

TESIS DOCTORAL

Análisis de Regresión Difusa: Nuevos Enfoques y Aplicaciones

Sergio Donoso Salgado

Granada, Octubre de 2006

Editor: Editorial de la Universidad de Granada
Autor: Sergio Donoso Salgado
D.L.: Gr. 2282 - 2006
ISBN: 978-84-338-4157-5



Análisis de Regresión Difusa: Nuevos Enfoques y Aplicaciones

memoria que presenta

Sergio Donoso Salgado

para optar al grado de

Doctor en Informática

Octubre de 2006

DIRECTORES

Dr. Nicolás Marín Ruiz

Dra. María Amparo Vila Miranda

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN
E INTELIGENCIA ARTIFICIAL

La memoria titulada “Análisis de Regresión Difusa: Nuevos Enfoques y Aplicaciones”, que presenta D. Sergio Donoso Salgado para optar al grado de Doctor, ha sido realizada en el Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada bajo la dirección de los Doctores Dr. Nicolás Marín Ruiz y Dra. María Amparo Vila Miranda.

Granada, Octubre de 2006.

El Doctorando | Los Directores

D. Sergio Donoso | Dr. Nicolás Marín | Dra. María Amparo Vila

Eran tres.
(Vino el día con sus hachas.)
Eran dos.
(Alas rastreras de plata.)
Era uno.
Era ninguno.
(Se quedó desnuda el agua.)
Federico Garcia Lorca
Poeta Español

Los cuatro puntos cardinales son tres: el Sur y el Norte.
Vicente Huidobro
Poeta chileno

Agradecimientos

Muchas personas me han apoyado en mi vida, en Chile y en Granada. Desde el punto de vista profesional, este trabajo ha sido posible gracias al apoyo permanente y afectuoso de mis dos directores de tesis, Amparo y Nicolás. Sin ellos, no habría tenido la oportunidad de desarrollarlo.

El Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada ha constituido un sitio magnífico en el que desarrollar mi trabajo y sus miembros siempre han estado disponibles para ayudarme en mi tarea.

Desde el punto de vista personal, deseo destacar la memoria de mi padre, que me ayudó en todas mis inquietudes intelectuales. Gracias también a mi madre por estar a mi lado siempre. Fue ella quién me invitó, aquel año de 1969, a ver la película *2001, Odisea del espacio*, donde escuchara por primera vez en mi vida la palabra computador y me fascinara por la aventura de la especie humana y la capacidad del hombre para cambiar el mundo.

Gracias al cariño de José Manuel, Paz y Sebastián.

Índice general

Índice de figuras

Índice de tablas

Motivación

La Inteligencia Artificial y la Estadística, áreas de la ciencia que se han desarrollado independientemente, tienen, sin embargo, muchos problemas que abordan en común.

Para problemas de predicción, tanto la Inteligencia Artificial como la Estadística, tienen un completo cuerpo teórico desarrollado, cada una con sus propias terminologías. Esto puede llevar a una incomunicación entre ellas. Las opiniones de cuándo es más conveniente la Inteligencia Artificial o la Estadística para resolver un determinado problema son diversas [Sar94, Alu01].

Sin embargo, no es difícil encontrar áreas donde convergen estas dos disciplinas. Por ejemplo, la Minería de Datos, orientada a extraer patrones predictivos ocultos en grandes bases de datos, es una de estas áreas.

La explosión de datos de la era digital, que con Internet se ha acelerado y puesto al alcance de muchas más personas, ha traído la necesidad de desarrollar técnicas para el análisis de nuevos tipos de datos y de enfrentar problemas nuevos y de una magnitud, por el volumen de datos, inimaginable hasta ahora. El marketing en las empresas; la información de los satélites en exploraciones mineras, agricultura, metereología, ...; el proyecto Genoma en biología y medicina, son clara muestra de esta nueva realidad.

Pero no sólo la cantidad de información aumenta, sino también ciertas particularidades como la aparición de *lagunas o omisiones* en los datos, la dispersión de las mediciones, la ambigüedad del significado del dato. ¿Cómo se puede aseverar ahora que una información es representativa, cuando es una muestra que se ha obtenido de Internet? ¿Se puede modelar igual la opinión de una encuesta que la medición en un laboratorio? Estas preguntas, creemos, son un desafío para la Estadística Probabilística que debiera renovarse para seguir aportando soluciones a las necesidades que plantea los problemas de Minería de Datos actuales.

Esta tesis se encuadra en un contexto donde la Estadística y la Inteligencia Artificial van de la mano: el Análisis de Regresión Difusa. Este tema se fundamenta en el Análisis de Regresión tradicional de la Estadística e intenta extender su aplicación a datos que pueden modelarse a través de subconjuntos difusos, una de las áreas de la Inteligencia Artificial que permite afrontar la imprecisión e incertidumbre que puede aparecer en los datos.

Nuestro estudio está orientado a formalizar, generalizar y extender las propuestas que se han formulado hasta la fecha sobre Regresión Difusa, con el claro objetivo de proporcionar un instrumental metodológico aplicable a problemas de determinación de factores y patrones, predicción, y validación de hipótesis en que los datos disponibles están afectados por imprecisión e incertidumbre.

Nuestra propuesta de Regresión Difusa, que tiene como primer objetivo explicar el comportamiento de una variable difusa, también intenta ampliar su aplicación a la predicción con nuevos valores de dicha variable. Como veremos a lo largo de esta memoria, nuestro trabajo de investigación se centra en modelar datos con funciones de pertenencia triangulares no simétricas, utilizar funciones de optimización flexibles, disponer de restricciones de acuerdo a las características de cada problema, posibilitar la selección de variables dado el carácter multivariante del análisis, y buscar la eliminación de las distorsiones que provienen de la multicolinealidad.

La memoria se organiza como sigue:

- El primer capítulo presenta el contexto científico en el que se encuadra este trabajo de investigación, describiendo cómo surge el análisis de Regresión Difusa a partir del uso conjunto de soluciones de Regresión Probabilística y la Teoría de Conjuntos Difusos.

- El segundo capítulo analiza los antecedentes sobre Regresión Difusa que se pueden encontrar en la literatura: se describen tanto propuestas posibilísticas como de mínimos cuadrados, así como propuestas basadas en otros enfoques. Este capítulo se completa estudiando con detalle el problema de las medidas de bondad del ajuste, presentando las medidas que se han utilizado hasta el momento y proponiendo un conjunto adicional de medidas procedentes de la Teoría de Conjuntos Difusos.

- El tercer capítulo introduce una parte esencial de la contribución de nuestro trabajo de investigación: una metodología para la aplicación del análisis del Regresión Difusa siguiendo un enfoque cuadrático y con una amplia variedad

de métodos que permite su uso flexible y adaptable al problema que se quiera resolver. El capítulo lo completa un estudio empírico de los métodos que se proponen y un análisis de su adecuación en función de los objetivos que se pretenden con la Regresión.

- Siguiendo con nuestra filosofía de buscar alternativas flexibles que permitan aplicar soluciones de Regresión Difusa en problemas diversos de predicción, en el capítulo 4 afrontamos el estudio de técnicas alternativas para llevar a cabo la selección de variables en problemas que lo demanden. Se describen distintos criterios que pueden guiar la selección en función del objetivo perseguido en el proceso.

- El capítulo 5 presenta extensiones de nuestras propuestas para aplicarlas a dos problemas conocidos de Regresión. Por un lado, se hace una propuesta para la aplicación de nuestros métodos en aquellos casos en los que se aprecia multicolinealidad en los datos. Por otro, desarrollamos el concepto de sistema de regresiones posibilísticas como solución aplicada al programa de la Regresión Ecológica.

- Por último, en el capítulo 6 presentamos las conclusiones principales que hemos extraído en el desarrollo del trabajo de investigación que se presenta en esta memoria, así como las posibles líneas de investigación futuras que identificamos como pendientes de afrontar tras la realización del mismo.

Capítulo 1

Introducción

Los dos antecedentes fundamentales de la Regresión Difusa son la Regresión Probabilistica y la Teoría de Conjuntos Difusos. A cada uno de ellos dedicamos una sección en este capítulo de introducción. La tercera sección está orientada a examinar la necesidad de la Regresión Difusa como enfoque específico del Análisis de Regresión.

1.1. Revisión de la Regresión Probabilística

Muchas herramientas han sido desarrolladas para el aprendizaje estadístico, buscando encontrar las relaciones subyacentes que hay entre los datos. Entre las principales herramientas estadísticas que persiguen este objetivo se encuentran la Clasificación, la Regresión y el Agrupamiento de Objetos[HTF01].

El problema general de regresión, como técnica utilizada para el aprendizaje supervisado de valores cuantitativos, consiste en encontrar la relación de una variable dependiente (de salida o endógena) con un conjunto de variables independientes (de entrada o exógenas). Formalmente, dado un conjunto de datos (x_i, y_i), para i=1,...,n, donde $x_i \epsilon R^m$ e y_i es el valor de salida correspondiente al vector x_i, y dada un función $f(x, A)$, se quiere encontrar el vector de parámetros A, tal que

$$y_i = f(x_i, A) \; para \; i = 1, ..., n \tag{1.1}$$

En general, no suele haber una solución exacta para la ecuación anterior debido a

la variabilidad de los datos del mundo real, a la infinidad de factores que se reflejan de cada dato y a la incertidumbre de muchas mediciones. Por este motivo, debe buscarse una manera de relajar el cumplimiento estricto de la igualdad.

La Regresión Probabilística, la más conocida y utilizada de entre las técnicas de regresión, relaja la relación (1.1) definiendo una función de pérdida, L, que mide cómo los errores de predicción entre y_i y $f(x_i, A)$ debieran penalizarse, con la idea de encontrar una solución *lo más aproximada posible* al cumplimiento de la igualdad. Una elección habitual de función de pérdida es la norma L_p:

$$L_p(y - f(x, A), x) = |y - f(x, A)|^p \tag{1.2}$$

para algún número positivo p. La función de pérdida más utilizada es el *Ajuste de Mínimos Cuadrados*, donde p=2, que presenta ventajas analíticas, ya que L_1 tiene el inconveniente de presentar discontinuidades en sus derivadas.

Lo más frecuente es que f sea una función lineal. En caso de no serlo, por lo general, se puede *linealizar* el modelo mediante transformaciones de las variables. Cuando es posible dicha linealización, se dice que se trata de un modelo *intrínsecamente lineal*.

El modelo clásico de Regresión Lineal, basado en la teoría de probabilidades, asume que se dispone de n observaciones independientes para las variables y, x_1, x_2, ..., x_m. Para cada observación se asume el modelo lineal siguiente:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_m x_{im} + \varepsilon_i \tag{1.3}$$

donde los valores β_i son parámetros no conocidos, y se hacen los siguientes supuestos sobre los residuos aleatorios ε_i:

I) El valor esperado de los residuos es cero.
II) La varianza de los residuos es una constante σ^2 no conocida.
III) La covarianza entre ε_i y ε_k, para $i \neq k$ es cero.

Un supuesto implícito, y a nuestro juicio muy relevante, es que los coeficientes son constantes, lo que significa que estructuralmente el sistema del cual proviene el modelo tiene un comportamiento muy determinístico y en el transcurso del tiempo en que se tomaron las mediciones que se quieren modelar no se produjeron modificaciones que pudieran alterar alguno de los parámetros del modelo.

En el caso de que m sea 1, hablamos de Regresión Lineal Simple, y si m es mayor que 1, hablamos de Regresión Lineal Múltiple.

El modelo (1.3) generalmente se expresa en forma matricial como:

$$y = x\beta + \varepsilon \tag{1.4}$$

donde y es el vector de dimensión n con las observaciones de la variable dependiente, x es la matriz n.(m+1) con las observaciones de las m variables independientes más una variable de unos, y β es un vector (m+1) con los parámetros no conocidos que hay que estimar (véase la Figura 1.1).

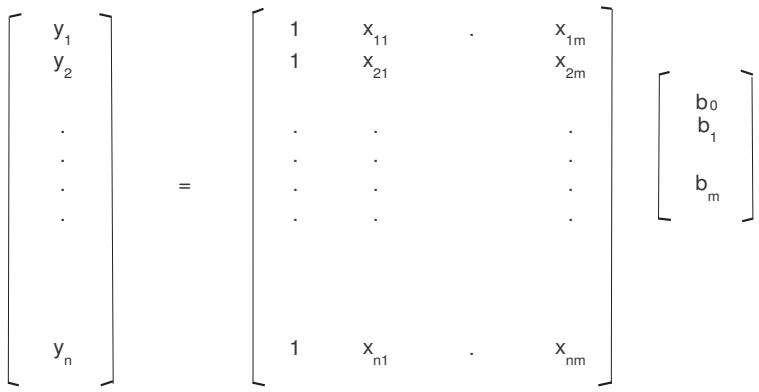

Figura 1.1: *Formulación matricial de la Regresión Probabilística Lineal*

Para encontrar una solución a este sistema se puede aplicar una función de pérdida, como se señalaba al comienzo. El método de estimación de mínimos cuadrados, que es el más utilizado, equivale a minimizar los valores de ε al cuadrado, con la expresión

$$(y - x\beta)'(y - x\beta) \tag{1.5}$$

cuya solución b, que estima a β, está dada por

$$b = (x'x)^{-1}x'y \tag{1.6}$$

Entre todos los estimadores insesgados de β, es decir, que cumplen la condición $E(b) = \beta$, el estimador de mínimos cuadrados es el de menor varianza. Este resultado

se conoce como Teorema de Gauss-Markov [HTF01]. Esta es una de las propiedades teóricas que ha hecho que el método de los mínimos cuadrados sea tan utilizado. Sin embargo, este resultado no garantiza que la varianza de los estimadores sea necesariamente pequeña, y se podría dar el caso de estimadores sesgados que tuvieran menor varianza que los calculados por mínimos cuadrados.

Una estimación para la varianza desconocida σ^2 es

$$\widehat{\sigma^2} = \frac{1}{n-m-1}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \tag{1.7}$$

donde $\widehat{y_i}$ se calcula a partir de (1.6) y se elige el denominador (n - m -1) en vez de n, para que $\widehat{\sigma^2}$ sea insesgado.

Con la ecuación (1.3) y la solución (1.6) es posible hacer una estimación puntual de y, $\widehat{y_i}$, sustituyendo los valores calculados de b, para un x_i determinado. Como esta estimación está sujeta a un determinado margen de error, representado por el componente aleatorio del modelo (1.4), se puede acompañar de un intervalo de confianza. Para estimar estos intervalos, y hacer tests de hipótesis del comportamiento de los coeficientes, se asume que los datos siguen una determinada distribución de probabilidades (por lo general, la distribución normal).

$$\varepsilon \sim N(0, \sigma^2) \tag{1.8}$$

y resulta la distribución para el estimador

$$b = \widehat{\beta} \sim N(\beta, (x'x)^{-1}\sigma^2) \tag{1.9}$$

Suponiendo esta distribución, un intervalo de confianza al $(1-\alpha)100$ por ciento para los valores b, viene dado por el intervalo [Buc04]:

$$\left[b - t_{\alpha/2}\sqrt[2]{\frac{\widehat{\sigma^2}}{(n-m-1)}}, b + t_{\alpha/2}\sqrt[2]{\frac{\widehat{\sigma^2}}{(n-m-1)}}\right] \tag{1.10}$$

donde $t_{\alpha/2}$ es el valor de la distribución t, con (n - m - 1) grados de libertad, de forma que la probabilidad de que una estimación particular se salga de este intervalo de confianza sea $\alpha/2$.

La Regresión Probabilística tiene su principal medida de bondad de ajuste en el Coeficiente de Determinación R^2, que se interpreta como la relación de la varianza explicada por la regresión en relación a la varianza total de la variable de salida.

La Regresión Probabilística requiere un número mínimo de observaciones y, a medida que aumentan los grados de libertad, también se hace más preciso el intervalo de confianza.

Sin lugar a dudas, la Regresión Probabilistica ha tenido un éxito notable en múltiples aplicaciones, dentro de sistemas acotados y con estricto cumplimiento de sus supuestos. No ocurre lo mismo cuando los sistemas son más complejos, presentan diversos tipos de incertidumbre y no se consideran los supuestos básicos.

Un tipo de variable instrumental, que ha sido muy útil dentro de los modelos probabilísticos, lo constituyen las llamadas *variables independientes binarias*. Estas variables sólo pueden tomar dos valores y pueden representar la presencia o ausencia de una cualidad.

Este modelo clásico de Regresión Probabilística presenta algunos problemas prácticos, como son:

I) Pueden existir errores de especificación del modelo, bien por omisión de una o más variables independientes, o bien porque la función elegida no es la adecuada.
II) Puede existir *multicolinealidad* entre las variables independientes de x. Es decir, una combinación lineal de las columnas de x puede ser cercana a 0, lo que hace que el cálculo de la matriz $(x'x)^{-1}$ sea inestable, aumentando considerablemente el valor absoluto de los coeficientes.
III) El término σ^2 puede no ser constante, y variar entre las diversas observaciones, lo que se denomina, *heterocedasticidad*.
IV) Los coeficientes β se suponen constantes, lo cual puede no ocurrir por diversas razones, como cambios de tendencia en el tiempo, poblaciones no homogéneas, etc..

El problema I) se constata ex-post haciendo un análisis de los residuos $y_i - \widehat{y_i}$. Si estos residuos no tienen un comportamiento aleatorio, entonces hay un error de especificación del modelo.

Se ha intentado abordar el problema II) con una alternativa de estimación sesgada, pero de menor varianza que los estimadores de mínimos cuadrados, conocida con el

nombre de *Regresión Ridge* [HK70], y que depende de un parámetro λ. Otros intentos para afrontar el problema de la multicolinealidad son soluciones como la *Regresión de Componentes Principales*, la *Regresión Lasso*, o la de *Mínimos Cuadrados Parciales*, cada una de las cuales tiene sus propios parámetros. En cualquier caso, la solución Ridge es el enfoque que produce una mayor continuidad en el comportamiento de los coeficientes para los diversos valores que puede tomar el parámetro[HTF01].

El problema IV) ha llevado a estudiar la regresión parcializada por tramos, para recalcular los parámetros para cada tramo. La *Regresión Ecológica*, donde se trata de modelar el comportamiento de grupos de personas en lugar de individuos, es otro caso en que, frecuentemente, no se cumple el supuesto de coeficientes constantes.

Una técnica utilizada dentro de la regresión probabilística es la *Selección de Variables* (o regresión paso a paso), orientada a excluir las variables independientes que no son relevantes para el modelo. De esta forma, se gana en simplicidad del modelo, al tener menos variables incorporadas, con un aumento de la incertidumbre muy pequeño.

Existe, en esta regresión paso a paso, el enfoque *hacia adelante* que va seleccionando una a una las variables que se incorporan al modelo, el enfoque *hacia atrás*, que, partiendo del conjunto de variables, va eliminando una a una las variables no relevantes al modelo, y el enfoque *por etapas*, que permite eliminar una variable que ha sido seleccionada en un paso anterior del proceso.

Una consideración que se puede tomar en cuenta es la *robustez de la regresión*, entendida en términos de que la estimación no dependa de ciertos puntos extremos (outliers). En la figura 1.2 se puede apreciar el efecto de un punto extremo sobre la estimación del modelo de regresión. En (a) no existen puntos extremos y la estimación se ajusta bien a los datos. Por el contrario, en (b), la curva de ajuste se distancia de la tendencia de la mayoría de los datos por la existencia de un punto extremo.

Por último, una situación específica de regresión que se tiene cuando y_i sólo puede tomar dos valores, por ejemplo, para un estudio de quiebras de empresas, donde y_i puede indicar si la empresa quebró o no quebró, se denomina *Regresión Logística*. En este caso, esta variable no sigue una distribución normal, sino una distribución binomial. Para ello, se hace una transformación del modelo estándar, quedando el modelo logístico como

$$\pi(x) = \frac{e^{x\beta}}{1 + e^{x\beta}} \tag{1.11}$$

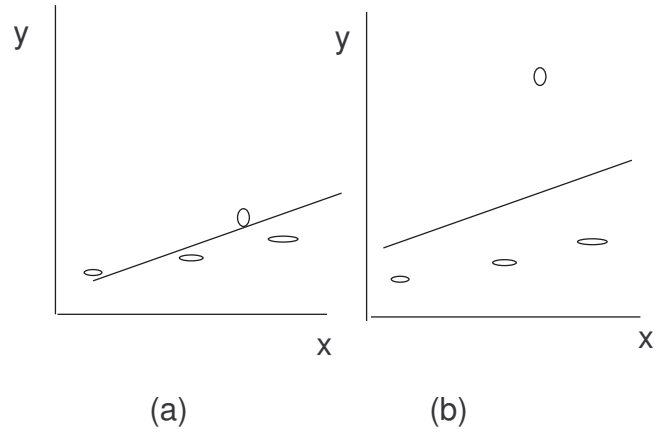

Figura 1.2: *Ejemplo de regresión no rubusta*

En estos casos, el método de estimación que se usa se denomina de *máxima verosimilitud*, y consiste en elegir los parámetros de un modelo probabilístico que tengan mayor probabilidad de ocurrir, de acuerdo a los datos disponibles [YC04].

1.2. Los Conjuntos Difusos y la Aritmética Difusa

Tras nuestra introducción del primero de los pilares sobre los que se fundamenta este trabajo de investigación, el Análisis de Regresión, nos disponemos ahora a presentar el segundo: la Teoría de Conjuntos Difusos[Zad65]. Esta teoría nos va a permitir trabajar con datos afectados por imprecisión e incertidumbre en nuestro análisis. En esta sección, se mostrarán los elementos básicos de la Teoría de Conjuntos Difusos, prestando especial atención a aquellos aspectos que tienen relación directa con las diversas modalidades en que se ha desarrollado la Regresión Difusa. También se introduce la notación que más frecuentemente se utiliza en esta teoría y que se usará en el resto de la memoria. Por último, se hará un breve resumen de la aplicación de esta teoría en la definición de la Aritmética Difusa, aritmética que es parte fundamental en la construcción de los modelos de Regresión Difusa.

1.2.1. Fundamentos de la Teoría de Conjuntos Difusos

Dos de las fuentes que *contaminan* normalmente la información son la *imprecisión* que tiene en su expresión y la *incertidumbre* que puede provocar la fuente que nos la proporciona. El ser humano se desenvuelve con sorprendente facilidad a la hora de manejar este tipo de información pero, sin embargo, cuesta explicar qué procedimientos sigue para ello.

El hallazgo de modelos matemáticos para poder hacer frente a información imperfecta ha sido un punto de gran interés en el mundo de la investigación, aportando teorías como la de la Probabilidad [Fel71], la de la Evidencia [Sch76], o la de los Factores de Certeza [SB75]. En 1965 Lotfi A. Zadeh propuso una de las herramientas más valiosas a la hora de trabajar con este tipo de información: la Teoría de Subconjuntos Difusos [Zad65]. Desde la aparición de esta teoría son incontables las aplicaciones que se han hecho de ella en el mundo de la investigación en general, y en particular en el área de las Ciencias de la Computación. Veamos algunos de sus conceptos fundamentales. Para un estudio más completo se puede consultar [TAT95, DP80a, KY95a].

El concepto de *conjunto difuso* se relaciona con una colección de objetos que pueden *pertenecer* a él con un cierto grado, desde un grado máximo de 1 para la completa pertenencia, a un grado mínimo de 0 para la no pertenencia, pasando por todos los valores intermedios.

Definición 1.1 (Subconjuntos difusos)**.** *Supongamos que $X = \{x\}$ es el conjunto de todos los posibles elementos respecto a un concepto (universo del discurso). Entonces un subconjunto difuso (o conjunto difuso, por simplicidad) A en X está definido como un conjunto de pares ordenados $(x, \mu_A(x))$, donde $x \epsilon X$ y $\mu_A(x)$ es el grado de pertenencia de x en A. $\mu_A : X \longrightarrow [0,1]$ es la función de pertenencia de A, y $\mu_A(x) \epsilon [0,1]$.*

Definición 1.2 (Negación)**.** *Se denomina negación a toda función $c : [0,1] \rightarrow [0,1]$ que verifique las siguientes propiedades:*

1. $c(0) = 1$ *y* $c(1) = 0$ ***(Frontera)***
2. $\alpha \leq \beta \Rightarrow c(\alpha) \geq c(\beta), \forall \alpha, \beta \in [0,1]$***(Monotonía)***

Para mejorar las negaciones, desde un punto de vista práctico, también se les exige que sean *continuas* e *involutivas* ($c(c(\alpha)) = \alpha, \forall \alpha \in [0,1]$).

Tabla 1.1: Negaciones habituales

Nombre	Definición
Estándar	$c(\alpha) = 1 - \alpha$
Umbral	$c(\alpha) = \begin{cases} 1 & \text{si } \alpha < umbral \\ 0 & \text{si } \alpha \geq umbral \end{cases}$

La tabla 1.1 muestra algunas funciones que se ajustan a la anterior definición. Mediante el uso de negaciones, el complemento de un conjunto difuso puede definirse de la siguiente manera:

$$\mu_{\overline{A}}(x) = c(\mu_A(x)) \tag{1.12}$$

Nosotros utilizaremos la negación estándar para calcular dicho complemento en esta memoria.

Definición 1.3. *(Cardinalidad escalar) La cardinalidad escalar de un conjunto difuso A en el conjunto finito X, se define como [LT72]*

$$|A| = \Sigma_{x \epsilon X} \mu_A(x) \tag{1.13}$$

Cuando X es finito, ($\{x_1, ..., x_n\}$), el conjunto difuso A se puede expresar como

$$A = \mu_A(x_1)/x_1 + ... + \mu_A(x_n)/x_n = \Sigma_{i=1}^{n} \mu_A(x_i)/x_i \tag{1.14}$$

En el caso de que X no sea finito, la notación es

$$A = \int_X \mu_A(x)/x \tag{1.15}$$

La Teoría de Conjuntos Difusos permite definir de diferente forma las operaciones de intersección y de unión entre conjuntos difusos, dando origen a diversas normas triangulares (t-normas) y t-conormas [KY95b].

Definición 1.4 (t-norma)**.** *Se denomina t-norma a toda función* $t : [0, 1] \times [0, 1] \rightarrow [0, 1]$ *que verifique las siguientes propiedades:*

1. $t(\alpha, 1) = \alpha, \forall\alpha \in [0, 1]$ *(Frontera)*
2. $\beta \leq \gamma \Rightarrow t(\alpha, \beta) \leq t(\alpha, \gamma), \forall\alpha, \beta, \gamma \in [0, 1]$*(Monotonía)*
3. $t(\alpha, \beta) = t(\beta, \alpha), \forall\alpha, \beta \in [0, 1]$ *(Commutativa)*
4. $t(\alpha, t(\beta, \gamma)) = t(t(\alpha, \beta), \gamma), \forall\alpha, \beta, \gamma \in [0, 1]$ *(Asociativa)*

La tabla 1.2 muestra algunas funciones que se ajustan a la anterior definición.

Tabla 1.2: Algunas t-normas habituales

Nombre	Definición
Mínimo	$t(\alpha, \beta) = min(\alpha, \beta)$
Producto	$t(\alpha, \beta) = \alpha \cdot \beta$
Lukasiewicz	$t(\alpha, \beta) = max(0, \alpha + \beta - 1)$

Mediante el uso de t-normas, la intersección de dos conjuntos difusos puede definirse de la siguiente manera:

$$\mu_{A\cap B}(x) = t(\mu_A(x), \mu_B(x)) \tag{1.16}$$

Para referirnos a funciones de esta familia utilizaremos la notación $\otimes$.

Definición 1.5 (t-conorma)**.** *Se denomina t-conorma a toda función* $u : [0, 1] \times [0, 1] \rightarrow [0, 1]$ *que verifique las siguientes propiedades:*

1. $u(\alpha, 0) = \alpha, \forall\alpha \in [0, 1]$ *(Frontera)*
2. $\beta \leq \gamma \Rightarrow u(\alpha, \beta) \leq u(\alpha, \gamma), \forall\alpha, \beta, \gamma \in [0, 1]$*(Monotonía)*
3. $u(\alpha, \beta) = u(\beta, \alpha), \forall\alpha, \beta \in [0, 1]$ *(Commutativa)*
4. $u(\alpha, u(\beta, \gamma)) = u(u(\alpha, \beta), \gamma)$ *(Asociativa)*

La tabla 1.3 muestra algunas funciones que se ajustan a la anterior definición.

Mediante el uso de t-conormas, la unión de dos conjuntos difusos puede definirse de la siguiente manera:

$$\mu_{A\cup B}(x) = u(\mu_A(x), \mu_B(x)) \tag{1.17}$$

Tabla 1.3: Algunas t-conormas habituales

Nombre	Definición
Máximo	$u(\alpha, \beta) = max(\alpha, \beta)$
Suma algebraica	$u(\alpha, \beta) = \alpha + \beta - \alpha\beta$
Lukasiewicz	$u(\alpha, \beta) = min(1, \alpha + \beta)$

Para referirnos a funciones de esta familia utilizaremos la notación $\oplus$.

En nuestro trabajo, se considerará como t-norma estándar (intersección) al mínimo y como t-conorma estándar (unión) al máximo.

Uno de los más importantes conceptos relativos a los conjuntos difusos es el concepto de α-corte.

Definición 1.6 (α-corte)**.** *Dado un conjunto difuso A definido en X y un número* $\alpha\epsilon[0, 1]$*, el* α*-corte* A^α *de A es el intervalo*

$$A^\alpha = \{x|\mu_A(x) \geq \alpha\}$$

Definición 1.7 (Soporte de un conjunto difuso)**.** *El soporte de un conjunto difuso A, dentro de un conjunto universal X, es el conjunto convencional (crisp) que contiene todos los elementos de X que tienen un grado de pertenencia mayor que 0 en A:*

$$sop(A) = \{x\epsilon X, \mu_A(x) > 0\} \tag{1.18}$$

Una importante característica de los conjuntos difusos definidos en R es su convexidad, similar a la convexidad clásica. Para que un conjunto A sea convexo, se requiere que todos sus $\alpha - cortes$ sean convexos y acotados.

Teorema 1.1. *A es un conjunto difuso convexo, si y solo si*

$$\mu_A(\lambda x_1 + (1 - \lambda)x_2) \geq min[\mu_A(x_1), \mu_A(x_2)] \tag{1.19}$$

para $x_1, x_2\ \epsilon R$ *y todo* $\lambda\epsilon\ [0, 1]$ *[KY95b].*

Gracias a la definición de los α-cortes se puede hacer otra representación de los conjuntos difusos muy útil para el manejo numérico de cantidades difusas. El llamado Teorema de la Descomposición establece que

$$A = \bigcup_{\alpha \varepsilon [0,1]} \alpha . A^{\alpha} \tag{1.20}$$

donde $\alpha . A^{\alpha}$ es un conjunto difuso en X cuya función de pertenencia es

$$\mu_{\alpha . A^{\alpha}}(x) = \begin{cases} \alpha & para\ x \in A^{\alpha} \\ 0 & para\ x \notin A^{\alpha} \end{cases}$$

Terminamos presentando la definición de número difuso.

Definición 1.8 (Número difuso)**.** *Sea A un conjunto difuso definido sobre el conjunto R de los números reales. Si los α-cortes de A son intervalos cerrados para todo $\alpha\ \varepsilon(0, 1]$ y el soporte de A es acotado, se dice que A es un número difuso.*

1.2.1.1. Funciones de Pertenencia

De la misma definición de los conjuntos difusos viene el término *funciones de pertenencia*. Las funciones de pertenencia pueden tener distintas características y formas. La Regresión Difusa se ha centrado en las llamadas *funciones de pertenencia LR*, que se construyen a partir de dos funciones L y R, que para nuestro propósito generalmente se asumen como dos funciones lineales y se denominan triangulares. Se definen como:

$$\mu_A(x) = \begin{cases} L(\frac{m-x}{a}) & a > 0, \forall x \leq m \\ R(\frac{x-m}{b}) & b > 0, \forall x \geq m \end{cases}$$

donde la función L es tal que: (1) L(-x)=L(x), (2) L(0)=1, y (3) L es creciente en $[0, +\infty]$, y las mismas condiciones para la función R [KY95b].

Estos conjuntos difusos LR son los más empleados por la Regresión Difusa, y la notación para uno de estos conjuntos difusos, A, será $(a, p_a, q_a)_{LR}$, donde a es el valor de mayor pertenencia, o *valor central*, p_a es la *extensión izquierda*, y q_a es la *extensión derecha*, de manera que una notación equivalente de tipo invervalar es: $[a - p_a, a, a + q_a]_{LR}$.

Para conjuntos difusos LR, el soporte para un α-corte viene dado por

$$S_A(\alpha) = [a - p_a L^{-1}(\alpha), a + q_a R^{-1}(\alpha)] \tag{1.21}$$

Los conjuntos LR más comúnmente usados son las funciones de pertenencia triangulares, en que las funciones de extensión se definen como $L(x) = R(x) = max\{0, 1 - x\}$.

Con las mismas funciones L y R, se definen las funciones de pertenecia LR trapezoidales (a,b,c,d), que son funciones LR en sus extensiones y tienen un intervalo de máxima pertenencia:

$$\mu_A(x) = \begin{cases} 0 & cuando\, x < a\, y\, x > d \\ L(\frac{a-x}{a-b}) & cuando\, a \leq x \leq b \\ 1 & cuando\, b \leq x \leq c \\ R(\frac{d-x}{d-c}) & cuando\, c \leq x \leq d \end{cases}$$

Otro tipo de funciones de pertenencia LR son las funciones de pertenencia normales $A = (a, \sigma)$, que tienen una similitud con la distribución de probabilidades normal o de Gauss y quedan definidas como:

$$\mu_A(x) = exp\left[-(\frac{x-a}{\sigma})^2\right]\ x\varepsilon R\ (\sigma > 0) \tag{1.22}$$

que puede ser generalizado como una función de pertenencia $A = (a, D_A)_e$ exponencial, definida por

$$\mu_A(x) = exp\left[-((x-a)' D_A^{-1}(x-a)\right]\ x\varepsilon R \tag{1.23}$$

Definir una función de pertenencia particular depende de cada contexto. Por ejemplo, la función de pertenencia que representa *temperatura alta* puede ser muy distinta dependiendo de las situaciones específicas. En muchos casos de Regresión Difusa, se ha recurrido a la familia de funciones de pertenencia triangulares y al disponerse de mediciones con un valor mínimo, un valor máximo y un valor relevante (por ejemplo, valor de cierre para el precio de una acción), se pueden precisar los parámetros de las funciones de pertenencia: $minimo = a - p_a$, $maximo = a + q_a$ y el valor relevante igual al valor central de la función de pertenencia.

La figura 1.3 muestra tres ejemplos de las funciones de pertenencia LR centradas en 0, definidas por la relación genérica $\mu = 1 - x^{1/r}$.

En caso de tener funciones de pertenencia no acotadas en un extremo, como es el caso de las funciones de pertenencia que representan expresiones como *muy alto* lo

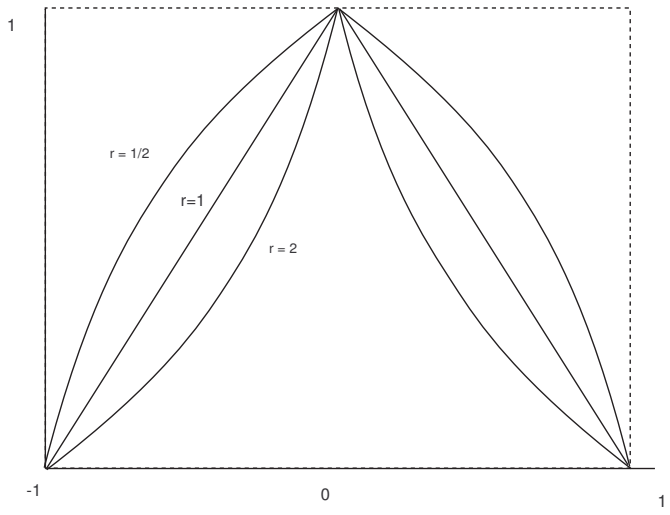

Figura 1.3: *Funciones de pertenencia: 1 - $x^{1/r}$*

que tiene sentido estimar es sólo la extensión acotada y el punto central, puesto que la otra extensión tiene valor de pertenencia constante 1.

Cuando un conjunto difuso representa conceptos lingüísticos tales como *muy pequeño*, *pequeño* y están interpretados en un contexto particular, su formalización es generalmente llamada variable lingüística [Zad75a, Zad75b, Zad75c] (ver figura 1.4).

Definición 1.9 (Variable lingüística)**.** *Una variable lingüística está caracterizada por una quíntupla (X, T(X),U,G,M), en la que:*

1. *X es el nombre de la variable.*
2. *T(X) es el conjunto de valores lingüísticos de X. Cuando los elementos de T(X) tienen una sola palabra se denominan términos atómicos. En caso contrario se habla de términos compuestos.*
3. *U es el universo de discurso de la variable.*
4. *G es una regla sintáctica (normalmente en forma de gramática) que determina la forma de generar valores de T(X).*
5. *M es una regla semántica que asocia a cada elemento de T(X) su significado. Para cada valor $L \in T(X)$, $M(L)$ será un subconjunto difuso de U.*

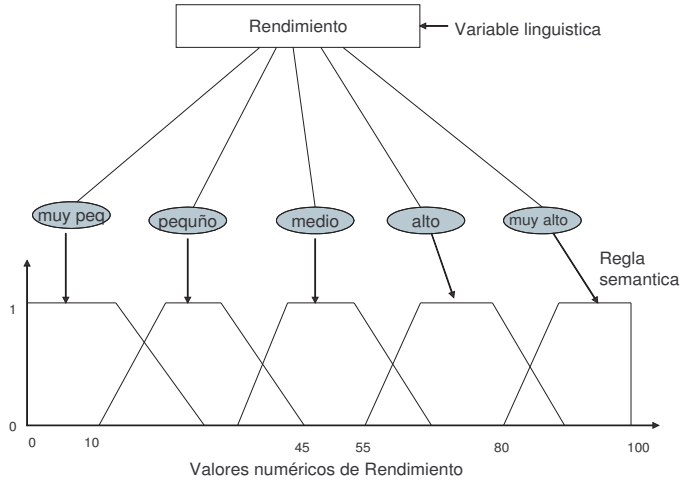

Figura 1.4: *Un ejemplo de variable linguística*

Las variables lingüísticas suelen definirse de forma simple, imponiendo algunas restricciones a la anterior quíntupla, de forma que resulten sencillas de manejar. Estas restricciones están normalmente establecidas sobre el número de términos y la forma de obtener la semántica de los términos generados.

1.2.1.2. Medidas Difusas

En nuestra memoria de investigación vamos a utilizar dos medidas de especial importancia: la medida de posibilidad y de necesidad. Estas medidas se definen axiomáticamente a partir del concepto de *medida difusa*.

Definición 1.10 (Medida difusa)**.** *Dado un conjunto universal X y una familia no vacía de subconjuntos de X, C, una medida difusa, g, se define en* $\{X, C\}$*, con las siguientes condiciones:*

I*)*

$$i) g(\emptyset) = 0 \; y \; g(X) = 1 \tag{1.24}$$

II*)*

$$ii) si \; A \subseteq B, \; entonces \; g(A) \leq g(B) \tag{1.25}$$

III*) Para una secuencia creciente* $A_1 \subset A_2 \subset ...$ *en C, si* $\bigcup_{i=1}^{\infty} A_i \epsilon C$*, entonces*

$$lim_{i\to\infty} g(A_i) = g(\bigcup_{i=1}^{\infty} A_i) \tag{1.26}$$

IV*) Para una secuencia decreciente* $A_1 \supset A_2 \supset ...$ *en C, si* $\bigcap_{i=1}^{\infty} A_i \epsilon C$*, entonces*

$$lim_{i\to\infty} g(A_i) = g(\bigcap_{i=1}^{\infty} A_i) \tag{1.27}$$

Definición 1.11 (Medida de necesidad)**.** *Sea Nec una medida difusa. Nec se llamará una medida de necesidad si*

$$Nec(\bigcap_{k\varepsilon K} A_k) = inf_{k\varepsilon K} Nec(A_k) \tag{1.28}$$

para toda familia de conjuntos $\{A_k | k\varepsilon K\}$ *en X, donde K es un archivo de indices arbitrario.*

Definición 1.12. *(Medida de posibilidad) Sea Pos una medida difusa. Pos será medida de posibilidad si cumple:*

$$Pos(\bigcup_{k\varepsilon K} A_k) = sup_{k\varepsilon K} Pos(A_k) \tag{1.29}$$

Esta definición da origen al nombre de *Regresión posibilística* en la mayoría de los Modelos de Regresión Difusa.

Dubois and Prade [DP83] propusieron los siguientes índices para comparar dos números difusos A y B:

$$Pos(A = B) = sup\{min\{A(x), B(x)\}, x\varepsilon\Re\} \tag{1.30}$$

$$Nes(A \subseteq B) = inf\{max\{1 - A(x), B(x)\}, x\varepsilon\Re\} \tag{1.31}$$

El primer índice indica el grado de posibilidad de que A sea igual a B, mientras el segundo índice señala el grado de necesidad en que A incluye B. Estos índices han sido ampliamente empleados en los diversos modelos de Regresión Difusa.

1.2.2. Aritmética Difusa

Con los elementos de la Teoría de Conjuntos Difusos definidos, pasamos ahora a presentar el *Principio de Extensión* [Zad75a] que permite extender las funciones de operaciones (aritméticas o algebraicas) desde los conjuntos tradicionales a los conjuntos difusos.

De forma sencilla, si $f : X \longrightarrow Y$ es una función (operación) sobre números reales, se puede extender esta operación sobre conjuntos difusos de la siguiente manera: siendo A y B conjuntos difusos con sus respectivas funciones de pertenencia $\mu_A(x)$ y $\mu_B(y)$, entonces la función (operación) $C = f(A, B)$ tiene la función de pertenencia $\mu_C(z)$ dada por la expresión:

$$\mu_C(z) = sup\,[min(\mu_A(x), \mu_B(y)) :\ x, y\varepsilon\Re,\ z = f(x, y)] \tag{1.32}$$

Cuando los conjuntos A y B son finitos, se puede reemplazar el supremo por el máximo.

Hay otras formas de formular el principio de extensión, fundamentalmente recurriendo a la función f^{-1}[KY95b].

Por ejemplo, las funciones máximo y mínimo de los números reales, que como tales definen la t-conorma y t-norma estándares, pueden extenderse a los conjuntos difusos como

$$MIN(A, B)(z) = sup\,[min(\mu_A(x), \mu_B(y)) :\ x, y\varepsilon\Re,\ z = min(x, y)] \tag{1.33}$$

$$MAX(A, B)(z) = sup\,[min(\mu_A(x), \mu_B(y)) :\ x, y\varepsilon\Re,\ z = max(x, y)] \tag{1.34}$$

que produce como resultado conjuntos difusos distintos, por lo general, a los resultantes de la función real.

El antecedente teórico de la operaciones de conjuntos difusos está en las operaciones entre intervalos. Considerando el Teorema de la Descomposición y que cada α-corte de un número difuso es un intervalo cerrado en los números reales para todo $\alpha\varepsilon(0, 1]$, se pueden definir las operaciones aritméticas en los números difusos en

términos de las operaciones aritméticas de sus α-corte (es decir, operaciones aritméticas en intervalos cerrados).

Una propiedad general de las operaciones aritméticas en intervalos cerrados las define como

$$[a,b] * [d,e] = \{f * g | a \leq f \leq b, d \leq g \leq e\} \tag{1.35}$$

, donde * puede representar la adición +, la resta -, la multiplicación , y la division /.

Excepto cuando $0\varepsilon[d,e]$, el resultado de una operación aritmética en intervalos cerrados es también un intervalo cerrado.

Las cuatro operaciones aritméticas en intervalos cerrados son definidas de la siguiente manera:

$$[a,b] + [d,e] = [a+d, b+e] \tag{1.36}$$

$$[a,b] - [d,e] = [a-d, b-d] \tag{1.37}$$

$$[a,b].[d,e] = [min(ad,ae,bd,be), max(ad,ae,bd,be)] \tag{1.38}$$

$$[a,b]/[d,e] \quad = \quad [a,b].[1/e, 1/d] =$$

$$[min(a/d, a/e, b/d, b/e), max(a/d, a/e, b/d, b/e)] \tag{1.39}$$

En consecuencia, la aritmética de los números difusos se puede fundamentar en la aritmética de los intervalos cerrados, como también se puede fundamentar en el principio de extensión.

