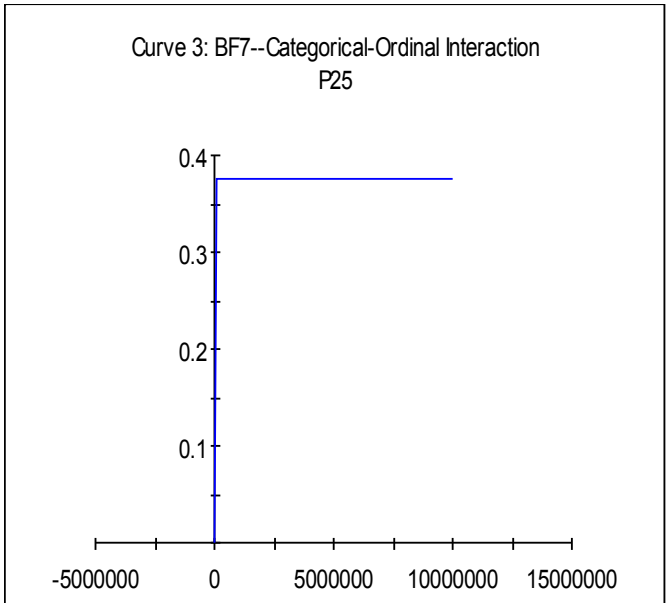


Por ejemplo, para el caso en que el IFPC es cero la contribución a la respuesta será aproximadamente de 0,34, tal como se muestra en la figura 4.1. En el modelo el coeficiente de la f_4 es de 0,000008. Por tanto, el valor de la constante de regresión aumenta levemente en 0,00000272 cuando el IFPC sea cero.

3. Al igual que la función basal 1 la función basal 5 esta formada por la variable categórica denominada categoría ocupacional y por la función basal 2 la cual esta conformada por la variable categórica abastecimiento de agua. La función tomara el valor 1 cuando el jefe de familia presente la categorías ocupacionales de familiar no remunerado, trabajador dependiente urbano, empleado del sector público o tiene una mejor actividad remunerada que su pareja y la vivienda de un jefe de familia es abastecida a través de la red pública o no provenga de la red pública pero la llave se encuentra dentro o fuera del sitio o se debe acarrear en caso contrario tomara el valor 0. El coeficiente de la función basal 5 es de -0.341, es decir el valor constante de la regresión decrece cuando las categorías ocupacionales son óptimas y cuando el abastecimiento del agua es en condiciones precarias.

4. La función basal 7 esta formada por la variable categórica denominada disponibilidad de tina o ducha y por la función basal 4 que considera la variable continúa IFPC. Cuando la vivienda del jefe de familia tiene tina o ducha de uso exclusivo y con agua caliente y su ingreso sea igual a cero la contribución será de 0,38 la cual permanecerá constante para los ingresos familiares per capita mayores a cero. Ver figura 4.2. El coeficiente de la función basal 7 es de -0,000008 por tanto, el valor de la constante de regresión decrece en - 0,00000304 cuando el ingreso familiar per cápita es cero y la vivienda del jefe de familia posee tina o ducha.

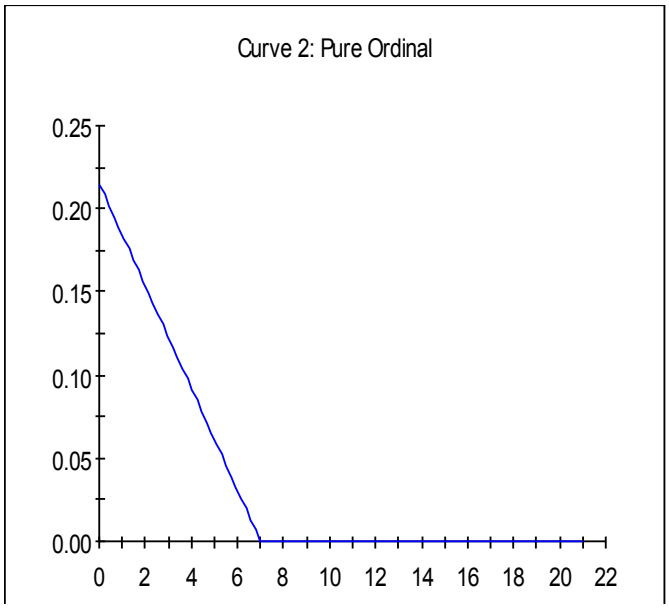
Figura 4.2 Contribución de la Función Basal 7