Definición 1.13 (Operaciones aritméticas difusas)**.** *Sean A y B numeros difusos y sea * el símbolo que representa una de los cuatro operaciones aritméticas básicas. Entonces, se define el conjunto difuso en $\Re$, A*B, definiendo sus α-cortes, $(A * B)^{\alpha}$ como*

$$(A * B)^{\alpha} = A^{\alpha} * B^{\alpha} \tag{1.40}$$

*con lo que se puede demostrar que A*B es también un número difuso.*

En el caso de funciones de pertenencia LR no simétricas, que se utilizarán fundamentalmente en el Análisis de Regresión Difusa, para sumar los números difusos $A = (a, p_a, q_a)_{LR}$ y $B = (b, p_b, q_b)_{LR}$, si se considera, primero, su extensión izquierda, para un cierto nivel de pertenencia w, se tiene la relación genérica [DP80b]

$$L((a - x)/p_a) = w = L((b - y)/p_b) \tag{1.41}$$

de donde se puede deducir una relación para la operación suma, en que aparece la función L:

$$x + y = a + b - (p_a + p_b)L^{-1}(w) \tag{1.42}$$

y resultará un número z, que define la función de pertenencia tal que

$$L\left(\frac{a + b - z}{p_a + p_b}\right) = w \tag{1.43}$$

y si de la misma manera se procede con la extensión derecha y la función R, se tiene, que la *suma* de estos dos números está dada por la expresión

$$(a, p_a, q_a)_{LR} + (b, p_b, q_b)_{LR} = (a + b, p_a + p_b, q_a + q_b)_{LR} \tag{1.44}$$

Esta relación se puede generalizar, para n números difusos LR, mediante

$$\sum_{i=1}^{n}(a_i, p_i, q_i)_{LR} = (\sum_{i=1}^{n} a_i, \sum_{i=1}^{n} p_i, \sum_{i=1}^{n} q_i)_{LR} \tag{1.45}$$

donde se supone que los n números difusos tienen la misma función de pertenencia LR. En caso que cada número difuso tenga las funciones de pertenencia L_i y R_i, entonces la en la última expresión se cambia LR por:

$$L = \left(\sum_{i=1}^{n} \frac{p_i}{\sum_{j=1}^{n} p_j} L_i^{-1}\right)^{-1} \tag{1.46}$$

$$R = \left(\sum_{i=1}^{n} \frac{q_i}{\sum_{j=1}^{n} q_j} R_i^{-1} \right)^{-1} \tag{1.47}$$

Para el *producto de un escalar por un número difuso*, siguiendo la misma lógica, se obtienen valores dependiendo del signo del escalar. Para el escalar positivo se tiene:

$$\lambda(a, p_a, q_a)_{LR} = (\lambda a, \lambda p_a, \lambda q_a)_{LR} \tag{1.48}$$

y para la multiplicación con un escalar negativo:

$$\lambda(a, p_a, q_a)_{LR} = (\lambda a, -\lambda q_a, -\lambda p_a)_{RL} \tag{1.49}$$

De las operaciones (1.44) y (1.49) queda definida la *sustracción* de dos números difusos:

$$(a, p_a, q_a)_{LR} - (b, p_b, q_b)_{LR} = (a - b, p_a + q_b, q_a + p_b)_{LR} \tag{1.50}$$

es decir, de la suma y de la resta de números difusos con función de pertenencia LR, resultan números difusos también con función de pertenencia LR.

No ocurre lo mismo con la *multiplicación*. Para números difusos positivos, la expresión (1.42) que se construyó para calcular la suma difusa, para la multiplicación difusa es

$$x.y = a.b - (ap_b + bp_a)L^{-1}(w) + p_a P_b (L^{-1}(w))^2 \tag{1.51}$$

Esta expresión, por lo general, no produce como resultado un número difuso de tipo LR. Sin embargo, si se descarta el término $p_a P_b (L^{-1}(w))^2$, considerando que p_a y p_b son pequeños en comparación con a y b, la solución de la multiplicación difusa es

$$(a, p_a, q_a)_{LR}.(b, p_b, q_b)_{LR} \approx (ab, ap_b + bp_a, aq_b + bq_a)_{LR} \tag{1.52}$$

Si A es negativo y B es positivo, la aproximación para la multiplicación difusa es

$$(a, p_a, q_a)_{RL}.(b, p_b, q_b)_{LR} \approx (ab, -aq_b + bp_a, -ap_b + bq_a)_{RL} \tag{1.53}$$

y si A y B son negativos, la aproximación para el cálculo del producto es

$$(a, p_a, q_a)_{RL}.(b, p_b, q_b)_{LR} \approx (ab, -aq_b - bq_a, -ap_b - bp_a)_{RL} \tag{1.54}$$

Para calcular el *inverso* de un A, número difuso LR positivo, se tiene que $\mu_{A^{-1}}(x)$ = $\mu_A(1/x)$, para $\forall x \neq 0$, y en general la función de pertenencia A^{-1} no es ni LR ni RL. Si se considera una vecindad de 1/a, se tiene que

$$\frac{1 - ax}{p_a x} \approx \frac{1/a - x}{p_a/a^2} \tag{1.55}$$

que sigue una función de pertenencia RL, por lo que se tiene la aproximación

$$(a, p_a, q_a)^{-1}_{LR} \approx (a^{-1}, q_a a^{-2}, p_a a^{-2})_{RL} \tag{1.56}$$

Si el número difuso A es negativo, se puede aplicar la relación $-(A^{-1}) = (-A)^{-1}$.

Considerando la *división* de dos números difusos como la multiplicación del primero por el inverso del segundo, se tiene la aproximación:

$$(a, p_a, q_a)_{LR} : (b, p_b, q_b)_{RL} \approx (a/b, aq_b + bp_a, ap_b + bq_a)_{LR} \tag{1.57}$$

La figura 1.5 muestra un ejemplo de aproximación a una función de pertenencia LR resultante de la división de dos conjuntos difusos LR [OS03].

Todo este desarrollo está basado en el Principio de Extensión que utiliza la t-norma triangular estándar en que la intersección entre A y B es el mínimo.

Otro trabajo de interés sobre aritmética difusa[Cho03] define la *Aritmética de Función Inversa* $L^{-1} - R^{-1}$ donde, por ejemplo, la adición es definida, para el α-corte h, como

$$A(h) \oplus B(h) = (L_A^{-1}(h) + L_B^{-1}(h), L_A^{-1} + R_B^{-1}(h), R_A^{-1}(h) + R_B^{-1}(h), R_A^{-1}(h) + R_B^{-1}) \tag{1.58}$$

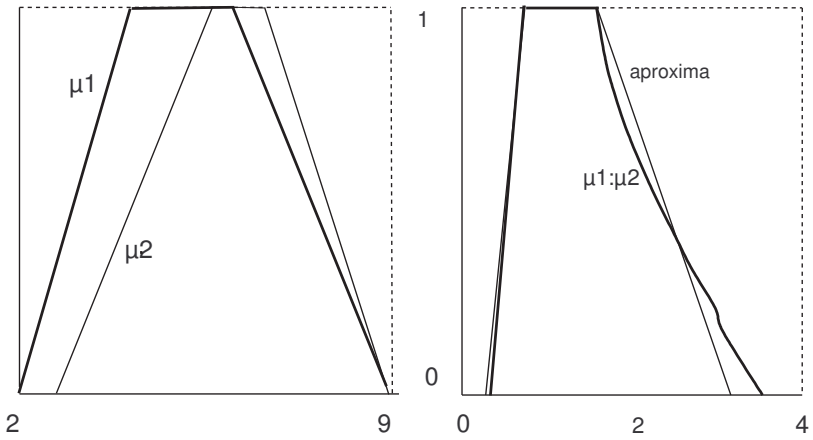

Figura 1.5: *Ejemplo de aproximación en division de 2 conjuntos difusos*

Para aplicaciones de Regresión Difusa, también se ha usado una aritmética difusa basada en la *t-norma más débil (MD)*,

$$T_{MD}(x,y) = \begin{cases} a & si\ b = 1 \\ b & si\ a = 1 \\ 0 & en\ otro\ caso \end{cases}$$

con la aplicación del principio de extensión en este nuevo contexto, que produce como resultados las siguientes operaciones difusas con la t-norma más débil [Hon06]:

$$(a, p_a, q_a)_{LR} \oplus_{MD} (b, p_b, q_b)_{LR} = (a+b, max(p_a, p_b), max(q_a, q_b))_{LR} \quad (1.59)$$

Para el producto *más débil difuso*, la expresión resultante es:

$$A \underset{MD}{\bigotimes} B = \begin{cases} (ab, max(p_a b, p_b a), max(q_a b, q_b a)_{LR} & para\ a, b > 0 \\ (ab, max(q_a b, q_b a), max(p_a b, p_b a))_{RL} & para\ a, b < 0 \\ (0, p_a b, q_a b)_{LR} & para\ a = 0, b > 0 \\ (0, -q_a b, -p_a b)_{RL} & para\ a = 0, b < 0 \\ (0, 0, 0) & para\ a = 0, b = 0 \end{cases}$$

De forma similar, también puede encontrarse en la literatura una aritmética difusa basada en la *t-norma débil acotada*[UW05].

Por último, otra forma que ha sido utilizada dentro de la Regresión Difusa, es el uso de desdifusificadores. La implementación de funciones de pertenencia no simétricas que utilizaremos más adelante en esta memoria, hace más interesante esta opción. Un trabajo reciente [KL05] del 2005, define como desdifusificador, una medida de entropía difusa, h, sobre un conjunto difuso con función de pertenencia triangular, A, resultando:

$$H(A) = \int_{a-p_a}^{a} h(\mu_A(x))p(x)dx + \int_{a}^{a+q_a} h(\mu_A(x))p(x)dx \tag{1.60}$$

con lo que se tiene, separando los dos términos de la expresión anterior, un indicador de entropía del lado izquierdo y otro del lado derecho.

Un método alternativo de desdifusificación utiliza la *aritmética difusa ponderada*[Cha01a], resultando un sistema de tres ecuaciones de mínimos cuadrados: extremo izquierdo, centro y extremo derecho. Por ejemplo, la *adición difusa ponderada* de A y B se calcula como

$$A + B = \frac{[\int_{\mu}(\mu_{A_L} + \mu_{B_L})\mu d\mu]_L + [\int_{\mu}(\mu_{A_R} + \mu_{B_R})\mu d\mu]_R}{\int \mu d\mu} \tag{1.61}$$

lo que resulta en

$$A + B = (a + b) + \frac{1}{6}[(q_a + q_b) - (p_a + p_b)] \tag{1.62}$$

lo que permite operar como con números reales para la resolución de su modelo difuso.

Para terminar debemos resaltar que la aritmética difusa cumple un doble papel en la definición de los modelos de Regresión Difusa:

- Por una parte permite relacionar diversos números difusos, fundamentalmente a través de la suma y la multiplicación por un escalar, para operar con los modelos lineales (o intrínsicamente lineales).
- Por otra parte, la aritmética difusa permite la creación de nuevos conjuntos a partir de conjuntos difusos existentes, que pueden incorporarse a un modelo de regresión. Por ejemplo, si se dispone de dos precios, podría crearse la variable *máximo* entre los dos precios, o la variable *multiplicación* de los precios.

1.3. Justificación de la Regresión Difusa

Hemos visto en este capítulo una breve descripción del Análisis de Regresión y de la Teoría de Conjuntos Difusos que fundamentan nuestro trabajo en el campo del Análisis de Regresión Difusa. Vamos a dedicar nuestro último esfuerzo en este capítulo señalando la necesidad de disponer de dicha técnica de análisis.

La primera justificación para la Regresión Difusa radica en que el manejo de cantidades afectadas por imprecisión e incertidumbre no está considerado en la Regresión Probabilística. Y sin embargo, hay muchas magnitudes cuantitativas que pueden representarse adecuadamente mediante números difusos: mediciones con márgenes de error; carga eléctrica requerida por una generadora por unidad de tiempo; valor de una acción en un día, una semana o un mes, en una bolsa de comercio; precio de las materias primas (cobre, oro, plata, azúcar, etc.) en los mercados internacionales donde se intercambian; valor de las monedas frente a otras monedas referenciales, en los mercados financieros; votación de un candidato en una elección con cédula única, en que los votos nulos pueden reflejar preferencias mal marcadas (en la reciente elección de primera vuelta presidencial en Perú, que se vota con cédula única, los votos nulos y blancos sumaron el 16 por ciento del total, y la diferencia entre el segundo y tercer candidato, para pasar a la segundo vuelta, fue de menos del 0.5 por ciento); la cifra indicada en un censo, que se obtiene después de que la información original ha sido leída, validada y corregida automáticamente, con asignaciones para los datos dudosos que estan basadas en supuestos y/o son asignaciones aleatorias; y un larguísimo etcétera.

Este tipo de datos, fundamentalmente cuantificables, pero con una imprecisión e incertidumbres provenientes de diversos orígenes, justifica la creación de un campo análitico propio para la Regresión Difusa.

Además, la Teoría de Conjuntos Difusos maneja nuevos modelos de medidas, como las medidas de posibilidad y de necesidad, que se traducen en nuevas visiones para afrontar el análisis de regresión.

Los números intervalares, que, como hemos visto, son una versión simplificada de los números difusos, también han sido considerados como números de interés para poder explicar su comportamiento [SIT04a]. Y la Regresión Difusa también maneja estos números intervalares como un caso particular.

La presencia de la incertidumbre, en la Regresión Probabilística queda plasmada en los intervalos de confianza, generalmente con un nivel del 95 %, lo que es atribuido

a factores aleatorios. Esta es la única forma de incertidumbre que maneja la Teoría de Probabilidades.

La aleatoriedad puede considerarse uno de los componentes de la vaguedad, junto a la imprecisión. Sin embargo, otras formas de incertidumbre se pueden encontrar en la ambigüedad, la incongruencia, problemas de especificación, factores no considerados por complejidad del sistema... En resumen, hay diversas fuentes de incertidumbre, entre las cuales la aleatoriedad es sólo una de ellas.

El hecho de que no sólo se considere la aleatoriedad como origen de incertidumbre, hace posible incorporar dentro de los modelos de Regresión Difusa, de forma natural, información que se dispone a priori de los problemas, que generalmente se traduce en restricciones a los modelos, y que ayudan a obtener estimaciones más ajustadas a la realidad.

Estas restricciones también se pueden incorporar en algunos modelos de Regresión Probabilística, pero con consecuencias para los supuestos de la fundamentación teórica de tales modelos, lo que no ocurre en la Regresión Difusa, donde la presencia de restricciones es cosustancial a su formulación, por no estar limitada su formulación a tantos supuestos como los de la Regresión Probabilística.

En el Análisis de Regresión Difusa, las desviaciones entre los valores de pertenencia observados y los valores de pertenencia estimados se asume que dependen de la incertidumbre de la estructura del modelo. En cambio en el Análisis de Regresión Lineal usual, las desviaciones se suponen causadas por errores, de origen aleatorio, en las observaciones.

Como ya indicamos en esta introducción, la relación (1.1) no se cumple, por lo general, para los números reales. Sin embargo, considerando números difusos, siempre es posible encontrar coeficientes, que con un cierto nivel de pertenencia, cumplan la relación. Esta nueva concepción se refleja en la figura (1.6).

En esta figura se aprecia que, en un sentido más tradicional, la igualdad entre el dato original y el valor estimado, se cumple para su valor central, con una pertenencia de *h*. Por otra parte, $Y^0 \subseteq f(x, A)^0$, con lo que cualquier valor que está incluido en el número difuso Y, también está incluido en el número difuso estimado. Esta situación se denominará estimación posibilística, habida cuenta de la medida de posibilidad que se ha presentado dentro de la Teoría de Conjuntos Difusos.

La medida de posibilidad representa un sentido amplio de reunión de todas las alternativas posibles, como medida antagónica a la medida de necesidad, que repre-

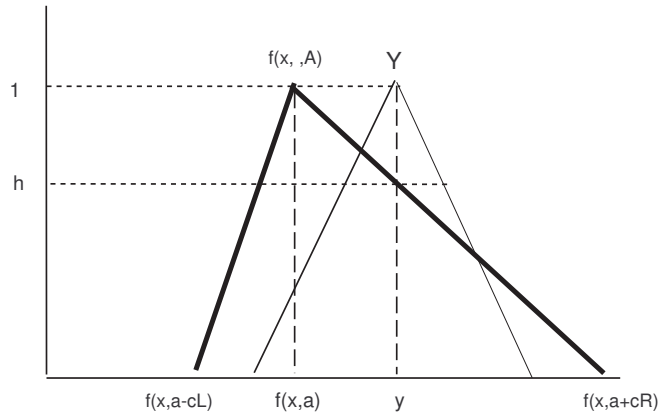

Figura 1.6: *Estimación posibilística de Regresión Difusa*

senta lo común a todas las alternativas posibles. Para datos intervalares, diversos autores han desarrollado, también, un enfoque de regresión basado en la medida de necesidad [TL98].

También hemos mencionado en esta introducción las dificultades de la regresión probabilística cuando se dispone de pocas observaciones, lo que va unido a la dificultad de comprobar la distribución asumida sobre los errores (¿se distribuyen realmente como distribuciones normales?). En esta situación, un modelo difuso, que puede incorporar un nivel de confianza posibilístico, es una alternativa ventajosa de modelado[KMK96].

Los datos precisos, que son un caso particular de datos difusos, también pueden ser modelados por la Regresión Difusa, en cuyo caso, puede considerarse las extensiones de los datos estimados como la alternativa a los intervalos de confianza de la regresión probabilística.

Para terminar, un breve repaso a la literatura referida a Regresión Difusa, nos lleva a encontrar aplicaciones de regresión difusa en el campo de la conductividad eléctrica [BBK88], riesgos para la salud [BBD91], ergonomia [CLK96] [BW97], manufactura [IKW03], ciclos económicos [WT02], funciones de costo [WL99], las cargas eléctricas [SREH97] [NZ99] [MKIK02] [SATEH02], las tasas de interés [dAT04], quiebras de empresas [TLss], paleontología [BPDN97], fiabilidad de software [Wat00], entre otros.

El estudio del comportamiento de los precios, que en diversos mercados compe-

titivos tienen permanentes variaciones, es otro campo donde la regresión difusa puede jugar un importante papel. En este sentido es de particular interés un estudio sobre la representación de la incertidumbre de variables económicas, con intervalos, conjuntos difusos y probabilidades [SJ02], que abre un campo auspicioso de aplicaciones.

Para terminar debemos señalar que, aunque en la literatura sólo hemos encontrado un caso de utilización de una variable lingüística como variable dentro de un modelo de regresión difusa [D'U03], y una reflexión sobre el tema [Wat01], pensamos que esta limitación se debe más a la visión que se ha tenido de la Regresión Difusa desde la perspectiva tradicional de la Regresión Cuantitativa, que a las potencialidades de incorporar categorías cualitativas dentro de la Regresión Difusa.

Capítulo 2

Regresión Difusa: Antecedentes y Medidas de Bondad del Ajuste

El capítulo anterior lo hemos dedicado a presentar las dos herramientas fundamentales sobre las que se sustenta este trabajo de investigación, a saber, el Análisis de Regresión y la Teoría de Conjuntos Difusos, de forma que hemos situado su contexto en el mundo de la Regresión Difusa. Ahora vamos a continuar nuestro estudio analizando con detalle las distintas propuestas que hay en la literatura al respecto de esta técnica de análisis de regresión capaz de trabajar con números difusos.

Bajo el título de *Regresión Difusa* se engloba todo método de regresión en el que alguna de sus variables y/o alguno de sus coeficientes sean números difusos.

Las primeras propuestas al respecto de la Regresión Difusa consideraban que la variable de salida era un conjunto difuso y la(s) variable(s) de entrada era(n) datos precisos. En este esquema, los coeficientes estimados eran difusos. Posteriormente se han presentado diversas propuestas donde se considera tanto la variable de salida como las variables de entrada como conjunto difusos.

Como veremos, en estos casos con múltiples variables difusas, tienen mucha importancia las consideraciones sobre aritmética difusa que fueron enunciadas en el capítulo anterior.

La primera sección está dedicada a presentar la Regresión Difusa Posibilística, origen de la Regresión Difusa, donde se encuentran fundamentalmente las aportaciones de Hideo Tanaka.

Tras la aparición de este modelo de regresión, se propusieron los métodos basados en el enfoque de los *mínimos cuadrados*. Dedicaremos también una sección a estudiar la Regresión Difusa de Mínimos Cuadrados.

Para terminar nuestro repaso a las distintas propuestas de Regresión Difusa, también se reseñan diversas propuestas de Regresión Difusa con otros enfoques que completan a los dos anteriormente indicados

Por último, y puesto que para formalizar el análisis de regresión es fundamental medir la calidad de las estimaciones, en este capítulo también incluimos una recopilación de las propuestas que se han presentado para medir la bondad de ajuste de las estimaciones de la regresión difusa. Dentro de este marco, como primera aportación de nuestro trabajo de investigación y, dado que nos movemos en un ambiente difuso, estudiamos algunas medidas de distancia entre números difusos que permiten también aseverar la bondad de un ajuste desde una perspectiva de la Teoría de Conjuntos Difusos.

2.1. Regresión Difusa Posibilística

En la relación (1.1), que presentamos en el capítulo anterior y que recordamos a continuación, define el problema de regresión en general. A partir de dicha formulación, se considerará que estamos ante un modelo de regresión difusa a partir del hecho de que el número y_i pueda ser un número difuso.

$$Y_i = f(x_i, A)\ para\ i = 1, ..., n$$

Supondremos pues que se tiene de partida un conjunto de n observaciones, donde los valores de entrada son precisos y están representados en la matriz X_{ij} (con i=1..n y j =1..m) de valores reales, y la variable de salida Y_i es imprecisa estando dados sus valores por funciones de pertenencia triangulares de parámetros (y_i, p_i, q_i).

En esta relación, si suponemos también que los coeficientes representados por A son números difusos, siempre se va a poder encontrar un valor para dichos coeficientes que permita cumplir la condición (recuérdese la figura 1.6), probablemente no con una pertenencia de uno, pero sí con una pertenencia mayor que cero. El objetivo que se ha trazado la Regresión Difusa es encontrar el o los coeficientes representados por A que tengan la menor incertidumbre posible.

Para obtener una solución, se considera que Y_i tiene una función de pertenencia de tipo LR, y que los coeficientes $A_j = (a_j, c_{Lj}, c_{Rj})$ también tienen una función de pertenencia $L_j R_j$.

En términos más generales, la relación general de regresión (1.1) debe cumplirse, en términos difusos, no sólo para el número difuso Y_i sino para sus h-cortes Y_i^h.

Para especificar las condiciones posibilísticas, y teniendo en consideración que para la mayor parte de este estudio f será una función lineal definida por

$$f(x, A) = A_0 + \sum_{j=0}^{m} A_j x_j \tag{2.1}$$

e Y_i tendrá una función de pertenencia no simétrica $(y_i, p_i, q_i)_{LR}$, las *restricciones posibilísticas*[Bar90] quedan representadas como

$$\sum_{j=0}^{m} a_j X_{ij} - L^{-1}(h) \sum_{j=0}^{m} c_{L_j} X_{ij} \leq y_i - L^{-1}(h) p_i \; para \; i = 1, ..., n \tag{2.2}$$

$$\sum_{j=0}^{m} a_j X_{ij} + R^{-1}(h) \sum_{j=0}^{m} c_{R_j} X_{ij} \geq y_i + R^{-1}(h) q_i \; para \; i = 1, ..., n \tag{2.3}$$

$$c_{L_j}, c_{R_j} \geq 0 \;\; para \;\; j = 0, ..., m \tag{2.4}$$

que pueden ser reescritas, tomando el caso más estudiado de funciones LR, como funciones de pertenencia triangulares (no necesariamente simétricas en nuestra presentación). En este caso, nuestras restricciones posibilísticas quedan definidas con las siguientes desigualdades:

$$\sum_{j=1}^{m} a_j X_{ij} - (1-h) \sum_{j=1}^{m} c_{L_j} X_{ij} \leq y_i - (1-h) p_i \;\; para \;\; i = 1, ..., n \tag{2.5}$$

$$\sum_{j=1}^{m} a_j X_{ij} + (1-h) \sum_{j=1}^{m} c_{R_j} X_{ij} \geq y_i + (1-h) q_i \; para \; i = 1, ..., n \tag{2.6}$$

Esta es la forma más habitual de plantear las restricciones posibilísticas de la Regresión Difusa. La figura 2.1 muestra un ejemplo de una función $L^{-1}(h)$ para dos valores de h.

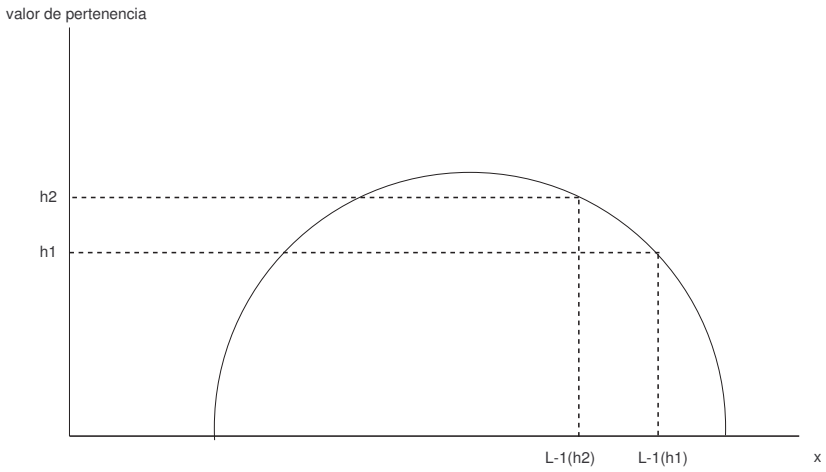

Figura 2.1: $L^{-1}(h)$ para diferentes valores de h

2.1.1. El Aporte de Tanaka

Se puede considerar que la Regresión Difusa aparece en la historia de los análisis de regresión en 1982 [TUA82], gracias a Hideo Tanaka y sus colaboradores (aunque existe una exposición previa de 1980 [TUA80]), pasando a constituir una nueva alternativa de regresión frente a las muchas metodologías de regresión que existían para trabajar con números precisos.

Los primeros intentos de Regresión Difusa están enfocados en base al principio posibilístico que acabamos de comentar, donde, recordemos, cada uno de los datos estimados $\widehat{y_i}$ contiene, en términos difusos, al dato original y_i (recuérdese la figura 1.6).

Dado este criterio posibilístico, que es el marco en que desarrolla la Regresión Difusa, Tanaka propone como criterio minimizar la incertidumbre de la estimación (usando funciones de pertenencia simétricas) expresada de la siguiente forma:

$$Min \sum_{i=1}^{n} \sum_{j=1}^{m} (c_{Lj} + c_{Rj}) \tag{2.7}$$

De esta manera queda formalizado un problema de programación lineal, con la función objetivo (2.7) y las restricciones posibilísticas (2.4)-(2.6).

Además, Tanaka y sus colaboradores sugirieron otras medidas para minimizar la incertidumbre[TSWA87]:

- Una primera, expresada como

$$\sum_{j=1}^{m} w_j c_j \tag{2.8}$$

  donde los coeficientes de las extensiones son ponderados por pesos w_i,

- y una segunda, formulada como

$$max_j\{c_j\} \tag{2.9}$$

  donde la máxima extensión es considerada la medida de incertidumbre del sistema.

Estas definiciones de incertidumbre tienen la limitación de que son dependientes de la escala de medición de las variables de entrada x. Por esa razón se ha hecho mucho más popular la siguiente función objetivo [Tan87]:

$$Min \sum_{i=1}^{n} \sum_{j=1}^{m} (c_{Lj} + c_{Rj})|x_{ij}| \tag{2.10}$$

El modelo que acabamos de formular ha sido el modelo de regresión más referenciado y utilizado en aplicaciones [IKW03], [HBS05], [WL99], [CA01], [BW97], [NZ99], etc..

En otro trabajo significativo del área, Bardossy [Bar90] sugiere el uso de otras medidas de incertidumbre de la estimación, dentro del enfoque posibilístico, como son:

- La incertidumbre es igual a la máxima amplitud de una extensión individual, es decir

$$V = max(max(c_{L_1}, c_{R_1}), ..., max(c_{L_m}, c_{R_m})) \tag{2.11}$$

- La incertidumbre se expresa como el promedio de las extensiones de los parámetros, es decir

$$V = \frac{1}{2m} \sum_{j=1}^{m} (c_{L_j} + c_{R_j}) \tag{2.12}$$

- La incertidumbre de la función difusa resultante en el dominio donde las variables independientes pueden tomar sus valores, es decir

$$V = \int_{x_m^-}^{x_m^+} ... \int_{x_1^-}^{x_1^+} \int_{-\infty}^{+\infty} sup\{min(\mu_{a_i}(a_i)); y = f(x,a)\} dy dx_1 ... dx_m \tag{2.13}$$

- La incertidumbre se mide como la función difusa resultante en los vectores x_i, es decir

$$V = \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{+\infty} sup\{min(\mu_{a_i}(a_i)); y = f(x,a)\} dy \tag{2.14}$$

Sin embargo, en la práctica, ninguna de estas sugerencias de medidas de incertidumbre han tenido difusión.

Tanaka ha seguido profundizando en el Análisis de Regresión Difuso con un enfoque posibilístico hasta el día de hoy [GT06], siempre considerando las variables de entrada como variables precisas, pero ha orientado su investigación, fundamentalmente, al uso de datos intervalares [TI92], [TKL96] [TL98], [TL99], [SIT04b].

En el esquema intervalar, Tanaka presenta tanto la regresión basada en el modelo de la *posibilidad*, como la regresión basada en el modelo de la *necesidad*, uniendo ambos enfoques, donde la extensión de un coeficiente del modelo de posibilidad (c_i^*) se relaciona con la extensión de un coeficiente del modelo de necesidad (c_{i*}) mediante la siguiente igualdad

$$c_i^* = c_{i*} + d_i \; con \; d_i \geq 0 \tag{2.15}$$

En su último trabajo [GT06], Tanaka se refiere a esta doble visión de la regresión difusa como *Modelos Duales de la Regresión Posibilística*. Aunque los modelos principales del trabajo consideran la función de optimización lineal 2.10, se presenta también un modelo con una función de optimización cuadrática sobre el parámetro anteriormente definido d_i. Además, se desarrolla una similitud entre esta regresión dual y los conjuntos aproximados (rough sets [PGBSZ95]).

La principal limitación de esta *Regresión Dual* es que la regresión basada en el modelo de la necesidad no siempre existe, por lo que hay que considerar un modelo polinomial para asegurar su existencia, que desvía la atención respecto al objetivo de simplicidad del modelo a construir.

En uno de estos trabajos (véase [TL98]), se presenta un enfoque de programación cuadrática para la regresión intervalar, en el que la incertidumbre de la estimación es medida como una distancia cuadrática. Este trabajo sirvió de primer estímulo para el desarrollo del enfoque de regresión difusa cuadrática que se presenta y evalúa en esta memoria.

Tanaka también ha propuesto un análisis de regresión posibilístico exponencial ([TIY95, TL99]). Este enfoque asume que los coeficientes A tienen función de pertenencia (o distribución de posibilidades) exponencial 1.23, con lo que hay una similitud con la distribución de probabilidades normales de la regresión probabilística. La matriz de covarianzas probabilística es denominada, para el enfoque posibilístico, *matriz de extensiones*.

Tanaka ha aplicado el análisis de regresión posibilístico a otras técnicas [HT90], como el GMDH (Group Method of Data Handling) que es una técnica para múltiples variables, para ir seleccionando un modelo polinomial por etapas.

Finalmente, es importante señalar que Tanaka ha aplicado el análisis de regresión posibilístico a casos prácticos concretos, como es el estudio de tensión en un puente en Japón [KTKF], donde las variables de entrada son los diversos factores de error que pueden existir en las especificaciones del puente. Otras aplicaciones se encuentran en [TSWA87, TUA82].

Sobre la elección del parámetro h, que se denomina *nivel de confianza*, y que define el α-corte en el cual se estimarán los coeficientes, Tanaka y Watawa [TW88] sugieren el valor 0 cuando hay una cantidad razonable de datos y un valor mayor

cuando hay pocos datos disponibles. Sin embargo, en la literatura hay ejemplos de utilización de h desde 0 hasta 0.9. El lector interesado puede consultar un extenso trabajo (véase [MK93]) que analiza el efecto de h. En términos prácticos, una elección por parte del analista que efectúa la estimación de un valor mayor para h, significa que las extensiones de los coeficientes estimados serán más amplias, por lo que la estimación presentará mayor incertidumbre.

2.1.2. Otros Modelos Basados en la Teoría de la Posibilidad

El planteamiento de Tanaka, que, como hemos visto, está orientado a minimizar la incertidumbre, no se preocupa por el comportamiento de la estimación de los valores centrales. Sin embargo, en la literatura se puede encontrar un modelo que incorpora una estimación específica de la tendencia central, fue desarrollado por Savic y Pedrycz [SP91, SP92]. Este modelo se desarrolla en dos faces:

- En la primera fase, se realiza un ajuste de mínimos cuadrados entre los valores x_i y los valores y_i -valores centrales de Y_i - con lo que se logran los valores modales a_i^* que son utilizados en la segunda fase.
- En la segunda fase, se utiliza el mismo criterio de vaguedad de la regresión difusa

$$Min \sum_{i=1}^{n} \sum_{j=1}^{m} (c_{Lj} + c_{Rj}) \tag{2.16}$$

sujeto a las condiciones posibilísticas

$$\sum_{j=1}^{m} a_j^* X_{ij} - (1-h) \sum_{j=1}^{m} c_{L_j} X_{ij} \leq y_i - (1-h) p_i \;\; para \;\; i = 1, ..., n \tag{2.17}$$

$$\sum_{j=1}^{m} a_j^* X_{ij} + (1-h) \sum_{j=1}^{m} c_{R_j} X_{ij} \geq y_i + (1-h) q_i \; para i = 1, ..., n \tag{2.18}$$

El modelo de Savic y Pedrycz tiene la virtud de disponer de una estimación que tiene una solución estándar para la tendencia central, pero conceptualmente no tiene la simplicidad de la solución de Tanaka.

Por otro lado, en diversos trabajos se han tomado en consideración los índices de posibilidad y necesidad de Dubois y Prade al comparar dos números difusos [DP83].

Sakawa y Yano [SY92] propusieron cuatro modelos de regresión posibilística:

I) Considerando la minimización de la función objetivo (2.10), se definen las restricciones basadas en el índice de posibilidad

$$Pos(y_i = ax_i) \geq h,\ para\ i = 1, ..., n \tag{2.19}$$

II) Considerando la minimización de la misma función objetivo, se definen las restricciones con el índice de necesidad

$$Nes(y_i \subset ax_i) \geq h,\ para\ i = 1, ..., n \tag{2.20}$$

III) Considerando la maximización de la función objetivo 2.10, se definen las restricciones

$$Nes(y_i \supset ax_i) \geq h,\ para\ i = 1, ..., n \tag{2.21}$$

IV) Por último, se puede considerar minimizar la función objetivo 2.10, bajo los conjuntos de restricciones

$$-y_i + \sum_{j=0}^{n} a_j x_{ij} \leq L^{-1}(h) \sum_{j=0}^{n} c_j |x_{ij}|\ para\ i = 1, ..., n \tag{2.22}$$

$$y_i - \sum_{j=0}^{n} a_j x_{ij} \leq L^{-1}(h) \sum_{j=0}^{n} c_j |x_{ij}|\ para\ i = 1, ..., n \tag{2.23}$$

Los autores plantean un modelo multiobjetivo para abordar estos cuatro problemas, puesto que junto a la función objetivo indicada, plantean maximizar el valor de h.

Este modelo ha sido criticado [RW96] por ser muy sensible a los puntos extremos y por producir, en ciertas condiciones, todos los estimadores como números crisp (precisos) [MNN04].

Más recientemente, otros autores [MNN04] enfocan la Regresión Posibilística desde el punto de vista del *riesgo*, usando las mismos índices de posibilidad y necesidad anteriores, pero incorporando una función objetivo cuadrática:

I) Problema de riesgo neutro

$$Min : D(h) = \sum_{i=1}^{n} \{(y_i + q_i) - (\widehat{y_i + q_i}) + (y_i - p_i) - (\widehat{y_i - p_i})\}^2 \quad (2.24)$$

$$s.a.\ (\widehat{y_i - p_i})(h) \leq (y_i + q_i)(h), (\widehat{y_i + q_i})(h) \geq (y_i - p_i)(h)\ para\ i = 1, ..., n \quad (2.25)$$

$$(\widehat{y_i + q_i}) - (\widehat{y_i - p_i}) \geq 0\ para\ i = 1, ..., n \quad (2.26)$$

II) Problema de riesgo adverso: minimiza el objetivo 2.24, sujeto a las restricciones

$$s.a.\ (\widehat{y_i - p_i})(h) \leq (y_i - p_i)(1 - h), (\widehat{y_i + q_i})(h) \geq (y_i + q_i)(1 - h)$$
$$para\ i = 1, ..., n \quad (2.27)$$

III) Problemas de búsqueda de riesgo: minimiza el objetivo cuadrático 2.24, sujeto a las restricciones

$$s.a.\ (\widehat{y_i + q_i})(1 - h) \leq (y_i - p_i)(h), (\widehat{y_i - p_i})(1 - h) \geq (y_i - p_i)(h)$$
$$para\ i = 1, ..., n \quad (2.28)$$

Cada uno de estos problemas constituye un problema de optimización en sí. Los autores desarrollan un algoritmo para determinar el nivel de h óptimo, entendiendo que a mayor h, mayor es la incertidumbre de la función objetivo. Estos modelos no son posibilísticos en un sentido estricto, pero sí garantizan que el valor central de los datos observados, y_i, se encuentre dentro del intervalo estimado $[\widehat{y_i - p_i}, \widehat{y_i + q_i}]$.

2.1.3. Críticas a la Regresión Difusa Posibilística

Hemos analizado los distintos modelos de Regresión Posibilística que se pueden encontrar en la literatura desde los trabajos iniciales de Tanaka hasta la actualidad. Aunque se trata de trabajos muy significativos y de gran valor en el mundo de la Regresión Difusa, la Regresión Posibilística no está exenta de críticas:

- Una de las críticas más recurentes sobre el enfoque posibilístico de la regresión difusa es el efecto de los puntos extremos (outliers) [Pet94]. En el gráfico (a) de la figura 2.2 es posible ver una estimación sin puntos extremos; en cambio en (b) se aprecia el efecto considerable que tiene en la ampliación del intervalo de estimación la aparición de un punto extremo.

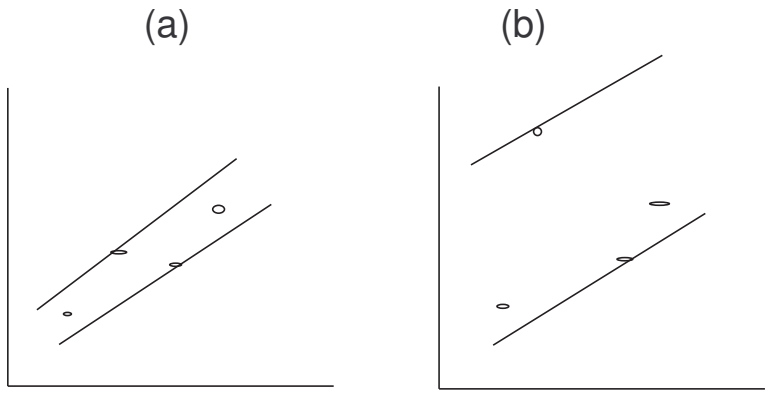

Figura 2.2: Efecto de un punto extremo en una estimación posibilística

Peters [Pet94] propuso un método intervalar para afrontar el problema causado por los puntos extremos, en el caso en que la variable y_i sea precisa. Sin embargo, esta propuesta no ha tenido mayor impacto en el desarrollo de alternativas de estimación. Como alternativa, Tsaur y Wang [TW99] y Chen [Che01] plantean metodologías para detectar puntos extremos.

Otro enfoque para afrontar el problema de los puntos extremos fue desarrollado por Ozelkan y Duckstein [OD00]. Este nuevo modelo incorpora una flexibilización en el criterio posibilístico, de modo que la diferencias entre los extremos (estimados y observados) deben ser menores que una cierta cantidad, denominada *variable de relajación*. Considerando las medidas de posibilidad y necesidad, se definen cuatro modelos multiobjetivo, que permiten manejar los casos de puntos extremos con flexibilidad, y donde la regresión de Tanaka y la regresión de mínimos cuadrados son casos particulares de estos modelos generales.

- Otra de las críticas a la formulación de Tanaka [Cel87] va en el sentido de que muchos de los coeficientes estimados A_i resultan ser números precisos (crisp). Una forma de abordar esta crítica fue ensayada por el mismo Tanaka,

proponiendo funciones de pertenencia cuadráticas para los coeficientes [TI91]. Otra propuesta de Tanaka, para la regresión intervalar, fue proponer una función objetivo cuadrática [TL98].

Esta crítica se relaciona con la idea de que la Regresión Difusa Posibilística no tiene una propuesta para interpretar el intervalo estimado por la regresión [WT00b].

Nos parece esta una crítica fundamental, en sus diversas variantes, porque el cálculo de números crisp (estimaciones con ambas extensiones cero) pone en duda, a nuestro entender, la interpretación más fecunda de la regresión difusa, y que puede formularse así: *un coeficiente de regresión difuso representa la incertidumbre que produce en el sistema la variable que acompaña a dicho coeficiente*. Es decir, el producto $A_j x_{ij}$ es la magnitud de la contribución de x_i a la incertidumbre de Y_i.

Ahora bien, si por problemas de modelado matemático, en este caso las características de la programación lineal, los datos resultan precisos, independientemente de la incertidumbre de las variables de entrada x_i, este modelado produce una grave limitación en la interpretación de una regresión difusa. Buena parte de nuestra investigación trata de solucionar este inconveniente (véase al capítulo 3).

Para ayudar a la interpretación del intervalo estimado de la regresión difusa, Wang y Tsaur [WT00b] construyen un índice de confianza de la estimación, en un sentido análogo al del coeficiente de determinación R^2 en la estadística probabilística, concluyendo que en la medida que este índice es más alto, mejor es la representación de $\widehat{y_i}$ para dar forma a y_i. Es decir, el valor central de la estimación tiene más *peso* para representar el valor central de los datos de entrada. Nos parece un avance el uso de una medida de confianza de la estimación, pero el problema de interpretación de la función de pertenencia estimada para cada observación queda pendiente.

Estos autores, con el objetivo de identificar un método de selección de variables, definen un índice de confianza parcial, para medir el efecto de una nueva variable en el modelo, lo que puede ser un camino que se debe considerar para mejorar la interpretación de los estimadores difusos de regresión.

- Otro problema que se ha planteado (véase [KL05]) al respecto de la Regresión Posibilística es que cuanto más datos se disponen sobre las variables, mayor será la extensión de los coeficientes estimados, es decir, mayor será la incer-

tidumbre de la estimación. Esta situación es efectiva, y en las próximas secciones se analizará cuándo es más conveniente realizar una regresión difusa incluyendo el enfoque posibilístico y cuándo no es recomendable incluirlo.

- Otros autores han señalado que la Regresión Lineal Difusa tiende a ser multicolineal cuando la cantidad de variables independientes aumenta [KMK96]. Hay dos estrategias para afrontar esta dificultad:

  I) Por un parte disminuir las variables independientes, para lo cual hay que realizar un proceso de selección de variables [WT00b]. Esta estrategia se analizará con detalle en un capítulo posterior de esta memoria.
  II) Por otra parte, manteniendo el conjunto de variables de entrada, también se puede buscar una estimación alternativa, que nosotros hemos centrado en el enfoque de Regresión Ridge, y que también desarrollamos en otro capítulo de esta tesis.

- Chang y Lee [CL94b] hacen una crítica a la Regresión Posibilística, en el sentido de que la tendencia de la estimación (tendencia central o tendencia modal) puede ser distinta a la tendencia de las extensiones de las estimaciones. Ponen un ejemplo en el que, mientras la tendencia central es ascendente, la tendencia de las extensiones, en cambio, es descendente. La regresión de Tanaka no permite hacer esta distinción entre dos tendencias y no puede estimarlas correctamente. Chan y Lee atribuyen esta situación a que el signo de las extensiones debe ser siempre positivo. Por esta razón proponen un modelo alternativo, en el que la restricción de signo de las restricciones 2.4 es sustituida por la restricción de que la extensión final de cada estimación $c'|X|$ debe ser mayor o igual a cero. Es decir, las extensiones de los coeficientes pueden ser negativas, pero la extensión de cada estimación debe ser positiva.

  La crítica es válida, en el sentido de la rigidez del comportamiento de las extensiones, lo que podría contribuir a la dificultad de la interpretación del intervalo estimado. Desde esta perspectiva, la contribución de Chang y Lee debe ser considerada como un aporte importante dentro de la literatura.

  Sin embargo, la estimación de extensiones negativas, y las posibles dificultades para utilizar el modelo estimado para fines predictivos, debieran ir acompañadas de las correspondientes explicaciones conceptuales, que no se entregan en el trabajo.

  Otra propuesta posibilística se encuentra en Soliman et al. [SATEH02]. En esta

propuesta, la función objetivo es una función lineal de las diferencias tanto de las extensiones, como del valor modal. Para funciones de pertenencia simétricas es:

$$J = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{m} a_i x_{ij} + p_i - \sum_{j=1}^{m} c_j x_{ij}) \tag{2.29}$$

Para funciones de pertenencia no simétricas se formula como:

$$J = \sum_{i=1}^{n}(4y_i - 4\sum_{j=1}^{m} a_i x_{ij} + p_i - \sum_{j=1}^{m} c_{L_j} x_{ij} + q_i - \sum_{j=1}^{m} c_{R_j} x_{ij}) \tag{2.30}$$

Esta idea que había sido anteriormente expuesta por Chang y Lee [CL94a].

2.2. Regresión Difusa de Mínimos Cuadrados

Como alternativa a los modelos posibilísticos se han desarrollado modelos enfocados principalmente a minimizar los cuadrados de los errores de estimación.

El primer trabajo que puede ser clasificado en este grupo es el de Celmins en 1987 [Cel87], aunque nosotros lo hemos encuadrado dentro de los modelos posibilísticos, puesto que considera las restricciones posibilísticas.

Más adelante, el estudio de Phill Diamond, en 1988 (véase [Dia88]), de amplia repercusión en la comunidad científica, establece una medida para la distancia entre dos números difusos LR, basada en la métrica L_2:

$$d(A,B)^2 = [a-b]^2 + [a - p_a - (b - p_b)]^2 + [a + q_a - (b + q_b)]^2 \tag{2.31}$$

con la cual define dos modelos, que analiza separadamente, con X e Y números difusos:

$$\begin{aligned} (F1):\ Y = a + bX\ a, b \varepsilon R \\ (F2):\ Y = A + bX\ b \varepsilon R, A\ numero\ difuso \end{aligned} \tag{2.32}$$

En estos modelos debe considerarse el signo de b para su resolución.