5. La función basal 9 esta formada por la variable categórica tipo de material del techo. Cuando material del techo de la vivienda del jefe de familia es de zinc o pizarreño sin cielo interior o de fonolita o paja, coirón, totora, caña o desecho tomara el valor 1 por lo tanto el valor constante de la regresión es incrementado en 0,19.

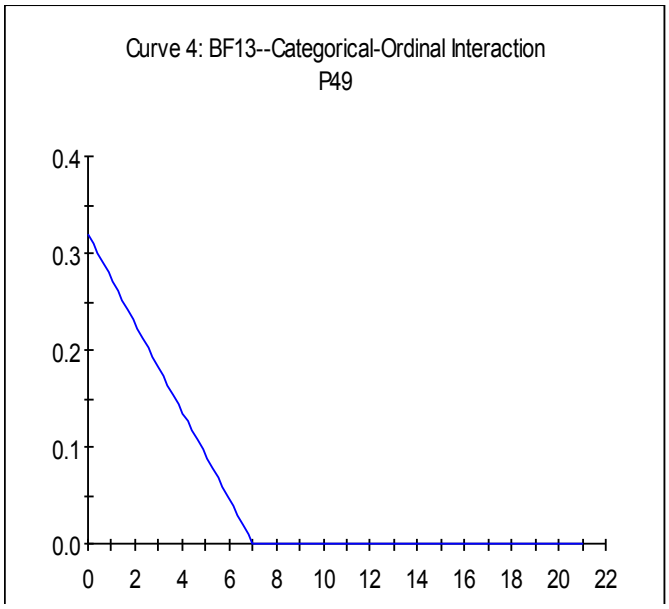
6. Cuando la escolaridad del jefe de familia (f_{12}) sea de 0 años de escolaridad la contribución de la variable al modelo será de 0,217 y la mínima contribución (0,031) será para aquellos jefes de familia cuya escolaridad es de 6 años. Ver figura 4.3. Por otra parte para los jefes de familia en que su escolaridad sea de 7 años y más la contribución será cero. En el modelo el coeficiente de la f_{12} es de 0,031. Así, el valor constante de la regresión aumenta en 0,007 cuando los años de escolaridad del jefe de familia es cero.

Figura 4.3 Contribución de la Función Basal 12



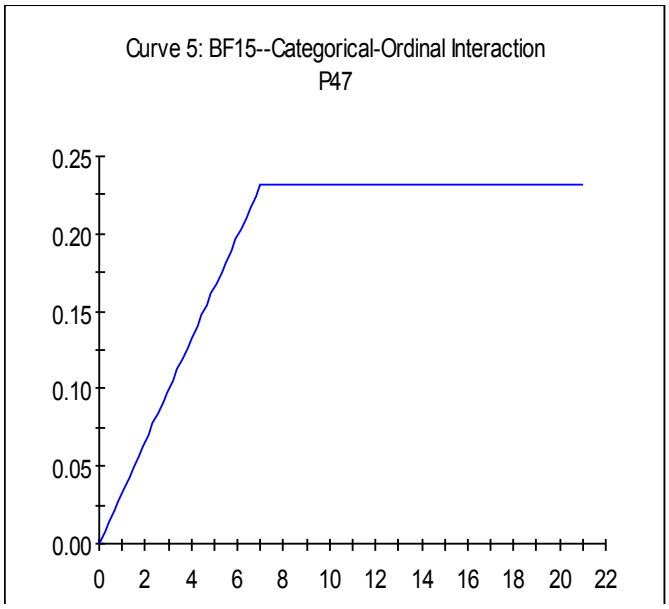
7. Cuando el jefe de familia no posee refrigerador (f_{13}) y además presenta 0 año de escolaridad la contribución será aproximadamente de 0,3. (Ver figura 4.4). Sin embargo, la contribución disminuirá a medida que aumentan los años de escolaridad siendo cero a partir de los 7 y mas años de escolaridad. En el modelo el coeficiente de la f_{13} es de 0,046. Así, el valor constante de la regresión aumenta en 0,014 cuando los años de escolaridad del jefe de familia es cero.