El tercer modelo desarrollado por Diamond, también de regresión simple, es:

$$Y = A + xB,\ donde\ x\varepsilon\Re,\ \text{y} A, B\ son\ numeros\ difusos \tag{2.33}$$

En otros trabajos se generalizan estos modelos para la regresión múltiple (véase [DK97]), abordando el problema de las posibles extensiones negativas con una diferencia de Hukuhara generalizada $B\Theta_H A$ a partir de la diferencia Hukuhara usual $B \sim_h A$ que se define, cuando existe, como:

$$[B \sim_h A]^\alpha = \{a\varepsilon\Re : [A]^\alpha + \{a\} \subseteq [B]^\alpha\},\ \alpha\varepsilon[0,1] \tag{2.34}$$

Esta medida, aplicada a dos números difusos triangulares simétricos (a, p_a) y (b, p_b), está bien definida si $p_b \geq p_a$ y resulta

$$B \sim_h A = (b - a, p_b - p_a)_\triangle \tag{2.35}$$

y la distancia generalizada extiende la definición al caso que $p_a > p_b$.

Este modelo de regresión múltiple ha tenido más un interés teórico por la definición de medidas y diferencias, que un interés práctico para la Regresión Difusa.

Una generalización de los modelos de Diamond fue desarrollada por Yang and Ko [YK97]. Posteriormente, investigadores de Hong et al. [HSD01] han aplicado la aritmética de la norma *más débil* a estos modelos, llegando a problemas de programación cuadrática sólo para resolver problemas de regresión simple.

Aunque durante mucho tiempo los trabajos de Celmics y Diamond fueron los únicos con el enfoque de mínimos cuadrados, luego vendría una gran cantidad de propuestas, algunas basadas en la minimización de los residuos al cuadrado, y otras concluyendo con las ecuaciones de mínimos cuadrados.

Kim y Bishu [KB98] realizaron una interesante propuesta en este sentido. En sus trabajos plantean en relación a las funciones de pertenencia triangulares, que estas funciones pueden ser determinadas por tres puntos, a saber, el valor central, el punto de inicio del intervalo a la izquierda y el punto de término del intervalo a la derecha. Luego si los tres puntos indicados del valor observado están muy cerca de los tres puntos correspondientes de los valores estimados, entonces sus funciones de pertenencia tendrán que ser muy similares. Con este objetivo llegan a tres ecuaciones

$$y_i - p_i = \sum_{j=0}^{m} (a_j - c_{Lj}) x_{ij} \tag{2.36}$$

$$y_i = \sum_{j=0}^{m} a_j x_{ij} \tag{2.37}$$

$$y_i + q_i = \sum_{j=0}^{m} (a_j - c_{Rj}) x_{ij} \tag{2.38}$$

que estiman, independientemente, como tres regresiones de mínimos cuadrados ordinarios. Con este enfoque, se logra obtener una estimación muy clara desde el punto de vista conceptual, utilizando el enfoque estándar de regresión.

Sin embargo, la limitación de este planteamiento aparece por realizar la estimación de forma independiente, puesto que no se garantiza que el comportamiento de las estimaciones cumpla las condiciones de que c_{Li} y c_{Ri} sean mayores o iguales a cero. Como veremos, el modelo cuadrático de regresión que se plantea en el siguiente capítulo puede ser visto como una generalización del modelo de Kim y Bishu, en que la estimación de las tres ecuaciones se efectúa simultáneamente, por lo que se garantiza que las estimaciones cumplan las restricciones propias de la definición de funciones de pertenencia.

Anteriormente, Chang había hecho su tesis de doctorado planteando la utilización de la aritmética difusa ponderada, para un modelo de regresión difusa [Cha01b], que fue publicada varios años después [Cha01a]. En estos trabajos, el autor llega a un sistema de ecuaciones normales muy similares a las de Kim y Bishu que adolece de la misma limitación: no se garantiza que las extensiones de los números estimados sean positivas.

Otro enfoque de mínimos cuadrados se encuentra en los trabajos realizados por P. D'Urso. En una primera publicación [DG00], este autor propone el uso de un modelo de mínimos cuadrados en el que las extensiones de la variable de salida dependen de la magnitud de la estimación de los centros, a través de un método recursivo.

Para funciones de pertenencia triangulares no simétricas, desarrolla el modelo

$$y = y^* + \varepsilon_y \; donde \; y^* = xa$$

$$p = p^* + \lambda \; donde \; p^* = c^* b + 1d$$
$$q = q^* + \rho \; donde \; q^* = c^* g + 1h \tag{2.39}$$

donde a, b, d, g y h son los parámetros de regresión que hay que estimar, mientras que $\varepsilon_c, \lambda, \rho$ son vectores de residuos. La función objetivo que se pretende minimizar es

$$(y - y^*)'(y - y^*)\pi_c + (p - p^*)'(p - p^*)\pi_p + (q - q^*)'(q - q^*)\pi_q \tag{2.40}$$

En esta ecuación, π_y, π_p y π_q son pesos arbitrarios positivos, puestos por el tomador de decisiones.

Este planteamiento general considera una solución para a, b, d, g y h que depende de esos mismos parámetros, por lo que se da una solución inicial y luego se repite recursivamente la búsqueda de solución. No se hace un estudio de convergencia de la solución propuesta, pudiendo ser poco eficiente, o converger a óptimos locales, lo que constituye una limitación de la propuesta.

Al respecto del mismo enfoque, D'Urso y Gastaldi [DG02] desarrollan un modelo para una regresión polinomial, donde se asume que cada coeficiente A_i depende de una función polinomial de la variable x_i. Es una alternativa que se puede considerar para el caso de regresiones polinomiales de orden alto.

Otro estudio de estimación de mínimos cuadrados con elementos difusos se puede encontrar en [KC03]. En este trabajo se calcula el término de error de los números estimados como otro número difuso, y se incorpora al modelo este término de error, estimándolo mediante un ranking de números difusos. Este método puede aplicarse a variables independientes que sean números difusos.

Para funciones de pertenencia normales, una propuesta de mínimos cuadrados se encuentra en el trabajo de Xu y Li [XL01]. En este trabajo se define la medida de distancia de dos número difusos como

$$\widetilde{d}(A, B) = (\int_0^1 f(x) d^2(A(x), B(x)) dx)^{\frac{1}{2}} \tag{2.41}$$

donde se tiene

$$d^2(A(x), B(x)) = (a_1(x) - b_1(x))^2 + (a_2(x) - b_2(x))^2$$

$$A(x) = [a_1(x), a_2(x)], B(x) = [b_1(x), b_2(x)] \tag{2.42}$$

y $f(x)$ es una función creciente en [0,1] con $f(0) = 0$ y $\int_0^1 f(x)dx = 1/2$. Para funciones de pertenencia normales resulta

$$\widetilde{d}^2(A, B) = (a - b)^2 + \frac{1}{2}(\sigma_A - \sigma_B)^2 \tag{2.43}$$

Por último, aunque no nos detendremos en ellos, otros trabajos interesantes donde se desarrolla un modelo difuso de mínimos cuadrados son [WT02] y [Wu03].

2.3. Otras Visiones de la Regresión Difusa

Para terminar nuestro análisis de la Regresión Difusa, vamos a presentar otras visiones que de la misma pueden encontrarse en la literatura.

La idea de aplicar los resultados de la teoría probabilística, con números precisos, para obtener resultados en una estructura difusa, incluyendo el análisis de regresión, ha sido desarrollada para distribuciones de probabilidades y predicción [Buc04]. Una de las posibles aplicaciones de esta línea de investigación es la construcción de *tests de hipótesis* [GP02].

Diversos estudios han analizado las series de tiempo con la técnica de la Regresión Difusa, campo de aplicación que no hemos incorporado a este memoria [TWY02]. También hay un ensayo sobre regresión difusa no paramétrica [CL99].

No hemos considerado en este trabajo de investigación entrar en el estudio de los *Árboles de Regresión*. Existe una aplicación de sistemas de poder en este área [MKIK02]. También existen propuestas interesantes en el uso de algoritmos genéticos [MP03, WY94].

Otros trabajos se han dedicado ha estudiar la realización de análisis de regresión difusa a través de redes neuronales [IT92, IN96, HbLZH98] o se han apoyado en el uso de ciertas redes, como las de función básica radial [CL99] o en el esquema de Máquina de Soporte de Vectores de Vapnik [Vap98], éstas últimas orientadas a la Regresión Ridge [HH04] y [HHA04].

Un caso especial de regresión, cuando se postulan distintos modelos por tramos de los datos, ha merecido algunos estudios para datos difusos [YTL99], [YTL01]. También hay trabajos sobre regresión difusa no-lineal [BBD93].

Terminaremos citando una aplicación de la Regresión Logística que nombrábamos en el capítulo de introducción como Regresión Difusa [TLss], donde se desarrolla un modelo específico para variables binarias.

2.4. Indices de Bondad del Ajuste

Pasamos ahora a centrarnos en otro aspecto fundamental dentro del campo de la Regresión Difusa que nos ocupa: el estudio de la bondad del ajuste obtenido. Para ello, vamos a analizar diferentes índices de bondad que se pueden utilizar para determinar la calidad de la estimación hecha gracias al proceso de regresión.

Consideramos este un aspecto fundamental para afrontar nuestro trabajo de investigación y poder evaluar los métodos que propondremos más adelante. De ahí que dediquemos una primera sección para repasar lo que se puede encontrar en la literatura al respecto de medidas de bondad del ajuste y, como primera propuesta de interés de nuestro trabajo, en la siguiente sección, estudiemos algunas medidas adicionales de interés.

2.4.1. Revisión de los Indices más Usados

No son muchos los estudios de Regresión Difusa que incorporan un estudio de bondad de ajuste de los métodos propuestos. Incluso un extenso estudio comparativo [CA01], no utiliza ningún índice de evaluación.

Probablemente, la aportación más relevante es la medida de Kim y Bishu [KB98], empleada también por Kao y Lin [KL05] y Kao y Chyu [KC02]. Esta medida tiene en cuenta la divergencia entre el valor observado y el valor estimado:

$$D_i = \int_{Sop_{Y_i} \bigcup Sop_{\widehat{Y}_i}} |\mu_{Y_i}(x) - \mu_{\widehat{Y}_i}(x)|dx \tag{2.44}$$

También se considera una medida de divergencia relativa, para cada una de las observaciones estimadas:

$$E_i = \frac{D_i}{\int_{Sop_{Y_i}} \mu_{Y_i}(x)dx} \tag{2.45}$$

En [Cha01b] se tiene como objetivo de la investigación construir índices de bondad de ajuste. Estos índices han sido aplicados en [MP03]. En esta propuesta se define un coeficiente de correlación híbrido, como

$$(HR)^2 = \frac{\sum_{i=1}^{n}(\widehat{y_i} - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{2.46}$$

donde, a partir de ahora, se considerarán $\widehat{y}$ los datos estimados y $\overline{y}$ los datos medios de todos los parámetros considerados.

Este indicador es idéntico al coeficiente de determinación de la regresión probabilística (R^2), con la diferencia que aquí se construye con la aritmética difusa ponderada. Sin embargo, el índice puede tomar valores mayores que 1 ([Cha01a] pág . 237, [MP03] pág. 437) lo que lo hace menos atractivo que otros índices que fluctúan entre 0 y 1.

También Chang define las siguientes medidas, usando la aritmética difusa ponderada:

Definición 2.1 (Media híbrida)**.**

$$\overline{Y} = \frac{\sum_{i=1^n Y_i}}{n} \tag{2.47}$$

Definición 2.2 (Desviación estándar híbrida)**.**

$$S_Y = \sqrt[2]{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \tag{2.48}$$

Definición 2.3 (Error estándar de la estimación hídrido)**.**

$$HS_e = \sqrt{\frac{1}{n-m-1}\sum_{i=1}^{n}(\widehat{Y_i} - Y_i)^2} \tag{2.49}$$

Por otro lado, Chang define un interesante *coeficiente de regresión parcial estandarizado* t_i, mediante la aritmética difusa ponderada y como un número preciso. Este autor sugiere que dicho valor representa el nivel de importancia que la variable que acompaña a cada coeficiente tiene en el modelo. Es decir, si $t_i > t_j$, entonces la variable X_i es más importante en el modelo que la variable X_j.

Definición 2.4 (Coeficiente de regresión parcial estandarizado)**.**

$$t_i = \widehat{A_i}\frac{S_x}{S_Y} \tag{2.50}$$

En los trabajos de D'Urso [DG02], se define un coeficiente de determinación, R^2, por analogía con el coeficiente de determinación probabilístico, que para funciones triangulares no simétricas es:

Definición 2.5 (Coeficiente de Determinación triangular)**.**

$$R^2 = \frac{(\widehat{y}-\overline{y})'(\widehat{y}-\overline{y}) + (\widehat{p}-\overline{p})'(\widehat{p}-\overline{p}) + (\widehat{q}-\overline{q})'(\widehat{q}-\overline{q})}{(y-\overline{y})'(y-\overline{y}) + (p-\overline{p})'(p-\overline{p}) + (q-\overline{q})'(q-\overline{q})} \tag{2.51}$$

donde $\overline{y}$, $\overline{p}$ y $\overline{q}$ son los promedios de y, p y q, respectivamente.

Para datos con funciones de pertenencia trapezoidales (c_1, c_2, p, q), se tiene el siguiente coeficiente:

Definición 2.6 (Coeficiente de Determinación trapezoidal)**.**

$$R^2 = \frac{(\widehat{c_1}-\overline{c_1})'(\widehat{c_1}-\overline{c_1}) + (\widehat{c_2}-\overline{c_2})'(\widehat{c_2}-\overline{c_2}) + (\widehat{p}-\overline{p})'(\widehat{p}-\overline{p}) + (\widehat{q}-\overline{q})'(\widehat{q}-\overline{q})}{(c_1-\overline{c_1})'(c_1-\overline{c_1}) + (c_2-\overline{c_2})'(c_2-\overline{c_2}) + (p-\overline{p})'(p-\overline{p}) + (q-\overline{q})'(q-\overline{q})} \tag{2.52}$$

donde $\overline{c_1}$, $\overline{c_2}$, $\overline{p}$ y $\overline{q}$ son los promedios de c_1, c_2, p y q, respectivamente.

El autor utiliza estas dos definiciones de R^2 para un proceso iterativo de ajuste polinomial, de manera que cuando R^2 no aumenta significativamente, se termina el proceso de incorporar nuevos términos polinomiales.

En [HBS05] se pueden encontrar tres medidas de bondad de ajuste:

I) El porcentaje promedio de intervalos contenidos en los intervalos estimados.
II) El porcentaje promedio de los intervalos estimados contenidos en los intervalos observados.
III) Una medida de similitud definida, gráficamente, como la intersección de la función de pertenencia del dato estimado y el dato observado, respecto a la unión de dichos dos números.

Esta última medida es una de las medidas que definiremos analíticamente y evaluaremos más adelante.

Una propuesta interesante es la de Wang y Tsuar [WT00b], en que se tiene en cuenta que la suma de cuadrados totales (SCT) se puede descomponer en la suma de cuadros de la regresión (SCR) más la suma de cuadrados del error (SCE), en forma similar al caso probabilístico. Para este fin se define

$$SCT = \sum_{i=1}^{n}(y_i - \widehat{y_i - p_i})^2 + \sum_{i=1}^{n}(\widehat{y_i + q_i} - y_i)^2 \tag{2.53}$$

$$SCR = \sum_{i=1}^{n}(\widehat{y_i} - \widehat{y_i - p_i})^2 + \sum_{i=1}^{n}(\widehat{y_i + q_i} - \widehat{y_i})^2 \tag{2.54}$$

$$SCE = 2\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2 \tag{2.55}$$

de forma que se puede definir un índice de confianza, IC, como

$$IC = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT} \tag{2.56}$$

Este resultado, que es similar al de la Regresión Probabilistica, sólo es válido cuando la función de pertenencia es simétrica, lo que le resta validez en general y, especialmente, para nuestro objetivo, de que las funciones de pertenencia de los coeficientes estimados tengan más flexibilidad.

Podemos indicar para terminar, que los trabajos sobre Regresión Difusa no han profundizado sobre el estudio de la calidad de la estimación, y la mayoría de las propuestas corresponden a casos específicos de desdifusificación o funciones de pertenencia simétricas. De ahí nuestra primera aportación en este trabajo de investigación, que pasamos a describir.

2.4.2. Una Propuesta de Medidas de Bondad de Ajuste

Para dimensionar la calidad del ajuste de cualquier regresión difusa, se deben definir medidas que muestren la similitud o divergencia entre los números observados y los números difusos estimados. La Regresión Probabilística utiliza como principales indicadores el *R-cuadrado* para indicar el porcentaje de la varianza de la variable dependiente que es explicado por la regresión, y el valor *t* para cada coeficiente estimado como una medida de la significancia de la variable correspondiente.

En el contexto de la Regresión Posibilística, como se vio en la sección anterior, se han desarrollado muy pocas medidas de bondad del ajuste, siendo la más conocida la medida de divergencia de Kim y Bishu, que tiene el grave inconveniente, que ya comentamos, de que no esta normalizada, puesto que el numerador es independiente del denominador. La otra medida propuesta en la literatura, el R^2 híbrido, sigue sin estar normalizado, puesto que no es posible determinar su valor máximo. Tampoco el R^2 tradicional probabilístico está normalizado en el ámbito difuso.

En este apartado, vamos a presentar un conjunto de medidas de bondad de ajuste, seis en total, que controlan diversos aspectos de la similitud entre dos números difusos, y que nos van a servir para evaluar la calidad de una estimación de regresión difusa. Luego de haberlas definido, se realizará una calificación de tales medidas, para saber hasta qué punto cumplen con su objetivo.

Para considerar, en el ámbito de las comparaciones entre números difusos, medidas de ajuste entre los datos originales de salida y los datos estimados por la regresión difusa, los indicadores que se definen parten de medidas de similitud (o divergencia) generales entre números difusos[ZCB87].

2.4.2.1. Indice de Bondad del Ajuste SIM_1

Si se considera el cardinal escalar de un conjunto difuso A como

$$|A| = \int \mu_A(x)dx \tag{2.57}$$

se puede definir el indicador de similitud S_i entre Y_i e $\widehat{Y}_i$, siguiente:

$$S_i = \begin{cases} \frac{|Y_i \bigcap \widehat{Y}_i|}{|Y_i \bigcup \widehat{Y}_i|} & if\ |Y_i \bigcup \widehat{Y}_i| > 0 \\ 0 & |Y_i \bigcup \widehat{Y}_i| = 0 \end{cases} \tag{2.58}$$

indicador que varía entre 0 (cuando las funciones de pertenencia son disjuntas) y 1 (cuando las funciones de pertenencia son idénticas), y que denominaremos, en general, S(A, B) para dos números difusos A y B.

Propiedad 2.1. *Sean A, B y C tres números difusos de tipo LR. Entonces*
(1) $0 \leq S(A, B) \leq 1$
(2) S(A, A) = 1

(3) S(A, B) = S(B, A)
(4) Si $A \subseteq B \subseteq C \Longrightarrow S(A,C) \leq S(A,B) \wedge (B,C)$.

Los primeros tres puntos son evidentes. Para demostrar el cuarto punto, consideraremos un α-corte determinado, en que A quedará representado por el intervalo $[a_1, a_2]$, B por $[b_1, b_2]$ y C por $[c_1, c_2]$. Entonces, por la premisa se cumple que $[a_1, a_2] \leq [b_1, b_2] \leq [c_1, c_2]$, y se tiene

$$S^{\alpha}(A,C) = \frac{a_2 - a_1}{c_2 - c_1} \tag{2.59}$$

$$S^{\alpha}(A,B) = \frac{a_2 - a_1}{b_2 - b_1} \tag{2.60}$$

$$S^{\alpha}(B,C) = \frac{b_2 - b_1}{c_2 - c_1} \tag{2.61}$$

y se cumple que $S^{\alpha}(A,C)$ es menor que $S^{\alpha}(A,B)$ porque $[b_1, b_2] \leq [c_1, c_2]$, y también se cumple que $S^{\alpha}(A,C)$ es menor que $S^{\alpha}(B,C)$ porque $[a_1, a_2] \leq [b_1, b_2]$, por lo que se cumple la cuarta condición para el α-corte elegido, y por integración, se cumple la condición para los tres números difusos.

Para el conjunto de los n datos de la regresión, se calcula la medida de similitud SIM_1 con el siguiente promedio:

$$SIM_1 = \frac{\sum_i^n S_i}{n} \tag{2.62}$$

que también varía entre 0 y 1. Una medida de la misma familia de indicadores es propuesta en el trabajo de Kim y Bishu[KB98], donde se define como medida de divergencia entre el valor observado y el valor estimado la siguiente formulación

$$D_i = \int_{S_{Y_i} \bigcup S_{\widehat{Y_i}}} |\mu_{Y_i}(x) - \mu_{\widehat{Y_i}}(x)| dx \tag{2.63}$$

donde se representa la diferencia de los valores de pertenencia entre las dos funciones de pertenencia, y S_{y_i} y $S_{\widehat{y_i}}$ son los soportes de $\mu_{Y_i}(x)$ y $\mu_{widetildeY_i}(x)$ respectivamente.

Kim y Bishu en 1998, y más tarde también Kao y Lin en 2005, proponen como medida de divergencia relativa a

$$E_i = \frac{D_i}{\int_{S_{Y_i}} \mu_{Y_i}(x)dx} \tag{2.64}$$

Esta medida toma valor 0 cuando los dos números difusos son idénticos, pero toma valores superiores a 1 cuando $Y_i \bigcap Y_i^{\sim} = \emptyset$ y $(p_i^{\sim} + q_i^{\sim} > p_i + q_i)$, por lo que el valor promedio

$$KB = KIM - BISHU = \frac{E_i}{n} \tag{2.65}$$

podría, teóricamente, ser mayor que 1, por lo que parece más recomendable tomar como indicador, la propuesta SIM_1. Más aun, cuando el soporte de Y_i es pequeño comparado con el soporte de $|Y_i - \widehat{Y}_i|$, puede tomar valores mucho mayores que uno, y no es extraño ver valores sobre 1.000.

Aunque las dos medidas estudiadas son distintas, existe una relación entre los términos de las definiciones empleadas, que se mostrará en la siguiente proposición:

Teorema 2.1. *El numerador de la medida de Kim y Bishu, D_i, es igual a la diferencia entre el denominador y el numerador de S_i*

$$\int_{S_Y \bigcup S_{\widehat{Y}}} |\mu_Y(x) - \mu_{\widehat{Y}}(x)|dx = |Y_i \bigcup \widehat{Y}_i| - |Y_i \bigcap \widehat{Y}_i| \tag{2.66}$$

Para demostrar la proposición anterior, se considera que $|\mu_Y(x) - \mu_{\widehat{Y}}(x)|$ puede tener, por tramos homogéneos, dos alternativas:

- $\mu_Y(x) \geq \mu_{\widehat{Y}}(x)$, donde se cumple

$$\int |\mu_Y(x) - \mu_{\widehat{Y}}(x)|dx = \int \mu_Y(x)dx - \int \mu_{\widehat{Y}}(x)dx = |Y_i \bigcup \widehat{Y}_i| - |Y_i \bigcap \widehat{Y}_i| \tag{2.67}$$

- y $\mu_Y(x) \leq \mu_{\widehat{Y}}(x)$, donde se cumple la relación

$$\int |\mu_Y(x) - \mu_{\widehat{Y}}(x)|dx = \int \mu_{\widehat{Y}}(x)dx - \int \mu_Y(x)dx = |Y_i \bigcup \widehat{Y}_i| - |Y_i \bigcap \widehat{Y}_i| \tag{2.68}$$

con lo que queda demostrada la igualdad de la proposición.

Desde un punto de vista teórico, tal vez SIM_1 sea el mejor indicador de bondad de ajuste, porque involucra a toda la función de pertenencia de los números difusos que evalúa. Sin embargo tiene el inconveniente, teórico, de que cuando no hay intersección entre dos números difusos, siempre vale 0. Es decir, tanto si los dos números están muy distantes, como si ambos números sólo tienen un punto común de intersección, el indicador devuelve el valor 0 y no discrimina entre ambas situaciones.

Desde un punto de vista práctico en el contexto de la Regresión Difusa, es difícil que SIM_1 se acerque a 1, por lo que será un indicador de magnitud más bien baja, por lo general, por debajo de 0.5.

2.4.2.2. Indice de Bondad del Ajuste SIM_2

Otras medidas de bondad de ajuste se pueden calcular con los intervalos de $Y_i = [y_i - p_i, y_i + q_i]$ y $\widehat{Y_i} = [\widehat{y_i} - \widehat{p_i}, \widehat{y_i} + \widehat{q_i}]$ de manera que se puede medir la similitud de los dos intervalos, considerados como funciones de pertenencia, a partir de la definición

$$\nabla([y_i - p_i, y_i + q_i], [\widehat{y_i} - \widehat{p_i}, \widehat{y_i} + \widehat{q_i}]) = \frac{|\widehat{y_i} - \widehat{p_i} - (y_i - p_i)| + |\widehat{y_i} + \widehat{q_i} - (y_i + q_i)|}{2(\beta_2 - \beta_1)} \tag{2.69}$$

donde

$$\beta_1 = min(\widehat{y_i} - \widehat{p_i},\ y_i - p_i) \tag{2.70}$$

$$\beta_2 = max(\widehat{y_i} + \widehat{q_i},\ y_i + q_i) \tag{2.71}$$

Al tomar la integral sobre todo el recorrido de los α-cortes, se tiene la definición

$$T(Y_i, \widehat{Y_i}) = \int_{\alpha=0}^{1} \nabla(\widehat{Y_i}^{\alpha},\ Y_i^{\alpha}) d\alpha \tag{2.72}$$

y se puede definir el indicador para el i-ésimo dato de test como

$$T_i = T(Y_i, \widehat{Y}_i) = \begin{cases} \frac{|\widehat{y}_i - \widehat{p}_i - (y_i - p_i)| + |\widehat{y}_i + \widehat{q}_i - (y_i + q_i)| + (|\widehat{p}_i - p_i| + |\widehat{q}_i - q_i|)}{2(\beta_2 - \beta_1)} & si\ \beta_2 - \beta_1 > 0 \\ 0 & si\ \beta_2 - \beta_1 = 0 \end{cases} \tag{2.73}$$

Propiedad 2.2. *Sean $A = (a_A, p_A, q_A)$, B, A' y B' cuatro números difusos de tipo LR. Entonces*
(1) $0 \leq T \leq 1$
(2) T(A, A) = 0
(3) $T(A, B) = T(B, A)$
(4) Si los números A y B están más distantes *que los números A' y B', luego* $T(A, B) > T(A', B')$.

Se dirá que A y B son más distantes que A' y B', si:

I) A=A' y $a_A - a_B = a_{A'} - a_{B'} + \Delta, con\ \Delta > 0$
II) $a_A - p_A - (a_B - p_B) = a_{A'} - p_{A'} - (a_{B'} - p_{B'}) + \Delta, con\ \Delta > 0$
III) $a_A + q_A - (a_B + q_B) = a_{A'} + q_{A'} - (a_{B'} + q_{B'}) + \Delta, con\ \Delta > 0$

Para demostrar (4), se escribirá $T(A', B') = \frac{N}{2D}$, y luego, como $T(A', B')$ es menor que uno, se tiene

$$T(A, B) = \frac{N + 2\Delta}{2D + 2\Delta} > \frac{N}{2D} = T(A', B') \tag{2.74}$$

La cualidad (4) anterior es importante para garantizar el sentido correcto del indicador T de divergencia. Sin embargo, cuando se tienen números difusos con funciones de pertenencia no simétricas, como es el requisito que hemos puesto a nuestros modelos, puede darse alguna situación contradictoria.

Por ejemplo, si suponemos dos números A y B próximos con $a_A + q_A > a_B + q_B$, y sólo se acorta la extensión derecha más extrema $a_A + q_A$, es decir:

I) $p_{A'} = p_{B'} = p_A = p_B$

II) $a_{A'} = a_{B'} = a_A = a_B$

III) $0 = q_{A'} < q_{B'} = q_A = qB$ (se considerará $a_{A'} + q_{A'} \geq a_B + q_B$ sólo para facilitar la demostración),

entonces $T(A', B') > T(A, B)$.

Para comprobar lo anterior, se escribirá $T(A, B) = \frac{N}{2D}$, de forma que se tiene

$$T(A', B') = \frac{N}{2(D - q_A)} > \frac{N}{2D} = T(A, B) \tag{2.75}$$

lo que indica que la no simetría de las funciones de pertenencia puede producir algunas inconsistencias al respecto del indicador T. Este resultado puede leerse en el sentido de que T es un indicador tanto de divergencia como de simetría, pero en nuestro contexto de buscar el ajuste o proximidad entre Y_i y $\widehat{Y}_i$ no es deseable.

Para el conjunto de los datos de una regresión se define el indicador de bondad de ajuste global

$$SIM_2 = \frac{\sum_i^n (1 - T_i)}{n} \tag{2.76}$$

Indicador varía entre 0 y 1.

2.4.2.3. Indice de Bondad del Ajuste SIM_3

La misma idea del indicador anterior T puede ser usada agregando la desviación de la tendencia central a la medida:

$$R_i = R(Y_i, \widehat{Y}_i) = \frac{|\widehat{y}_i - \widehat{p}_i - (y_i - p_i)| + |\widehat{y}_i + \widehat{q}_i - (y_i + q_i)| + (|\widehat{y}_i - y_i|)}{3(\beta_2 - \beta_1)} \tag{2.77}$$

Propiedad 2.3. *Sean $A = (a_A, p_A, q_A)$, B, A' y B' cuatro números difusos de tipo LR. Entonces*
(1) $0 \leq R \leq 1$
(2) R(A, A) = 0
(3) $R(A, B) = R(B, A)$
(4) Si los números A y B están más distantes *que los números A' y B', luego* $R(A, B) > R(A', B')$.

Se dirá que A y B son más distantes que A' y B', si:

I) A=A' y $a_A - a_B = a_{A'} - a_{B'} + \Delta, con\ \Delta > 0$

II) $a_A - p_A - (a_B - p_B) = a_{A'} - p_{A'} - (a_{B'} - p_{B'}) + \Delta, con\ \Delta > 0$
III) $a_A + q_A - (a_B + q_B) = a_{A'} + q_{A'} - (a_{B'} + q_{B'}) + \Delta, con\ \Delta > 0$

Para demostrar (4), se escribirá $R(A', B') = \frac{N}{3D}$, y, luego, como $R(A', B')$ y $R(A, B)$ son menores que uno, se tiene

$$R(A, B) = \frac{N + 3\Delta}{3(D + \Delta)} > \frac{N}{3D} = R(A', B') \tag{2.78}$$

Es decir R, al igual que T, como medida de divergencia, tiene el sentido correcto de aumentar de valor cuando los números difusos se separan.

En el contraejemplo que se colocó para T, en caso de aumentar la asimetría de un número difuso, se tiene, escribiendo $R(A, B) = \frac{N}{3D}$, lo siguiente

$$R(A', B') = \frac{N - q_A}{3(D - q_A)} \tag{2.79}$$

Es decir, el numerador disminuye en q_A mientras que el denominador disminuye en $3q_A$. Sin embargo, dada la relación que existe entre N y D, si $D >> N$, que es la situación más interesante porque indica que A y B están próximos, entonces

$$R(A', B') = \frac{N - q_A}{3(D - q_A)} < \frac{N}{3D} = R(A, B) \tag{2.80}$$

lo que produce un resultado *deseado*; si $D \approx N$, se tiene que

$$R(A', B') = \frac{N - q_A}{3(D - q_A)} \approx \frac{N}{3D} = R(A, B) \tag{2.81}$$

y, finalmente, si $D < N$, que es la situación en que los números se encuentran muy distantes, y por lo tanto, la menos importante, se tiene

$$R(A', B') = \frac{N - q_A}{3(D - q_A)} > \frac{N}{3D} = R(A, B) \tag{2.82}$$

que es el caso contradictorio con el *sentido* del indicador, pero que es el menos relevante, porque los números observados y estimados debieran estar muy distantes.

Con esta definición de R_i, se construye un tercer índice de bondad de ajuste, que varía entre 0 y 1:

$$SIM_3 = \frac{\sum_i^n (1 - R_i)}{n} \tag{2.83}$$

2.4.2.4. Indice de Bondad del Ajuste SIM_4

Otra medida de similitud basada en la métrica de Hausdorff esta dada por la relación

$$U_i = U(Y_i, \widehat{Y}_i) = \begin{cases} \frac{max(|\widehat{y}_i - \widehat{p}_i - (y_i - p_i)|,\ |\widehat{y}_i + \widehat{q}_i - (y_i + q_i)|)}{(\beta_2 - \beta_1)} & if\ \beta_2 - \beta_1 > 0 \\ 1 & if\ \beta_2 - \beta_1 = 0 \end{cases} \tag{2.84}$$

Propiedad 2.4. *Sean $A = (a_A, p_A, q_A)$, B, A' y B' cuatro números difusos de tipo LR. Entonces*
(1) $0 \leq U \leq 1$
(2) U(A, A) = 0
(3) $U(A, B) = U(B, A)$
(4) Si los números A y B están más distantes *que los números A' y B', luego* $U(A, B) > U(A', B')$.

Se dirá que A y B son más distantes que A' y B', si:

I) A=A' y $a_A - a_B = a_{A'} - a_{B'} + \Delta, con\ \Delta > 0$
II) $a_A - p_A - (a_B - p_B) = a_{A'} - p_{A'} - (a_{B'} - p_{B'}) + \Delta, con\ \Delta > 0$
III) $a_A + q_A - (a_B + q_B) = a_{A'} + q_{A'} - (a_{B'} + q_{B'}) + \Delta, con\ \Delta > 0$

Para demostrar (4), se escribirá $U(A', B') = \frac{N}{D}$, y luego, como $U(A', B')$ y $U(A, B)$ son menores que uno, se tiene

$$U(A, B) = \frac{N + \Delta}{D + \Delta} > \frac{N}{3D} = U(A', B') \tag{2.85}$$

Este resultado confirma el *sentido corrector* de la medición del ajuste por parte de U_i.

Sin embargo, el mismo contraejemplo utilizado anteriormente, en que, si suponemos dos números A y B próximos con $a_A + q_A > a_B + q_B$, y sólo se acorta la extensión derecha más extrema $a_A + q_A$, es decir:

I) $p_{A'} = p_{B'} = p_A = p_B$
II) $a_{A'} = a_{B'} = a_A = a_B$
III) $0 = q_{A'} < q_{B'} = q_A = qB$ (se considerará $a_{A'} + q_{A'} \geq a_B + q_B$ sólo para facilitar la demostración),

entonces $U(A', B') > U(A, B)$.

Para comprobar lo anterior, se escribirá $U(A, B) = \frac{N}{D}$, de forma que se tiene,

$$U(A', B') = \frac{N}{(D - q_A)} > \frac{N}{D} = T(A, B) \tag{2.86}$$

Esto indica que la no simetría de las funciones de pertenencia puede producir algunas inconsistencias en la medición del indicador U_i, ya que en un contexto en que el ajuste o proximidad entre Y_i y $\widehat{Y}_i$ *mejora*, porque un extremo de la función de pertenencia se acerca al valor central de ambas funciones de pertenencia, el indicador muestra una mayor divergencia.

Considerando U_i para el conjunto de n observaciones, se crea otro índice de bondad de ajuste, que fluctúa entre 0, cuando los n números observados se encuentra muy distantes de los n números estimados, y 1, cuando las funciones de pertenencia de las n parejas de números difusos son iguales. Su formulación es la siguiente:

$$SIM_4 = \frac{\sum_i^n (1 - U_i)}{n} \tag{2.87}$$

2.4.2.5. Indice de Bondad del Ajuste SIM_5

Para terminar nuestro estudio, nos centramos ahora en una medida basada sólo en un punto de las funciones de pertenencia: el supremo de la intersección. Esta medida se define como

$$V_i = sup(\mu_{Y_i \bigcap \widehat{Y}_i}(x)), \tag{2.88}$$

y calcula un valor de pertenencia (entre 0 y 1) para el dato i. Dados dos conjuntos difusos $A \subset B$, se tiene que el valor V_i de ellos es 1, por lo que las propiedades del indicador de bondad de ajuste primero que presentamos en nuestro estudio también son aplicables a este índice.

La determinación de V_i como cruce de dos funciones de pertenencia se ubica entre los dos puntos centrales del número observado y del número difuso estimado, por lo que tiene una directa relación con dichos dos puntos y es, por lo tanto, un indicador aproximado del ajuste de la tendencia central.

Por lo general, el comportamiento de los extremos no afecta a V_i, especialmente si consideramos funciones de pertenencia no simétricas. Por este motivo, en un contexto de estimación posibilística, puede resultar distinto a los índices de bondad de ajuste definidos anteriormente.

Sin embargo, en una estimación no posibilística, en que las extensiones estimadas son más pequeñas, puede ocurrir que V_i sea menor que el producido en la estimación posibilística, dado que ante soportes mucho más pequeños, el supremo de la intersección de las funciones de pertenencia no garantiza ser alto aunque los puntos centrales estén próximos.

Tomando la definición de V_i, puede extenderse para al conjunto de datos de la regresión, a otra medida de bondad de ajuste, que varía entre 0 y 1:

$$SIM_5 = \frac{\sum_i^n (V_i)}{n} \tag{2.89}$$

2.4.2.6. Medida de Bondad del Ajuste de la Tendencia Central

Para medir la calidad del ajuste de la tendencia central, se conoce de la Regresión Probabilistica el *coeficiente de determinación*, llamado también *R-cuadrado*, que varía entre 0 y 1, y es un indicador muy utilizado.

Esta medida parte del concepto de que la varianza total de las observaciones de la variable dependiente se puede descomponer en la suma de la varianza explicada más la varianza no explicada, interpretándose el coeficiente de determinación como la proporción de la varianza total que es explicada por la regresión. Formalmente tenemos que

$$R^2 = \frac{Varianza\ \ explicada}{Varianza\ \ total} = \frac{\sum(\widehat{y_i} - \overline{y})^2}{\sum(y_i - \overline{y})^2} \tag{2.90}$$

Chang propone, como medida equivalente, el llamado *coeficiente de correlación híbrido*, que generalmente toma valores entre 0 y1, pero puede sobrepasar este máximo, como se aprecia en los ejemplos del artículo [Cha01b]. La fórmula de Chang es

idéntica a la del coeficiente de determinación, pero en vez de operar con la aritmética real, ocupa la *aritmética difusa ponderada*. La fórmula del coeficiente de correlación híbrido 2.46 no nos parece adecuada para la regresión difusa, porque puede tomar valores bastante mayores a 1.

Lo mismo ocurre con la propuesta de D'Urso, que es una generalización de la propuesta de Chang para R^2, (2.51 y 2.52), porque sólo garantiza que son iguales o mayores que cero, pero no puede determinase una cota superior. Además, ambos casos comparan con el número estimado, y sus extensiones, con los promedios de los datos observados, y los promedios de sus extensiones, lo que no nos parece lo más riguroso.

En nuestro caso, proponemos considerar un nuevo indicador R^2 de tendencia central, cuya principal característica es que, a medida que las diferencias cuadráticas entre el valor central observado y el valor central estimado tiende a cero, el indicador tenderá a uno:

$$R^2_{difuso} = max\Big(0,\ 1 - \frac{\sum_i^n (y_i - \widehat{y_i})^2}{\sum_i^n (y_i - \overline{y})^2}\Big) \tag{2.91}$$

donde $\overline{y}$ es el promedio de las observaciones y_i. Este nuevo indicador toma valores entre 0 y 1.

Como el denominador que aparece en la expresión es el mismo que el denominador del coeficiente de determinación probabilístico, se pude mantener la interpretación de éste, en el sentido de que R^2_{difuso} es una medida de proporción de la parte de la variación (cuadrática) de los y_i que es explicada por la regresión. Por ejemplo, si R^2_{difuso} resulta 0.7, lo interpretaremos como que la regresión explica el 70 % de la variación de los datos centrales y_i.

Otra medida que es interesante considerar, es el aporte que realiza a un modelo de regresión difusa la inclusión de una variable adicional. Para este fin, se tomará a R^2_{difuso} como una referencia del aporte del conjunto de variables a la regresión.

Se notará $R^2_{difuso}(X_1, ..., X_k)$ al coeficiente R^2_{difuso} cuando el modelo tiene como variables de entrada a $X_1, ..., X_k$; se escribirá $RP^2_{difuso}(X_k|X_1, ..., X_{k-1})$ el indicador del aporte de incorporar X_k al modelo, dado que ya estaban presentes en el modelo X_1 a X_{k-1}. Esta definiciones operan para un coeficiente de confianza h determinado y parámetros de ponderación k_1 y k_2 específicos. Por tanto, este indicador *R cuadrado parcial* queda definido analíticamente como:

$$RP^2_{difuso}(X_k|X_1,...,X_{k-1}) = R^2_{difuso}(X_1,...,X_k) - R^2_{difuso}(X_1,...,X_{k-1}) \tag{2.92}$$

donde se define $R^2_{difuso}(\emptyset) = 0$.

Que una variable de entrada tenga un coeficinete parcial pequeño, no significa, necesariamente, que no pueda ser una variable significativa en el modelo de regresión. Esta situación será comentada con mayor detalle en el capítulo de regresión difusa con multicolinealidad en los datos de entrada.

En resumen, las seis medidas de confiabilidad de la regresión definidas en este sección, están normalizadas (varían entre 0 y 1), con 1 significando total ajuste y 0 total divergencia. Sus sentidos son los siguientes:

- SIM_1 pondera las diferencias entre las distribuciones de posibilidad de Y_i e $\widehat{Y_i}$ incluyendo la totalidad de las funciones de pertenencia.
- SIM_2 mide las diferencias en el soporte, tanto del punto central como sus dos extensiones, entre los valores de salida y sus respectivas estimaciones.
- SIM_3 mide las diferencias tanto de las extensiones como de la tendencia central.
- SIM_4 mide la diferencia máxima de las extensiones de los datos de entrada con sus respectivas estimaciones.
- SIM_5 mide la proximidad de las funciones de pertenencia con un solo punto, el supremo de la intersección.
- R^2_{difuso} mide las diferencias cuadráticas del valor central observado con el valor central estimado.

La tabla 2.1 resume su formulación.

2.4.2.7. Un Indice de Bondad del Ajuste Integrado: SIM

De las medidas de bondad del ajuste presentadas, SIM_1 es la que mejor representa, teóricamente, el parecido entre los números difusos que se comparan, al incorporar en su integridad las funciones de pertenencia de ambos números.

Tabla 2.1: Resumen de las medidas de calidad del ajuste

Indice	Formulacón
SIM_1	$\frac{\sum_i^n \frac{\lvert Y_i \bigcap \widehat{Y_i}\rvert}{\lvert Y_i \bigcup \widehat{Y_i}\rvert}}{n}$
SIM_2	$\frac{\sum_i^n (1 - \frac{\lvert\widehat{y_i}-\widehat{p_i}-(y_i-p_i)\rvert+\lvert\widehat{y_i}+\widehat{q_i}-(y_i+q_i)\rvert+(\lvert\widehat{p_i}-p_i\rvert+\lvert\widehat{q_i}-q_i\rvert)}{2(\beta_2-\beta_1))}}{n}$
SIM_3	$\frac{\sum_i^n (1 - \frac{\lvert\widehat{y_i}-\widehat{p_i}-(y_i-p_i)\rvert+\lvert\widehat{y_i}+\widehat{q_i}-(y_i+q_i)\rvert+(\lvert\widehat{y_i}-y_i\rvert)}{3(\beta_2-\beta_1)})}{n}$
SIM_4	$\frac{\sum_i^n (1 - \frac{max(\lvert\widehat{y_i}-\widehat{p_i}-(y_i-p_i)\rvert,\ \lvert\widehat{y_i}+\widehat{q_i}-(y_i+q_i)\rvert)}{(\beta_2-\beta_1)})}{n}$
SIM_5	$\frac{\sum_i^n (sup(\mu_{Y_i \bigcap \widehat{Y_i}}(x)))}{n}$
R^2_{difuso}	$max\left(0,\ 1 - \frac{\sum_i^n (y_i-\widehat{y}_i)^2}{\sum_i^n (y_i-\overline{y})^2}\right)$

Sin embargo, la forma aritmética de la medida en sí, el cociente entre la intersección y la unión de ambos conjuntos, con la particularidad de la forma concreta de las funciones de pertenencia triangulares utilizadas en este trabajo de investigación, podría presentar algunos resultados del cálculo concreto del indicador que hicieran necesario una revisión de esta medida. Para estudiar esta inquietud, se mostrarán tres situaciones diferentes donde SIM_1 tiene el mismo valor.

En el gráfico 2.3 se aprecian dos comparaciones entre Y_i e $\widehat{Y_i}$ en que el valor SIM_1 tiene el mismo valor de 0.17. Si bien no se puede plantear que la estimación (a) se ajusta mejor o peor que la estimación (b) – dependerá en cada problema específico cuál situación es más valorada por el análisis (o si es indiferente), se plantea la duda del comportamiento del indicador en situaciones intermedias entre ambas.

Por otra parte, en el gráfico 2.4 se vuelven a repetir dos estimaciones en que el indicador SIM_1 tiene el valor de 0.17.

En este segundo caso, la extensión derecha de la estimación parte junto a la extensión derecha del dato original (a), pero se va alargando hasta ser muy superior en magnitud a la del dato original (b).

Si analizamos en mayor detalle este gráfico, se verá que los indicadores que reflejan el ajuste de los extremos (SIM_2, SIM_3 y SIM_4) tenderán a tener un valor más alto en la situación (a) que en la situación (b), mientras el indicador SIM_5, supremo de la intersección de ambas funciones de pertenencia, tendrá un valor más alto en la gráfico (b) que en el gráfico (a).