Figura 4.4 Contribución de la Función Basal 13



8. Si el sitio es propio sin deudas o el sitio es propio sin deudas atrasadas y escolaridad es de 7 años la contribución al modelo será de aproximadamente de 0,24 la cual permanecerá constante para los 8 y mas años de escolaridad. En tanto que para los jefes de familias con menos de 7 años de escolaridad y que cuenten con sitio propio sin deudas o el sitio es propio sin deudas atrasadas la contribución declinara siendo esta cero para jefes de familia con cero años de escolaridad. En el modelo el coeficiente de la f_{15} es de -0,033. Así, el valor constante de la regresión decrece en 0,008 cuando los años de escolaridad del jefe de familia es de 7 años y la tenencia de sitio es propio.

Figura 4.5 Contribución de la Función Basal 15



4.7 Comparación MARS y Análisis Discriminante Basado en Distancias

Los resultados del modelo II obtenido por MARS y los resultados obtenidos por análisis discriminante Basado en Distancia se presentan en la tabla 4.29.

Tabla 4.29 Indicadores de Bondad para los métodos MARS y Análisis Discriminante Basado en Distancia

Método	Sensibilidad	Especificidad	Total Correctas (%)
MARS (modelo II)	95,48	88,76	90,1
Discriminante B.D	100,0	68,0	74,4

5 DISCUSION Y CONCLUSION

En el presente trabajo se aplicó las metodologías Multivariate Adaptive Regression Splines (MARS) y Análisis Discriminante Basado en Distancias (DB).

La aplicación del Análisis Discriminante Basado en Distancias genera una matriz de clasificaciones de los objetos sin generar un modelo explicito matemático lo que implica que cada vez que se desee evaluar la incorporación de una familia al Programa Chile Solidario será necesario ejecutar el programa estadístico para estos fines. Por otra parte, tampoco es posible conocer la jerarquización de las variables involucradas en el proceso de clasificación para de esta manera evaluar su pertinencia socioeconómica en el modelo.

Los resultados de este análisis nos indican que la sensibilidad es de un 100 % es decir, del total de familias seleccionadas por MIDEPLAN como beneficiarias del Programa Chile Solidario, el DB las predice como beneficiarias. En tanto, la especificidad es de un 68 % es decir un 32 % de las familias seleccionadas por MIDEPLAN como no beneficiarias son predichas como beneficiarias

A través de la aplicación de MARS, como se muestra en el capitulo anterior, se implementaron cuatro modelos de los cuales se selecciona el

modelo II por su mejor bondad de ajuste y por la plausibilidad de las funciones bases. Los resultados indican que la sensibilidad fue de un 95, 5 % es decir, del 100 % de los beneficiarios seleccionados por MIDEPLAN la aplicación de MARS excluye solo un 4,5%. En cuanto a la especificidad esta es de un 88,8 % es decir un 11,2 % de las familias seleccionadas como no beneficiarias por MIDEPLAN son predichas por MARS como beneficiarias.

Los resultados del modelo óptimo de MARS se compararon con los resultados obtenidos a través del análisis discriminante basado en distancia (DB). Al comparar los resultados de ambas metodologías en relación a la clasificación de las familias como beneficiarias del Programa Chile Solidario podemos señalar que los resultados de MARS son mejores en cuanto a la especificidad y al porcentaje total de correcta clasificación.

A diferencia del Análisis Discriminante Basado en Distancias, MARS proporciona un modelo matemático explicito el cual ayuda a entender la relación no lineal, describe las interacciones utilizadas y puede unir agrupación de categorías de las variables que tienen efectos similares sobre la variable dependiente permitiendo de esta forma conocer las características particulares que definen la clasificación de los diferentes objetos.