En resumen, de los cinco indicadores que hemos definido, basados en la Teoría

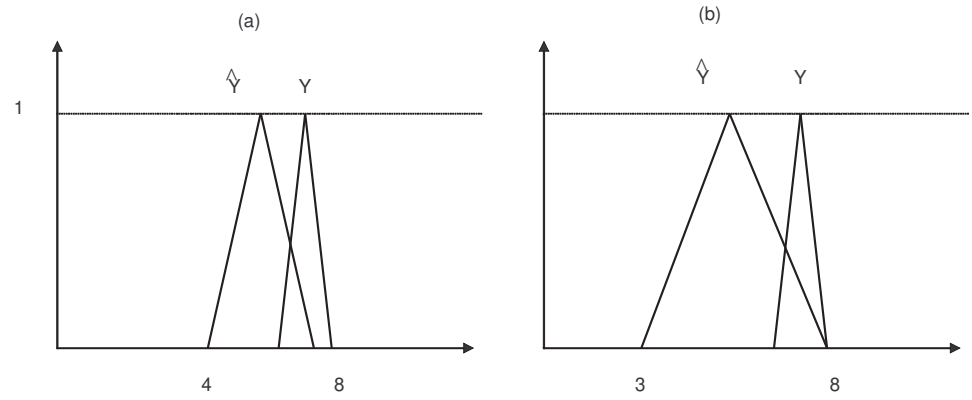

Figura 2.3: Equivalencia del índice SIM_1 para distintas funciones de pertenencia (1)

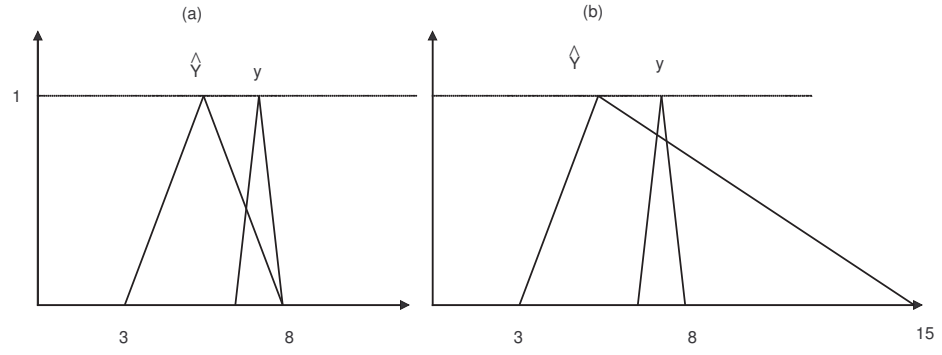

Figura 2.4: Equivalencia del índice SIM_1 para distintas funciones de pertenencia (2)

de Conjuntos Difusos, SIM_1 valora con una igualdad ambas situaciones del gráfico 2.4, mientras otras medidas valoran más la situación (a) y una la situación (b).

Desde nuestro punto de vista, en cambio, basado en una concepción más tradicional de la regresión, que es similar a la que se desprende de la literatura de la Regresión Difusa, y que se expresa en la forma de la función objetivo de minimizar tanto la tendencia central como las diferencias de las extensiones, en este ejemplo es preferible la situación (a) a la situación (b), ya que la diferencia en la tendencia central es igual en ambas gráficas, pero las diferencias de las extensiones es mucho menor en el caso (a).

Del gráfico 2.4 se desprende también otra situación importante. Se plantea la duda de si el índice es constante durante todo el cambio de magnitud de la extensión derecha, desde (a) - extensión de largo 3 - hasta (b) - extensión de largo 10 -, o si el índice tiene otro tipo de comportamiento más variable.

Esta duda se resuelve en la figura 2.5 en que se muestra el indice SIM_1 en función del tamaño de la extensión derecha de la estimación, a partir de 0. Se observa que el indicador aumenta hasta un máximo, para luego empezar a disminuir lentamente. Lo que queremos destacar, es que el máximo de SIM_1 no se produce cuando coinciden ambos extremos derechos, que corresponde a la extensión de 3 para la estimación, sino SIM_1 continua aumentando a medida que aumenta la extensión, hasta el valor máximo que corresponde a la extensión de 4, y que justo cuando la extensión alcanza el valor 6 se iguala el valor de SIM_1 con el que tenía con la extensión derecha 3.

Sin embargo, nuestro concepto de ajuste tiene implícito que cuando los extremos de ambos datos están más próximos, entonces el ajuste es *mejor*, por lo que la cuantificación que se produce en este ejemplo, cuando el tamaño de extensión aumenta de 3 hasta 6, es una situación no deseada en la cuantificación del ajuste.

Para afrontar esta ambigüedad, sería preferible un indicador que, para este ejemplo, entre los valores 3 y 6 de extensión derecha, disminuyera, o se mantuviera constante.

Para lograr esta finalidad se propone un indicador integrado de similitud, entre los índices SIM_1 y SIM_2, que queda definido por

$$SIM = \frac{SIM_1 + SIM_2}{2} \tag{2.93}$$

y cuyo comportamiento para el ejemplo que se representa en el gráfico 2.4 se muestra

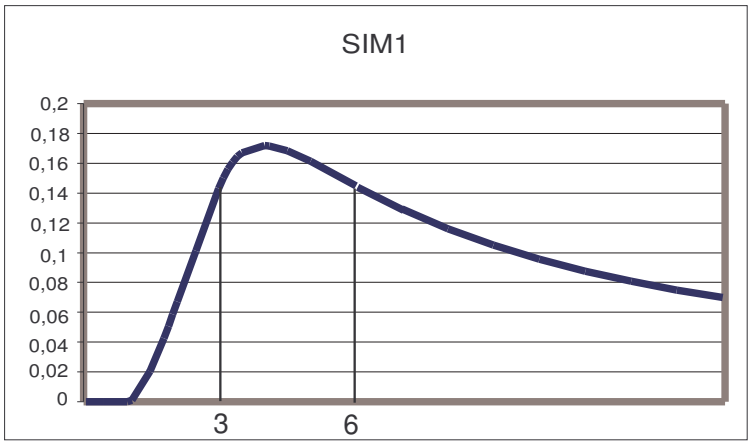

Figura 2.5: Indice SIM_1 variando la extensión derecha de la estimación

en el siguiente gráfico 2.6.

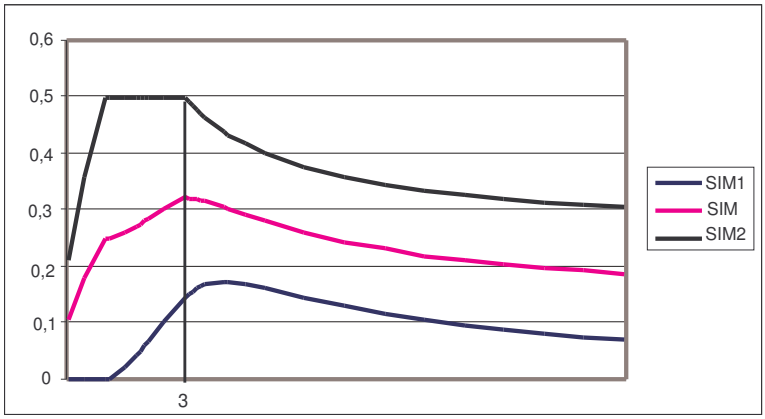

Figura 2.6: SIM_1, SIM_2 y SIM variando la extensión derecha de la estimación (1)

En este gráfico se aprecia que al llegar a 3 la extensión, el crecimiento de SIM se detiene, empezando a disminuir rápidamente.

Un ejemplo parecido, pero con distinta extensión derecha, se presenta en el gráfico 2.7, en que el nuevo índice SIM permanece, prácticamente, constante, a partir de la extensión derecha de 3 hasta el valor 4.3, para empezar a disminuir. En realidad, entre 3 y 3.5, el índice SIM experimenta un pequeño aumento.

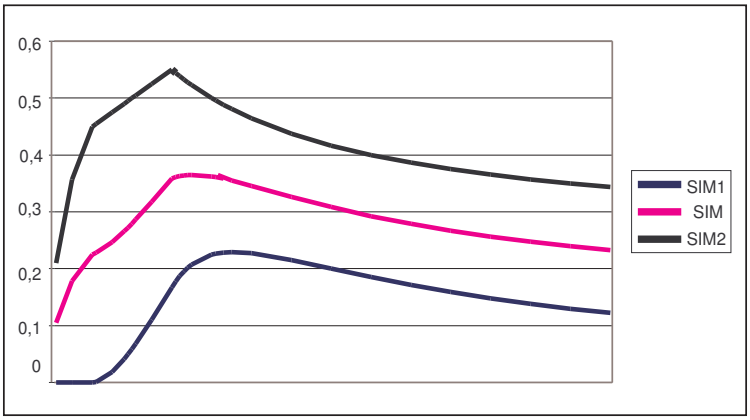

Figura 2.7: SIM_1, SIM_2 y SIM variando la extensión derecha de la estimación (2)

Otra situación que corrige el indicador SIM, cuando consideramos la regresión no posibilística, es la distancia entre dos funciones de pertenencia disjuntas, como la que aparecía en la figura 2.8.

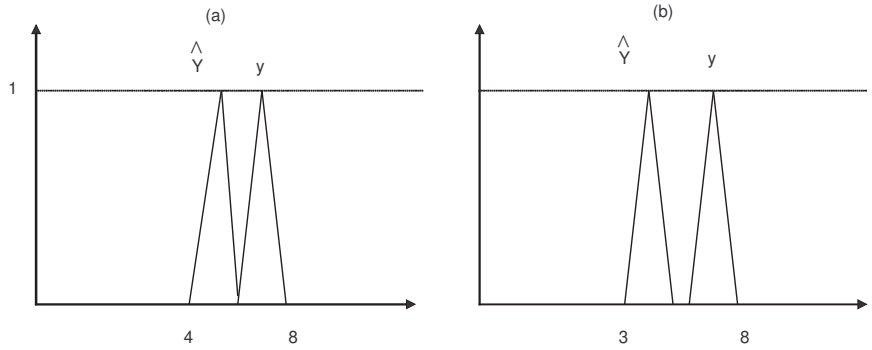

Figura 2.8: Dos estimaciones con el mismo indicador SIM_1=0

En esta figura, se muestran dos estimaciones distintas, y en ambas SIM_1 toma el valor 0, cuando nuestro *sentido* del ajuste de regresión nos indica que la estimación (a) es mejor ajuste que la estimación (b), y debiera tener un mejor indicador que la segunda.

En la figura 2.9 se aprecia como el indicador SIM resuelve esta ambigüedad. En este gráfico se muestran los valores de SIM_1, SIM_2 y SIM en función de la

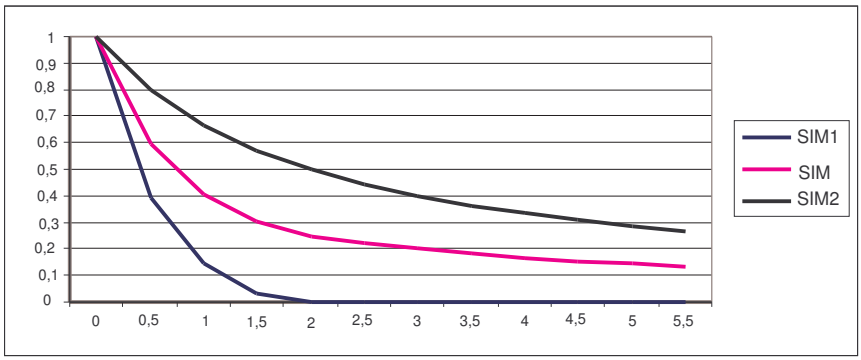

Figura 2.9: SIM_1, SIM_2 y SIM distanciando el centro de la estimación

distancia entre los centros de dos funciones de pertenencia simétricas. Cuando ambos centros coinciden, los indicadores toman el valor 1, cuando los centros empiezan a alejarse, los indicadores disminuyen sus valores hasta que los valores comparados tienen sólo un punto en común. En ese momento SIM_1 llega a cero y así continua aunque la diferencia de los centros aumente. Mientras tanto SIM_2 continua disminuyendo su valor. Se aprecia que el comportamiento de SIM es una función decreciente, que tiende asintóticamente a cero.

Capítulo 3

Métodos de Regresión Difusa

En este capítulo afrontamos una de las principales partes de nuestro trabajo de investigación, con la proposición y estudio de distintos métodos para afrontar el análisis de Regresión Difusa. Se verá en detalle el enfoque posibilístico de la Regresión Difusa, primero en su versión lineal y luego en su versión cuadrática.

Como veremos, al proponer una nueva función objetivo que incluye una minimización más precisa de las extensiones, se logrará definir un modelo que puede operar tanto con las restricciones posibilísticas como sin ellas.

De igual manera, trataremos de compatibilizar el criterio de minimización de las extensiones de las estimaciones, *criterio de incertidumbre*, junto al criterio de minimización de mínimos cuadrados de la tendencia central o modal, fusionando en un único modelo ambos objetivos.

De esta manera, el modelo que se propone de Regresión Difusa Lineal, compatibiliza los enfoques posibilístico y no posibilístico de la Regresión Difusa, al mismo tiempo que también compatibiliza los objetivos de minimización de la incertidumbre con el objetivo de mínimos cuadrados.

En el estudio que a continuación comenzamos, se proponen, dentro de este nuevo enfoque, diversos modelos basados en índices construidos a partir de la Teoría de la Posibilidad.

Concretamente, se proponen siete nuevos métodos de Regresión Difusa, para los cuales se hace una experimentación, que es medida mediante los indicadores de bondad de ajuste que hemos analizado y definido al final del anterior capítulo, con el

objetivo de determinar las particularidades de cada modelo, y poder concluir con recomendaciones precisas sobre su implementación.

3.1. Enfoque Lineal para la Regresión Difusa

Ya hemos comentado que para enfrentar el objetivo general de la regresion (1.1), la Regresión Difusa propone una alternativa flexible para cumplir con la ecuación.

Al describir el aporte de Tanaka a la Regresión Difusa, hemos mostrado las restricciones (2.4)-(2.6) para el modelo en que los coeficientes son funciones triangulares no simétricas.

En tal caso, las funciones de pertenencia de cada coeficiente A_i tienen la forma

$$\mu_{A_i}(x) = \begin{cases} 0 & \text{if } x \leq a_i - c_{L_i} \\ 1 - \frac{a_i - x}{c_{L_i}} & \text{if } a_i - c_{L_i} \leq x \leq a_i \\ 1 - \frac{x - a_i}{c_{R_i}} & \text{if } a_i \leq x \leq a_i + c_{R_i} \\ 0 & \text{if } a_i + c_{R_i} \leq x \end{cases}$$

y Tanaka [TUA82] demuestra que, aplicando el Principio de Extensión de Zadeh, la función de pertenencia del valor i-esimo estimado por la regresión es

$$\mu_{\widetilde{y_i}}(y) = \begin{cases} 1 - \frac{y - \sum a_j X_{ij}}{\sum c_{R_i} |X_{ij}|} & \text{if } \sum a_j X_{ij} \leq y \leq \sum a_j X_{ij} + \sum c_{R_i} |X_{ij}| \\ 1 - \frac{\sum a_j X_{ij} - y}{\sum c_{L_i} |X_{ij}|} & \text{if } \sum a_j X_{ij} - \sum c_{L_i} |X_{ij}| \leq y \leq \sum a_j X_{ij} \\ 0 & \text{en otro caso} \end{cases}$$

Para no considerar el signo en el producto escalar de un número difuso, se añade la condición

$$x_{ij} \geq 0 \text{ para } i = 1, ..., n \; j = 1, ..., m \tag{3.1}$$

que no es restrictiva desde el punto de vista conceptual, ya que se puede transformar cualquier variable x con valores negativos, en otra variable con todos los valores positivos, sin más que sumar el máximo valor negativo a cada término de la matriz x.

Por lo general, la literatura sobre Regresión Difusa ha considerado el uso de funciones LR simétricas. En esta trabajo de investigación se asume siempre el uso de

funciones de pertenencia no simétricas. Esta incorporación, que sólo es una generalización sobre las funciones de pertenencia simétricas, significa agregar una ventaja sobre la Regresión Probabilística. Esta última, al calcular los intervalos de confianza de toda estimación puntual, siempre supone una distribución de probabilidades simétrica, mientras que en el caso difuso, las estimaciones de números difusos a partir de los coeficientes difusos de regresión, que equivalen a los intervalos de confianza probabilísticos por tener incorporada la incertidumbre, se les permite ajustar con mayor libertad la estimación al no tener que ser, necesariamente, simétricos.

En consecuencia, sólo tomando en cuenta el objetivo que debe satisfacer la regresión, es decir, que se cumpla la ecuación (1.1), se han llegado a definir las condiciones posibilísticas, que forman un conjunto de 2n inecuaciones. Para el caso lineal y con funciones de pertenencia no simétricas LR, estas condiciones quedan reflejadas en (2.5)-(2.6). Además, en el sentido de la Teoría de Conjuntos Difusos, hay un valor de mayor pertenencia, que corresponde a las estimaciones sin extensión, que se denominará *tendencia central* de la estimación.

Para cada Y_i, el objetivo que se debe tener para calcular cada valor estimado $\widehat{Y_i}$ no es que su valor de pertenencia para y_i (tendencia central de Y_i) sea lo más alto posible, porque dadas las características de las funciones LR, eso sólo se logra con extensiones muy amplias. Por el contrario, el objetivo es que se cumpla el principio posibilístico de que $Y_i \subseteq \widehat{Y_i}$.

También se indicó que la función objetivo (2.10) representa la función lineal mejorada de la regresión de Tanaka, que nosotros extendemos a funciones de pertenencia no simétricas, generándose un problema de programación lineal con el cual se obtienen las estimaciones para los coeficientes A_i.

De esta manera, configuramos el primer modelo de Regresión Difusa como sigue.

Modelo 3.1. *Modelo de Regresión Difusa Lineal, LIN.*

Función objetivo:

$$Min \sum_{i=1}^{n} \sum_{j=1}^{m} (c_{Lj} + c_{Rj}) x_{ij} \tag{3.2}$$

Restricciones:

$$\sum_{j=1}^{m} a_j X_{ij} - (1-h) \sum_{j=1}^{m} c_{L_j} X_{ij} \leq y_i - (1-h) p_i \;\; para \;\; i = 1, ..., n \tag{3.3}$$

$$\sum_{j=1}^{m} a_j X_{ij} + (1-h)\sum_{j=1}^{m} c_{R_j} X_{ij} \geq y_i + (1-h)q_i \; para \; i = 1, ..., n \tag{3.4}$$

$$c_{L_j}, c_{R_j} \geq 0 \;\; para \;\; j = 1, ..., m \tag{3.5}$$

Tabla 3.1: Datos de ejemplo: precios de la acción de Copec

$Celulosa$	$Cobre$	p_i	y_i	q_i
472.5	72.59	62.5	37.5	2352
465.0	72.41	28.4	86.6	2228
510.0	67.31	44.2	30.8	2534
510.0	67.47	21.9	38.1	2471
499.0	71.63	39.1	20.9	2409
490.0	73.52	32.6	17.4	2402
490.0	72.31	71.1	38.9	2471
480.0	72.93	50.8	29.2	2550
480.0	76.21	50.3	39.7	2480
480.0	76.55	87.4	88.6	2586
481.3	76.54	53.7	61.2	2688
504.0	75.81	49.7	70.3	2749
552.4	76.87	156.4	228.5	3556
534.0	76.66	111.6	168.5	3551
531.7	75.71	133.8	95.2	3603
521.1	77.99	60.8	89.2	3560
516.5	80.93	131.7	98.3	3751
511.6	80.00	107.9	242.0	3957
520.5	81.01	209.5	245.4	4179
527.7	81.84	202.9	167.1	4352
534.3	85.32	267.8	252.2	4437
545.6	92.66	309.1	185.9	4389
550.0	93.64	314.4	125.6	4214

En la Tabla 3.1 aparecen los datos del ejemplo que consideraremos en este capitulo, que corresponden al precio quincenal de la acción de la compañía chilena COPEC, con dos variables de entrada (precio de la celulosa y precio del cobre), cuya estimación, mediante programación lineal y considerando un nivel de confianza $h = 0$, que se analizará en todos los ejemplos de este capítulo, produce los resultados de la tabla 3.2.

Tabla 3.2: Estimaciones del ejemplo de COPEC con el modelo LIN

$Coeficiente$	a_i	c_L	c_R
constante	-5853	0	0
$Celulosa$	9.8	0	0
$Cobre$	55.5	8.9	8.4

De acuerdo a estos resultado, a título de ejemplo, se muestra en la figura 3.1, el valor de la acción para el dato 21, y la estimación del modelo lineal para el mismo dato.

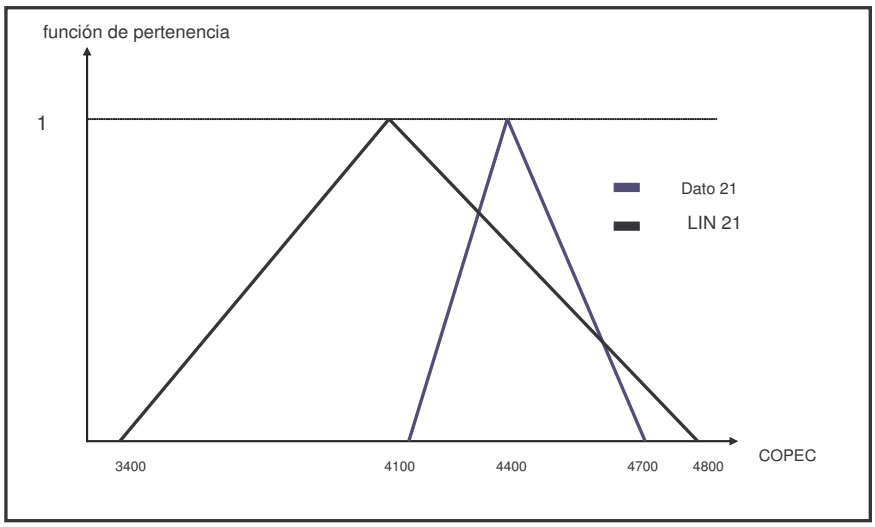

Figura 3.1: Función de pertenencia del dato 21 y su estimación por el modelo LIN

Una crítica a esta definición de función objetivo lineal de la Regresión Difusa es que, en muchas aplicaciones, produce muchos coeficientes A_i precisos (crisp). Esta consecuencia puede deducirse de la forma del espacio de soluciones factibles que generan las restricciones posibilísticas en su forma lineal, que se describe en la figura 3.2.

En esta figura se representa el espacio de soluciones factibles para la estimación de una de las extensiones de dos coeficientes, c_1 de A_1 y c_2 de A_2. Como se ve, las soluciones al minimizar las extensiones mediante funciones lineales se encontrarán, en general, en el punto en que la recta que limita el espacio de soluciones factibles corta el eje c_1, o el eje c_2, con lo que una de las extensiones será crisp. Por esta razón, en general, una de las dos extensiones será igual a cero.

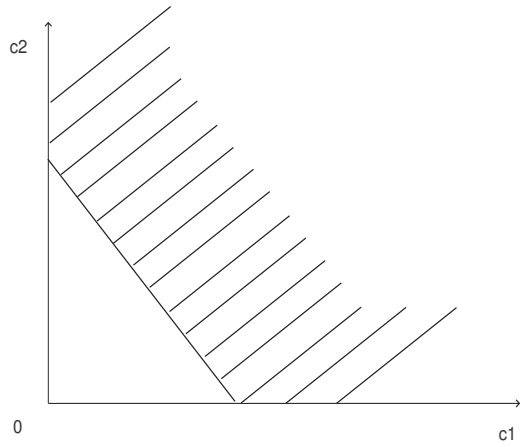

Figura 3.2: Soluciones factibles para la Regresión Posibilística Lineal

En este modelo, con función objetivo lineal que minimiza las extensiones de las estimaciones, no es posible omitir un criterio de restricción, como las restricciones posibilísticas que se definieron, ya que sin ellas, las extensiones resultan cero.

Como mencionamos en capítulos anteriores, una crítica a la Regresión Difusa (y en particular a la Regresión Difusa Posibilística) es que no produce una interpretación del intervalo estimado.

Para atender esta crítica, sería fácil afirmar que cada coeficiente, de acuerdo al objetivo inicial de la función objetivo ($\sum_j A_j x_j$), representa la magnitud de la contribución que hace cada variable a la intertidumbre de la estimación. Sin embargo, nos parece que esta apreciación no es verdadera, debido a la distorsión que provocan las restricciones a las que se somete la función objetivo.

Las restricciones posbilísticas producen, en la práctica, que las extensiones sean determinadas, principalmente, por los puntos extremos de los datos. En la figura 3.3, se aprecia que sólo cuatro observaciones, dentro de una gran cantidad de datos, son las que determinan las caraterísticas del intervalo estimado, para un h – nivel de confianza – dado.

En consecuencia, la estimación de las extensiones en la Regresión Posibilística, está dada fundamentalmente por las características de los puntos extremos y, por tanto, su interpretación debe estar ligada a ellos. Para un nivel de confianza dado, existe una alta posibilidad de que cualquier dato nuevo que se pueda recoger se encuentre

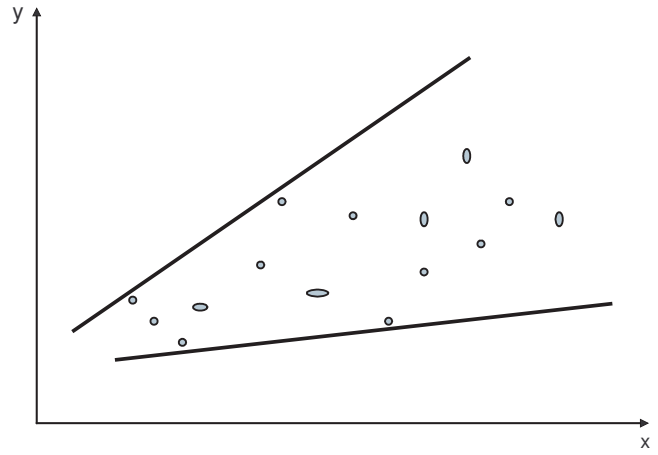

Figura 3.3: Determinación del intervalo estimado por puntos extremos

dentro del intervalo estimado.

Esta interpretación hace posible comparar el intervalo posibilístico estimado, con el intervalo de confianza de la Regresión Probabilística: dado un cierto porcentaje de confianza, generalmente un 95 %, se estima que dentro de ese margen, los nuevos datos deberían estar en tal intervalo.

Por lo tanto, los coeficientes estimados A_i por la Regresión Posibilística, no indican la contribución de cada variable a la incertidumbre total de la estimación, sino que el conjunto de ellos permite definir un intervalo que, dentro del nivel de confianza elegido, permite asegurar que normalmente nuevos datos se encontrarán dentro del intervalo posibilístico.

Se verá en la siguiente sección un nuevo enfoque en el que la función objetivo lineal es reemplazada por otra función objetivo de tipo cuadrático que tiende a ser más precisa en el criterio de minimización de las extensiones e integra el criterio de los mínimos cuadrados.

3.2. Enfoque Cuadrático para la Regresión Difusa

En esta sección se presentarán diversos modelos de regresión difusa, basados en un enfoque de programación cuadrática, que constituyen un *nuevo enfoque para la Regresión Difusa*.

El objetivo que se persigue es doble: tener una estimación más precisa de las extensiones al mismo tiempo que se incorpora un criterio de ajuste de la tendencia central.

Como verermos, el hecho de asumir funciones de pertenencia no simétricas hace posible que el doble objetivo que se persigue pueda ser armonizado exitosamente.

Por otra parte, al asumir funciones objetivo cuadráticas, es más fácil que la estimación óptima, dentro del espacio de soluciones factibles que se muestra en 3.2, no coincida con uno de los vértices y los coeficientes estimados reflejen de mejor manera la incertidumbre que cada variable aporta al sistema.

Se presentan primero los modelos basados en el criterio posibilístico y a continuación, en una segunda sección, los modelos basados en otras medidas de la Teoría de la Posibilidad.

3.2.1. Modelos con Restricciones Posibilísticas

El primero en proponer, para la regresión intervalar, una función objetivo cuadrática fue Tanaka [TL98]. Extendiendo esa propuesta para los número difusos Y y los coeficientes difusos A que nos ocupan, la incertidumbre se define como la suma al cuadrado de las extensiones de los números estimados. De esta forma, la función objetivo para la minimización de la incertidumbre queda reflejada, para funciones de pertenencia LR no simétricas, en la relación

$$J = c_L x^{'} x c_L^{'} + c_R x^{'} x c_R^{'} \tag{3.6}$$

Esta definición de incertidumbre, que suma las extensiones de las estimaciones al cuadrado, en forma similar a la suma de los residuos de la regresión de mínimos cuadrados, tiene como contrapartida la existencia de las restricciones posibilísticas, puesto que, sin estas, la minimización de las extensiones daría cero. Con esta nueva definición, sólo se pasa de la suma lineal de las extensiones a la suma cuadrática de ellas, basándonos en el supuesto de que el número de extensiones estimadas que resultan crisp disminuye con el nuevo enfoque.

A este motivo tenemos que unir la ventaja de que, como se verá en el capítulo de datos multicolineales, el enfoque cuadrático permite mayor flexibilidad para acomodar la función objetivo a nuevos requerimientos.

Hemos mencionado en la introducción que varios métodos de regresión difusa

publicados por diversos autores se basan en un objetivo de mínimos cuadrados en torno a la tendencia central de los números difusos de salida. El enfoque cuadrático que aquí proponemos, sin embargo, se concentra exclusivamente en la minimización de la incertidumbre, dejando fuera cualquier objetivo de minimización relativo a la tendencia central, es decir, objetivos respecto al punto en que la función de pertenencia toma valor 1.

Para considerar también el objetivo de tendencia central, se puede recurrir a la regresión probabilística más conocida, y considerar la función objetivo de la minimización de los mínimos cuadrados, cuya expresión es:

$$J = \sum_{i=1}^{n} (y_i - x_i a^{'})^2 \tag{3.7}$$

e integrar el objetivo de minimizar la intertidumbre con el objetivo de minimizar la desviación respecto a la tendencia central en una sola función a minimización, a través de la relación:

$$J = k_1 \sum_{i=1}^{n} (y_i - x_i a^{'})^2 + k_2 (c_L x^{'} x c_L^{'} + c_R x^{'} x c_R^{'}) \tag{3.8}$$

Esta relación incluye las constantes de peso k_1 y k_2 que permiten introducir un criterio a priori a la optimización, ponderando bien el criterio de la minimización de la desviación en torno a la tendencia central, o bien el criterio de la minimización de la incertidumbre.

Teniendo esto en cuenta, pasamos a formalizar el siguiente modelo de Regresión Difusa.

Modelo 3.2. *Modelo de regresión difusa MTE, Modelo de Tanaka Extendido.*

Función objetivo:

$$J = k_1 \sum_{i=1}^{n} (y_i - x_i a^{'})^2 + k_2 (c_L x^{'} x c_L^{'} + c_R x^{'} x c_R^{'}) \tag{3.9}$$

Restricciones:

$$\sum_{j=1}^{m} a_j X_{ij} - (1-h) \sum_{j=1}^{m} c_{L_j} X_{ij} \leq y_i - (1-h) p_i \;\; para \;\; i = 1, ..., n \tag{3.10}$$

$$\sum_{j=1}^{m} a_j X_{ij} + (1-h)\sum_{j=1}^{m} c_{R_j} X_{ij} \geq y_i + (1-h)q_i \; para \; i = 1, ..., n \tag{3.11}$$

$$c_{L_j}, c_{R_j} \geq 0 \;\; para \;\; j = 1, ..., m \tag{3.12}$$

Este modelo, junto a las restricciones posibilísticas, constituye un problema de programación cuadrática que tiene por objetivo estimar los número difusos que constituyen los estimadores del modelo de regresión. Estos número difusos son triangulares no simétricos $\widehat{A_i}$ en el intervalo $[\widehat{a_i} - \widehat{c_{L_i}}, \widehat{a_i}, \widehat{a_i} + \widehat{c_{R_i}}]$ y, por lo tanto, hay que estimar el triple de valores que el número de variables de entrada, ya que por cada variable de entrada se estiman los valores los tres coeficientes que definen el subconjunto difuso (a_i, c_{L_i}, c_{R_i}).

La forma general de un problema de programación cuadrática es

$$\begin{aligned} &Min \; \frac{1}{2}(w'Hw) + f'w \\ &sujeto \; a \; Aw \leq b \end{aligned} \tag{3.13}$$

que aplicada al modelo propuesto queda definido por

$$w = (a_1, ..., a_m, c_{L_1}, ..., c_{L_m}, c_{R_1}, ..., c_{R_m}) \tag{3.14}$$

$$\mathbf{H} = \begin{pmatrix} 2k_1x'x & 0 & 0 \\ 0 & 2k_2x'x & 0 \\ 0 & 0 & 2k_2x'x \end{pmatrix}$$

$$\mathbf{f} = \begin{bmatrix} -2k_1x'y & 0 & 0 \end{bmatrix}$$

$$\mathbf{A} = \begin{pmatrix} -x & 0 & -(1-h)x \\ x & -(1-h)x & 0 \\ 0 & diag(-1) & 0 \\ 0 & 0 & diag(-1) \end{pmatrix}$$

y

$$\mathbf{b} = \begin{pmatrix} -(y + (1-h)q) \\ y - (1-h)p \\ 0 \\ 0 \end{pmatrix}$$

Al aplicar los datos del ejemplo de la acción de COPEC, a este nuevo modelo, para algunas combinaciones de k_1 y k_2, se tienen los resultados que se muestran en la tabla 3.3.

Tabla 3.3: Estimaciones del modelo MTE para el ejemplo COPEC)

k_1	k_2	$Coeficiente$	a_i	c_L	c_R
0	1	constante	-5853	0	0
		$Celulosa$	9.8	0	0
		$Cobre$	55.3	8.7	8.7
1	1	constante	-7827	0	0
		$Celulosa$	15.7	0	0
		$Cobre$	40.5	7.9	8.5
1	0	constante	-9324	0	0
		$Celulosa$	15.4	0	0
		$Cobre$	60.0	9.7	10.6

La expresión (3.6), que se repite en la función objetivo (3.8), toma como objetivo que las extensiones de las estimaciones sean mínimas (minimiza los valores $c_{L_i} y c_{R_i}$). Por lo tanto, nuevamente, esta minimización sólo tiene sentido en el ámbito de las restricciones posibilísticas puesto que, si no existieran, el mínimo de las extensiones sería cero.

El criterio de minimización de la incertidumbre hasta aquí considerado en la Regresión Difusa Posibilística, no considera la magnitud de las desviaciones entre las extensiones de salida y las extensiones estimadas. Como los datos conocidos Yi también tienen extensiones, sería más preciso minimizar las diferencias de los extremos de los número difusos entre, por una parte, los datos iniciales, $y_i - p_i$ y $y_i + q_i$, y, por otra, los datos resultantes de la estimación. Esto queda reflejado en la siguiente nueva función objetivo:

$$J = k_1 \sum_{i=1}^{n} (y_i - a^{'} x_i)^2 + k_2 (\sum_{i=1}^{n} (y_i - p_i - (a^{'} - c_L^{'}) x_i)^2 + \sum_{i=1}^{n} (y_i + q_i - (a^{'} + c_R^{'}) x_i)^2) \tag{3.15}$$

Esta función objetivo, que incorpora un nueva medida de incertidumbre y permite combinar dos visiones de la Regresión Difusa, junto a las restricciones posibilísticas (2.3) a (2.5), constituye un problema de programación cuadrática que será el modelo principal de regresión difusa posibilística de este trabajo de investigación.

Modelo 3.3. *Modelo de regresión difusa MCP, Modelo Cuadrático Posibilístico.*

Función objetivo:

$$J = k_1 \sum_{i=1}^{n} (y_i - a^{'} x_i)^2 + k_2 (\sum_{i=1}^{n} (y_i - p_i - (a^{'} - c_L^{'}) x_i)^2 + \sum_{i=1}^{n} (y_i + q_i - (a^{'} + c_R^{'}) x_i)^2) \tag{3.16}$$

Restricciones:

$$\sum_{j=1}^{m} a_j X_{ij} - (1-h) \sum_{j=1}^{m} c_{L_j} X_{ij} \leq y_i - (1-h) p_i \;\; para \;\; i = 1, ..., n \tag{3.17}$$

$$\sum_{j=1}^{m} a_j X_{ij} + (1-h) \sum_{j=1}^{m} c_{R_j} X_{ij} \geq y_i + (1-h) q_i \; para \; i = 1, ..., n \tag{3.18}$$

$$c_{L_j}, c_{R_j} \geq 0 \;\; para \;\; j = 1, ..., m \tag{3.19}$$

Los resultados que se obtienen con este modelo, para el ejemplo que estamos considerando, se muestran en la tabla 5.12.

La estimación de este modelo con parámetros $k_1 = 1$ y $k_2 = 1$ para la observación 21, se encuentra reflejada en el gráfico 3.4.

Teorema 3.1. *El modelo cuadrático propuesto es invariante a cambios de escala en los datos de entrada.*

Es decir, si A = $(A_1, ..., A_l, ..., A_k)$ es la solución del modelo planteado, en caso de que se reemplace X_l por $10^s X_l$ (s es un número entero), entonces $A^* = (A_1^*, ..., A_l^*, ...A_k^*)$ es la solución del problema con los nuevos datos con

Tabla 3.4: Estimaciones del modelo MCP (datos: acción de Copec)

k_1	k_2	$Coeficiente$	a_i	c_L	c_R
0	1	constante	-7968	0	0
		$Celulosa$	16.2	0	0
		$Cobre$	38.0	6.3	11.1
1	1	constante	-8000	0	0
		$Celulosa$	16.2	0	0
		$Cobre$	38.0	6.3	11.1
1	0	constante	-9456	0	0
		$Celulosa$	15.4	0	0
		$Cobre$	61.8	10.0	10.2

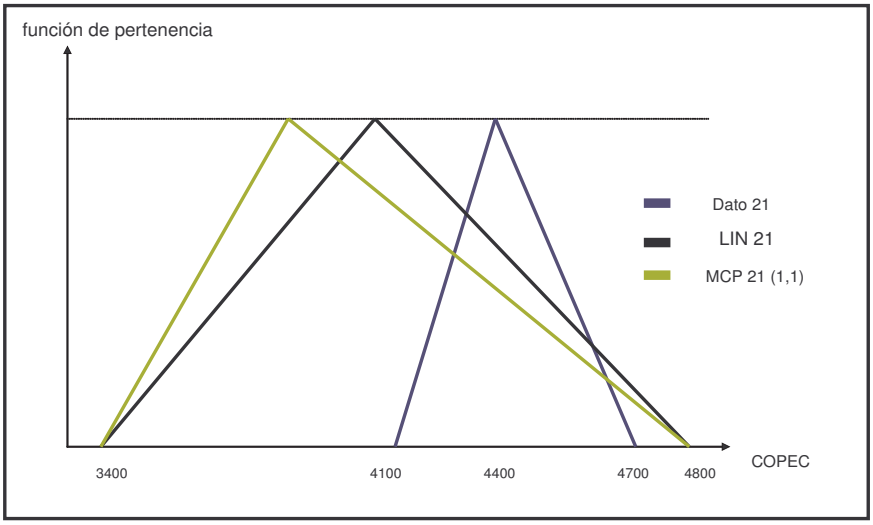

Figura 3.4: Función de pertenencia del dato 21 y su estimación por los modelos MCP y LIN

$$A_i = A_i^* \ \ para \ \ i \neq l \ y \ \ A_l^* = 10^{-s} A_l \tag{3.20}$$

donde

$A_l^* = 10^{-s}(a_l, c_{L_l}, c_{R_l}) = (10^{-s}a_l, 10^{-s}c_{L_l}, 10^{-s}c_{R_l})$

Para demostrar esta proposición se puede descomponer la función objetivo en sus tres términos y demostrar en cada término que se cumple la proposición; de igual manera las restricciones se analizan en dos grupos, las cotas posibilísticas superiores y las cotas posibilísticas inferiores, y se demostrará la proposición en cada grupo.

El término de la función objetivo

$$\sum_{i=1}^{n}((y_i + q_i) - (\sum_{j=1}^{k}(a_j + c_{R_j})x_{ij}))^2 \tag{3.21}$$

que cumple con el conjunto inicial de valores x, se compara con el término de la función objetivo de los nuevos datos x^*:

$$\sum_{i=1}^{n}((y_i + q_i) - (\sum_{j=1}^{k}(a_j^* + c_{R_j}^*)x_{ij}^*))^2 \tag{3.22}$$

que se puede desagregar como

$$\sum_{i=1}^{n}((y_i + q_i) - (\sum_{j\neq l}^{k}(a_j^* + c_{R_j}^*)x_{ij}^* + (a_l^* + c_{R_l}^*)x_{il}^*))^2 \tag{3.23}$$

donde se pueden reemplazar los valores x_{ij}^* por su equivalente en términos de x_{ij}, quedando la expresión anterior como

$$\sum_{i=1}^{n}((y_i + q_i) - (\sum_{j\neq l}^{k}(a_j^* + c_{R_j}^*)x_{ij} + (a_l^* + c_{R_l}^*)10^s X_{il}))^2 \tag{3.24}$$

igual a

$$\sum_{i=1}^{n}((y_i + q_i) - (\sum_{j\neq l}^{k}(a_j^* + c_{R_j}^*)X_{ij} + (10^s a_l^* + 10^s c_{R_l}^*)X_{il}))^2 \tag{3.25}$$

Si se reemplazan los supuestos para el término l: $a_l^* = 10^{-s}a_l$ y $c_{R_l}^* = 10^{-s}c_{R_l}^*$, se tiene

$$\sum_{i=1}^{n}((y_i + q_i) - (\sum_{j\neq l}^{k}(a_j^* + c_{R_j}^*)X_{ij} + (a_l + c_{R_l})X_{il}))^2 \tag{3.26}$$

que si se compara con la expresión (3.21), se observa que se cumple la igualdad de los dos términos, con $a_j = a_j^*$ y $c_{R_j} = c_{R_j}^*$ para $j \neq l$.

Con esto se demuestra que para el término considerado de la función objetivo, la solución no se ve afectada por cambios en la escala de medición.

De la misma forma se demuestra esta proposición para el término que representa la extensión izquierda de los datos en la función objetivo.

Centrémonos ahora en el tercer término de la función objetivo:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{m} a_j X_{ij})^2 \tag{3.27}$$

el cual puede descomponerse, separando el término l y desarrollando el cuadrado, en

$$\Big(\sum_{i=1}^{n} y_i^2 - 2\sum_{i=1}^{n} y_i(\sum_{j\neq l}^{m} a_j X_{ij} + a_l X_{il}) +$$

$$\sum_{i=1}^{n}\Big[(\sum_{j\neq l}^{m} a_j X_{ij})^2 + 2(\sum_{j\neq l}^{m} a_j X_{ij})(a_l X_{il}) + (a_l X_{il})^2)\Big]\Big) \tag{3.28}$$

El nuevo problema tiene solución A^*, y la expresión equivalente para esta segunda solución es:

$$\Big(\sum_{i=1}^{n} y_i^2 - 2\sum_{i=1}^{n} y_i(\sum_{j\neq l}^{m} a_j^* X_{ij}^* + a_l^* X_{il}^*) +$$

$$\sum_{i=1}^{n}\Big[(\sum_{j\neq l}^{m} a_j^* X_{ij}^*)^2 + 2(\sum_{j\neq l}^{m} a_j^* X_{ij}^*)(a_l^* X_{il}^*) + (a_l^* X_{il}^*)^2)\Big]\Big) \tag{3.29}$$

Si se reemplaza x* por los datos originales x en el término (3.29), se tiene

$$\Big(\sum_{i=1}^{n} y_i^2 - 2\sum_i y_i(\sum_{j\neq l} a_i^* X_{ij} + a_l^* 10^s X_{il}) +$$

$$\sum_{i=1}^{n}\Big[(\sum_{j\neq l}^{m} a_j^* X_{ij})^2 + 2(\sum_{j\neq l}^{m} a_j^* X_{ij})(a_l^* 10^s X_{il}) + (a_l^* 10^s X_{il})^2)\Big]\Big) \tag{3.30}$$

donde se pude sustituir a_l^* en función de a_l como $a_l^* = 10^{-s} a_l$:

$$\Big(\sum_{i=1}^{n} y_i^2 - 2\sum_i y_i(\sum_{j\neq l} a_i^* X_{ij} + +10^{-s} a_l 10^s X_{il})$$

$$+\sum_{i=1}^{n}\Big[(\sum_{j\neq l}^{m} a_j^* X_{ij})^2 + 2(\sum_{j\neq l}^{m} a_j^* X_{ij})(10^{-s} a_l 10^s X_{il}) + (10^{-s} a_l 10^s X_{il})^2)\Big]\Big) \tag{3.31}$$

lo que es equivalente a

$$\Big(\sum_{i=1}^{n} y_i^2 - 2\sum_i y_i(\sum_{j\neq l} a_i^* X_{ij} + a_l X_{il})$$

$$+\sum_{i=1}^{n}\Big[(\sum_{j\neq l}^{m} a_j^* X_{ij})^2 + 2(\sum_{j\neq l}^{m} a_j^* X_{ij})(a_l X_{il}) + (a_l X_{il})^2)\Big]\Big) \tag{3.32}$$

que, si se compara con el expresión (3.28), se cumple que $a_i = a_i^*$ para $i \neq l$, y el coeficiente l, $a_l^* = 10^{-s} a_l$, que es lo que se quería demostrar.

Con esto se ha demostrado para los tres términos de la función objetivo, la similitud de las soluciones con los datos x e x^*.