En conclusión, MARS proporciona una estimación que es flexible debido a que, a diferencia de los análisis estadísticos convencionales, permite generar puntos de corte (nodos) tanto para las variables continuas como discretas. Estos puntos de corte, identificados en las funciones basales de los resultados

del modelo, permiten a los agentes que toman decisiones conocer el comportamiento de los individuos u objetos clasificados en base sus atributos específicos. Así por ejemplo, para la variable años de escolaridad el punto de corte definido por la función basal es de 7 años es decir, para los jefes de familias que presentan 7 o menos años de escolaridad serán valorados como posibles beneficiarios. Este resultado es muy concordante con la política implementada el año 2003 por el ministerio de educación de garantizar la obligación y gratuidad de 12 años de escolaridad señalando que la educación es un factor determinante para salir de la pobreza. También el modelo nos indica que para la variable categoría ocupacional los puntos de corte quedan definidos en las mejores ocupaciones es decir, para aquellos jefes de familia que tengan un buen trabajo serán valorados como posibles no beneficiarios. Otro resultado que esta en línea con las políticas que apuntan a un "trabajo decente". Este comportamiento es un resultado empírico que no requiere de la intervención arbitraria del modelador, por tanto, se sustenta exclusivamente en la información contenida en las variables de análisis.

Así MARS permite estudiar a partir de las relaciones entregadas por las funciones base y sus nodos las variables de importancia en la clasificación de jefes de familia. También es posible contrarrestar la información con políticas y programas sociales en ejecución o definir cambios o implementar nuevas políticas si así se evalúa.

Finalmente, señalar que instrumentos más flexibles como MARS entregan información adicional que no solo es útil para clasificar como en

este caso, si no que además permiten orientar políticas públicas destinadas a poblaciones especificas.

6 BIBLIOGRAFIA

Alvarado, S. (2002). Predicción de Calidad del Aire para Material Particulado PM10 en la estación Pudahuel de la Red de Monitoreo MACAM-2, Comparación de Dos Modelos Predictivos, Tesis para optar al Grado de Magíster en Bioestadística. Universidad de Chile.

Castañeda, T. y Lindert, K. (2005). Designing and Implementing Household Targeting Systems: Lessons from Latin American and the United States. The World Bank, Washington, D.C.

Coaday, D., Grosh, M. y Hodinott, J. (2004). La Focalización de las Transferencias en los Países en Desarrollo: Revisión de Lecciones y Experiencias. The World Bank, Washington, D.C.

Cuadras, C.M. (2004). Análisis Multivariante, Apuntes de Clases. Universidad de Barcelona.

Cuadras C.M. (2004). Métodos Multivariantes Basados en Distancias, Apuntes de Clases. Universidad de Barcelona.

De Cáceres M. (2005). Ginkgo User' s Manual versión 1.4. Departamento de Biología Vegetal, Unidad de Botánica, Universidad de Barcelona.

De Cáceres M., Oliva F. y Font X. (2003). Ginkgo, un Programa de Análisis Multivariante Orientado a la Clasificación Basada en Distancias. Congreso Nacional de Estadística e Investigación Operativa, Universidad de Barcelona.

De Cáceres M., Font X., García R. y Oliva F. (2001). VEGANA, un Paquete de Programas para la Gestión y Análisis de Datos Ecológicos. VII

congreso Nacional de la Asociación Española de Ecología Terrestre, España.

Grosh, M., Baker. Proxy Means Tests for Targeting Social Programs. (1995). The World Bank, Washington, D.C.

Hastie, T., Tibshirani, R. y Friedman, J. (2001). The Elements of Statistical Learning. Springer.

Irigoyen, I. Cuadras, C.M y Arenas C. (2003). Nuevo Método de Agrupación Jerárquica Basado en Distancias. LX Conferencia Española de Biometría, La Coruña, España.

Larrañaga, O. (2005). Focalización de Programas en Chile: El Sistema CAS. Serie de Informes Sobre Redes de Protección Social. Banco Mundial.

Legovini, A. (1999). Targeting Methods for Social Programs. Banco Inter-Americano de Desarrollo.