Con respecto a las restricciones, para el conjunto de ellas que representa la extensión derecha de los datos, se tiene

$$\sum_{j=1}^{m} a_j X_{ij} + R^{-1}(h) \sum_{j=1}^{m} c_{R_j} X_{ij} \geq y_i + R^{-1}(h) q_i \; para \; i = 1, ..., n \qquad (3.33)$$

que cumplen el primer conjunto de datos de entrada X_{ij}. Se puede considerar las mismas restricciones para el segundo conjunto de datos

$$\sum_{j=1}^{m} a_j^* X_{ij}^* + R^{-1}(h) \sum_{j=1}^{m} c_{R_j}^* X_{ij}^* \geq y_i + R^{-1}(h) q_i \; para \; i = 1, ..., n \qquad (3.34)$$

que se puede descomponer, separando el término correspondiente a X_l^* como

$$\sum_{j \neq l}^{k} a_j^* X_{ij}^* + a_l^* X_{il}^* + R^{-1}(h)(\sum_{j \neq l}^{k} c_{R_j}^* X_{ij}^* + c_{R_l}^* X_{il}^*) \geq y_i + R^{-1}(h) q_i \; para \; i = 1, ..., n \qquad (3.35)$$

donde reemplazando X_{ij}^* por su equivalente en términos de X_{ij}, queda

$$\sum_{j \neq l}^{k} a_j^* X_{ij} + a_l^* 10^s X_{il} +$$

$$R^{-1}(h)(\sum_{j \neq l}^{k} c_{R_j}^* X_{ij} + c_{R_l}^* 10^s X_{il}) \geq y_i + R^{-1}(h) q_i \; para \; i = 1, ..., n \qquad (3.36)$$

y reemplazando las equivalencias $a_j^* = 10^{-s} a_j$ y $c_{R_j}^* = 10^{-s} c_{R_j}^*$, queda la expresión

$$\sum_{j \neq l}^{k} a_j^* X_{ij} + a_l X_{il} + R^{-1}(h)(\sum_{j \neq l}^{k} c_{R_j}^* X_{ij} + c_{R_l} X_{il}) \geq y_i + R^{-1}(h) q_i \; para \; i = 1, ..., n \qquad (3.37)$$

que corresponde a la expresión (3.33) con $a_j = a_j^*$ y $c_{R_j} = c_{R_j}^*$ para $j \neq l$.

De la misma manera se puede proceder con el otro conjunto de restricciones correspondientes a la extensión izquierda de los datos.

Con lo expuesto anteriormente, la proposición queda demostrada.

Este nuevo modelo, dado con la función objetivo (3.15), que pretende ser más preciso en la estimación de las extensiones de los coeficientes, tiene la diferencia con el modelo anterior de que no requiere, necesariamente, operar en junto a las restricciones posibilísticas, puesto que ahora no se minimizan las extensiones de los datos estimados, sino la diferencia entre las extensiones de los datos de salida y los datos estimados.

Por esta circunstancia, se definirá un nuevo modelo dentro del enfoque cuadrático, que tendrá la particularidad de ser no posibilístico, y que estará formado por la función objetivo (3.15) y las restricciones que aseguren que los coeficientes no tienen extensiones negativas.

Modelo 3.4. *Modelo de regresión difusa MNP, modelo no posibilístico.*

Función objetivo:

$$J = k_1 \sum_{i=1}^{n} (y_i - a^{'} x_i)^2 + k_2 (\sum_{i=1}^{n} (y_i - p_i - (a^{'} - c_L^{'}) x_i)^2 + \sum_{i=1}^{n} (y_i + q_i - (a^{'} + c_R^{'}) x_i)^2) \tag{3.38}$$

Restricciones:

$$c_{L_j}, c_{R_j} \geq 0 \ \ para \ \ j = 1, ..., m \tag{3.39}$$

Para el ejemplo de la acción de Copec, las estimaciones con este nuevo modelo se muestran en la tabla 3.5.

Se constata, para este ejemplo, que las variaciones producidas en las estimaciones (por las diversos combinaciones de k_1 y k_2) para los valores a_j, c_{L_j} y c_{R_j} son menores, comparadas con los modelos probados anteriormente.

En la figura 3.5 se muestran las funciones de pertenencia de la estimación del dato 21 con este modelo y $k_1 = 1$ y $k_2 = 2$, junto al dato original y a la misma estimación con los modelos anteriormente mostrados.

Las magnitudes de los coeficientes de la tendencia central en este modelo no posibilístico son similares a las del modelo posibilístico. Sin embargo, las desviaciones

Tabla 3.5: Estimaciones del modelo MNP (datos: accion de Copec)

k_1	k_2	$Coeficiente$	a_i	c_L	c_R
0	1	constante	-9322	0	0
		$Celulosa$	15.6	0	0
		$Cobre$	59.2	1.54	1.42
1	1	constante	-9367	0	0
		$Celulosa$	15.5	0	0
		$Cobre$	60.1	1.54	1.42
1	0	constante	-9457	0	0
		$Celulosa$	15.4	0	0
		$Cobre$	61.9	1.54	1.42

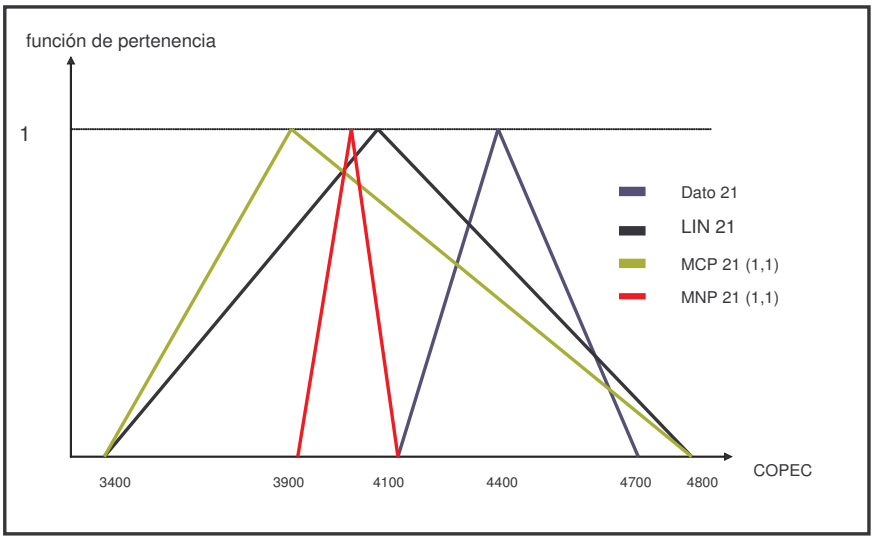

Figura 3.5: Funciones de pert. para dato 21 y modelos estimados

de las extensiones disminuyen en más de un 50 %, lo que era esperable al disminuir las restricciones.

También es posible incorporar otras restricciones distintas de las restricciones posibilísticas. Por ejemplo, si se incorpora el criterio del índice de posibilidad de Dubois y Prade, de que dos números difusos sean iguales, entonces el modelo estimado tiene como solución la misma tabla 3.5 de la estimación no posibilística. Es decir, las restricciones del índice de posibilidad de la igualdad del número observado y el número estimado no implican una restricción adicional al modelo, para este ejemplo.

Una consecuencia del enfoque cuadrático, al incorporar los términos $(a^{'} - c_{L}^{'})$ y $(a^{'} + c_{R}^{'})$ al cuadrado, es que produce restricciones que no son lineales para los valores a_i, c_{L_i} y c_{R_i}, ya que descomponiendo estas expresiones se tienen términos de la forma $(a_i * c_{L_i}$ y $(a_i * c_{R_i}$. Para poder mantener la estructura de un problema de programación cuadrática, se resolverá el problema de minimización (3.15) considerando cinco coeficientes a estimar por cada una de las k variables de entrada (a saber, a_i, c_{L_i}, c_{R_i}, $(a_i - c_{L_i})$ y $(a_i + c_{R_i})$), en lugar de los tres coeficientes con que se puede resolver el problema (3.8) (que son a_i, c_{L_i}, c_{R_i}).

3.2.2. Modelos con otras Restricciones Basadas en la Teoría de la Posibilidad

Las restricciones posibilísticas, que han sido el referente principal de la Regresión Difusa, pueden ser consideradas como restricciones muy amplias, produciendo un intervalo estimado muy grande. Por este motivo resulta también conveniente contar con alternativas menos restrictivas.

Un conjunto de restricciones alternativas puede darse con el índice de Posibilidad de que dos números difusos sean iguales:

$$Pos(A = B) = sup_{x \epsilon \Re} min\{\mu_A(x), \mu_B(x)\} \tag{3.40}$$

que, como se indicó, es equivalente a dos desigualdades (2.25 y 2.26), que se representan en la figura 3.6.

En esta figura, la parte (a) muestra que la estimación de un valor y_i no puede tener una función de pertenencia cuyo extremo izquierdo sea mayor al extremo derecho de y_i. Se permite un espacio de soluciones factibles, para la función de pertenencia, que

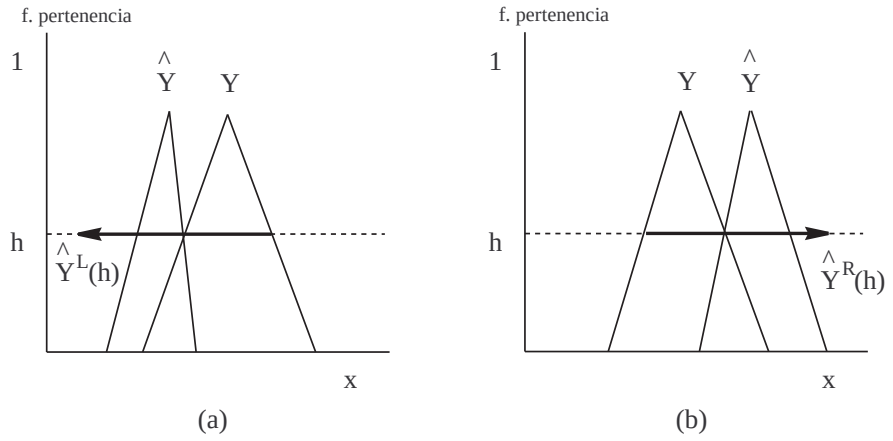

Figura 3.6: Espacio de soluciones factibles del índice de Posibilidad($\widehat{Y} = Y$)

se representa con la línea gruesa para un nivel de confianza h.

Para la extensión derecha, se debe cumplir la condición reflejada en la parte (b), es decir, que el extremo derecho de la estimación no debe ser menor que el extremo izquierdo de y_i. En otras palabras, el extremo izquierdo de la estimación puede fluctuar entre $(-\infty, y_i + q_i(h)]$ mientras el extremo derecho de la estimación puede fluctuar entre $[y_i - p_i(h), +\infty)$.

Teniendo esto en cuenta, podemos definir un nuevo modelo de regresión como sigue:

Modelo 3.5. *Modelo de regresión difusa POS_1, con medida de igualdad de la posibilidad.*

Función objetivo:

$$J = k_1 \sum_{i=1}^{n} (y_i - a^{'} x_i)^2 + k_2 (\sum_{i=1}^{n} (y_i - p_i - (a^{'} - c_L^{'}) x_i)^2 + \sum_{i=1}^{n} (y_i + q_i - (a^{'} + c_R^{'}) x_i)^2) \tag{3.41}$$

Restricciones:

$$\widehat{y_i} - (1-h)\widehat{p_i} \leq y_i + (1-h)q_i) \; para \; i = 1, ..., n \tag{3.42}$$

$$\widehat{y_i} + (1-h)\widehat{q_i} \geq y_i - (1-h)p_i \; para \; i = 1, ..., n \tag{3.43}$$

$$c_{L_j}, c_{R_j} \geq 0 \ \ para \ \ j = 1, ..., m \tag{3.44}$$

Para este Modelo basado en la medida de igualdad de la posibilidad, los resultados para el ejemplo de de la acción de COPEC se muestran en la tabla 3.6. Se puede apreciar que para la estimación de mínimos cuadrados ($k_1 = 1$ y $k_2 = 0$), ofrece la misma estimación central que el modelo no posibilístico, aunque las extensiones son considerablemente mayores.

Tabla 3.6: Estimaciones del modelo POS_1 para el ejemplo de Copec)

k_1	k_2	$Coeficiente$	a_i	c_L	c_R
0	1	constante	-8230	0	0
		$Celulosa$	13.4	0	0
		$Cobre$	59.5	4.1	5.74
1	1	constante	-8461	0	0
		$Celulosa$	14.1	0	0
		$Cobre$	57.8	4.1	5.77
1	0	constante	-9457	0	0
		$Celulosa$	15.4	0	0
		$Cobre$	61.9	5.36	5,64

Otro índice que describimos al repasar la Teoría de la Posibilidad es el índice de la necesidad que un número difuso contenga a otro número difuso, $Nes(\widehat{Y_i} \supset Y_i)$.

Este índice queda reflejado en las restricciones

$$\widehat{y_i - p_i}(h) \leq (y_i - p_i)(1 - h) \tag{3.45}$$

$$\widehat{y_i + q_i}(h) \geq (y_i + q_i)(1 - h) \tag{3.46}$$

Nótese que $(y_i + q_i)(1 - h)$ es igual a $Y_i + |L^{-1}(1 - h)|q_i$, lo que resulta, para la funciones triangulares, $Y_i + hq_i$.

Con esta idea, podemos construir los siguientes nuevos modelos de Regresión.

Modelo 3.6. *Modelo de regresión difusa NES_1, con índice de necesidad de contener el valor observado.*

Función objetivo:

$$J = k_1 \sum_{i=1}^{n} (y_i - a^{'} x_i)^2 + k_2 (\sum_{i=1}^{n} (y_i - p_i - (a^{'} - c_L^{'}) x_i)^2 + \sum_{i=1}^{n} (y_i + q_i - (a^{'} + c_R^{'}) x_i)^2) \tag{3.47}$$

Restricciones:

$$\widehat{y_i} - (1-h)\widehat{p_i} \leq y_i - (h)p_i \tag{3.48}$$

$$\widehat{y_i} + (1-h)\widehat{q_i} \geq y_i + (h)q_i \tag{3.49}$$

$$c_{L_j}, c_{R_j} \geq 0 \ \ para \ \ j = 1, ..., m \tag{3.50}$$

Las restricciones de este índice quedan reflejadas en la figura 3.7, para un α-corte h. Se muestra, con la línea gruesa, el espacio de alternativas factibles de $\widehat{y_i - p_i}(h)$ y de $\widehat{y_i + q_i}(h)$, que es más amplio que el espacio de las alternativas posibilísticas. Dubois y Prade lo interpretan como el grado de certeza de que, dado un valor de Y_i, ese mismo valor lo pueda toma $\widehat{Y_i}$.

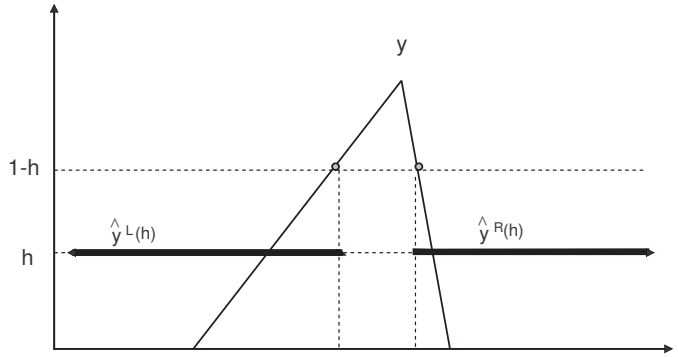

Figura 3.7: Espacio de soluciones de $\widehat{Y_i}$ con Nes($\widehat{Y_i} \supset Y_i$)

El modelo, con este índice de necesidad en las restricciones, tiene las soluciones, para el ejemplo de la acción de COPEC, dadas por la tabla 3.7.

Modelo 3.7. *Modelo de regresión difusa NES_2, con índice de necesidad de contener el valor estimado.*

Tabla 3.7: Estimaciones del modelo NES_1 para el ejemplo de Copec

k_1	k_2	$Coeficiente$	a_i	c_L	c_R
0	1	constante	-9242	0	0
		$Celulosa$	17.4	0	0
		$Cobre$	46.1	4.9	7.6
1	1	constante	-9242	0	0
		$Celulosa$	17.4	0	0
		$Cobre$	46.1	4.9	7.6
1	0	constante	-9415	0	0
		$Celulosa$	15.4	0	0
		$Cobre$	61.9	6.7	7.2

Función objetivo:

$$J = k_1 \sum_{i=1}^{n} (y_i - a^{'} x_i)^2 + k_2 (\sum_{i=1}^{n} (y_i - p_i - (a^{'} - c_L^{'}) x_i)^2 + \sum_{i=1}^{n} (y_i + q_i - (a^{'} + c_R^{'}) x_i)^2) \tag{3.51}$$

Restricciones:

$$\widehat{y_i} + (h)\widehat{q_i} \leq y_i - (1-h)p_i) \; para \; i = 1, ..., n \tag{3.52}$$

$$\widehat{y_i} - (h)\widehat{p_i} \geq y_i - (1-h)p_i \; para \; i = 1, ..., n \tag{3.53}$$

$$c_{L_j}, c_{R_j} \geq 0 \;\; para \;\; j = 1, ..., m \tag{3.54}$$

El índice de necesidad de que $\widehat{Y_i} \subset Y_i$, que está representado en las restricciones (2.28), puede ser visto a través del espacio de soluciones factibles para $\widehat{Y_i}(1-h)$, que se muestra en la figura 3.8 a través de la línea gruesa.

Estas restricciones pueden interpretarse, como el grado de certeza de que, dado un valor de $\widehat{Y_i}$, ese mismo valor lo pueda tomar Y_i.

Para el ejemplo de la acción de COPEC, que hemos estimado con otros modelos, resulta que no existen soluciones factibles para el modelo NES_2.

Con los dos últimos índices, se puede definir el índice de necesidad que $\widehat{Y_i}$ sea igual a Y_i [DP88], el cual se define como

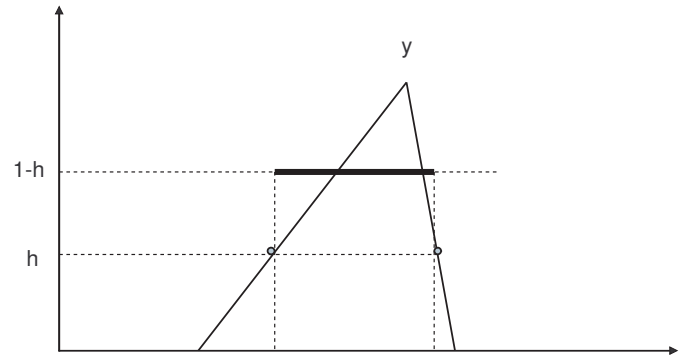

Figura 3.8: Espacio de soluciones de $\widehat{Y}_i(1-h)$ con Nes($\widehat{Y}_i \subset Y_i$)

$$Nes(\widehat{Y_i = Y_i}) = min(Nes(\widehat{Y}_i \supset Y_i), Nes(\widehat{Y}_i \subset Y_i)) \tag{3.55}$$

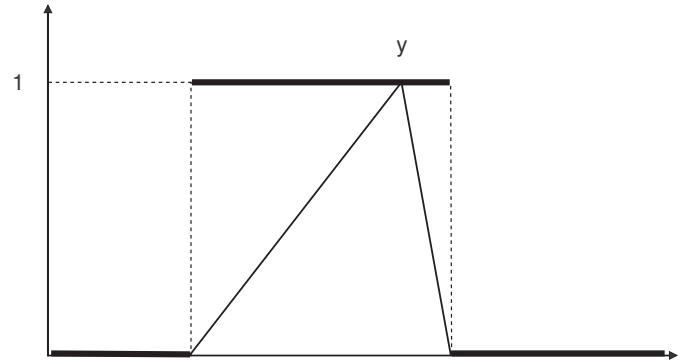

Figura 3.9: Espacio de soluciones de $\widehat{Y}_i$ con Nes($\widehat{Y}_i = Y_i$)para h=0

La figura 3.9 muestra, con la línea gruesa, el espacio de soluciones factibles para $\widehat{Y}_i$, considerando el índice de igualdad de la necesidad, para h=0.

Este último modelo se formula como sigue:

Modelo 3.8. *Modelo de regresión difusa NES_3, con índice de necesidad de la igualdad.*

Función objetivo:

$$J = k_1 \sum_{i=1}^{n} (y_i - a^{'} x_i)^2 + k_2 (\sum_{i=1}^{n} (y_i - p_i - (a^{'} - c_L^{'}) x_i)^2 + \sum_{i=1}^{n} (y_i + q_i - (a^{'} + c_R^{'}) x_i)^2) \tag{3.56}$$

Restricciones:

$$\widehat{y_i} + (h)\widehat{q_i} \leq y_i - (1-h)p_i) \; para \; i = 1, ..., n \tag{3.57}$$

$$\widehat{y_i} - (h)\widehat{p_i} \geq y_i - (1-h)p_i \; para \; i = 1, ..., n \tag{3.58}$$

$$\widehat{y_i} - (1-h)\widehat{p_i} \leq y_i - (h)p_i \tag{3.59}$$

$$\widehat{y_i} + (1-h)\widehat{q_i} \geq y_i + (h)q_i \tag{3.60}$$

$$c_{L_j}, c_{R_j} \geq 0 \;\; para \;\; j = 1, ..., m \tag{3.61}$$

Para los datos de nuestro ejemplo, como era de esperar, no tiene solución factible, ya que el modelo anterior, no la tenía.

3.2.3. Panorama de los Modelos Presentados

Para terminar esta sección, vamos a mostrar un resumen representando las restricciones, la influencia de las extensiones en los coeficientes estimados y las funciones objetivos, para cada uno de los modelos introducidos. Este resumen puede verse en el esquema de la figura 3.10.

Las restricciones tienen un papel central en el efecto de los modelos de regresión difusa:

- En la parte superior de la tabla, están las restricciones *más difusas*, lo que significa que las estimaciones tienen extensiones más amplias. En nuestra terminología, esto significa que los valores estimados presentan más incertidumbre. Por lo tanto, las estimaciones son menos arriesgadas, dado que la incertidumbre del sistema ya esta incorporada en la definición de las restricciones.

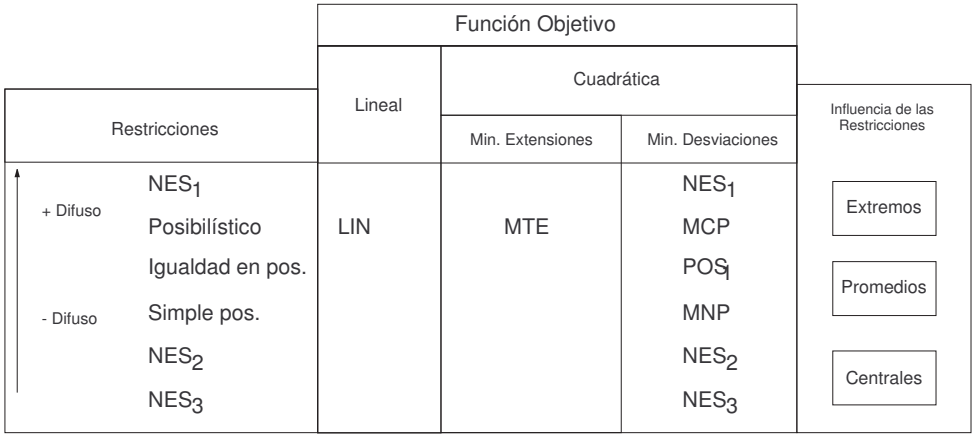

Figura 3.10: Resumen de métodos de regresión difusa lineal

- Por otra parte, las estimaciones *menos difusas*, con extensiones más pequeñas, son de mayor riesgo, en el sentido de la posible aparición de valores excepcionales. En cualquier caso son más precisas, puesto que no están influidas por los puntos extremos.

3.3. Experimentación con los Métodos de Regresión Difusa

En la sección anterior hemos presentado un conjunto de modelos que constituyen nuestra propuesta para el análisis de Regresión Difusa. Para completar nuestro estudio, procedemos ahora a analizar empíricamente la bondad de las estimaciones conseguidas con estos modelos. Para ello, vamos a utilizar los indicadores de bondad del ajuste que se describieron en el capítulo anterior.

Calcularemos, por tanto, el valor de estos indicadores para todos los modelos, con diversos conjuntos de datos, y con diversas metodologías con el objetivo de discriminar entre los diversos métodos de acuerdo con los valores producidos por los indicadores, tratando de concluir qué métodos resultan más adecuados para utilizar.

Hay que tener presente, que los factores que pueden afectar una estimación de regresión difusa son diversos. Por ejemplo:

- El tamaño de las extensiones respecto al valor central.
- Simetría o no simetría de las funciones de pertenencia.

- Cantidad de observaciones del archivo de aprendizaje.
- Presencia de multicolinealidad en las variables de entrada.
- Problemas de especificación en la función de regresión.
- Regularidad o no de las funciones L_i y R_i de las funciones de pertenencia de cada dato respecto a los otros datos.
- Presencia de datos extremos que, como se dijo anteriormente, afecta considerablemente a la estimación posibilística.
- Cantidad de variables de entrada.
- Nivel de correlación entre las variables de entrada y la variable de salida.
- Efecto producido por la posible omisión de variables en el modelo de regresión.

Como se ve, no es posible hacer una sola experimentación que pueda reunir tal cantidad de factores. Pensamos, sobretodo, en el efecto conjunto que se puede producir por dos o más de los factores enunciados. Por esta razón, se presentarán diversas situaciones que nos han parecido lo más representativas para el caso de Regresión Difusa, para poder evaluar los ocho métodos considerados en este trabajo de investigación.

3.3.1. Análisis del Ejemplo del Precio de COPEC

A lo largo de la presentación realizada en este capítulo, para un ejemplo de dos variables de entrada para el precio de la acción de la empresa COPEC, se han ido mostrando los resultados que se obtienen de cada uno de los métodos de regresión propuestos. Estos valores están en las tablas 3.2 a 3.7 presentadas en la sección anterior.

Procedemos ahora a calcular la magnitud de los indicadores para diversos valores de los parámetros h, k_1 y k_2 para los métodos: LIN, MTE, MCP, MNP, POS_1, NES_1, NES_2 y NES_3. Recuérdese que para el ejemplo de COPEC, los métodos con resultados son los seis primeros.

En la Tabla 3.8 se muestra el cálculo de los 6 indicadores de bondad de ajuste propuestos, para los dos métodos basados en los modelos de Tanaka, a saber, LIN y

Tabla 3.8: Indicadores de bondad de ajuste, LIN y MTE (datos acción de Copec)

Mét.	h	k_1	k_2	R^2_{difuso}	RH^2	SIM_1	SIM_2	SIM_3	SIM_4	SIM_5
LIN	0			0.7725	0.5416	0.0968	0.3611	0.6295	0.3113	0.5407
MTE	0	0	1	0.7659	0.5770	0.0966	0.3653	0.6271	0.2997	0.5176
	0	1	1	0.7993	0.5790	0.1008	0.3652	0.6345	0.3094	0.5452
	0	1	0	0.8469	0.8206	0.1031	0.3502	0.6565	0.3967	0.7093
	0.5	0	1	0.7926	0.6204	0.0780	0.3155	0.6477	0.3967	0.7319
	0.5	1	1	0.8053	0.6303	0.0784	0.3159	0.6493	0.3980	0.7390
	0.5	1	0	0.8448	0.7669	0.0738	0.3098	0.6595	0.4223	0.8095

Tabla 3.9: Indicadores de bondad de ajuste, MCP y MNP (datos acción de Copec)

Mét.	h	k_1	k_2	R^2_{difuso}	RH^2	SIM_1	SIM_2	SIM_3	SIM_4	SIM_5
MCP	0	0	1	0.8145	0.5532	0.1041	0.3651	0.6422	0.3103	0.5760
	0	1	1	0.8219	0.5812	0.1058	0.3642	0.6450	0.3163	0.5949
	0	1	0	0.8471	0.8469	0.1022	0.3488	0.6575	0.4024	0.7179
MCP	0.5	0	1	0.7323	0.7263	0.0779	0.3154	0.6416	0.4165	0.7290
	0.5	1	1	0.7823	0.6627	0.0781	0.3154	0.6462	0.4060	0.7320
	0.5	1	0	0.8471	0.8462	0.0706	0.3069	0.6614	0.4416	0.8291
MNP	0	0	1	0.8468	0.8232	0.1048	0.4655	0.5180	0.4247	0.2096
	0	1	1	0.8470	0.83112	0.1152	0.4672	0.5201	0.4268	0.2144
	0	1	0	0.8471	0.8471	0.1228	0.4701	0.5242	0.4313	0.2240
MNP	0.5	0	1	0.8469	0.8232	0.1048	0.4655	0.5180	0.4247	0.2097
	0.5	1	1	0.8470	0.8311	0.1152	0.4672	0.5201	0.4268	0.2144
	0.5	1	0	0.8471	0.8471	0.1228	0.4701	0.5242	0.4313	0.2239

MTE, y para dos valores de h [1]. También está incorporado el indicador R^2 híbrido de Chang (RH^2).

Si comparamos los resultados entre h=0 y h=0.5, se aprecia un comportamiento dispar entre los indicadores: mejoran los índices SIM_4 y SIM_5, y bajan su valor los índice SIM_1 y SIM_2. En cambio para SIM_3 las diferencias son menores. Esto significa que el paso de h=0 a h=0.5, aumentando las extensiones de los estimadores, tiene efectos diversos sobre el concepto de *ajuste difuso*.

Centrémonos ahora en los métodos MCP y MNP. La Tabla 3.9 muestra los valores de los indicadores de bondad del ajuste para estos métodos.

Como podemos apreciar, los indicadores de bondad de ajuste de los métodos cuadráticos MCP y MNP (3.9) presentan un comportamiento similar al caso anterior, a pesar del curioso comportamiento del indicador SIM_5. Este indicador baja considerablemente su valor con el método no posibilístico, lo que significa que cuando el soporte de $\widehat{Y_i}$ disminuya, el valor máximo de intersección de SIM_5 generalmente

[1] LIN no pudo calcularse para $h = 0{,}5$ en MATLAB, por una gran diferencia de tamaño entre números en el cálculo

Tabla 3.10: Indicadores de bondad de ajuste, POS_1 y NES_1 (datos acción de Copec)

Mét.	h	k_1	k_2	R^2_{difuso}	RH^2	SIM_1	SIM_2	SIM_3	SIM_4	SIM_5
POS_1	0	0	1	0.8395	0.7054	0.1114	0.4010	0.6311	0.3283	0.4840
	0	1	1	0.8421	0.7236	0.1123	0.4005	0.6333	0.3324	0.4962
	0	1	0	0.8471	0.8470	0.1175	0.3904	0.6429	0.3764	0.5760
POS_1	0.5	0	1	0.7869	0.7347	0.0910	0.3315	0.6431	0.4038	0.6952
	0.5	1	1	0.8117	0.7003	0.0915	0.3316	0.6470	0.3946	0.7013
	0.5	1	0	0.8471	0.8465	0.0855	0.3226	0.6599	0.4249	0.7858
NES_1	0	0	1	0.8346	0.6642	0.1099	0.3878	0.6366	0.3200	0.5251
	0	1	1	0.8384	0.6857	0.1112	0.3873	0.6389	0.3244	0.5381
	0	1	0	0.8471	0.8469	0.1238	0.3754	0.6491	0.3829	0.6306
NES_1	0.5	0	1	0.7323	0.7263	0.0779	0.3154	0.6416	0.4165	0.7291
	0.5	1	1	0.7823	0.6628	0.0781	0.3154	0.6462	0.4059	0.7324
	0.5	1	0	0.8471	0.8462	0.0706	0.3069	0.6613	0.4416	0.8291

disminuirá, bajando el valor de este indicador. Esta es una de las debilidades de SIM_5 al tomar como referencia sólo un punto de la función de pertenencia.

Para todos los modelos que estamos evaluando, el indicador de divergencia de Kim Bishu, sale superior a 100, con valores superiores a 400 en algunos casos. Esto nos hace pensar que este no es un indicador apropiado para medir la bondad de ajuste de una estimación de Regresión Difusa. Hay una relación directa entre valores de SIM_1 pequeños, como resulta para estos ejemplos, en que se aproxima al valor 0.1, con valores del indicador KB muy alto, lo que resalta las cualidades de SIM_1.

Veamos, por último, los dos métodos que nos restan.

Para lo métodos POS_1 y NES_1 basados en índices de posibilidad y necesidad, se han obtenidos los valores que se muestran en la tabal 3.10 (el método NES_2 resultó con un espacio de soluciones factibles vacío). Se puede observar un comportamiento que no difiere con los de los métodos anteriores. En todos los casos, cuando $k_1 = 1$ y $k_2 = 0$ (es decir, cuando se realiza la optimización de mínimos cuadrados) el valor de R^2_{difuso} es igual, y es el más alto valor de R^2_{difuso} encontrado.

Se destaca que el indicador SIM_5 por lo general, experimenta una significativa mejoría con una leve mejora en R^2_{difuso}, salvo en el método MNP, lo que es un indicio de la fuerte relación que puede existir entre estos dos indicadores. Mientras R^2_{difuso} no considera las características de las funciones de pertenencia de los números comparados (puesto que mide una relación exclusivamente con los valores centrales, o de mayor pertenencia, de cada número difuso), SIM_5 es el supremo de la intersección de ambas funciones de pertenencia, que, aunque no necesariamente, tiene una fuerte relación con la calidad del ajuste de la tendencia central, que mide R^2_{difuso}.

Se puede apreciar en estos ejemplos, que R^2_{difuso} tiene un comportamiento más estable (varía de 0.73 a 0.84) que el R^2 híbrido (varía entre 0.55 y 0.84), lo que muestra otra ventaja de tomarlo como referencia para el ajuste del valor central.

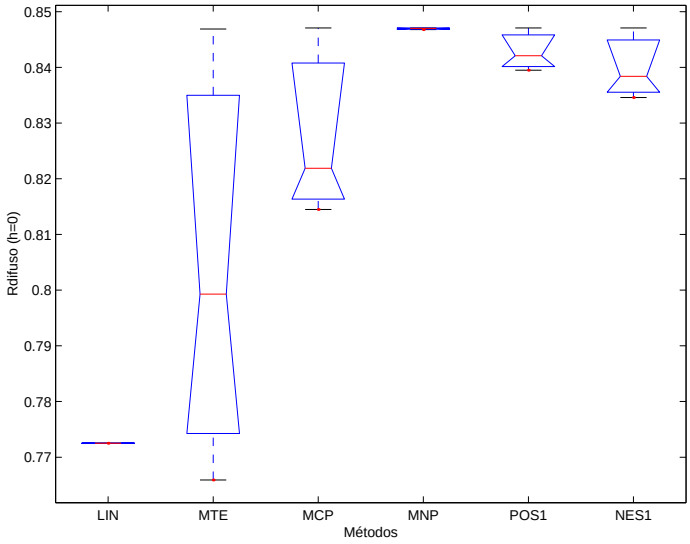

Figura 3.11: Comparación de medias, R^2_{difuso} (Ejemplo COPEC)

La figura 3.11 muestra las medias y sus dispersiones del valor R^2_{difuso} para el caso en que el parámetro h=0, para cada uno de los seis métodos.

Salvo para el método Lineal, se destaca que para la combinación de parámetros $k_1 = 1$ y $k_2 = 0$, que corresponde a la optimización de mínimos cuadrados, todos los métodos tienden a igualarse al valor superior de la gráfica (algo menos de 0.85). Ya en esta gráfica, se aprecia que los R^2_{difuso} para LIN y MTE son inferiores al resto de los métodos.

En la figura 3.12, ofrecemos la misma gráfica para el indicador de similitud SIM. Podemos encontrar una diferenciación mucho más clara, donde el método MNP obtiene los valores más altos, los métodos POS_1 y NES_1 un promedio intermedio, y LIN, MTE y MCP un promedio menor.

Esta misma clasificación en tres categorías de los métodos coincide con el tamaño de las extensiones estimadas para los coeficientes. Si se promedian las extensiones estimadas de las tablas 3.2 a 3.7, distintas de cero, se aprecia que para MNP el tamaño de la extensión es 1.5, después viene POS_1 con una extensión de 5, NES_1 con una extensión de 6, y LIN, MTE y MCP con una extensión promedio del orden de 9.

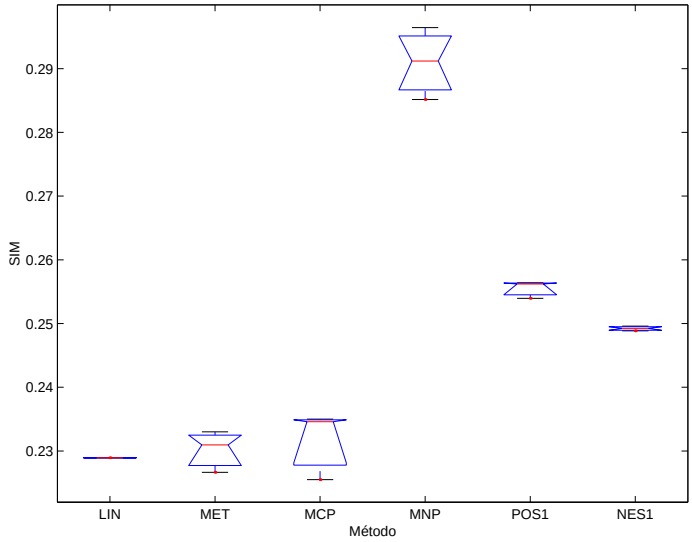

Figura 3.12: Comparación de medias, SIM (Ejemplo COPEC)

Con respecto al indicador SIM_3, la gráfica 3.13 es una muestra de la variabilidad de los indicadores: en este caso, MNP tiene claramente un valor menor, mientras todos los otros métodos tienen un valor prácticamente similar.

Este resultado que cambia totalmente en el gráfico 3.14 que representa los valores obtenidos para el indicador SIM_4. Como se puede apreciar, en este caso el método MNP no está tan distanciado del resto, por lo que el análisis de varianza no concluye que haya una diferenciación significativa entre los métodos.

Curiosamente, vuelven a invertirse con los valores de SIM_5; se produce una diferenciación significativa entre el método MNP no posibilístico y el resto de los métodos calculados, cuyos resultados pueden visualizarse en el gráfico 3.15. Esta último resultado es totalmente lógico, dada la estructura del indicador. Como MNP es el método que estima soportes más pequeños, el supremo de la intersección de las dos funciones de pertenencia generalmente será menor para MNP que para el resto de los métodos.

En cualquier caso, estos resultados no son contradictorios con las propuestas teóricas, dado que el método MNP es el mejor valorado a tenor de los valores de SIM, que es el más representativo de los indicadores de incertidumbre, a los valores de R^1_{difuso}, que es el más representativo como indicador del ajuste central, y a las particularidades de los datos: el promedio de las variables de salida es 8 y el promedio

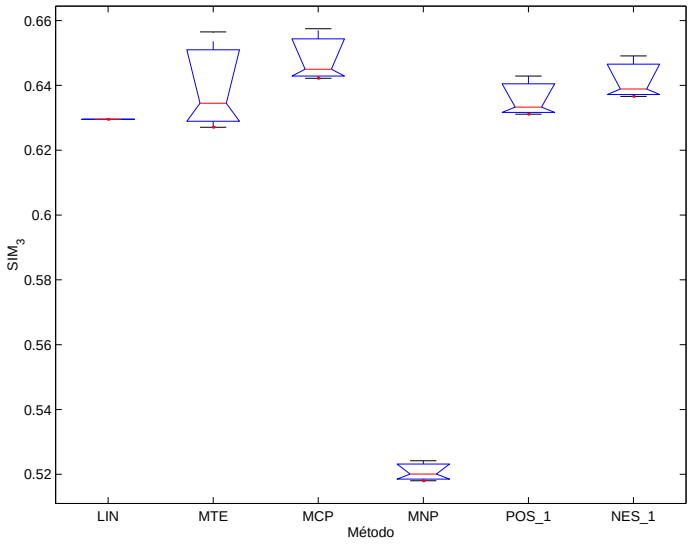

Figura 3.13: Comparación de medias, SIM_3 (Ejemplo COPEC)

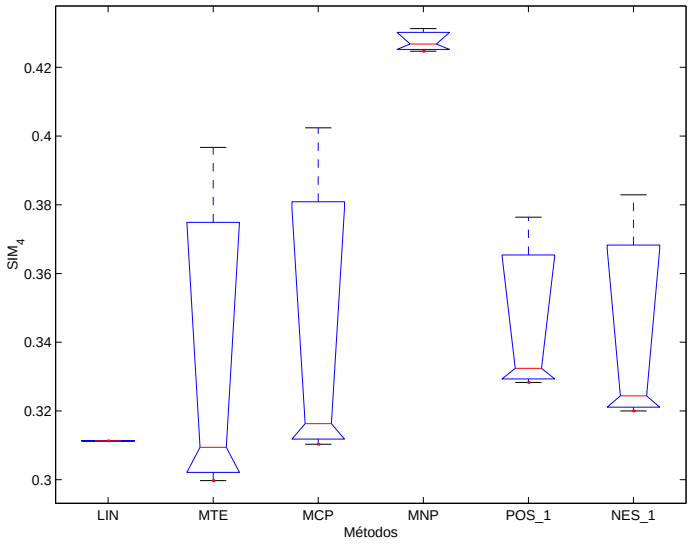

Figura 3.14: Comparación de medias, SIM_4 (Ejemplo COPEC)

de las extensiones es 0.02, por lo que las extensiones son muy pequeñas y, al estimar el método MNP extensiones también pequeñas, produce soportes que pueden diferir de los soportes de entrada.

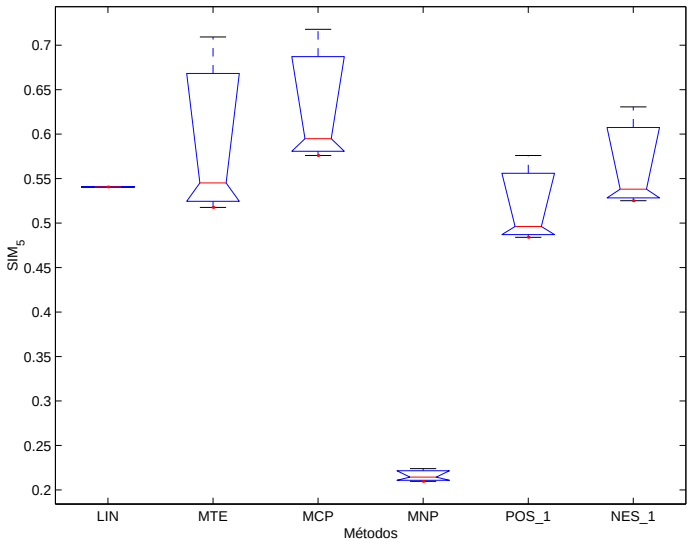

Figura 3.15: Comparación de medias, SIM_5 (Ejemplo COPEC)

3.3.2. Análisis de Ejemplos con Datos de Prueba

La literatura de la Regresión Difusa sólo registra una aplicación donde se hayan utilizado datos de prueba para comprobar la calidad del ajuste de los datos estimados [TSWA87]. Nosotros vamos a continuar nuestro análisis empírico de los modelos utilizando también esta alternativa.

Se procederá a hacer dos mediciones de los indicadores: primero con los datos de aprendizaje, y luego sólo con los datos de prueba. Se considerarán para cada método los parámetros $k_1 = 1$ y $k_2 = 1$ con h=0.

Utilizaremos dos conjuntos de datos:

- Uno con 3 variables de entrada y sin presencia de multicolinealidad.
- Otro con 2 variables de entrada y con presencia de multicolinealidad en las variables de entrada.

3.3.2.1. Datos sin Multicolinealidad en las Variables de Entrada

Se considerará un archivo de datos de aprendizaje (3.11) con cuatro variables de entrada, incluyendo la constante, y funciones de pertenencia simétricas en la variable de salida. El promedio de la Y es 45 y el promedio de las extensiones es 4.2. La matriz de correlación lineal de los valores centrales, entre todas las variables, se muestra en la tabla 3.12.

Tabla 3.11: Datos de aprendizaje (ejemplo de 4 variables)

y_i	p_i	x_1	x_2	x_3
9.1805	0.7378	1	1	5
14.8832	2.4530	1	4	1
23.2739	1.5254	3	2	3
51.6023	10.7212	4	16	4
31.7517	2.5626	5	3	5
31.8340	3.6719	6	5	2
39.1279	4.0739	7	4	1
44.5992	4.5205	8	6	2
40.8313	3.1918	9	1	5
62.9655	4.7751	10	7	10
71.0323	6.0332	11	8	2
51.0070	5.2949	12	3	4
58.7960	4.5384	13	4	8
72.5302	3.9585	14	6	1
68.5829	4.6948	15	2	1

Tabla 3.12: Correlación de los datos de aprendizaje (ejemplo de 4 variables)

Variable	y	x_1	x_2	x_3
y	1.00	0.89	0.38	0.08
x_1	0.89	1.00	-0.032	0.06
x_2	0.38	-0.034	1.00	0.03
x_3	0.08	0.06	0.03	1.00

Para los datos de prueba de la Tabla 3.13, la media de Y es 58, la media de las extensiones 5.4 y la matriz de correlaciones se muestra en la Tabla 3.14.

Tal vez la principal diferencia de los datos de prueba respecto a los datos de aprendizaje, consista en que las correlaciones entre los datos centrales aumentaron respecto a los datos de aprendizaje, especialmente la correlación entre la salida y x_1, que es de 0.96. Esto plantea la hipótesis de que el ajuste con los datos de prueba debería ser mejor.