Lewis, P.A y Stevens, J.G. (1991). Nonlinear Modeling of Time Series Using Multivariate Adaptive Regression Splines (MARS). Journal of the American Statistical Association, Vol. 86, No 416, Applications and Case Studies.

MARS User Guide (2001). Salford Systems.

MIDEPLAN. (2003). Estadísticas de Pobreza. Online. Internet. *www.mideplan.cl*

MIDEPLAN. (2000). La Ficha CAS como Instrumento de Focalización de Programas Sociales.

MIDEPLAN. (2000). Metodología Encuesta CASEN. Online. Internet. www.mideplan.cl

Peña, D. (2002). Análisis de Datos Multivariantes. Mc Graw Hill.

Muñoz, J. and Felicísimo, A. (2004). Comparación of Statistical Methods Commonly Used in Predictive Modelling. Journal of Vegetation Science 15: 285-292.

Silva, C., Alvarado, S., Montaño, R. y Pérez, P. (2003). Modelamiento de la Contaminación Atmosférica por Partículas: Comparación de Cuatro Procedimientos Predictivos en Santiago, Chile. Biomatemática XII: 113-117.

Silva, C. (2004). Modelos Multivariados, Apuntes de Clases, Escuela de Salud Pública, Universidad de Chile.

Skoufias, E., Davis, B. y Behrman, J. (1999). Evaluación de la Selección de Beneficiarios en el Programa de Educación, Salud y Alimentación (Progresa), México.

Stokes, H. y Lattyak, W. (2005). Multivariate Adaptive Regresión Spline (MARS) Modeling Using the B34S. ProSeries Econometric System and SCA WorkBench, University of Illinois at Chicago.

Valdivia, M., Dammert, A. (2001). Focalizando las Transferencias Públicas en el Perú: Evaluando Instrumentos de Identificación del Nivel Socio-Económico de los Individuos/Hogares, Grupo de Análisis para el Desarrollo, Perú.

Xiong, R. and Meullenet, J.F (2004). Aplication of Multivariate Adaptive Regression Splines (MARS) to the Preference Mapping of Cheese Sticks. Jounal of Food Science, Vol. 69, Nr.4.

Anexo 1

Descripción de la Ficha CAS

La Ficha CAS-2 contiene 50 preguntas distribuidas en nueve secciones:

- Sección 0: Datos generales. Identifica la región, comuna, provincia, unidad vecinal y domicilio de los entrevistados.

- Sección 1: Protección Ambiental. Registra la información relativa al tipo de material utilizado en los muros exteriores, en el piso y en el techo de la vivienda.

- Sección 2: Hacinamiento. Se recoge aquí la información respecto del número y uso de las piezas ocupadas de la vivienda.

- Sección 3: Saneamiento y Confort. Registra información relativa al tipo de abastecimiento de agua en la vivienda, al sistema de eliminación de excretas, a la disponibilidad de tina o ducha, y a la existencia de suministro eléctrico.

- Sección 4: Identificación de los Residentes. Registra la identificación personal de los encuestados, como son nombre completo, sexo, fecha de nacimiento, relación de parentesco con el jefe de familia, y familia y hogar de pertenencia.

- Sección 5: Ocupación e Ingresos. Esta sección, dirigida a todos los residentes de la vivienda mayores de 14 años, registra la actividad ocupacional que desarrollan las personas, y el monto y periodicidad de los ingresos que perciben.

- Sección 6: Subsidios Monetarios. Registra información respecto de qué subsidio monetario estatal percibe la persona, en caso que perciba alguno.
- Sección 7: Educación. Registra los años de estudio aprobados por cada residente de la vivienda de 6 años de edad y más.

- Sección 8: Patrimonio. Aporta información respecto de la propiedad de las familias, como son televisor, refrigerador, calefont y tipo de uso del sitio en que viven.

Anexo 2

Factores, Subfactores y Variables Utilizadas en el Cálculo del Índice CAS

Factor (1)	Coef. 3 Ponderacion (2)	Subfactor	Coef. 2 Ponderación Subfactor (CSF)	Variables	Coef. 1 Ponderación Variable (CV)
Vivienda	0,26	Protección Ambiental	0,40	Muro Piso Techo	0,35 0,35 0,30
		Hacinamiento	0,22	Dormitorios/Personas Vivienda	1,0
		Saneamiento y Confort	0,38	Agua Elimin. Excretas Tina – Ducha	0,35 0,30 0,35
Educación	0,25			Años de estudios Jefe de Familia	1,0
Ocupación	0,22			Categ. ocupac. más alta de la pareja	1,0
Ingresos / Patrimonio (3)	0,27	Ingreso	0,43	Ingreso Familiar per cápita	1,0
		Sitio	0,13	Propiedad Sitio	1,0
		Equipamiento	0,44	Refrigerador Calefont	0,50 0,50

Anexo 3

Descripción de Homologación CAS-CASEN

Homologación Familia

Una de las diferencias entre la encuesta CASEN y la ficha CAS es el concepto de familia. En la CASEN el núcleo familiar es un subconjunto del hogar que se conforma por la presencia de una pareja legal o de hecho, con hijos solteros dependan o no económicamente de la pareja y que no formen otro núcleo. En la ficha CAS, sin embargo, la familia es aquella constituida por una persona o grupo de personas con o sin vínculos de parentesco, que tienen la intención de convivir juntos de un modo permanente y donde cada uno de sus integrantes es reconocido como tal por el jefe de familia.

Homologación Vivienda

Sub factor Protección Ambiental: En este subfactor se incluye a las variables muro, piso y techo de la Ficha CAS II. Todas ellas están disponible en la encuesta CASEN, excepto las categorías CAS denominadas mixto aceptable y mixto deficiente.

Subfactor Hacinamiento: Se calculó como la razón entre el número de dormitorios (uso exclusivo) de una vivienda y el número de personas que habita dicha vivienda.

Subfactor Confort y Saneamiento: en este ítem se considera a las variables abastecimiento de agua, eliminación de excretas y existencia de ducha. La homologación de la variable CAS " abastecimiento de agua" se construyó a través de la variables CASEN fuente del agua y sistema de distribución.

Para homologar la variable "eliminación de excretas" se utilizaron las variables números de baños en la vivienda y sistema de eliminación de

excretas de la CASEN para así poder inferir las condiciones de régimen "exclusivo" y "compartido" que aparece incorporada en la ficha CAS II.

Para definir si una vivienda posee o no ducha con las condiciones que establece la ficha CAS II fue necesario combinar las variables CASEN: número de baños; sistema de distribución de aguas y disponibilidad de agua caliente.

Homologación Educación

El puntaje para el factor educación se obtiene a partir de la variable años de escolaridad para el jefe de familia disponible en CASEN.

Homologación Ocupación

Dado que la ficha CAS II presenta una categorización distinta a la entregada por CASEN, fue necesario homologar esta variable considerando la ocupación principal, el oficio y la rama del jefe de familia o de la cónyuge cuando corresponda. Adicionalmente también se homologaron todas aquellas personas que están cesantes o aquellas que no realizan ninguna actividad.

Homologación Patrimonio

Subfactor Ingresos: Para calcular el ingreso per cápita se consideró el ingreso autónomo de toda la familia (sin corrección) dividido por el número de personas que conforman dicha familia.

Subfactor Sitio: Esta variable se refiere a la posesión del sitio donde se ubica la vivienda. Esta variable está definida en CASEN pero con distinta codificación que la ficha CAS II, puesto que el primer caso no considera si el pago del sitio está en mora. En consecuencia, se utilizó la variable pago del crédito hipotecario con el fin de encontrar una homologación para las categorías "propio sin deudas atrasadas" y "propio con deudas atrasadas".

Subfactor Equipamiento: este ítem considera la existencia de calefont y refrigerador en la vivienda. Ambas variables se encuentran disponibles en la encuesta CASEN.