Los resultados obtenidos confirman esa hipótesis, especialmente para R^2_{difuso}. Vamos a mostrar las gráficas que describen los datos de nuestros indicadores para los

Tabla 3.13: Datos de prueba (ejemplo de 4 variables)

y_i	p_i	x_1	x_2	x_3
12.9012	0.6176	1	1	5
14.7087	2.5644	1	4	1
40.9805	8.0305	3	12	3
35.1852	4.0764	4	6	4
66.9324	8.9306	5	13	15
41.4250	4.2752	6	5	2
53.8336	10.4695	7	14	1
52.2490	4.0796	8	6	2
64.8209	8.2330	9	11	5
33.9784	4.4945	1	7	10
97.7453	2.3391	21	1	2
104.5764	5.4428	22	3	14
115.2609	7.0588	23	4	8
130.3135	10.1672	24	12	1
8.0645	0.7515	1	1	1

Tabla 3.14: Correlación de los datos de aprendizaje (ejemplo de 4 variables)

Variable	y	x_1	x_2	x_3
y	1.00	0.96	0.38	0.26
x_1	0.96	1.00	-0.05	0.14
x_2	0.20	-0.05	1.00	0.07
x_3	0.26	0.14	0.07	1.00

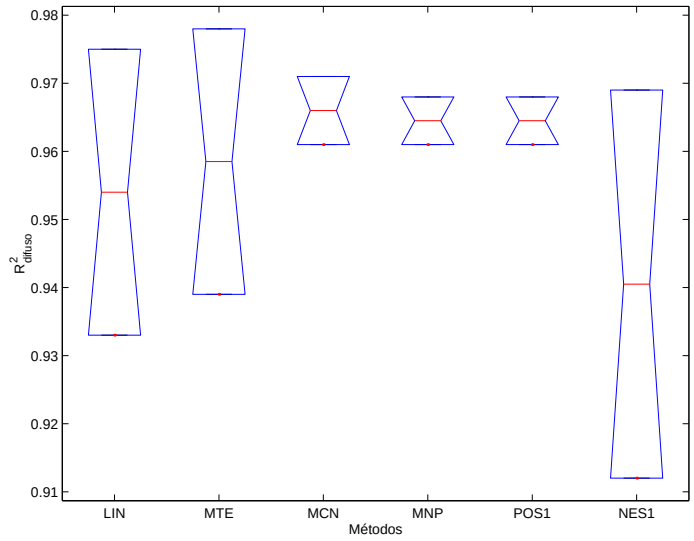

Figura 3.16: Comparación de medias, R^2_{difuso} (Ejemplo sin multicolinealidad)

seis métodos, incluyendo las dos observaciones para cada método: el valor con los datos de aprendizaje y el valor con los datos de prueba.

Hechas las estimaciones con los ocho métodos, volvió a resultar que los métodos NES_2 y NES_3 no tienen soluciones factibles. Para el indicador de tendencia central 3.16 se aprecia que los métodos MCE, MNP y POS_1 tienen valores muy parecidos además de ser los que muestran un mejor promedio.

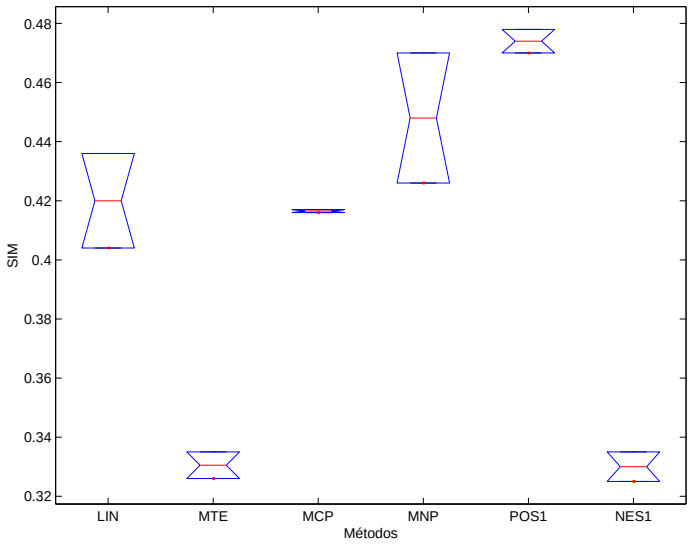

Figura 3.17: Comparación de medias, SIM (Ejemplo sin multicolinealidad)

Para el indicador integrado SIM (véase figura 3.17, vuelven a tener un mejor rendimiento los métodos MCE, MNP y POS_1, en este caso, junto al método LIN.

Para SIM_3, representado en la gráfica 3.18, el peor valorado es el método MNP. Sin embargo, la diferencia con los otros métodos es mucho menor que la que se representaba en el primer ejemplo (veáse la figura 3.13).

Lo mismo ocurre con SIM_5: en el primer ejemplo (figura 3.15) mostraba a MNP muy por debajo del resto de los indicadores; ahora (veáse la figura 3.20), siendo MNP y POS_1 los métodos menos valorados, sus magnitudes no son tan distantes del resto de los métodos.

Finalmente, para el indicador SIM_1, que en el primer ejemplo (figura 3.14) mostraba solamente a MNP como el método con mejor valoración, en este nuevo ejemplo, cuyos datos se muestran en la figura 3.19, son MNP y POS_1 los métodos que se distancian del resto.

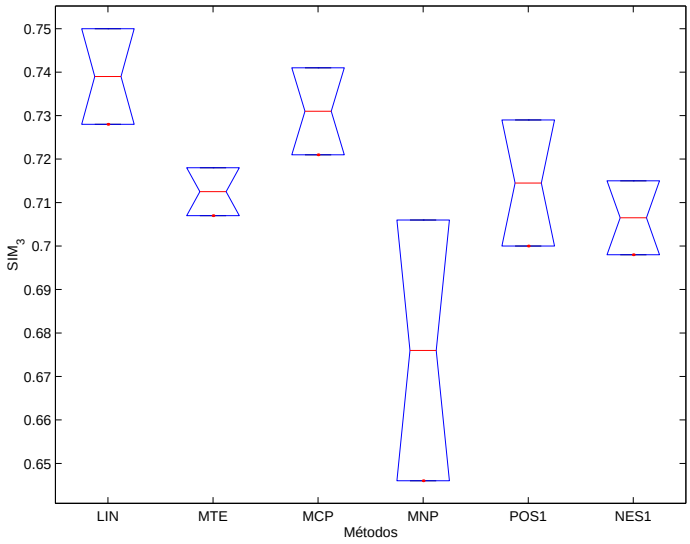

Figura 3.18: Comparación de medias, SIM_3 (Ejemplo sin multicolinealidad)

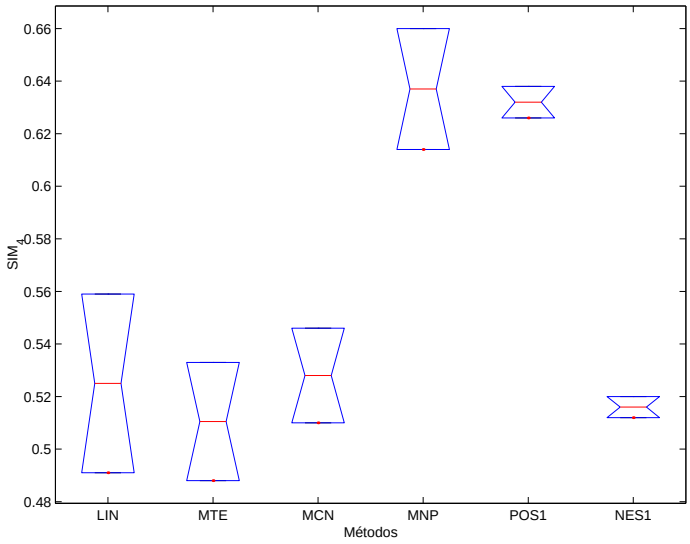

Figura 3.19: Comparación de medias, SIM_4 (Ejemplo sin multicolinealidad)

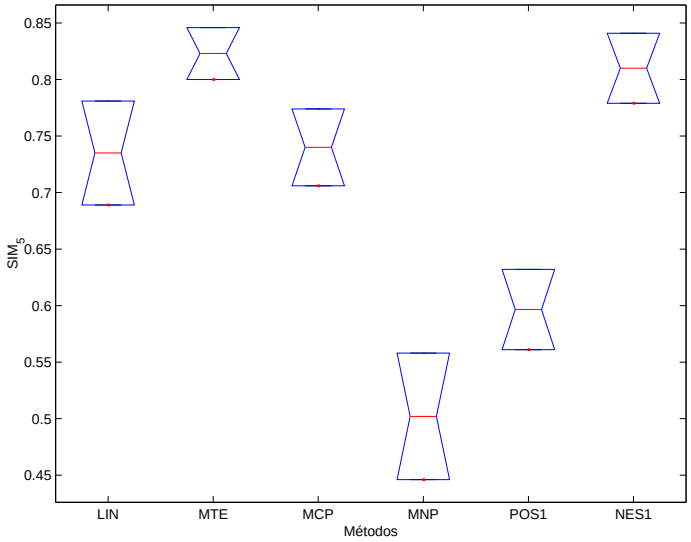

Figura 3.20: Comparación de medias, SIM_5 (Ejemplo sin multicolinealidad)

3.3.2.2. Datos con Multicolinealidad en las Variables de Entrada

Una situación que se va a analizar en detalle en el capítulo de Regresión Ridge que afrontaremos más adelante en esta memoria, es la presencia de datos multicolineales en una Regresión Difusa. En presencia de datos afectados por multicolinealidad, de momento, no se sabe lo que ocurre en la precisión de las estimaciones y en la cuantificación de los indicadores que la miden.

Una hipótesis que se puede aventurar, es que en caso de multicolinealidad en los datos de entrada, las estimaciones aplicadas a los datos de prueba debieran ser peores que las estimaciones con los datos originales.

Para nuestra experimentación, se ha considerado el conjunto de datos de aprendizaje que se muestran en la Tabla 3.15. Este conjunto de datos tiene un promedio de 29 para la variable de salida y un promedio de las extensiones igual a 4. Los coeficientes de correlación lineal se muestran en la Tabla 3.16, donde se aprecia que la correlación entre las dos variables de entrada es 0.88.

Los datos de prueba se muestran en la Tabla 3.17. Tienen un promedio de 26 en su salida, 3.4 en sus extensiones y su matriz de correlaciones se muestra en la Tabla 3.18.

En relación a los resultados obtenidos en nuestra experimentación, hemos com-

Tabla 3.15: Datos de aprendizaje (ejemplo con multicolinealidad)

y_i	p_i	x_1	x_2
28.9613	4.5829	2.0832	0.9501
21.0799	2.5245	1.2498	0.2311
34.7512	3.8784	1.7658	0.6068
25.6300	4.9939	1.9895	0.4860
35.1855	3.9761	2.4058	0.8913
35.9903	4.2061	2.0773	0.7621
26.2268	5.0932	1.8397	0.4565
15.2024	1.5377	0.5654	0.0185
26.2208	5.1689	1.9904	0.8214
30.6663	4.3713	1.6400	0.4447
32.4267	3.9020	2.1776	0.6154
28.7184	3.7449	1.7432	0.7919
37.9540	3.8122	2.6875	0.9218
31.3361	3.7343	1.9861	0.7382
20.8663	3.2793	1.2154	0.1763

Tabla 3.16: Correlación de datos de aprendizaje (ejemplo con multicolinealidad)

Variable	y	x_1	x_2
y	1.00	0.87	0.80
x_1	0.87	1.00	0.88
x_2	0.80	0.88	1.00

Tabla 3.17: Datos de prueba (ejemplo con multicolinealidad)

y_i	p_i	x_1	x_2
23.48	2.69	1.38	0.40
39.83	6.31	2.74	0.93
36.95	4.33	2.42	0.91
24.38	2.99	1.19	0.41
37.12	5.94	2.13	0.89
16.07	1.46	0.31	0.05
25.47	3.32	1.45	0.35
36.49	4.34	2.07	0.81
18.65	1.85	0.56	0.01
14.50	1.76	0.45	0.13
23.91	3.21	1.13	0.20
19.41	2.49	0.81	0.19
25.52	3.75	2.17	0.60
26.38	3.38	1.44	0.27
19.61	3.16	1.02	0.19

Tabla 3.18: Correlación de datos de prueba (ejemplo con multicolinealidad)

Variable	y	x_1	x_2
y	1.00	0.93	0.95
x_1	0.93	1.00	0.95
x_2	0.95	0.95	1.00

probado que los indicadores con los datos de aprendizaje tienen mejores valores que los obtenidos con los datos de prueba, donde se aprecia una dispersión bastante grande en los indicadores.

Para el indicador R^2_{difuso}, cuyos resultados, mediante una gráfica de comparación de medias, pueden verse en la figura 3.21, el valor fluctúa entre 0.60 y 0.76. Se aprecia un mismo comportamiento para MNP y POS_1, no habiendo un método que se destaque por sobre los otros.

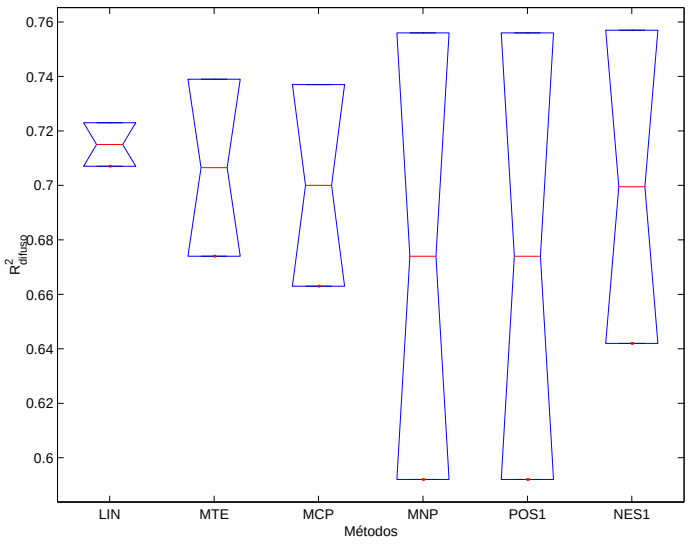

Figura 3.21: Comparación de medias, R^2_{difuso} (Ejemplo con multicolinealidad)

En cuanto al comportamiento de los métodos con el indicador SIM (figura 3.22), MTE y NES_1 son los menos valorados, mientras que LIN y MCP presentan una situación intermedia. MNP Y POS_1, que vuelven a tener un mismo comportamiento, tienen un alto valor para los datos de aprendizaje, pero un muy bajo valor para los datos de prueba, con lo que no permite concluir que sean los más recomendables en situación de multicolinealidad.

De los otros indicadores, que tienen un comportamiento muy similar, el SIM_3, que se muestra en la figura 3.23, vuelve a señalar que MNP y POS_1 tienen mejor comportamiento para los datos de aprendizaje, y el peor valor para los datos de prueba. Los otros cuatro métodos no presentan diferencias en su valoración.

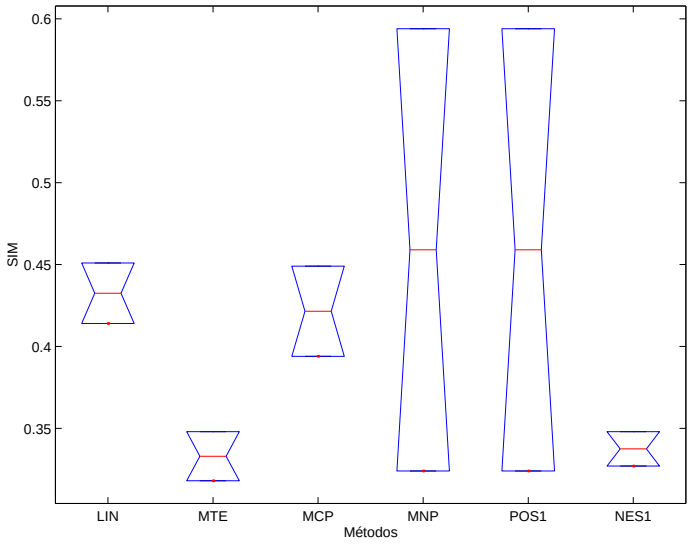

Figura 3.22: Comparación de medias, SIM (Ejemplo con multicolinealidad)

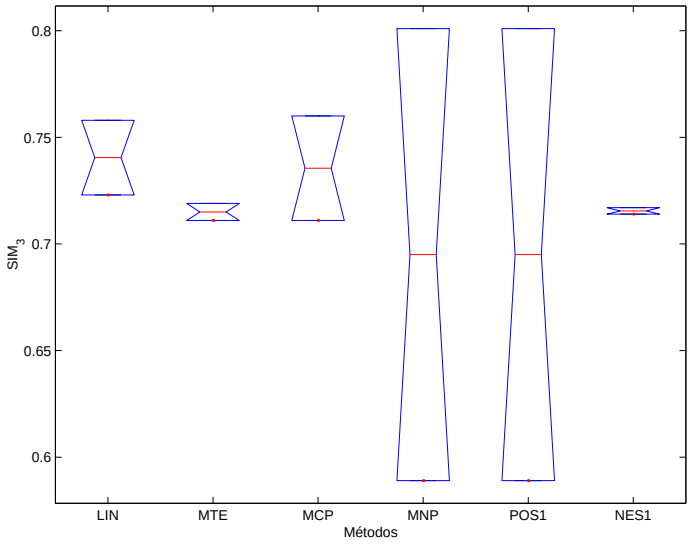

Figura 3.23: Comparación de medias, SIM_3 (Ejemplo con multicolinealidad)

3.3.3. Análisis de Ejemplos de la Literatura

El siguiente paso en nuestro análisis experimental se centra en estudiar el comportamiento de nuestros métodos al trabajar sobre conjuntos de datos de uso frecuente en la literatura para evaluar técnicas de Regresión Difusa. Hemos elegido cuatro conjuntos de datos muy utilizados. Se agrega un quinto ejemplo, en que el indicador de ajuste central llega a tomar el valor 0, por lo que es un caso interesante de estudio.

- El primer conjunto de datos puede verse en la Tabla 3.19. Los datos provienen de [TL98] y consisten en ocho observaciones y tres coeficientes a estimar. La variable de salida tiene función de pertenencia triangular simétrica y el valor de R^2_{difuso} es alrededor de 0.96.

Tabla 3.19: Datos del ejemplo 1 de la literatura de Regresión Difusa

X_1	X_2	p_i	y_i	q_i
1	1	7,50	22,50	7,50
2	4	8,75	28,75	8,75
3	9	10,00	25,00	10,00
4	16	17,50	42,50	17,50
5	25	15,00	40,00	15,00
6	36	12,50	52,50	12,50
7	49	20,00	75,00	20,00
8	64	15,00	85,00	15,00

- El segundo conjunto de datos está utilizado en [Cha01b], fue introducido por Tanaka y es la aplicación más estimada en los estudios de Regresión Difusa. Consiste en 15 observaciones, seis coeficientes a estimar (incluyendo la constante), y la función de pertenencia de la salida es triangular simétrica. El valor de R^2_{difuso} es alrededor de 0,99 (ver tabla 3.20).

- El tercer fichero se encuentra en [Cha01b], con ocho observaciones, dos coeficientes a estimar y funciones de pertenencia triangulares no simétricas (ver Tabla 3.21). R^2_{difuso} es del orden de 0.70.

- El cuarto fichero también ha sido aplicado en varias comparaciones y se puede ver en [KC03]. Sólo se estiman dos coeficientes, con cinco datos (ver tabla 3.22). R^2_{difuso} es aproximadamente 0.80.

- Finalmente se ha agregado un quinto conjunto de datos, que no ha sido publicado anteriormente. Se presenta porque diversas estimaciones producen un R^2_{difuso} igual a cero, lo que indica que la imprecisión tiene una forma muy

Tabla 3.20: Datos del ejemplo 2 de la literatura de Regresión Difusa

x_1	x_2	x_3	x_4	x_5	p_i	y_i	q_i
1	38,09	36,43	5	1	500	6060	500
1	62,10	26,50	6	1	50	7100	50
1	63,76	38,09	7	1	400	8080	400
1	74,52	17,50	8	1	150	8260	150
1	75,38	41,10	7	2	750	8650	750
2	52,99	26,49	4	2	450	8520	450
2	62,93	26,49	5	2	700	9170	700
2	72,04	33,12	6	3	200	10310	200
2	76,12	43,06	7	2	600	10920	600
2	90,26	42,64	7	2	100	12030	100
3	85,70	31,33	6	3	350	13940	350
3	95,27	27,64	6	3	250	14200	250
3	105,98	27,64	6	3	300	16320	300
3	79,25	66,81	6	3	500	16320	500
3	120,50	32,25	6	3	650	15990	650

Tabla 3.21: Datos del ejemplo 3 de la literatura de Regresión Difusa

X_1	X_2	p_i	y_i	q_i
1	2	1,20	14,00	1,60
1	4	1,60	16,00	1,90
1	6	1,40	14,00	1,70
1	8	1,20	18,00	1,80
1	10	2,30	18,00	1,90
1	12	1,50	22,00	1,70
1	14	2,70	18,00	1,80
1	16	1,10	22,00	1,90

Tabla 3.22: Datos del ejemplo 4 de la literatura de Regresión Difusa

X_1	X_2	p_i	y_i	q_i
1	1	1,80	1,50	1,80
1	2	2,30	2,10	2,30
1	3	2,60	1,94	2,60
1	4	2,60	13,50	2,60
1	5	2,40	13,00	2,40

particular, lo que no se observa en ningún otro caso de estudio. Además, refleja una síntesis estadística en que los valores de salida varían entre cero y un máximo, de modo que, por lo general, p_i es similar a y_i y el valor máximo puede ser alto, por lo que q_i es mayor que y_i (ver tabla 3.23). Hay nueve observaciones y tres coeficientes a estimar y funciones de pertenencia triangulares no simétricas.

Tabla 3.23: Datos del ejemplo 5

x_1	x_2	p_i	y_i	q_i
1	2	1,50	1,50	2,50
1	3	2,10	2,10	3,90
1	4	1,94	1,94	5,06
1	5	2,00	2,00	5,00
1	6	0,90	1,90	0,10
2	2	2,00	2,00	3,00
2	3	1,50	1,50	4,50
2	4	1,39	1,39	4,61
2	5	1,56	1,56	2,44

Para compatibilizar los indicadores de todos estos ejemplos, los resultados de las 5 estimaciones se presentan en los gráficos de las figuras siguientes, de manera que en cada estimación la mejor valoración tiene un magnitud 1.

Para el indicador R^2_{difuso} las diferencias son menores, salvo para el ejemplo 5, en que sólo MNP y POS_1 tiene un valor mayor que cero.

El gráfico 3.24 muestra que los métodos MCP, MNP y POS_1 presentan el mejor promedio del indicador, mientras LIN es claramente el indicador más bajo.

Más interesante es el gráfico 3.25, que muestra el indicador SIM para las cinco estimaciones. El mejor valor lo presentan MNP y POS_1, que por tratarse la experimentación con conjuntos con pocas observaciones, entregan la mismas soluciones, seguidos por el método MCP, que tiene un promedio superior a 0.95. LIN, MTE y NES_1 presentan valores más pequeños.

3.3.4. Experimentación según la Cantidad de Observaciones

Es bien conocido que la Regresión Posibilística es superior a la Regresión Probabilística cuando se estima con pocas observaciones y que, en cambio, la Represión Probabilística es superior en presencia de muchas observaciones [KMK96].

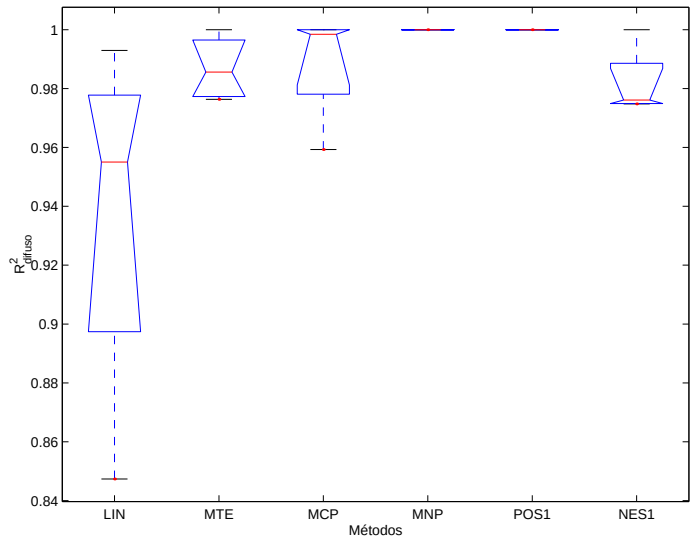

Figura 3.24: Comparación de medias, R^2_{difuso} (Ejemplos de otras publicaciones)

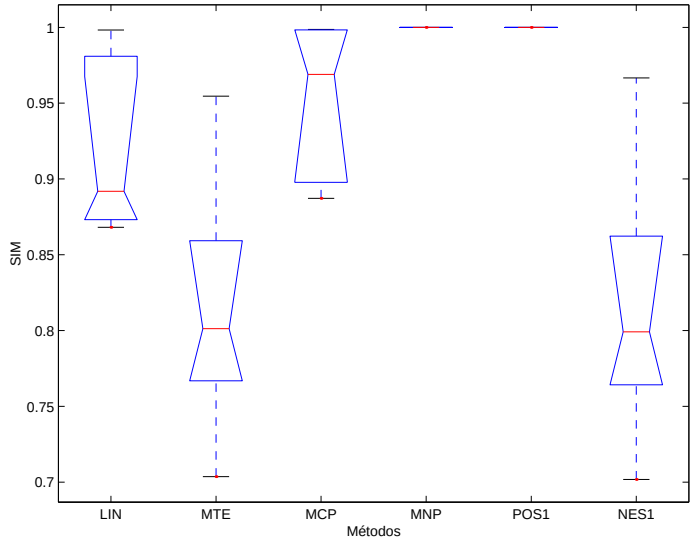

Figura 3.25: Comparación de medias, SIM (Ejemplos de otras publicaciones)

Tabla 3.24: Datos según cantidad de observaciones (15 datos)

x_1	x_2	x_3	x_4	x_5	y_i	p_i
0.63	0.44	0.02	0.52	0.23	5.69	2.66
0.09	0.03	0.06	0.85	0.02	2.19	0.98
0.97	0.72	0.29	0.78	0.75	11.44	1.52
0.13	0.26	0.71	0.54	0.57	6.64	2.49
0.01	0.58	0.07	0.01	0.12	3.14	0.63
0.98	0.01	0.25	0.76	0.98	8.46	3.01
0.67	0.68	0.37	0.24	0.97	10.04	2.43
0.26	0.12	0.01	0.66	0.45	4.16	1.12
0.15	0.12	0.95	0.10	0.07	4.54	0.58
0.47	0.54	0.26	0.05	0.02	5.31	0.49
0.54	0.76	0.75	0.17	0.23	9.53	1.60
0.86	0.00	0.82	0.21	0.68	8.62	2.64
0.36	0.07	0.91	0.25	0.64	6.45	2.70
0.41	0.03	0.71	0.51	0.84	6.61	3.25
0.40	0.24	0.56	0.12	0.19	5.32	1.05

Terminamos nuestro análisis experimental haciendo un estudio en el que datos provenientes de un mismo modelo serán estimados con 15 observaciones, con 150 observaciones y con 300 observaciones; y se medirán los indicadores de bondad de ajuste. Se generaran los datos de acuerdo a un modelo que contempla alteraciones aleatorias en los coeficientes. No se presentará el efecto de puntos extremos dado que las alteraciones aleatorias están acotadas, ya que la presencia de puntos extremos podría alterar las conclusiones de las comparaciones.

El modelo tiene 5 variables de entrada, cada variable con valores entre 0 y 1 aleatorios, como se observa en el la Tabla 3.24, y la función de pertenencia de la variable de salida es simétrica.

Los resultados obtenidos muestran que, para R^2_{difuso}, no hay diferencias significativas entre las tres estimaciones. Esto indica que las variaciones en el ajuste central no están influidas por la cantidad de datos, en el caso de funciones de pertenencia no simétricas para el valor estimado que se ha propuesto.

En cambio, para el indicador SIM, se produce siempre una bajada de la valoración cuando se compara la estimación de 15 datos con la de 150 datos. El gráfico 3.26 muestra los dos valores obtenidos para cada método, donde el valor superior es con 15 datos y el valor inferior es el correspondiente a la estimación con 150 datos.

Mientras que para el método MTE la diferencia es de un 19 %, para el método NRS_1 la diferencia es de 18 %, para el método LIN la diferencia es del 17 % y para el método MCP la diferencia es del 15 %. En todos estos métodos se manifiesta

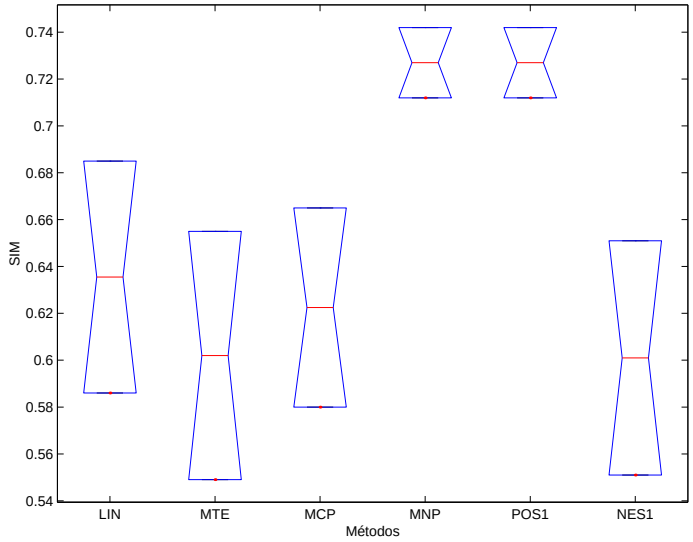

Figura 3.26: Comparación de medias, SIM (Según cantidad de observaciones)

abiertamente el efecto negativo que produce el aumento del número de observaciones.

Por otra parte, para MNP y POS_1, la diferencia es de un 4 %, por lo que el efecto del aumento de las observaciones es mucho más leve que en los otros métodos.

En otras palabras, cuando se estima con 15 observaciones, MNP tiene para SIM un valor de 0.74, que baja a 0.71 con 150 observaciones. En cambio, MCP baja de 0.67 con 15 observaciones a 0.58 con 150 observaciones.

Si se aumenta el número de observaciones para la estimación a 300, con el mismo modelo de generación de datos, y se repiten las estimaciones para cada método se obtienen nuevas conclusiones.

Para SIM, MNP se obtiene el valor 0.718. Es decir, casi el mismo que con 150 observaciones (incluso algo mejorado). Mientras tanto, MCP obtiene un valor de 0.57, bajando muy levemente respecto al anterior valor de 0.58. Esto estaría indicando que, como no hay presencia de puntos extremos, SIM se muestra estable a partir de una cierta cantidad de observaciones.

Lo mismo ocurre con SIM_3 (figura 3.27) y SIM_4 (figura 3.28): MNP y POS_1 son los mas estables y mejor valorados.

Para el indicador SIM_5, ocurre lo mismo que con R^2_{difuso}. No se produce un cambio de magnitud del indicador en función de la cantidad de observaciones, lo

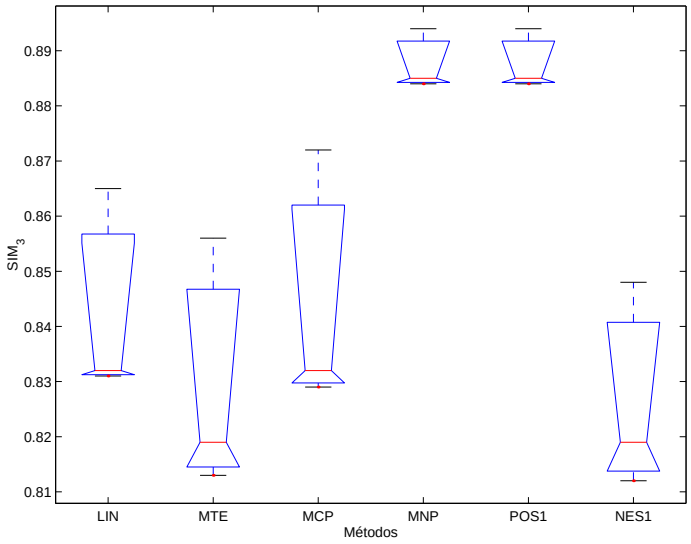

Figura 3.27: Comparación de medias, SIM_3 (Según cantidad de observaciones)

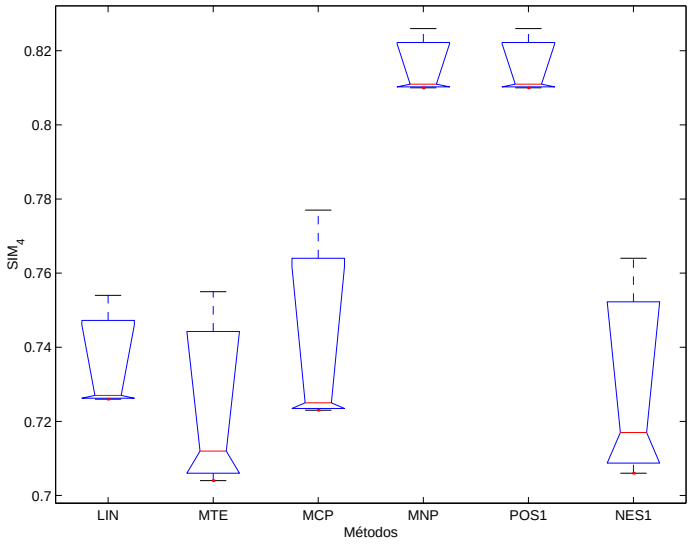

Figura 3.28: Comparación de medias, SIM_4 (Según cantidad de observaciones)

que se aprecia en la figura 3.29. El indicador fluctua entre 0.9 y 0.96 para todos los métodos y todas las estimaciones.

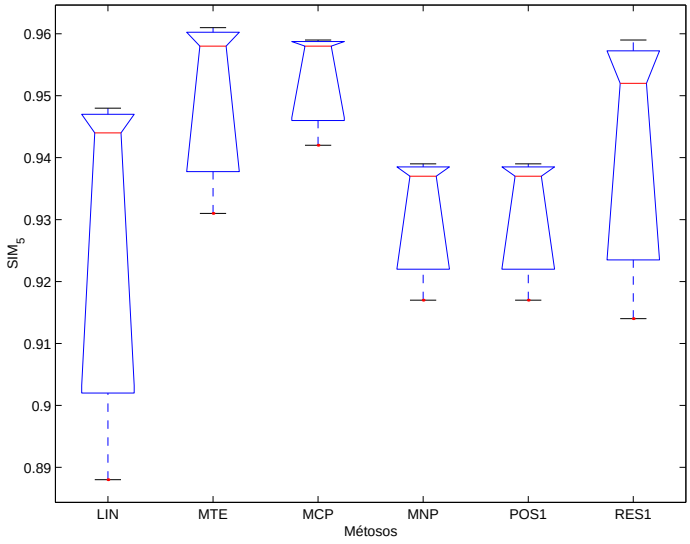

Figura 3.29: Comparación de medias, SIM_5 (Según cantidad de observaciones)

3.3.5. Resumen de los Resultados de la Experimentación

Para el conjunto de experimentos que se han realizado, y que se han cuantificado con el conjunto de indicadores de bondad del ajuste presentado, se observa una cierta tendencia en los resultados que se resumen en la figura 3.30. Los casilleros en blanco en el gráfico reflejan resultados similares entre diversos métodos.

En esta representación se aprecia que para el indicador SIM, que reúne las mejores cualidades de SIM_1 y SIM_2, y para SIM_4, es indiscutible la superioridad de los métodos MNP y POS_1, que en muchos casos, obtienen las mismas estimaciones.

Para los indicadores SIM_3 y SIM_5, los resultados son más matizados, incorporándose, además de los dos anteriores, los métodos MCP, MTE y LIN en algunas de las experimentaciones.

Para R^2_{difuso}, la conclusión más importante es que los métodos con nuestra función objetivo cuadrática resultan los más recomendables, especialmente el método MNP.

	Indicadores				
Datos	R^2_{difuso}	SIM	SIM_3	SIM_4	SIM_5
Ejemplo COPEC	**MNP**	**MNP**	MCP	**MNP**	MCP
Chequeo sin multic	MCP	POS_1	LIN	**MNP**	MTE
Chequeo con multic		**MNP** POS1	MCP	**MNP** POS1	**MNP** POS1
Ejemplos literatura	**MNP** POS1	**MNP** POS1		**MNP** POS1	
Según cantidad de datos		**MNP** POS1	**MNP** POS1	**MNP** POS1	

Figura 3.30: Resumen de los resultados del estudio de métodos

3.4. Conclusiones sobre Métodos de Regresión Difusa

Las propuestas sobre Regresión Difusa que se han propuesto en la literatura científica, que hemos revisado en el segundo capítulo de esta memoria, han estado cruzadas por dos consideraciones:

(a) Según las restricciones, han existido propuestas posibilísticas y propuestas no posbilísticas.
(b) Según la función objetivo, se ha propuesto la minimización de la incertidumbre o se ha propuesto la minimización del ajuste central (mínimos cuadrados).

Esta doble separación, cuyos exponentes más destacados son el enfoque posibilístico con minimización de la incertidumbre (Tanaka), y en el enfoque no posibilístico con minimización del ajuste central, ha sido superada con nuestra propuesta de una optimización de una función cuadrática, al integrar en una sola función objetivo las dos orientaciones anteriores y al permitir una diversidad de restricciones que van desde la posibilística a la no posibilística.

De la experimentación realizada se desprende que los métodos LIN y MTE son los que presentan valores más bajos para la mayoría de los indicadores y la mayoría de los conjuntos de datos. Tomando en cuenta que el método LIN es el que ha sido más utilizado en aplicaciones prácticas, pensamos que al incorporar esta nueva

metodología de Regresión Difusa, se podrá disponer de un método que mejorará los resultados en nuevas aplicaciones.

Por otra parte, los métodos NES_2 y NES_3 resultaron no factibles en la mayoría de las experimentaciones, y el método NES_1 tiene un comportamiento que se asemeja mucho al del método MTE (por ejemplo en 3.27), por lo que no presenta ventajas frente a los métodos mejor valorados.

Respecto a los otros tres métodos, MCE, MNP y POS_1, aunque MNP es el que en más ocasiones presenta la mejor valoración (ver figuras 3.12 y 3.25), pensamos que, dependiendo del propósito del tomador de decisiones, será el método que le sea más apropiado emplear. Sólo en el caso de disponer de una gran cantidad de observaciones, en que MCE no es el más recomendable (por ejemplo, ver figuras 3.27 y 3.28), en otras situaciones la elección dependerá del objetivo que se persiga.

Cuando se busca una estimación en que los riesgos estén bien controlados, el método MCP será el apropiado. A su vez, MNP será el enfoque en que se busca mayor precisión en la implementación del modelo, sin importar mucho los riesgos que se asumen. El método POS_1 refleja una situación intermedia entre los dos situaciones anteriores.

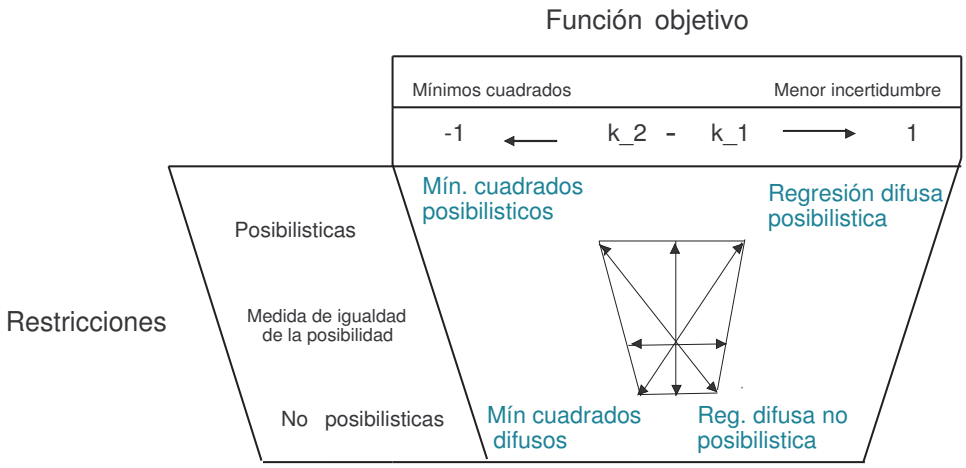

Figura 3.31: Propuesta del enfoque cuadrático de la regresión difusa

Estas conclusiones se han resumido en el gráfico de la figura 3.31, el cual tiene una forma de trapecio invertido. Esto es debido a que las estimaciones posibilísticas permiten un mayor margen de elección entre el objetivo de mínimos cuadrados y el

objetivo de minimización de las extensiones de las estimaciones. Mientras tanto, las estimaciones no posibilísticas, al poner menos condiciones a las extensiones, acercan ambos objetivos a resultados mucho más parecidos.

El resumen de nuestra propuesta de Regresión Difusa con funciones de optimización cuadráticas, donde por el lado de la función objetivo se puede hablar de un continuo de soluciones (dependiendo de los parámetros elegidos), se encuentra reflejado en las flechas del gráfico de la figura 3.31. Se puede observar la gradualidad de las alternativas: por el lado de las restricciones, son fundamentalmente tres los niveles propuestos; en la función objetivo, depende de la elección de los parámetros k_1 y k_2.

Las expresiones *mínimos cuadrados difusos* corresponden a una estimación de mínimos cuadrados (en nuestro modelo $k_1 = 1$ y $k_2 = 0$), con extensiones no posibilísticas. Mientras tanto *mínimos cuadrados posibilísticos* corresponde a una estimación de mínimos cuadrados con extensiones posibilísticas.

Con respecto a los indicadores, que hemos evaluado en detalle, se ha notado una distinta valoración de las estimaciones para los diversos métodos. Por una parte, los indicadores SIM y SIM_4 tienen un comportamiento bastante similar y son los que muestran una valoración más acertada en la línea de *predicción difusa*. En cambio, SIM_3 se parece mucho más a SIM_5, con un comportamiento disímil para los diversos experimentos. Por último, R^2_{difuso} es un indicador que presenta un tercer comportamiento, como reflejo del ajuste de la tendencia central.

Capítulo 4

Selección Automática de Variables

Siguiendo con nuestro objetivo general de búsqueda de una metodología flexible y adaptable para aplicar Regresión Difusa en problemas de predicción en minería de datos, nos centramos en este capítulo en estudiar y aportar una nueva metodología que permita la selección de variables en aquellos problemas que vienen descritos por un gran número de variables independientes.

La selección de un subconjunto de variables de entrada entre todas las variables disponibles ha sido un tema de gran interés para la investigación en la regresión múltiple clásica de mínimos cuadrados. En general, el objetivo es definir un modelo con el menor número de variables independientes; por tanto, cuando se dispone de un gran número de variables de entrada, la finalidad es seleccionar un subconjunto representativo de tales variables. Conviene disponer del menor número de variables de entrada posible, tanto por la simplificación y facilidad de comprensión del modelo resultante, como por la disminución de tiempo de proceso que puede significar tener un modelo con una gran cantidad de variables.

4.1. Descripción del Problema

En la regresión clásica, para la selección de variables se han planteado tres métodos distintos:

- *Regresión hacia adelante* (forward), que consiste en partir con un modelo con una sola variable independiente, y en los pasos sucesivos se añaden una a una

las otras variables relevantes.

- *Regresión paso a paso* (stepwise), similar a la regresión hacia adelante, pero con el añadido de que, en una etapa determinada del proceso de selección, se realiza un chequeo para la posible eliminación de variables ya seleccionadas.
- *Regresión hacia atrás* (backward), que consiste en comenzar con todas las variables independientes dentro del modelo, e ir eliminando variables no significativas en cada una de las etapas siguientes.

El problema de selección de variables en el ámbito difuso se puede describir así: Se dispone de una variable difusa Y=(y,p,q) con función de pertenencia triangular no simétrica, y se tienen $x_1, ..., x_m$ variables exactas como posibles variables explicativas de Y, todas con n mediciones. *¿Cuáles de estas variables son más relevantes para construir un modelo explicativo?*

Wang y Tsaur [WT00a] presentan un método para la incorporación paso a paso de variables de entrada en un modelo de regresión difusa, siguiendo el esquema de la regresión hacia adelante de la regresión clásica. Plantean un doble criterio que se debe minimizar para evaluar la incorporación de variables a un modelo de regresión difusa múltiple. Estos dos criterios son:

I) La borrosidad del modelo o amplitud de la estimación, medida como la suma para los n datos de las diferencias $(\widehat{y_i}+\widehat{q_i}-(\widehat{y_i}-\widehat{p_i}))$, que se pueden simplificar a $(\widehat{q_i}+\widehat{p_i})$.

II) La suma de los residuos al cuadrado, es decir, la diferencia entre el valor observado y el valor estimado, que es el criterio clásico de la regresión estadística que trata de minimizar con el método de los mínimos cuadrados.

En caso de que un solo modelo no minimice ambos criterios, es decir, no hay un único modelo dominante para cada criterio, los autores plantean ocupar el criterio de *la menor resistencia posible*, a través de una distancia asociada a cada criterio: el modelo que promedie una menor distancia será el modelo elegido.

Sobre qué criterios se deben utilizar en el proceso de selección de variables, se pueden aplicar muchas medidas o indicadores. Por ejemplo, para el mismo enfoque de los autores Wang y Tsaur, se podrían plantear los criterios alternativos:

I) Las diferencias $|y_i + q_i - (\widehat{y_i} + \widehat{q_i})| + |y_i - p_i - (\widehat{y_i} - \widehat{p_i})|$, por ser un criterio

más general que funciona mejor tanto para modelos posibilísticos como no posibilísticos.

II) Las diferencias $|y_i - \widehat{y_i}|$, que consideran la distancia real entre los dos centros de los números difusos, y no la distancia cuadrática, que podría tener un mayor efecto distorsionador.

Dada la gran cantidad de criterios que pueden utilizarse, en nuestra investigación, hemos considerado razonable elegir una criterio de minimización de la incertidumbre y otro criterio de minimización de la desviación central. Sin embargo, siempre que sea posible, es recomendable elegir un solo criterio que permita discriminar en la selección de las variables que se deben incorporar al modelo de Regresión Difusa.

En la siguiente sección se define un procedimiento general de selección de variables para la Regresión Difusa, que incluye el criterio paso a paso de reconsideración de variables ya seleccionadas en pasos anteriores. Luego se presenta, para la regresión de tipo posibilístico, la proposición un solo criterio de selección de variables: el indicador integrado de bondad de ajuste SIM que ya conocemos.