Anexo 4

Índice CAS de Corte para cada Región del País

Región	Índice CAS de corte
Tarapacá (I)	489
Antofagasta (II)	510
Atacama (III)	501
Coquimbo (IV)	463
Valparaíso (V)	494
O'Higgins (VI)	471
Maule (VII)	468
Bío Bío (VIII)	472
La Araucanía (IX)	457
Los Lagos (X)	462
Aysén (XI)	464
Magallanes (XII)	510
Metropolitana	503

Fuente: MIDEPLAN, 2002

Anexo 5

Total de Familias Seleccionadas por Año

Región	Total Familias	Año 2002	Año 2003	Año 2004	Año 2005
Tarapacá (I)	5.633	2.178	1.408	1.408	639
Antofagasta (II)	7.137	1.099	2.642	2.220	1.176
Atacama (III)	6.559	2.198	1.744	1.562	1.055
Coquimbo (IV)	6.163	2.151	1.958	1.435	619
Valparaíso (V)	18.285	6.198	4.731	4.469	2.887
O'Higgins (VI)	7.630	3.609	1.670	2.114	237
Maule (VII)	15.080	5.667	4.619	3.535	1.259
Bío Bío (VIII)	40.142	7.400	10.102	13.998	8.642
La Araucanía (IX)	26.998	4.725	8.445	8.201	5.627
Los Lagos (X)	24.218	5.501	7.072	7.609	4.036
Aisén (XI)	998	542	456	0	0
Magallanes (XII)	1.088	686	402	0	0
Metropolitana	49.467	14.101	15.069	13.255	7.042
País	**209.398**	**56.055**	**60.318**	**06**	**33.219**
% cobertura anual		26.8%	28.8%	28.6%	15.9%

Fuente: MIDEPLAN, 2002

Anexo 6

Distribución de la Muestra de Entrenamiento por Comunas de la V Región según Programa Chile Solidario

COMUNA	No beneficiario	Beneficiario	Total
LA LIGUA	35	11	46
PETORCA	26	13	39
CABILDO	20	5	25
LOS ANDES	19	6	25
SAN ESTEBAN	34	12	46
CALLE LARGA	8	3	11
RINCONADA	23	7	30
SAN FELIPE	35	0	35
PUTAENDO	16	11	27
SANTA MARIA	9	1	10
PANQUEHUE	23	4	27
LLAY-LLAY	18	14	32
CATEMU	27	12	39
QUILLOTA	26	7	33
CALERA	42	8	50
NOGALES	29	4	33
HIJUELAS	25	13	38
LIMACHE	23	9	32
OLMUE	26	12	38
VALPARAISO	39	4	43
VIÑA DEL MAR	40	1	41
QUINTERO	29	6	35
PUCHUNCAVI	29	8	37
QUILPUE	38	5	43
VILLA ALEMANA	41	2	43
CASABLANCA	23	4	27
CON -CON	10	0	10
SAN ANTONIO	28	5	33
CARTAGENA	25	5	30
EL TABO	5	3	8
EL QUISCO	30	4	34
Total	801	199	1000

Anexo 7

Distribución de la Muestra de Validación por Comunas de la V Región según Programa Chile Solidario

COMUNA	No beneficiario	Beneficiario	Total
LA LIGUA	29	5	34
PETORCA	26	11	37
CABILDO	26	8	34
LOS ANDES	34	6	40
SAN ESTEBAN	25	7	32
CALLE LARGA	8	4	12
RINCONADA	26	10	36
SAN FELIPE	29	6	35
PUTAENDO	33	13	46
SANTA MARIA	8	1	9
PANQUEHUE	21	6	27
LLAY-LLAY	26	10	36
CATEMU	22	12	34
QUILLOTA	37	2	39
CALERA	28	6	34
NOGALES	28	7	35
HIJUELAS	21	12	33
LIMACHE	34	9	43
OLMUE	18	6	24
VALPARAISO	32	4	36
VIÑA DEL MAR	46	3	49
QUINTERO	25	5	30
PUCHUNCAVI	27	8	35
QUILPUE	33	3	36
VILLA ALEMANA	32	4	36
CASABLANCA	23	4	27
CON –CON	7	1	8
SAN ANTONIO	37	6	43
CARTAGENA	25	15	40
EL TABO	8	0	8
EL QUISCO	27	5	32
Total	801	199	1000