A continuación, para cualquier regresión difusa, se proponen dos criterios de selección, uno basado en el ajuste central y el otro en las extensiones. Finalmente, se aplican estos criterios a un caso real, en que se evalúan los dos tipos de criterios indicados.

4.2. Procedimiento General de Selección de Variables

En nuestra investigación, hemos tratado de definir un algoritmo de selección de variables para la Regresión Difusa que sea lo más general posible, adaptable a diferentes combinaciones de criterios de selección.

Como vamos a ver, se ha desarrollado un procedimiento general, independiente del número de criterios que se consideren en el proceso de selección de variables. Esto no significa que se abandone la conveniencia del principio de *parsimonia* – mientras más criterios se incorporen, mejor es el selección, sino que consideramos, por el contrario, que mientras menos criterios se incorporen a nuestro objetivo de selección, más claro e interpretable es el procedimiento.

Un elemento adicional que se incorpora en este procedimiento, es la posibilidad de reconsiderar la inclusión de una variable en el modelo, lo que en la regresión clásica se denomina paso a paso (stepwise).

Para explicar esta reconsideración, mostraremos una situación de ejemplo considerando el aporte de las variables medido con el indicador R^2_{difuso}:

- Si se tienen 3 variables, puede ocurrir que la variable 1, sea la mejor variable para añadir al modelo en el primer paso, ($R^2_{difuso}(x_1) = 0{,}5$), frente a las otras dos variables con ($R^2_{difuso}(x_2) = 0{,}45, R^2_{difuso}(x3) = 0{,}3$).
- El segundo paso, nos lleva a elegir entre x_2 y x_3 para que seleccionar la segunda variable en el modelo, y se elige x_3 con un valor para el indicador $R^2_{difuso}(x_1, x_3) = 0{,}65$.
- Sin embargo, en este punto de la selección, podría ocurrir que el valor de $R^2_{difuso}(x2, x3)$ sea igual a 0.7. Entonces se *reconsidera* la incorporación de x_1, y nuestro mejor modelo con dos variables es con x_2 y x_3.

En nuestro sencillo ejemplo, esta aparente inconsistencia en el proceso de selección paso a paso, se explica debido a que la correlación entre x_1 y x_2 es alta, puesto que si no existiera esta alta correlación, esta situación de reconsideración de variables dentro del modelo no debiera producirse.

Para el caso de que un conjunto de variables no sea el dominante en todos los criterios, hemos definido una medida de distancia que determinará qué conjunto de variables es más razonable elegir.

Dicho esto, el procedimiento general de selección de variables para la regresión difusa es el siguiente:

1 Se consideran k criterios de selección $(cr_1, ..., cr_k)$. Estos criterios de selección deben tener el *sentido* de que a menor valor, mejor es la valoración del modelo. También se considera una ponderación (valoración) de cada criterio, λ_k, de manera que la suma de las ponderaciones sea 1.
Por último, se definen los conjuntos siguientes que se emplearán en el algoritmo de selección:

   - $NI = \{x_1, ..., x_m\}$ con todas las variables que no han seleccionado para añadir al modelo de regresión.
   - $I = \varnothing$ que contiene las variables seleccionadas para el modelo de regresión.

2 Seguidamente, se elige una medida inicial de cada criterio, $CR_1^0, ..., CR_k^0$, en un valor muy alto, que supere los posibles valores que puede tomar el criterio. Por ejemplo si el criterio puede variar entre 0 y 1, se toma el valor 1.
También se elige una cota s de manera que, si el aporte de una variable al modelo es menor que s, se para el proceso de selección de variables. Por ejemplo, se puede tomar s=0.1 (Wang y Tsaur toman en su ejemplo un valor de 0.3).

3 Si $NI = \varnothing$, se termina el proceso de selección de variables, y el conjunto I es el conjunto de variables elegidas en la selección.

4 En otro caso, (Paso p), para cada x_j de NI, se realiza la regresión difusa con las variables $I \bigcup \{x_j\}$, y se evalúan los criterios $cr_1, ..., cr_k$, obteniéndose los valores $CR_1^{p(j)}, ..., CR_k^{p(j)}$.

- Si para cada criterio k, el mínimo entre los $CR_k^{p(j)}$ corresponde a la misma variable j, que será denominado como CR_k^p, se elige la variable x_j como candidata a ingresar en el modelo.

  Esta situación, de encontrar una única variable j, siempre ocurrirá si hay sólo un criterio en consideración.

- Si no hay un criterio predominante, para cada criterio, se calculan los $CR_k^{p(j)}$ mínimos, que denominaremos $CR_k^{p(min)}$ y se calculan, para normalizar las diferencias, los valores

$$d_k^j = \frac{CR_k^{p(min)}}{CR_k^{p(j)}} \tag{4.1}$$

  y se totalizan estas distancias en un solo indicador

$$L^j = \sum_k \lambda_i (1 - d_k^j) \tag{4.2}$$

  eligiéndose la variable x_j con un menor L^j , como variable candidata para ingresar en el modelo.

5 Se comprueba si el aporte de la variable candidata es significativo para el modelo. De no serlo, el proceso de selección de nuevas variables se termina.
Para hacer esta comprobación, se calcula

$$t = \frac{\frac{CR_1^{p+1}}{CR_1^p} + ... + \frac{CR_k^{p+1}}{CR_k^p}}{k} \tag{4.3}$$

y si (1-t) es menor que s, entonces se termina el proceso de selección por ser no significativo el aporte de la nueva variable candidata.

6 Se hace una *reconsideración* para confirmar si la variable candidata no implica tener que retirar una variable anteriormente incorporada.
Para ello se calcula la correlación lineal entre x_k y las variables en el conjunto I y, con la variable de mayor correlación absoluta x_c, se calcula $CR_k^{p-1(*)}$ incorporando a x_k y excluyendo a x_c, como variables independientes, para los k criterios.
Se deciden las variables finalmente seleccionados como:

- x_j ingresa al modelo si CR_k^{p-1} es mejor que $CR_k^{p-1(*)}$
  y se redefinen los conjuntos

$$I = I \bigcup \{x_j\} \tag{4.4}$$

$$NI = NI - \{x_j\} \tag{4.5}$$

- Si el antiguo óptimo CR_k^{p-1} es peor criterio que el nuevo valor $CR_k^{p-1(*)}$, la variable x_c es excluida del modelo y se reemplaza por x_j, con lo que los conjuntos de variables quedan conformados por

$$I = I - \{x_c\} \bigcup \{x_j\} \tag{4.6}$$

$$NI = NI - \{x_j\} \bigcup \{x_c\} \tag{4.7}$$

7 Se vuelve al punto 3.

A continuación, en las siguientes secciones, este procedimiento será aplicado a distintos conjuntos de criterios de selección.

4.3. Selección de Variables para Modelos Posibilísticos

Para los modelos de regresión difusa de tipo posibilístico que se definieron en el capítulo anterior (LIN, MTE, MCP, POS_1), se considerará que el criterio de selección de variables más recomendable, cuando se quiera utilizar un único criterio,

debe tener en cuenta nuestra proposición de bondad de ajuste integrada SIM (véase apartado 2.4.2.7), que reúne diversas cualidades deseables de la estimación.

Este indicador es el único que incorpora en forma global las funciones de pertenencia de los números involucrados ((SIM_1)), y tiene una corrección para el ajuste de las extensiones ((SIM_2)). Por este motive, consideramos que reúne, con la mejor precisión para un solo indicador, tanto la medición de la incertidumbre como la magnitud del ajuste de la tendencia central. Como hemos usado el sentido general de que la medición del criterio debe disminuir al mejorar el modelo, se deberá considerar como criterio el valor 1 - SIM, en vez de SIM.

Veamos un ejemplo de su aplicación. En el ejemplo, se dispone de cinco variables. Todas las regresiones incluirán, además, el término constante. Los datos se muestran en la tabla 4.1, junto a la variable Y, formada por el valor central y por la extensión simétrica p.

Se calculará el valor SIM para los cuatro modelos de tipo posibilístico que consideraremos, con los parámetros h=0, $k_1 = 1$ y $k_2 = 1$ y el valor de cota de significancia s=0.1.

El primer paso es estimar los modelos que contienen sólo una variable, además de la constante.

Se aprecian los resultados obtenidos para SIM en la tabla 4.2, donde se observa que, para todos los métodos, la mejor variable para ser añadida al modelo es la variable 5. Nótese que, en general, la progresión en el ordenamiento de SIM es la misma, lo que habitualmente debiera ocurrir, aunque no es algo garantizado.

Luego, para el único criterio que estamos empleando, 1 - SIM, en el paso 1, el valor CR_1^1 es 1 - 0.408, 0.592, si tomamos como referencia el método MCP.

El segundo paso es calcular todas las combinaciones con 2 variables, a partir de $I = \{x_5\}$.

En la tabla 4.3 se aprecian los resultados de las 4 posibles combinaciones de variables. Se comprueba que los valores de SIM son, por lo general, mayores que los de la tabla 4.2. La variable que presenta mejor indicador es la variable 2, para el método MCP, por lo que es la segunda variable elegida para el modelo.

Para el criterio que se definió, 1 - SIM, en el paso 2, el valor CR_1^2 es 1 - 0.502, 0.498, (método MCP), debiéndose calcular la razón del valor t, que resulta 0.498/0.592. Por tanto, 1-t es mayor que s, y continua el proceso de selección de variables.

Tabla 4.1: Datos del ejemplo 1 de selección de variables

y	p	x_1	x_2	x_3	x_4	x_5
5.8294	0.6829	0.3172	0.9302	0.6821	1.3353	1.5646
3.8135	0.4778	0.8202	1.0053	0.1445	0.5288	0.3147
6.1790	0.8059	0.6380	1.0986	1.2423	1.4619	0.7381
4.7988	0.6363	1.0863	0.8667	0.4782	0.4856	0.8820
6.4726	0.8032	0.7493	1.2248	0.8105	1.6309	1.0571
6.7535	0.8513	0.6609	1.0081	1.0990	1.7087	1.2768
5.7381	0.7276	0.9913	0.5503	0.5465	1.4321	1.2179
3.2304	0.4241	0.5310	0.2982	0.4798	0.8068	0.1145
4.9479	0.5948	0.4637	0.7161	0.5364	1.2340	0.9977
5.8565	0.7914	0.8135	0.7371	1.2437	0.9289	1.1334
6.1173	0.8087	1.0337	1.0080	0.9362	0.8099	1.3296
5.1741	0.6324	0.2713	0.6708	0.8782	1.4103	0.9436
6.5872	0.8606	0.9742	1.2592	1.0448	1.2126	1.0963
6.0067	0.7379	0.6141	0.9050	0.7578	1.6749	1.0549
4.5635	0.5985	0.8878	0.4509	0.5341	0.8973	0.7934
4.3162	0.5407	0.6318	0.8879	0.4589	0.8048	0.5329
7.2111	0.9335	1.0068	0.9736	1.1169	1.6503	1.4635
6.7319	0.8739	0.7203	0.8634	1.2868	1.1557	1.7057
5.1177	0.6318	0.4350	0.8818	0.7654	0.8372	1.1983
7.1948	0.8822	0.4737	1.0292	1.1536	1.7987	1.7396
3.1667	0.3621	0.2118	0.8212	0.2423	0.7394	0.1519
4.9167	0.6293	0.7944	1.1179	0.5824	0.5542	0.8678
6.7263	0.8538	0.5612	0.8937	1.2502	1.6266	1.3945
4.1191	0.5350	0.5448	0.8846	0.6865	0.6383	0.3648
2.5568	0.3293	0.1950	0.2364	0.5408	0.2683	0.3163
4.8355	0.6282	0.7623	1.0723	0.6843	0.4035	0.9132
4.4436	0.5793	0.4415	0.3622	0.9083	0.7996	0.9320
5.6351	0.7093	1.0519	0.5279	0.4059	1.2145	1.4349
5.1306	0.6830	0.9402	1.0000	0.7592	0.6340	0.7973
4.2046	0.5026	0.6568	0.8281	0.1644	0.7676	0.7878

Tabla 4.2: Selección de variables, etapa 1, SIM (ejemplo 1)

$Variables$	LIN	MPC	MTE	POS_1
1	0.372	0.318	0.245	0.372
2	0.337	0.338	0.260	0.375
3	0.359	0.364	0.269	0.429
4	0.376	0.384	0.288	0.502
5	0.405	0.408	0.306	0.544

Tabla 4.3: Selección de variables, etapa 2, SIM (ejemplo 1)

$Variables$	LIN	MPC	MTE	POS_1
1, 5	0.408	0.416	0.317	0.539
2, 5	0.498	0.502	0.392	0.565
3, 5	0.459	0.464	0.354	0.596
4, 5	0.486	0.488	0.383	0.538

La tercera etapa consiste en la combinación de las variables que no han sido incorporadas al modelo aún, con las variables x_2 y x_5. Los resultados del indicador SIM se muestran en la tabla 4.4.

Tabla 4.4: Selección de variables, etapa 3, SIM (ejemplo 1)

$Variables$	LIN	MPC	MTE	POS_1
1, 2, 5	0.514	0.515	0.400	0.566
2, 3, 5	0.553	0.556	0.437	0.635
2, 4, 5	0.559	0.557	0.429	0.656

En esta tabla, se aprecia que para los métodos lineal y cuadrático posibilístico, el mayor SIM se obtiene con las variables x_2, x_4 y x_5, mientas que con el método de Tanaka Extendido y Posibilístico 1, el mayor SIM lo entregan las variables x_2, x_3 y x_5.

A pesar de que la diferencia en el método MCP es tan pequeña, se elige la variable x_4 como candidata, dado que se ha tomado a MCP como método de referencia. Ahora corresponde comprobar si hay que cambiar alguna de las dos primeras variables en el modelo. Para ello, se calcula la correlación entre la variable 4 y las variables 2 y 5, que es 0.25 y 0.63 respectivamente, por lo que se evaluará una ecuación sin la variable más correlacionada, es decir la variable x_5. Los valores SIM para el modelo con las variables x_2 y x_4 son 0.410 para MCP, menor que 0.502 que había sido elegido en la tabla 4.3, por lo que se confirma a la variables 4 dentro del modelo.

Para el criterio que se está empleando, 1 - SIM, el valor CR_1^3 es 1 - 0.557, 0.443, (método MCP), quedando la razón para el valor t como 0.443/0.498. Como se ve, 1-t es levemente mayor que s, y continua el proceso de selección de variables.

Como vemos en la tabla 4.5, no hay duda en eligir como candidata a la variable x_3. Para comprobar si hay que eliminar alguna variables, se calcula la correlación lineal entre 3 y las variables 2, 4, y 5, que es 0.3, 0.5 y 0.61, por lo que la variables más correlacionada es la variable 5. Los valores SIM para el modelo sin la variables x_5, es decir, con x_2, x_3 y x_4, son menores a los obtenidos en la tabla 4.4 para x_2, x_4 y x_5, por lo que el modelo queda formado, tras esta etapa, con las variables x_2, x_3, x_4 y x_5, lo que se comprueba con el criterio de término.

Tabla 4.5: Selección de variables, etapa 4, SIM (ejemplo 1)

$Variables$	LIN	MPC	MTE	POS_1
1, 2, 4, 5	0.622	0.620	0.514	0.698
2, 3, 4, 5	0.677	0.671	0.550	0.719

Para el criterio de término, el valor CR_1^4 es 1 - 0.671, 0.329, (método MCP), quedando la razón para el valor t como 0.329/0.443. 1-t es mayor que s, y continua el proceso de selección de variables, y solo resta incorporar la variable 1 en la última etapa, si su aporte resulta significativo.

4.4. Criterios de Selección de Variables en la Regresión Difusa

En esta sección nos centramos en el estudio de los criterios que se pueden utilizar para guiar el proceso de selección de variables. En ese contexto, partiendo de la propuesta existente de Wang y Tsaur, proponemos como alternativa nuestros propios índices, para que el tomador de decisiones pueda elegir en función del objetivo perseguido en el proceso de selección.

Los modelos de regresión difusa que se han propuesto en esta memoria tienen un doble objetivo: minimizar la incertidumbre y minimizar las diferencias del ajuste central.

Como hemos nombrado al principio de este capítulo, Wang y Tsaur presentaron dos criterios de selección que tienen el mismo sentido anterior. Pero la formalización de esos dos criterios, nos parece que puede ser mejorada.

Siguiendo este doble objetivo de la Regresión Difusa, se proponen dos criterios para evaluar la inclusión de variables en un modelo de regresión difusa:

1 El criterio basado en las diferencias en las extensiones:

$$inc = \sum_{i=1}^{n}(|y_i + q_i - (\widehat{y_i} + \widehat{q_i})| + |y_i - p_i - (\widehat{y_i} - \widehat{p_i})|) \tag{4.8}$$

2 El criterio basado en las diferencias de la tendencia central:

$$dif = \sum_{i=1}^{n} |y_i - \widehat{y_i}| \tag{4.9}$$

Se aplicará este doble criterio en un segundo ejemplo de evaluación, con 7 variables (datos en la tabla 4.6). Se eligen las poderaciones λ_i igual a 0.5 y s=0.1. La primera etapa se muestra en la tabla 4.7 para tres modelos de regresión difusa: el

modelo no posibilístico, el modelo cuadrático posibilístico y el modelo posibilístico 1, con los parámetros de ponderación $k_1 = 1$ y $k_2 = 1$ y el nivel de confianza h=0.

Tabla 4.6: Datos del ejemplo 2 de selección de variables

y	p	x_2	x_3	x_4	x_5	x_6	x_7
5.2237	0.7498	0.9226	0.9817	1.0139	0.4276	0.5398	0.3380
5.8182	0.8440	0.9071	0.6515	0.8787	0.6914	0.8534	0.8361
5.4101	0.7478	0.5191	0.8863	1.0186	0.4516	1.0043	0.5302
6.1249	0.9046	1.0668	1.0068	1.1476	0.5020	0.6947	0.7070
4.6256	0.6886	0.5037	0.1364	1.0273	0.3219	0.9071	0.7292
7.4505	1.1009	0.9907	0.5418	1.1538	1.3258	1.0240	1.4143
4.5720	0.6471	0.6844	0.6279	0.7326	0.6080	0.4368	0.4823
6.0294	0.9096	0.8373	0.2276	1.0579	1.2661	0.4696	1.1710
5.5552	0.8296	0.7586	0.6299	0.6633	0.4859	0.6986	1.3188
4.1365	0.5868	0.7621	0.2924	0.6407	0.6163	0.4965	0.3286
8.2014	1.1030	0.9194	1.3020	0.7902	1.2586	1.8124	1.1188
5.7291	0.7447	0.8613	0.8039	0.3534	1.0562	1.1510	0.5033
5.8000	0.8402	0.8502	0.4370	0.7253	0.7923	0.9691	1.0261
6.3879	0.8963	0.7934	1.1268	0.3890	1.2451	0.4408	1.3928
5.8040	0.7975	1.1492	0.9913	0.4173	1.2486	0.3930	0.6046
6.7551	0.9823	1.0529	1.1973	0.8058	0.4260	1.0639	1.2092
5.4367	0.7204	0.3950	0.8558	0.9050	0.9058	0.9078	0.4673
6.6785	0.8984	0.8272	0.6820	0.6432	1.3758	1.3149	0.8354
6.8488	0.9403	0.4229	0.6993	0.6673	1.3368	1.2584	1.4641
4.4672	0.6217	0.2617	0.5134	0.9201	0.9371	0.2670	0.5680
7.5735	1.1265	1.1850	1.0451	0.5848	0.8823	0.9546	1.9218
8.4372	1.1886	1.0346	1.2285	1.1432	1.3303	1.4295	1.2712
8.1768	1.1055	0.3321	1.1805	1.0236	1.4968	1.6217	1.5220
6.9626	1.0192	0.7724	0.7493	0.9723	0.7209	1.2631	1.4846
5.0629	0.6604	0.4923	0.8496	0.3704	0.6619	1.0102	0.6784
8.2512	1.2170	1.2796	0.9422	0.7975	1.2856	1.1046	1.8417
3.8255	0.5721	0.3703	0.3136	0.5493	0.1610	0.4557	0.9757
5.9244	0.7700	0.2683	0.8353	0.6061	1.3456	0.9677	0.9014
4.6645	0.6569	0.4825	0.6739	0.6814	0.9101	0.1758	0.7407
6.1093	0.8910	1.1246	0.8540	1.1513	0.9448	0.5099	0.5245

Como se ve, los resultados obtenidos no son concluyentes para seleccionar entre las variables x_6 y x_7:

- Para el método MNP tanto el indicador inc como el dif, tienen mejor valor para la variable x_7, pero nuestra referencia SIM_1 tiene mejor valor para la variable x_6. Para el método MCP y el indicador dif, lo recomendable es seleccionar la variable x_7, mientras para el indicador inc, el valor más bajo lo presenta la variable x_6, mientras que el SIM_1 es preferible la variable 6.
- Para el método POS_1, el criterio dif recomienda seleccionar la variable x_7, mientras el criterio inc recomienda la variable x_6. El criterio de referencia

Tabla 4.7: Selección de variables, etapa 1, (ejemplo 2)

	MNP	MNP	MCP	MCP	POS_1	POS_1
Variables	dif	inc	dif	inc	dif	inc
2	26.8	53.6	28.9	129.9	28.1	61.3
3	25.3	50.6	25.1	98.1	25.3	50.8
4	30.4	60.9	30.7	124.2	31.2	63.3
5	23.9	47.8	23.9	96.9	23.9	47.7
6	20.6	41.3	20.6	93.3	20.6	41.4
7	20.2	40.6	20.2	112.5	20.2	50.8

SIM_1 tiene mejor valor para la variable x_7.

En resumen, entre las 6 variables, las posibles candidatas en nuestro primer paso son las variables x_6 y x_7, y debe recurrirse a la solución de controversias (punto 6 del procedimiento general).

En todo caso, está claro que la selección de variables es dependiente del método de regresión difusa que se haya elegido, porque tanto el comportamiento de la tendencia central como el de las extensiones (incertidumbre) dependen del método elegido.

Para MCP, como método de referencia de esta selección se tiene en la tabla 4.8 el cálculo de resolución de controversia.

Tabla 4.8: Elección entre x_6 y x_7, etapa 1, MCP (ejemplo 2)

Criterio	x_6	x_7
cr1	20.6	20.2
cr2	93.3	112.5
d1	0.98	1
d2	1	0.83
L	0.001	0.085

Por el principio de *mínima resistencia* para el método MCP, se elige la variable x_6 en este paso, y lo mismo ocurre con el método POS_1.

Para seguir con el ejemplo, se seleccionará en esta primera etapa la variable 6. Los resultados de este paso 2 se muestran en la tabla 4.9.

En este segundo paso, no hay mayor duda sobre la variable que debe ser seleccionada, porque por los dos criterios, y para los métodos MNP y POS_1, el mejor desempeño lo tiene la combinación de variables x_6 y x_7. Para el método MCP, en cambio, no hay un criterio predominante, porque para el primer criterio la mejor combinación es también x_6 y x_7, pero para el segundo criterio la mejor combinación

Tabla 4.9: Selección de variables, etapa 2, (ejemplo 2)

	MNP	MNP	MCP	MCP	POS_1	POS_1
Variables	dif	inc	dif	inc	dif	inc
2, 6	16.8	33.6	16.9	67.0	16.7	33.6
3, 6	16,4	32.9	17.2	82.5	16.4	32.9
4, 6	19.3	38.7	21.5	89.0	19.3	39.0
5, 6	17.4	34.7	17.5	70.5	17.4	34.8
6, 7	14.1	28.2	15.9	70.9	14.1	29.2

se obtiene con x_2 y x_6.

Seleccionando en el paso dos la variable x_7, se aminora la ambigüedad que se había presentado en el paso 1 entre las variables x_6 o x_7, puesto que si en el paso uno se hubiera elegido la variable x_7, posiblemente le hubiera correspondido el turno a la variable x_6 en el segundo paso.

El paso tres corresponde a las combinaciones posibles con las variables 6 y 7, cuyos indicadores pueden verse en la tabla 4.10.

Tabla 4.10: Selección de variables, etapa 3, (ejemplo 2)

	MNP	MNP	MCP	MCP	POS_1	POS_1
Variables	dif	inc	dif	inc	dif	inc
2, 6, 7	12.3	24.6	12.5	42.6	12.3	24.6
3, 6, 7	11.0	22.1	11.1	53.9	11.0	22.1
4, 6, 7	13.1	26.2	13.8	68.4	13.1	26.3
5, 6, 7	12.6	25.2	13.3	49.3	12.6	25.2

En el paso 3, se observa que la variable candidata a ingresar al modelo es la variable x_3, puesto que para los modelos MNP y POS_1 es la más conveniente por ambos criterios. En cambio, para el modelo MCP, para el criterio uno es recomendable la variable x_3 pero, por el criterio 2, es más recomendable la variable x_2.

Se realiza el proceso de reconsideración de variables, donde para la variable elegida x_3, se calcula su correlación lineal con las variables x_6 (0.44) y x_7 (0.25). Según estos valores, se excluye a la variable de mayor correlación, x_6, y se realiza la regresión con las variables x_3 y x_7, lo que se muestra en la tabla 4.11.

Tabla 4.11: Selección de variables, etapa 3.1, (ejemplo 2)

	MNP	MNP	MCP	MCP	POS_1	POS_1
Variables	dif	inc	dif	inc	dif	inc
3, 7	14.7	29.5	15.3	61.3	14.7	29.5

Si se compara 4.9 con 4.11, se observa que la combinación original, x_6 y x_7, es

más recomendable que la combinación alternativa x_3 y x_7.

La cuarta etapa, consiste en realizar todas las regresiones con las variables x_3, x_6 y x_7. Los resultados se muestran en la tabla 4.12.

Tabla 4.12: Selección de variables, etapa 4, (ejemplo 2)

	MNP	MNP	MCP	MCP	POS_1	POS_1
Variables	dif	inc	dif	inc	dif	inc
2, 3, 6, 7	9.4	18.8	9.7	36.0	9.4	18.8
3, 4, 6, 7	10.1	20.2	10.1	40.2	10.1	20.2
3, 5, 6, 7	9.1	18.2	9.5	37.1	9.1	18.2

En esta etapa, se elige la variable x_5, que es la recomendable para los MNP y POS_1, mientras para el método MCP, con el criterio dif habría que elegir la x_5, mientras con el criterio inc habría que elegir la x_2.

Dada la dualidad de preferencias para la selección que se produce en el método MCP, nos parece interesante examinar este paso si se hubiera utilizado otro criterio para la selección de variables.

En la tabla 4.13 aparecen los resultados, para los tres métodos considerados, y tres combinaciones de parámetros para cada uno de ellos, para el criterio empleado en la sección anterior, SIM.

Tabla 4.13: Selección de variables, etapa 4, criterio SIM (ejemplo 2)

	MNP	MNP	MNP	MCP	MCP	MCP	POS_1	POS_1	POS_1
Variables	$k_1 = 0$	$k_s = 1$	$k_2 = 0$	$k_1 = 0$	$k_s = 1$	$k_2 = 0$	$k_1 = 0$	$k_s = 1$	$k_2 = 0$
2, 3, 6, 7	0.6731	0.6731	0.6731	0.5929	0.5924	0.5888	0.6731	0.6731	0.6731
3, 4, 6, 7	0.6538	0.6538	0.6538	0.5722	0.5721	0.5694	0.6538	0.6538	0.6538
3, 5, 6, 7	0.6860	0.6860	0.6860	0.5880	0.5861	0.5913	0.6860	0.6860	0.6860

Se aprecia, en primer lugar, que los resultados, en cada método, son bastante parecidos para las combinaciones de variables posibles. Sin embargo, para el método MNP siempre es mejor la selección de la variable x_5 y para el método POS_1 también siempre es mejor la selección de la variable x_5. En relación al método MCP, la combinación de parámetros $k_1 = 0$ y $k_2 = 1$, y la combinación $k_1 = 1$ y $k_2 = 1$, tienen su mejor valor con la variable x_2, mientras que para la combinación $k_1 = 1$ y $k_2 = 0$, la mejor variable para seleccionar es la variable x_5.

Para seguir profundizando en este ejemplo, consideraremos un nuevo criterio, R^2_{difuso}, que es el indicador que hemos seleccionado para representar el ajuste de la tendencia central de la regresión difusa.

En la tabla 4.14 se muestran los resultados tomando como criterio de evaluación para la selección de variables del índice R^2_{difuso}. Destacan en esta tabla:

Tabla 4.14: Selección de variables, etapa 4, criterio R^2_{difuso}(ejemplo 2)

	MNP	MNP	MNP	MCP	MCP	MCP	POS_1	POS_1	POS_1
Variables	$k_1 = 0$	$k_s = 1$	$k_2 = 0$	$k_1 = 0$	$k_s = 1$	$k_2 = 0$	$k_1 = 0$	$k_s = 1$	$k_2 = 0$
2, 3, 6, 7	0.9027	0.9027	0.9027	0.8987	0.9007	0.9027	0.9027	0.9027	0.9027
3, 4, 6, 7	0.8930	0.8930	0.8930	0.8926	0.8927	0.8930	0.8930	0.8930	0.8930
3, 5, 6, 7	0.9063	0.9063	0.9063	0.9040	0.9046	0.9063	0.9063	0.9063	0.9063

- Todos los valores son muy homogéneos, con un valor mínimo de 0.8926 y un valor máximo de 0.9063. Para la primera combinación, la fluctuación es entre 0.8987 y 0.9027; para la segunda, entre 0.8926 y 0.8930; y para la tercera combinación de variables, entre 0.9040 y 0.9063. En todos los casos, menos de un 1 % de variación entre el valor más pequeño y el valor máximo.
- A pesar de la proximidad de los valores indicados, el criterio de selección es siempre el mismo: en primer lugar la variable x_5, en segundo lugar la variable x_2, y en último lugar la variable x_4.
- En los tres métodos de regresión difusa de la tabla 4.14, la última combinación de parámetros $k_1 = 1$ y $k_2 = 0$, corresponde a la solución de mínimos cuadrados, y por lo tanto, el valor de R^2_{difuso} es el mismo para los tres métodos.

En consecuencia, utilizando el criterio de R^2_{difuso}, que sólo pondera la calidad del ajuste central de la estimación, se repite lo que se había presentado en en la tabla 4.12 con el criterio dif, que también pondera únicamente la calidad del ajuste de la tendencia central. Ambos producen la misma selección y en el mismo orden, para todos los métodos.

Después de haber hecho esta reflexión y análisis de los criterios de selección, continuemos ahora con el procedimiento general de selección de variables, en su paso 4. Tenemos que controlar el efecto del ingreso de la variable x_5 con las anteriores variables en el modelo, calculándose la correlación entre la nueva variable y el resto, que es 0.32 con la variable x_3, 0.38 con la variable x_6 y 0.39 con la variable x_7; luego se excluye a la variable x_7 en la ecuación de regresión de re-calculo del paso 4.

En la tabla 4.15 se muestra el resultado de la nueva regresión y, si se compara con la regresión de las variables x_3, x_6 y x_7 de la tabla 4.10, se ve que la variable x_7 debe continuar como variable seleccionada para el modelo de regresión.

Tabla 4.15: Selección de variables, etapa 4.1, (ejemplo 2)

	MNP	MNP	MCP	MCP	POS_1	POS_1
Variables	dif	inc	dif	inc	dif	inc
3, 5, 6	14.6	29.2	14.6	53.2	14.6	29.2

En el paso cinco, se hacen las regresiones con las variables x_3, x_5, x_6 y x_7. En la tabla 4.16 se muestran las dos regresiones posibles, para los tres métodos considerados.

Tabla 4.16: Selección de variables, etapa 5, (ejemplo 2)

	MNP	MNP	MCP	MCP	POS_1	POS_1
Variables	dif	inc	dif	inc	dif	inc
2, 3, 5, 6, 7	6.0	12.1	6.0	19.0	6.0	12.1
3, 4, 5, 6, 7	6.6	13.3	6.6	19.3	6.6	13.3

En esta etapa, la variable elegida es la variable x_2, ya que para los dos criterios es preponderante en los tres métodos. La variable con la que se tiene que comprobar la permanencia en el modelo es la variable x_7, resultado la regresión que se muestra en la tabla 4.17.

Tabla 4.17: Selección de variables, etapa 5.1, (ejemplo 2)

	MNP	MNP	MCP	MCP	POS_1	POS_1
Variables	dif	inc	dif	inc	dif	inc
2, 3, 5, 6	10.3	20.7	10.4	37.1	10.3	20.7

Con esta regresión, se confirman las variables seleccionadas, y el quinto paso termina con las variables x_2, x_3, x_5, x_6 y x_7 en el conjunto I.

Se aprecia en este ejemplo, que en las primera etapas, los tres métodos considerados daban resultados diferentes, pero a medida que se avanzó en el proceso y, por lo tanto, las estimaciones se hicieron más ajustadas al valor observado, el método no posibilístico produjo los mismos resultados que el método posibilístico 1, lo que significa que la restricción de la igualdad de la posibilidad, no restringe el espacio de soluciones factibles óptimo cuando la estimación es cercana al valor observado.

Asímismo, se comprueba que el método cuadrático posibilístico es que el produce las extensiones más amplias.

4.5. Ejemplo Real de Selección de Variables

Tomando como base las propuestas de las dos secciones anteriores, en las cuales se propuso un criterio único para la regresión difusa posibilística, y un doble criterio para la regresión difusa, se aplicarán esas propuestas al ejemplo concreto que hemos estamos considerando en nuestro trabajo de investigación, que es la determinación de los factores que influyen en el precio de las acciones de la empresa chilena COPEC.

COPEC es una de las mayores empresas chilenas, propiedad del principal conglomerado empresarial del país, el grupo Angellini, y su rubro principal de actividad es la producción de celulosa para la exportación (en estos momentos se está transformando en el primer productor mundial de celulosa) y, en segundo lugar, la distribución de petróleo dentro de Chile. Tiene subsidiarias para otros servicios, pero de menor importancia. Se considerará el precio quincenal de la acción durante un año.

Para determinar los factores más relevantes en el precio de la acción de COPEC, se partirá de un conjunto de siete variables, todas relevantes para la economía chilena y los productos comercializados por COPEC, y se determinará cuáles de ellas son relevantes para el objetivo de explicar el precio de la acción. Las variables que formarán parte del proceso de selección son:

- x_1 (Tasa) Tasa de interés LIBO a 180 días, que influiría en los gastos financieros de COPEC.
- x_2 (Dolar) El precio del dólar en Chile, que influye tanto en los retornos de las exportaciones de COPEC como en los costos de las importaciones.
- x_3 (Euro) La relación entre el dólar y el euro, dado el papel que ha empezado a jugar el euro como moneda relevante en las transacciones en los mercados internacionales, terminando con el virtual monopolio del dólar.
- x_4 (Cell) El valor de la tonelada de celulosa en el mercado internacional.
- x_5 (Cobre o Cu) El valor de una libra de cobre en el mercado internacional, dada la relevancia del cobre en la marcha de la economía chilena.
- x_6 (Petr) Precio del petróleo en el mercado internacional.
- x_7 (Cell$) Precio de la celulosa en pesos chilenos, es decir, el producto de la variable x_2 por la variable x_4.

Se han seleccionado estas siete variables generales sobre la economía chilena e internacional. No se han agregado variables finas de la empresa, porque el objetivo no es predecir el precio de la acción, sino encontrar los factores de la estructura económica chilena más relevantes para esta empresa.

Se considerarán los parámetros $k_1 = 1$ y $k_2 = 1$, y el valor de cota de ingreso de nuevas variables $s = 0{,}1$. Se mostrará tanto el valor de SIM que propusimos para la regresión posibilística, como los dos criterios para la regresión difusa en general. Si no hay contradicción entre los valores dif e inc, no se recurrirá a arbitraje.

Se realizará la selección mostrando los indicadores de tres de los métodos de regresión propuestos: el método cuadrático posibilístico (MCP), el método no posibilístico (MNP) y el método posibilístico 1 basado en la medida de igualdad de la posibilidad (POS_1). Los datos de referencia serán tomados del método MCP.

Para el primer paso de la selección de variables, los resultados se muestran en la tabla 4.18.

Tabla 4.18: Selección de variables, etapa 1, (ejemplo COPEC)

	MNP	MNP	MNP	MCP	MCP	MCP	POS_1	POS_1	POS_1
Variables	dif	inc	SIM	dif	inc	SIM	dif	inc	SIM
Tasa	3.46	6.91	0.208	3.56	11.48	0.200	3.46	9.26	0.196
Dolar	4.17	8.34	0.145	4.46	15.81	0.180	4.71	13.18	0.177
Euro	2.76	5.51	0.218	2.75	10.09	0.211	2.79	8.05	0.211
Cell	2.44	4.95	0.238	2.48	10.62	0.211	2.47	8.55	0.210
Cobre	3.07	6.27	0.145	3.07	8.86	0.211	3.07	7.40	0.203
Petr	5.29	10.55	0.097	5.29	15.12	0.179	5.29	12.79	0.177
$Cell_pes$	4.72	9.40	0.137	5.00	14.47	0.182	4.95	12.13	0.180

Los resultados de este primer paso no son concluyentes a primera vista. Para el método MNP, no hay dudas de que la variable Cell es la variable que debiera seleccionarse, dado que para los tres criterios considerados, este variable obtiene el mejor valor.

Sin embargo, para el metodo MCP, para dif la mejor variable es Cell, para inc la mejor variable es Cobre, y para SIM hay un empate entre las variables Euro, Cell y Cobre.

Para el método MCP, para resolver la controversia para las tres variables recientemente nombradas, se muestran los cálculos en la tabla 4.19. Por el principio de *mínima resistencia*, menor L, se elige la variable Cell en este primer paso.

La misma discrepancia se produce para el método POS_1, ya que para el criterio

Tabla 4.19: Elección entre Euro, Cell y Cobre; etapa 1; método MCP

Criterio	Euro	Cell	Cobre
dif	2.75	2.48	3.07
inc	10.09	10.62	8.86
$1 - SIM$	0.789	0.789	0.789
d1	0.902	1.000	0.808
d2	0.878	0.834	1.000
d3	1.000	1.000	1.000
L	0.073	0.055	0.064

dif la mejor variable es Cell, para el criterio inc la mejor variable es Cu y para el criterio SIM la mejor variable es Euro.

Esta discrepancia en la elección de la mejor variable se somete al procedimiento de resolución de controversias por *mínima resistencia*, cuyos valores podemos observar en la tabla 4.20. A la vista de estos resultados, también para el método POS_1 la variable elegida en la primera selección debería ser Cell.

Tabla 4.20: Elección entre Euro, Cell y Cobre, etapa 1, POS_1

Criterio	Euro	Cell	Cobre
dif	2.79	2.47	3.07
inc	8.05	8.55	7.40
$1 - SIM$	0.789	0.790	0.797
d1	0.885	1.000	0.805
d2	0.919	0.865	1.000
d3	1.000	0.999	0.990
L	0.065	0.045	0.068

Dado que en esta ocasión se ha producido la situación muy particular de que tres variables tenían la opción de ser elegidas por tener el mejor valor en alguno de los criterios, se va a analizar qué hubiera ocurrido si estuviéramos analizando la selección con el criterio R^2_{difuso}.

Como se aprecia en la tabla 4.21, las tres variables que se han considerado como posibles candidatas a la primera selección, Euro, Cell y Cobre, son las tres que presentan más altos valores de R^2_{difuso}. Sin embargo, sin duda alguna, según este criterio, para ningún método ni parametrización, la mejor variable para seleccionar es la variable Cell. Reafirma esta valoración lo que había ocurrido anteriormente cuando se consideró por primera vez el criterio de R^2_{difuso}: se trata de un criterio certero, preciso, y con resultados similares a los otros criterios considerados.

Pasando al paso 2, los resultados de los criterios de valoración se muestran en la tabla 4.22, donde en todos los casos la regresión puede formarse con $I = \{Cell =$

Tabla 4.21: Selección de variables, etapa 1, R^2_{difuso} (ejemplo COPEC)

	MNP	MNP	MNP	MCP	MCP	MCP	POS_1	POS_1	POS_1
Var.	$k_1 = 0$	$k_s = 1$	$k_2 = 0$	$k_1 = 0$	$k_s = 1$	$k_2 = 0$	$k_1 = 0$	$k_s = 1$	$k_2 = 0$
Tasa	0.475	0.475	0.475	0.449	0.457	0.475	0.470	0.470	0.475
Dolar	0.252	0.252	0.253	0.200	0.215	0.253	0.130	0.169	0.253
Euro	0.671	0.671	0.672	0.670	0.670	0.672	0.668	0.669	0.672
Cell	0.704	0.704	0.704	0.703	0.704	0.704	0.704	0.704	0.704
Cobre	0.659	0.660	0.660	0.660	0.660	0.660	0.659	0.659	0.660
Petr	0.013	0.013	0.013	0.013	0.013	0.013	0.012	0.012	0.013
Cell$	0.093	0.094	0.094	0.060	0.067	0.094	0.068	0.075	0.094

$x_4\}$.

Tabla 4.22: Selección de variables, etapa 2, (ejemplo COPEC)

	MNP	MNP	MNP	MCP	MCP	MCP	POS_1	POS_1	POS_1
Variables	dif	inc	SIM	dif	inc	SIM	dif	inc	SIM
Tasa, Cell	2.43	4.86	0.238	2.57	8.52	0.222	2.42	6.91	0.227
Dolar, Cell	2.41	4.90	0.226	2.43	10.02	0.217	2.43	7.72	0.228
Euro, Cell	2.29	4.62	0.218	2.45	8.01	0.227	2.36	5.76	0.237
Cobre, Cell	1.75	3.56	0.284	2.00	7.73	0.234	1.83	5.38	0.256
Petr, Cell	2.28	4.62	0.238	2.35	10.39	0.213	2.27	8.27	0.224
Cell$, Cell	2.40	4.86	0.235	2.43	9.95	0.218	2.43	7.69	0.229

En este segundo paso, no hay ninguna duda sobre la variable que debe ser seleccionada, porque por los tres criterios, para los tres métodos, señalan la selección como nueva variable al Cobre.

Se destaca que Cobre era una de las tres variables que estuvieron en la controversia para elegir la primera variable, y que resultó en segundo lugar en dicho momento, por lo que su selección en la segunda etapa consolida la convicción que las dos variables elegidas son las que mejor explican el comportamiento de y_i.

Los criterios óptimos de esta etapa, tomando en consideración el método MCP, son $CR^2_1 = 2{,}00$, $CR^2_2 = 7{,}73$ y $CR^2_3 = 0{,}766$. Comparando con los valores CR^1, se tiene

$$t = \frac{\frac{2{,}00}{2{,}48} + \frac{7{,}73}{10{,}62} + \frac{0{,}766}{0{,}789}}{3} = 0{,}835$$

lo que da un valor 1-t de 0.165, que es mayor que el 0.1 puesto inicialmente como cota, por lo que se continua el proceso de selección de variables.

El paso tres corresponde a las combinaciones posibles con las variables Cell y

Cu, cuyos indicadores pueden verse en la tabla 4.23.

Tabla 4.23: Selección de variables, etapa 3, (ejemplo COPEC)

	MNP	MNP	MNP	MCP	MCP	MCP	POS_1	POS_1	POS_1
Variables	dif	inc	SIM	dif	inc	SIM	dif	inc	SIM
Tasa, Cu, Cell	1.73	3.58	0.259	1.98	6.64	0.245	1.90	4.70	0.258
Dolar,Cu, Cell	1.75	3.56	0.295	1.90	7.62	0.235	1.90	5.22	0.248
Euro, Cu, Cell	1.73	3.56	0.251	1.85	6.92	0.241	1.77	4.74	0.257
Pet, Cu, Cell	1.73	3.53	0.300	1.97	7.65	0.235	1.86	5.19	0.249
Cell$, Cu, Cell	1.74	3.56	0.297	1.89	7.60	0.235	1.89	5.18	0.249

En el paso 3, se observa que no hay una variable claramente candidata para ingresar a conjunto I. Para el método MCP, la variable elegida debería salir de una controversia entre Tasa (mejor valor para los criterios inc y SIM) y Euro (mejor valor para dis); mientras, para POS_1 la variable candidata sería Euro y para el método MNP la variable candidata es Petr.

Si se desarrolla el cálculo de la controversia para el método MCP, L para Tasa es 0.022 mientas que L para Euro es 0.015, por lo que la variable candidata para ingresar es Euro, por el principio de *mínima resistencia*.

El cálculo del valor t, para esta nueva ecuación es

$$t = \frac{\frac{1{,}85}{2{,}00} + \frac{6{,}92}{7{,}73} + \frac{0{,}759}{0{,}766}}{3} = 0{,}937$$

por lo que 1-t es menor que 0.1 y se para el proceso de selección de variables.

Se realiza el proceso de reconsideración de variables, donde para la variable elegida Euro, se calcula su correlación lineal con las variables Cell (-0.77) y Cu (- 0.67). A la vista de estos valores, se excluye a la variable de mayor correlación, Cell, en el proceso de reconsideración, lo que se muestra en la tabla 4.24, que confirma la presencia de la variable Cell en el conjunto de variables seleccionadas.

Tabla 4.24: Selección de variables, etapa 3.1, (ejemplo COPEC)

	MNP	MNP	MNP	MCP	MCP	MCP	POS_1	POS_1	POS_1
Variables	dif	inc	SIM	dif	inc	SIM	dif	inc	SIM
Euro, Cobre	2.22	4.61	0.197	2.34	8.30	0.226	2.28	6.36	0.233

Hasta aquí llega el proceso de selección de variables, incorporando sólo a Celulosa y Cobre como las variables seleccionadas en la ecuación de regresión y dejando el resto de las variables como poco relevantes en el modelo explicativo del precio de la acción de la empresa COPEC.

4.6. Conclusiones sobre Selección de Variables

Como parte de esta memoria, hemos presentando diversos métodos alternativos de regresión difusa, y se ha dejado la opción para que el tomador de decisiones pueda elegir alguna de dichas alternativas, según sus preferencias.

En el caso de la metodología de selección automática de variables para la regresión difusa, ha resultado el mismo esquema: dentro de los diversos métodos de regresión difusa, el tomador de decisiones tiene la opción de elegir el criterio por el cual quiere efectuar esta selección.

Y este criterio puede ser único, o puede ser múltiple. Para un solo criterio, se recomienda la alternativa de los índices SIM y R^2_{difuso}. Y para múltiples criterios, se pueden elegir criterios basados en el ajuste del valor central (como serían los criterios R^2_{difuso}, dif u otra función de pérdida L) o criterios basados en las funciones de pertenencia de los números comparados, principalmente en las diferencias de sus extensiones (como son los criterios SIM, inc o $(\widehat{q} + \widehat{p})$).

Con ello se configura un procedimiento automático de selección de variables, con las siguientes opciones de decisión:

- La decisión de elegir el o los métodos de regresión difusa que se emplearán (MCP, POS_1, MNP, NES_1, LIN). Se recomienda como primera selección la regresión posibilística MCP.
- La decisión de los parámetros de ponderación para el método de regresión difusa elegido (k_1, k_2). Se recomienda como opción por defecto $k_1 = 1$ y $k_2 = 1$.
- La decisión del criterio o los criterios, para la selección de las variables. Se recomienda, para un criterio, la opción por defecto SIM y, para dos criterios, SIM y R^2_{difuso}.
- El parámetro de control para determinar la cota a partir de la cual no se añaden nuevas variables al conjunto de variables de entrada (valor s). Se recomienda s=0.05 cuando se elige un único criterio de selección, y s=0.10 cuando se eligen varios criterios.

Las mejoras que presenta esta metodología de selección de variables con respecto a la conocida [WT00a], se pueden resumir en:

- Se puede elegir entre muchos criterios de selección.
- Se puede operar con un único criterio de selección.
- Se incorpora la cualidad de reconsiderar una variable seleccionada en un paso previo, con la idea de disponer, en todo momento, del conjunto de variables más calificado.
- Se proponen criterios de selección de variables creados en este trabajo de investigación, como el R^2_{difuso} y el indicador de similitud SIM. Se dispone de una sola medida de resolución de controversias, de manera que se hace completamente automático este procedimiento.
- También se propone un valor de cota de inclusión más realista.

Capítulo 5

Extensiones y Aplicaciones

Hemos desarrollado hasta aquí una metodología para aplicar regresión difusa flexible y adaptable al problema que se pretende resolver. Vamos a dedicar ahora este capítulo a estudiar dos problemas abiertos dentro del campo de la regresión: la multicolinealidad y la regresión ecológica.

En esta línea, en primer lugar, se presenta una nueva metodología para aplicar nuestra propuesta de regresión difusa en situaciones con datos de entrada multicolineales. Luego se aplica esta metodología a dos ejemplos.

En segundo lugar, se presenta un caso de aplicación de regresión difusa a estudios sociológicos y demográficos, en el contexto de la llamada *regresión ecológica*. Este problema de regresión, que siendo, en apariencia, relativamente simple, puesto que se ha aplicado a tablas de datos 2x2, no ha tenido hasta ahora una solución teórica satisfactoria. Como veremos, para este tipo de aplicaciones, se desarrolla una propuesta consistente en un sistema de ecuaciones de regresión difusa, que se estiman simultáneamente.

5.1. Regresión Difusa con Datos Multicolineales

El primer ejemplo de regresión difusa en el trabajo germinal de Tanaka[TUA82] es un caso que presenta una alta multicolinealidad en los datos de entrada.

Entendemos por *multicolinealidad* la existencia de una alta correlación entre las variables de entrada (columnas de la matriz X). Esta correlación queda reflejada en

la diferencia entre los valores propios de la matriz producto de los datos de entrada $X'X$; en caso de multicolinealidad, esta matriz presenta una significativa diferencia entre el valor propio más alto y el valor propio más bajo.

En el ejemplo citado de Tanaka, el valor propio más pequeño es 0.4 mientras el más grande es 114561, es decir, más de 250.000 veces mayor que el más pequeño, lo que indica, sin lugar a dudas, la presencia de multicolinealidad.

La aplicación de técnicas de *selección de variables* como las descritas en el Capítulo 4, es una alternativa para eliminar variables que no tienen una aportación significativa en el modelo y que pueden producir multicolinealidad. Sin embargo, muchas veces, el conjunto de variables de entrada constituye una unidad que no es conveniente reducir. Por ejemplo, las variables de entrada X pueden representar diversas cualidades de población, representando el conjunto X al universo poblacional. O las variables X pueden representar una desagregación de categorías de un atributo determinado, por ejemplo nivel educacional, en que el efecto simultáneo de todas esas variables permite estimar características por nivel educacional. En estos casos parece más adecuado tratar de eliminar el efecto distorsionador de la multicolinealidad por otros medios diferentes que no impliquen la eliminación de variables.

En el contexto de la Regresión Probabilística, se han propuesto diversos métodos para afrontar la multicolinealidad. Los principales de ellos constituyen estimadores sesgados de los coeficientes. Entre estos métodos, se encuentran la *Regresión Ridge*, la *Regresión de Componentes Principales*, la *Regresión Lasso* y la *Regresión de Mínimos Cuadrados Parciales* [HTF01]. Sin embargo, la Regresión Ridge es la que presenta una mejor regularidad en el proceso de estimación, en el ajuste de los estimadores, lo que hace más atractivo su uso.

Existe en la literatura una propuesta de Regresión Difusa Ridge de Hong et al. [HH04, HHA04], basada en un enfoque dual a través de máquinas vectoriales [Vap98], en la cual se estima un coeficiente por observación, y no un coeficiente por variable de entrada. Esto tiene el inconveniente de la alta cantidad de coeficientes y de no poder interpretar dichos coeficientes.

Proponemos ahora, en el ámbito del enfoque cuadrático de la Regresión Difusa, métodos de Regresión Ridge Difusa, que permiten estimar un modelo lineal aminorando los problemas ocasionados por la multicolinealidad en la matriz X. También se una normalización de los datos para la estimación Ridge difusa, dado que los resultados dependen de la unidad de medida de las variables. Para evaluar los métodos propuestos, se realiza una experimentación para diversas características de datos.

5.1.1. Nuevo Esquema de Regresión Difusa Ridge

La idea intuitiva de la Regresión Ridge[1] consiste en disminuir el tamaño de los estimadores, porque los estimadores con alta multicolinealidad tienen una alta varianza, cuya consecuencia son estimadores con valores absolutos muy altos.

Es decir, si los coeficientes verdaderos de una regresión, A_i, de centro a_i, tienen un tamaño determinado, que podemos calcular en la suma de sus valores absolutos como

$$L = \sum_{j=1}^{m} |a_i| \tag{5.1}$$

entonces la expresión equivalente resultante de la estimación de regresión es

$$L_{est} = \sum_{j=1}^{m} |\widehat{a_i}| \tag{5.2}$$

con el resultante, en presencia de multicolinealidad, de que

$$L_{est} > L \tag{5.3}$$

En la literatura de la Regresión Difusa se ha propuesto un modelo de Regresión Ridge basado en el concepto de *vector machine* [HH04, HHA04], y que consiste en estimar el modelo dual de regresión. Es decir, las variables ocupan el lugar de los datos y los datos el lugar de las variables. Sin embargo, como generalmente el número de datos es mayor que el número de variables, esto lleva a operar con una cantidad de variables muy grande, y los coeficientes estimados no pueden interpretarse como el aporte de cada variable, quedándose sin una interpretación semántica clara.

En la Regresión Probabilística, donde la solución de mínimos cuadrados se expresa, en términos matriciales, como:

[1] El término Ridge proviene del inglés, y significa literalmente *cresta*, lo cual no tiene una analogía concreta con el método mismo. Esta palabra, Ridge, ha sido ampliamente utilizada en la literatura sobre multicolinealidad y aprendizaje estadístico. No hay una palabra alternativa que se esté empleando en castellano, aunque en una oportunidad, una profesora francesa la tradujo como *regresión acotada*, nombre que refleja el sentido de acotar o achicar el tamaño de los estimadores. Sin embargo, en la actualidad existen varios métodos alternativos que tienen por objeto contraer el valor absoluto de los estimadores, por lo que preferimos mantener su denominación inglesa

$$\widehat{\beta_{MC}} = (X'X)^{-1}X'Y \tag{5.4}$$

la solución Ridge [HK70], por la presencia de multicolinealidad, *corrige* la estimación 5.4 con un factor, λ, que se suma a la diagonal de la matriz $X'X$:

$$\hat{\beta}_{\lambda} = (X'X + \lambda I)^{-1}X'Y \tag{5.5}$$

Esta idea de penalización también es aplicada en las redes neuronales, donde es conocida como *weight decay*.

Esta incorporación del término λ al cálculo de los estimadores tiene como consecuencia que la matriz $(X'X + \lambda I)$, con $\lambda > 0$, tiene valores propios menos extremos que los de la matriz (X'X), con lo que la varianza de los estimadores $\widehat{\beta}_{\lambda}$ es menor que la varianza de los estimadores $\widehat{\beta}_{MC}$.

Esto se traduce en que el error debido a la varianza de los estimadores disminuye. Cuando λ empieza a aumentar cerca de cero, el error de la varianza disminuye considerablemente, para luego ir disminuyendo en forma más suave. Sin embargo, al mismo tiempo, dado que esta estimación Ridge es sesgada, al aumentar λ, a partir de 0, va aumentando el error debido al sesgo de la estimación.

El objetivo para determinar el valor λ óptimo es encontrar el valor en que la suma de ambos errores, es decir, la suma de los errores producidos por la alta varianza de los estimadores y el error por el sesgo de los estimadores, sea mínimo. Esto no es posible de determinar explícitamente debido a que depende de los valores teóricos de los coeficientes A_i y de la varianza de los términos de error, que son desconocidos.

La formulación probabilística de los estimadores Ridge, reflejada en la solución 5.5, no es independiente de la escala de medición de las variables de entrada, por lo que, en general, se aplica esta estimación normalizando previamente la matriz X, como una matriz con media 0 y desviación estándar 1.

En nuestro caso, en base a nuestro modelo de Regresión Difusa con optimización cuadrática, donde también aparece el elemento $(X'X)$, proponemos utilizar una lógica similar a la de la regresión Ridge probabilística. Nuestro modelo de regresión Ridge difuso consiste en añadir a la función objetivo propuesta para el método MCP, el término cuadrático, en el nivel de coeficiente central, siguiente

$$\lambda(a'a) = \lambda \sum_{j=1}^{j=m} a_j^2 \tag{5.6}$$

donde se hace variar λ desde 0, aumentándolo con un cierto incremento, hasta que los coeficientes A_i se muestren estables; específicamente hasta que el valor 5.2 sea prácticamente constante.

Dada la forma cuadrática del término definido por la ecuación (5.6), esta propuesta de regresión Ridge difusa resulta armoniosa con los otros modelos de regresión difusa propuestos en esta memoria. Y esta forma cuadrática (en el coeficiente a_j) no es arbitraria, sino que es cosustancial a su formulación inicial [HK70], dado que la función objetivo de mínimos cuadrados también es cuadrática en a_i. De esta forma, se logra relacionar el término λ de la regresión Ridge con la matriz $(X'X)$ de los datos de entrada, que tiene incorporado el efecto de multicolinealidad, resultando la estimación (5.5).

El incremento de λ va a depender de la escala de medición de X. Para X normalizados, λ varía, generalmente, de 0 a 1, con un incremento de 0.1, aunque para algunos problemas puede tomar valores mayores a 1. El gráfico de los valores $\hat{\beta}_\lambda$ al variar los valores de λ se denomina traza Ridge, y permite determinar gráficamente el valor de λ cuando los estimadores se estabilizan, es decir, cuando la magnitud (5.2) tiende a permanecer constante.

Para abordar de forma integral la multicolinealidad en los modelos difusos de regresión, se propone ampliar el objetivo inicial representado en el término (5.6) –que solamente apunta al efecto de la tendencia central- por un objetivo que incluya, además, a las extensiones de los coeficientes. Para que esta función objetivo sea generalizada, en el sentido de poder discriminar entre el efecto de la tendencia central y el efecto de las extensiones, se introducirá una ponderación para la tendencia central y otra ponderación para las extensiones. De esta manera, el efecto anti-multicolinealidad en la función objetivo queda reflejado en el término

$$\lambda \sum_{j=1}^{j=m} k_3 a_j^2 + k_4((a_j - c_{L_j})^2 + (a_j + c_{R_j}^2) \tag{5.7}$$

donde λ irá aumentando a partir de 0, y k_3 y k_4 tomarán valores, dados por el analista, entre 0 y 1.

Al igual que en el caso de los modelos del enfoque cuadrático de regresión difusa, en los que una función objetivo puede ir acompañada de varios conjuntos de restricciones alternativos, los modelos de regresión difusa Ridge se completarán con uno de los siguientes conjuntos de restricciones:

1 Restricciones posibilísticas, que dependen del nivel de confianza h definido por el análisis de datos, que para funciones de pertenencia no simétricas triangulares, son las siguientes:

$$\sum_{j=1}^{m} a_j X_{ij} - (1-h)\sum_{j=1}^{m} c_{L_j} X_{ij} \leq y_i - (1-h)p_i \;\; para \;\; i = 1, ..., n \quad (5.8)$$

$$\sum_{j=1}^{m} a_j X_{ij} + (1-h)\sum_{j=1}^{m} c_{R_j} X_{ij} \geq y_i + (1-h)q_i \;\; para\, i = 1, ..., n \quad (5.9)$$

2 Restricciones sobre la medida de igualdad en la posibilidad, que también dependen del nivel de confianza h, y que para funciones de pertenencia no simétricas triangulares son las siguientes:

$$\widehat{y_i} - (1-h)\widehat{p_i} \leq y_i + (1-h)q_i \;\; para\, i = 1, ..., n \quad (5.10)$$

$$\widehat{y_i} + (1-h)\widehat{q_i} \geq y_i - (1-h)p_i \;\; para\, i = 1, ..., n \quad (5.11)$$

$$c_{L_j}, c_{R_j} \geq 0 \;\; j = 1, ..., m \quad (5.12)$$

3 Restricciones de simple positividad, que exige que las extensiones de los coeficientes A_i sean mayor o igual a cero:

$$c_{L_j}, c_{R_j} \geq 0 \;\; j = 1, ..., m \quad (5.13)$$

A partir de estas especificaciones del concepto de *restricciones de tipo difuso*, se definirán tres modelos de regresión difusa Ridge, que se presentan del caso más particular al modelo más general.

El primer modelo de regresión Ridge queda definido como el problema de optimización siguiente:

Modelo 5.1. *Modelo de Regresión Ridge Difuso - RRD*

Función objetivo:

$$k_1 \sum_{i=1}^{n}(y_i - a^{'} x_i)^2 + k_2(\sum_{i=1}^{n}(y_i - p_i - (a^{'} - c_L^{'})x_i)^2 +$$

$$\sum_{i=1}^{n}(y_i + q_i - (a^{'} + c_R^{'})x_i)^2) + \lambda(a'a) \tag{5.14}$$

Restricciones: Un conjunto de restricciones de tipo difuso

Para un valor λ específico, la estimación de este modelo la se denominará RRD_λ.

Este modelo de regresión difusa Ridge asume que la multicolinealidad se encuentra presente sólo en la tendencia central de los datos.

Para ampliar la influencia de la multicolinealidad a las extensiones de los datos, $a_j - c_{L_j}$ y $a_j + c_{R_j}$ se presenta un segundo modelo.

Modelo 5.2. *(Modelo de Regresión Ridge Ampliado - RRA)*

Función objetivo:

$$k_1 \sum_{i=1}^{n}(y_i - a^{'} x_i)^2 + k_2(\sum_{i=1}^{n}(y_i - p_i - (a^{'} - c_L^{'})x_i)^2 +$$

$$\sum_{i=1}^{n}(y_i + q_i - (a^{'} + c_R^{'})x_i)^2) + \lambda \sum_{j=1}^{j=m}(k_3 a_j^2 + k_4((a_j - c_{L_j})^2 + (a_j + c_{R_j})^2)) \tag{5.15}$$

Restricciones:

Un conjunto de restricciones de tipo difuso.

Una estimación con el modelo ampliado la denominaremos $RRA_\lambda(k_3, k_4)$, dependiendo del parámetro Ridge λ y los parámetros de multicolinealidad central k_3 y de multicolinealidad de las extensiones k_4.

Luego, el modelo RRD es un caso particular de RRA, que cumple

$$RRD_\lambda = RRA_\lambda(1, 0) \tag{5.16}$$

Un enfoque más general aún, consiste en transformar la constante λ, en un parámetro diferenciado para cada coeficiente, $\lambda = (\lambda_1, ..., \lambda_m)$, de manera que el parámetro

Ridge que se agrega a la función objetivo sea variable, dependiente, por ejemplo, de cada variable de entrada. Teniendo esto en cuenta, se define el modelo siguiente:

Modelo 5.3. *(Modelo de Regresión Ridge Generalizado - RRG)*

Función objetivo:

$$k_1 \sum_{i=1}^{n}(y_i - a^{'} x_i)^2 + k_2(\sum_{i=1}^{n}(y_i - p_i - (a^{'} - c_L^{'})x_i)^2 +$$

$$\sum_{i=1}^{n}(y_i + q_i - (a^{'} + c_R^{'})x_i)^2) + \sum_{j=1}^{j=m} \lambda_j (k_3 a_j^2 + k_4((a_j - c_{L_j})^2 + (a_j + c_{R_j})^2)) \quad (5.17)$$

Restricciones:

Un conjunto de restriccones de tipo difuso.

Una estimación con este modelo será identificada como $RRG_{\lambda_j}(k_3, k_4)$.

El término que refleja la corrección Ridge a la estimación puede ser representado en la función objetivo por la suma de dos matrices diagonales $[\lambda_j k_3 a_j^2]_{jj}$ y $[\lambda_j k_4 (c_{L_j}^2 + c_{R_j}^2)]_{jj}$, lo que permite visualizar una generalización, en que estos términos diagonales se transforman en una matriz cuadrada mxm, postulando un término de ajuste de la multicolinealidad de la forma aDa' y cEc', en el que el vector *a* representa los coeficientes centrales y el vector *c* los coeficientes de las extensiones, y *D* y *E* son matrices no necesariamente diagonales.

Dada la orientación eminentemente práctica de nuestro trabajo de investigación, no se entrará a estudiar las formas que podrían tomar las matrices *D* y *E*, lo que se dejará para investigaciones futuras.

5.1.2. Normalización de los Datos para la Regresión Difusa Ridge

Una dificultad que presenta el tratamiento que le hemos dado al objetivo de eliminar el efecto de la multicolinealidad en las estimaciones de regresión difusa es que, al igual que ocurre en el caso Ridge probabilístico, el modelo resultante no es invariante a cambios de escala de las variables de entrada X. Es decir, cada escala de medición que se use para las variables X_i producirá estimaciones distintas de los coeficientes difusos Ridge; dicho cambio también afecta las características de la matriz X'X que interviene en la función objetivo cuadrática de los tres modelos propuestos.

Además, si no se normalizara el sistema, la escala de variación del parámetro λ sería especial para cada caso, ya que su efecto está relacionado con las magnitudes de la citada matriz X'X. Lo ideal es tener un parámetro λ que variara entre 0 y 1, con un incremento de 0.1 o 0.05, lo que permite un procedimiento más estandarizado de la estimación de los coeficientes. Por tanto, es conveniente encontrar un sistema de referencia que permita estimar un $\widehat{A}_\lambda$ que sea lo más estandarizado posible. Una opción sería considerar la tradicional estandarización de cada x_i con media 0 y varianza 1.

Una consecuencia de elegir un X estandarizado, que se notará como X^N, es que la estimación de los modelos

$$Y = \sum_{i=0}^{k} A_i X_i^N \quad con\ X_0 = 1 \tag{5.18}$$

y

$$Y = \sum_{i=1}^{k} A_i X_i^N \quad es\ decir\ X_0 = 0 \tag{5.19}$$

producen la misma estimación de los número difusos A_i, dado que el término que influye en la estimación, aparte de X'X, es el producto $X^{N\prime}Y$. Este producto es equivalente a $X^{N\prime}(Y - \overline{Y})$, puesto que, como $\sum X_i^N = 0$, se cumple

$$X^{N\prime}(Y - \overline{Y}) = X^{N\prime}Y - \overline{Y}\sum X_i^N = X^{N\prime}Y \tag{5.20}$$

ya que al elegir X^N se tiene que $\widehat{A_0} = \overline{Y}$.

Sin embargo, la normalización tradicional no es aplicable directamente a nuestro modelo, porque produce valores x_{ij} menores que 0 y esto es incompatible con la condición impuesta de que todos los datos de entrada son no negativos, limitación impuesta para facilitar las operaciones de la aritmética difusa.

Para determinar un deflactor que sirva para normalización de las variables de entrada y de la variable de salida, se podría utilizar sólo la desviación estándar de cada x_i y de la variable de salida, con lo que se mantiene cierta similitud con la regresión Ridge probabilística normalizada. Pero con esta normalización resultarían muchos datos normalizados mayores que uno, por lo que los valores posibles de λ^N podrían llegar a alcanzar cifras bastante más altas que 1.

Por esta razón, se propone como criterio de normalización de las variables de entrada y de la variable de salida, que tengan un valor máximo de 1. También se podría considerar el criterio que el valor máximo sea $\frac{1}{n}$.

Con este principio, los nuevos datos normalizados, para realizar las estimaciones difusas Ridge, son

$$y_i^N = y_i/\sigma_Y \quad para\ i = 1, ..., n \tag{5.21}$$

y

$$x_{ij}^N = x_{ij}/\sigma_{X_j} \quad para\ i = 1, ..., n\ y\ j = 1, ..., m \tag{5.22}$$

donde

$$\sigma_Y = max_i\{y_i\} \tag{5.23}$$

es el deflactor propuesto para normalizar la variable de salida Y, y

$$\sigma_{X_j} = max_i\{x_{ij}\}\ para\ j = 1, ..., m \tag{5.24}$$

es el deflactor para normalizar la j-ésima variable de entrada.

Las cifras con extensiones $y_i + (1-h))q_i$ e $Y_i - (1-h)p_i$, correspondientes a α-cortes de Y, también serán normalizadas con el factor σ_Y.

Para establecer una relación entre el modelo Ridge difuso RRD con los datos originales y el modelo resultante con los datos normalizados, se tiene el siguiente teorema:

Teorema 5.1. *Relación de los modelos RRD y RRG y la estimación normalizada.*

Si se estima el modelo RRD_λ con los datos X e Y normalizados, X^N y Y^N, la solución equivalente con los datos X e Y originales esta dada por el modelo RRG(1,0) con las siguientes equialencias entre los dos modelos:

$$A_j = \frac{\sigma_Y}{\sigma_{X_j}} A_j^N\ para\ j = 1, ..., m \tag{5.25}$$

y

$$\lambda_j = \frac{\lambda^N}{\sigma^2_{x_j}} \; para \; j = 1, ..., m \tag{5.26}$$

Para demostrar este teorema, analizaremos primero las restricciones 5.8, que para los datos normalizados tienen la forma:

$$\sum_{j=1}^{m} a_j^N x_{ij}^N - (1-h) \sum_{j=1}^{m} c_{Lj}^N x_{ij}^N \leq (y_i - (1-h)p_i)^N \;\; para \;\; i = 1, ..., n \tag{5.27}$$

que es equivalente, reemplazando los valores X^N e Y^N, a

$$\sum_{j=1}^{m} a_j^N (x_{ij}/\sigma_{X_j}) - (1-h) \sum_{j=1}^{m} c_{Lj}^N (x_{ij}/\sigma_{X_j}) \leq (y_i - (1-h)p_i)/\sigma_Y \;\; para \; i = 1, ..., n \tag{5.28}$$

y reagrupando términos

$$\sum_{j=1}^{m} a_j^N (\sigma_Y/\sigma_{X_j}) x_{ij} - (1-h) \sum_{j=1}^{m} c_{Lj}^N (\sigma_Y/\sigma_{X_j}) x_{ij} \leq (y_i - (1-h)p_i) \;\; para \; i = 1, ..., n \tag{5.29}$$

Luego, comparando la ecuación 5.29 con la restricción 5.8 del modelo original, resulta que los coeficientes estimados por el modelo normalizados, para que sean equivalentes al modelo inicial, se necesita hacer la equivalencia

$$A_j = \frac{A_j^N \sigma_Y}{\sigma_{Xj}} \;\; para \; j = 1, ..., m \tag{5.30}$$

que es lo indicado en el Teorema. El resto de las restricciones, 5.9, tienen el mismo comportamiento.

Respecto a la función objetivo, a partir de la definición 5.17, se tiene que los datos normalizados tendrían la función objetivo siguiente, que se muestra separada en sus tres términos:

$$F =_1^N = \sum_{i=1}^{n} (y_i/\sigma_Y - \sum_{j=1}^{j=m} a_j^{N'} X_j/\sigma_{X_j})^2 \tag{5.31}$$

mientras el segundo termino, FO_2^N, está dado por

$$\sum_{i=1}^{n}((y_i - p_i)/\sigma_Y - (a^N - c_L^N)X_j/\sigma_{X_j})^2 + ((y_i + q_i)/\sigma_Y - (a^N + c_R^N)X_j/\sigma_{X_j})^2 \tag{5.32}$$

y el tercer término por

$$\lambda(a^{N'}a^N) \tag{5.33}$$

El primer término, FO_1^N, reagrupando de acuerdo con los datos originales X e Y, se expresa como

$$FO_1^N = (1/\sigma_Y^2)\sum_{i=1}^{n}(y_i/ - a^N(\sigma_Y/\sigma_{X_j})X_j/)^2 = (\frac{1}{\sigma_Y^2})FO_1 \tag{5.34}$$

donde se vuelve a encontrar la equivalencia entre el modelo original y el modelo normalizado, con el agregado de una constante $(\frac{1}{\sigma_Y^2})$.

Para el segundo término, se tiene

$$(1/\sigma_Y^2)\sum_{i=1}^{n}((y_i - p_i) - (a^N - c_L^N)\sigma_Y/\sigma_{X_j}X_j)^2 + ((y_i + q_i) - (a^N + c_R^N)\sigma_Y/\sigma_{X_j}X_j)^2 \tag{5.35}$$

que se puede reducir a

$$FO_2^N = (\frac{1}{\sigma_Y^2})FO_2 \tag{5.36}$$

donde para los estimadores $A = (a, c_L, c_R)$ se vuelve a tener la misma relación que con los estimadores A^N, considerando el término constante adicional $(\frac{1}{\sigma_Y^2})$.

$$FO_1^N = (1/\sigma_Y^2)\sum_{i=1}^{n}(y_i/ - a^N(\sigma_Y/\sigma_{X_j})X_j/)^2 = (\frac{1}{\sigma_Y^2})FO_1 \tag{5.37}$$

.

Para el tercer término de la función objetivo

$$FO_3^N = \lambda(a^{N'}a^N) = \lambda \sum_{j=1}^{j=m} a_j^{N^2} \tag{5.38}$$

puede reescribirse, usando la equivalencia 5.30 para reemplazar el coeficiente normalizado A^N en función del coeficiente no estandarizado A, como

$$\lambda^N \sum_{j=1}^{j=m} (a_j \sigma_{x_j}/\sigma_Y)(a_j \sigma_{x_j}/\sigma_{Y_j})^2 = (\frac{1}{\sigma_Y^2}) \sum_{j=1}^{j=m} F_j a_j^2 \tag{5.39}$$

con el factor $\sigma_{x_j}^2$, que apareció en la ecuación anterior, reflejado en F_j

$$F_j = \frac{\lambda^N}{\sigma_{x_j}^2} \tag{5.40}$$

donde la presencia de los términos F_j hacen a esta expresión diferente al modelo RRD con los datos normalizados, y lo transforman en un modelo RRG, donde λ es reemplazado por los F_j.

Este teorema permite conocer la relación entre la estimación normalizada y la estimación con los datos originales. Además, demuestra que el modelo RRG no es sólo una construcción teórica, sino que puede contribuir a encontrar soluciones más adecuadas.

Corolario: La estimación del modelo $RRA_\lambda(k_3, k_4)$ con los datos normalizados, representada como $A_j^N(\lambda)$, es equivalente a la estimación del modelo $RRG_{\lambda_j}(k_3, k_4)$ con los datos no normalizados, presentada como $A_j(\lambda)$, con las relaciones

$$A_j = \frac{\sigma_Y}{\sigma_{X_j}} A_j^N \; para \; j = 1, ..., m \; y \; k_3, k_4 \; dados \tag{5.41}$$

y

$$\lambda_j = \frac{\lambda^N}{\sigma_{x_j}^2} \; para \; j = 1, ..., m \; y \; k_3, k_4 \; dados \tag{5.42}$$

en que λ^N variará, usualmente, entre 0 y 1.

La siguiente definición tiene por objetivo precisar una estimación Ridge difusa que se entenderá equivalente a la invarianza de la escala de medición de los datos, que se calcula a partir de los datos en su forma normalizada.

Definición 5.1. *Estimación Ridge Difusa*

Se define la estimación de la regresión ridge difusa X en Y, $A_j(\lambda)(k_3, k_4)$, a partir del modelo $RRA(k_3, k_4)$, como la estimación calculada con los datos normalizados, con λ = 0, 0.1, 0.2, ..., 1, ... y que entrega los estimadores dados por la relación

$$A_j(\lambda)(k_3, k_4) = \frac{\sigma_Y}{\sigma_{X_j}} A_j^N(\lambda)(k_3, k_4) \; para \; j = 1, ..., m \; y \; k_3, k_4 \; dados \qquad (5.43)$$

Para elegir el valor de λ, en forma visual, se utiliza el gráfico denominado traza ridge, y en forma práctica se han sugerido varios métodos, siendo uno de los más usados el de *validación cruzada*[HTF01].

No es recomendable utilizar los indicadores de bondad de ajuste que se han definido en esta memoria para seleccionar el mejor valor de λ, debido a que dichos indicadores sólo miden el efecto conjunto de los estimadores $\widehat{A}_i(\lambda)$, y el problema de la multicolinealidad se refleja en la magnitud individual de dichos valores. Al *corregir* con λ las estimaciones, es posible interpretar adecuadamente el significado de los coeficientes.

Idealmente, si hay suficientes datos, la forma de validar un modelo es dejando un conjunto de datos de prueba, separados de los datos de la estimación. Dado que muchas veces sólo se dispone de un conjunto limitado de datos, el método de validación cruzada en k grupos, es usado para elegir el valor λ más adecuado.

Este procedimiento parte dividiendo el conjunto de datos en K grupos de igual tamaño. Para la K-ésima parte, se ajusta el modelo con los otros k-1 grupos de datos y se calcula el error de predicción del modelo estimado con la k-ésima parte de los datos.

Formalmente, sea $k = \{1, ..., n\} \rightarrow \{1, K\}$ la función de índices que indica la partición a la cual la observación i es asignada aleatoriamente. Se define como $f^k(x)$ la función estimada, calculada sin la parte k de los datos. Entonces, la estimación de predicción de error de la validación cruzada está dada por

$$CV = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f^{k(i)}(x_i)) \qquad (5.44)$$

donde L es la función de medición de pérdida del ajuste de regresión, que será, por lo general, el valor absoluto de la diferencia, o el cuadrado de la diferencia.

Cuando los datos son pocos, como en el caso del ejemplo presentado por Tanaka, que tiene 15 observaciones, se aplica este método con la cantidad de grupos K igual a n.

Cuando se tiene un modelo paramétrico, como es el caso de λ en la regresión Ridge, el indicador CV dependerá de dicho parámetro y se elige un λ que minimice la función de error $CV(\lambda)$.

5.1.3. La Práctica de la Regresión Ridge Difusa

El ejemplo presentado por Tanaka en su trabajo inicial [TUA82], referido a la estimación del precio de una vivienda, es un caso que presenta una clara multicolinealidad entre los datos de entrada. Vamos a analizar con detalle este ejemplo.

De igual manera se harán simulaciones, con valores conocidos de los parámetros, para evaluar los métodos difusos Ridge que hemos propuesto.

5.1.3.1. El ejemplo de Tanaka

En el ejemplo mencionado de Tanaka hay una correlación lineal de 0.91 entre dos de las cinco variables de entrada, lo que es otro indicador del alto grado de multicolinealidad en los datos de entrada.

Vamos a mostrar los resultados de las estimaciones, a través de la traza Ridge, variando λ entre 0 y 1. Todas las estimaciones están realizadas con el nivel de confianza h igual a cero y los parámetros $k_1 = 1$ y $k_2 = 1$.

En la figura 5.1 se pueden observar las estimaciones del coeficiente central para el modelo RRA(1, 1) tomando como restricciones el conjunto de simple positividad.

En el gráfico, se observa claramente que tres coeficientes, que con el método de mínimos cuadrados ($\lambda = 0$) tienen valores centrales negativos, pasan a tener valores positivos significativos, aunque la magnitud total de los coeficientes disminuye poco.

En cambio, en la misma estimación, el comportamiento de las extensiones c_L tiene pocas variaciones, como se aprecia en la figura 5.2. Los coeficientes de la extensión derecha c_R presentan un comportamiento similar a c_L.

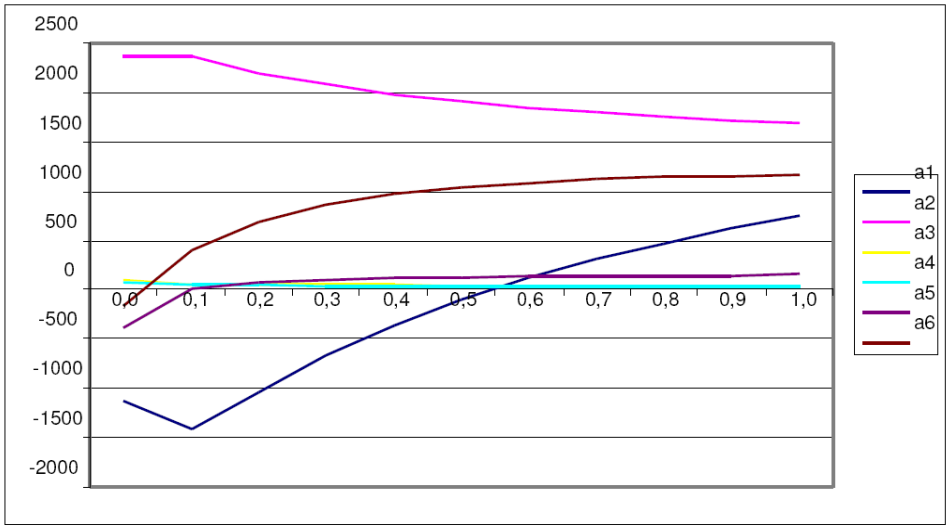

Figura 5.1: Modelo RRA(1,1) normalizado, coef. a_j no posibilísticos (Ej. Tanaka)

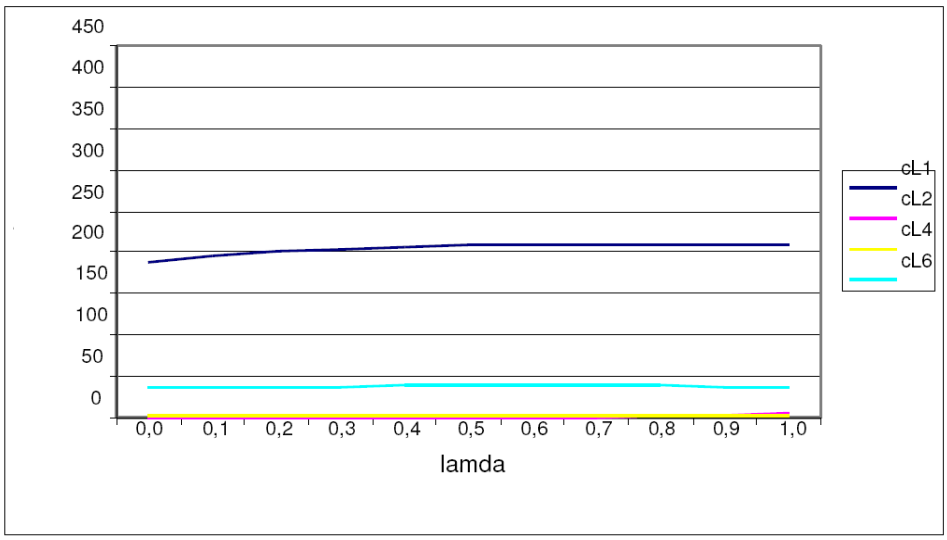

Figura 5.2: Modelo RRA(1,1) normalizado, coef. c_L no posibilísticos (Ej. Tanaka)

Los resultados para el mismo modelo, pero con restricciones posibilísticas (modelo MCP), se aprecian en el gráfico de la figura 5.3. La tendencia del comportamiento de los coeficientes es la misma, pero la magnitud del cambio es menor, por lo que hay un coeficiente a_6 que permanece con valor negativo.

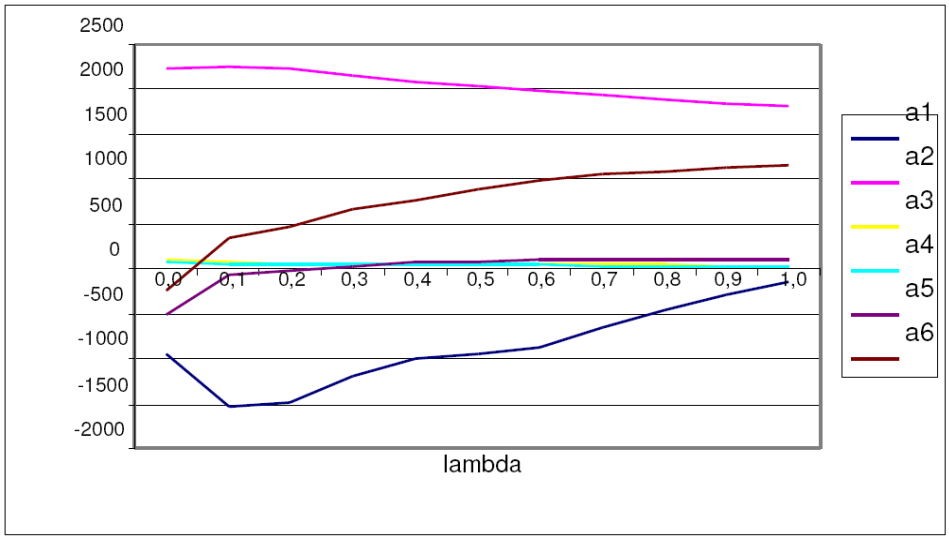

Figura 5.3: Modelo RRA(1,1) normalizado, coef. a_j posibilísticos (Ej. Tanaka)

Para este modelo posibilístico, en el caso de las extensiones, los coeficientes tienen mayor variabilidad, aumentando levemente c_L cuando aumenta λ, y disminuyendo significativamente c_R en el mismo caso.

5.1.3.2. Un Caso de Simulación

Para terminar nuestro estudio de sobre la regresión Ridge difusa, vamos a trabajar ahora sobre un conjunto de datos con 25 observaciones y cuatro variables de entrada más la constante.

La matriz de correlaciones entre las variables de entrada y salida se muestran en la tabla 5.1. Como se observa son muy altas, tanto entre las variables de entrada, como entre ellas y el valor central de la variable de salida.

Los valores propios de la matriz $X'X$ varían entre 0.16 a 925, por lo que la razón es de 1 a 5780 entre los valores extremos.

Tabla 5.1: Correlaciones lineales, ejemplo simulado

Variable	y	x_2	x_3	x_4	x_5
y	1	0.77	0.92	0.91	0.92
x_2	0.77	1.00	0.90	0.85	0.84
x_3	0.92	0.90	1.00	0.98	0.95
x_4	0.91	0.85	0.98	1.00	0.98
x_5	0.92	0.84	0.95	0.98	1.00

Vamos a hacer la estimación con el modelo RRA, con restricciones no posibilísticas, con $k_1 = 1$ y $k_2 = 0$. Es decir, la tendencia central se estimará como el método de mínimos cuadrados y, de acuerdo a la definición de estimación Ridge difusa (con los datos normalizados).

Tabla 5.2: Estimación Ridge RRD, restricciones NP, ejemplo simulado

Coef.	Real	Min. Cuadrados	$\lambda = 0{,}1$	$\lambda = 0{,}2$	$\lambda = 0{,}3$	$\lambda = 0{,}4$
a_1	1	-6.4	-3.1	1.5	5.1	7.9
a_2	3.5	-51.6	-9.9	2.4	10.2	15.5
a_3	40.6	102.9	39.9	32.9	29.6	27.5
a_4	6.6	-38.4	7.3	9.5	9.9	9.8
a_5	8.8	28	11.2	9.4	8.8	7.9

Los resultados de la tendencia central se muestran en la tabla 5.2, en la que aparece para cada coeficiente central, el valor del promedio simple de los valores simulados de entrada, es decir el promedio simple de los valores reales, la estimación de mínimos cuadrados, y cuatro estimaciones Ridge, con diversos valores λ.

En la tabla, se observa que los valores reales son todos positivos. Sin embargo, la estimación equivalente de mínimos cuadrados entrega tres coeficientes negativos. El valor absoluto de los cinco coeficientes reales es 60.5, mientras que el valor absoluto de los cinco valores estimados por mínimo cuadrados es de 227.3, lo que muestra la inflación en la magnitud de los coeficientes debido a la multicolinealiadad. En cambio, al aumentar λ disminuye este valor absoluto y, para $\lambda = 0{,}2$, la suma de los cinco coeficientes es 55.7 y *los cinco coeficientes son positivos*.

Los resultados para la tendencia central que se muestran en la tabla 5.2, se muestran en la traza Ridge de la figura 5.4.

Respecto a la elección del valor λ más adecuado, el gráfico de la figura 5.5 muestra la variación del método de validaciones cruzadas, con K=25, variando λ entre 0 y 0.4.

Según este método, el valor más recomendable –el valor mínimo de CV – se

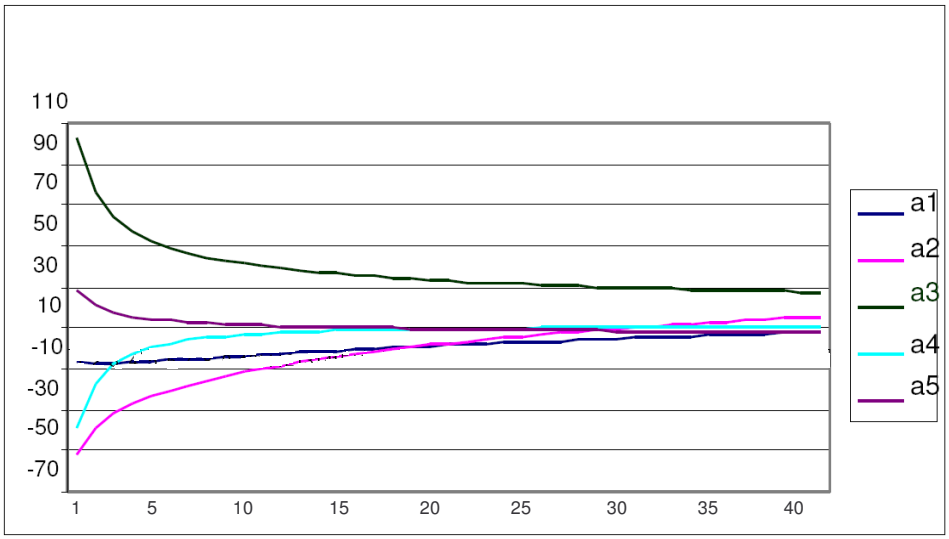

Figura 5.4: Modelo RRD normalizado, coef. a_i no posibilísticos. (Ej. simulación)

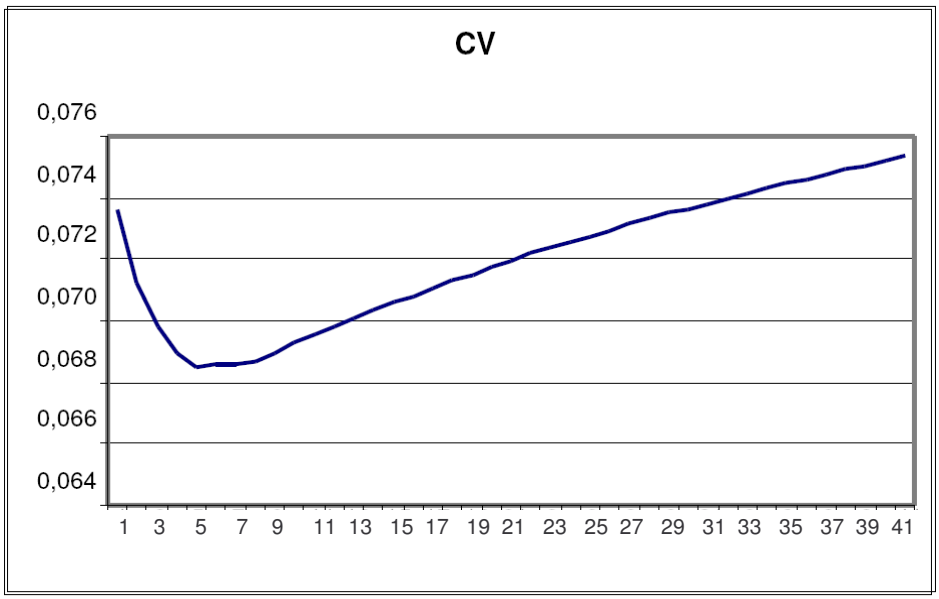

Figura 5.5: Validación cruzada, K=25, a_i no posibilísticos (Ej. simulación)

encuentra cuando $\lambda = 0{,}6$.

Como en este ejemplo, por haberse simulado los datos, se dispone de un valor muy aproximado al valor real de los parámetros, si se calculan las variaciones relativas de los coeficientes respecto a su valor real, se tiene el gráfico de la figura 5.6.

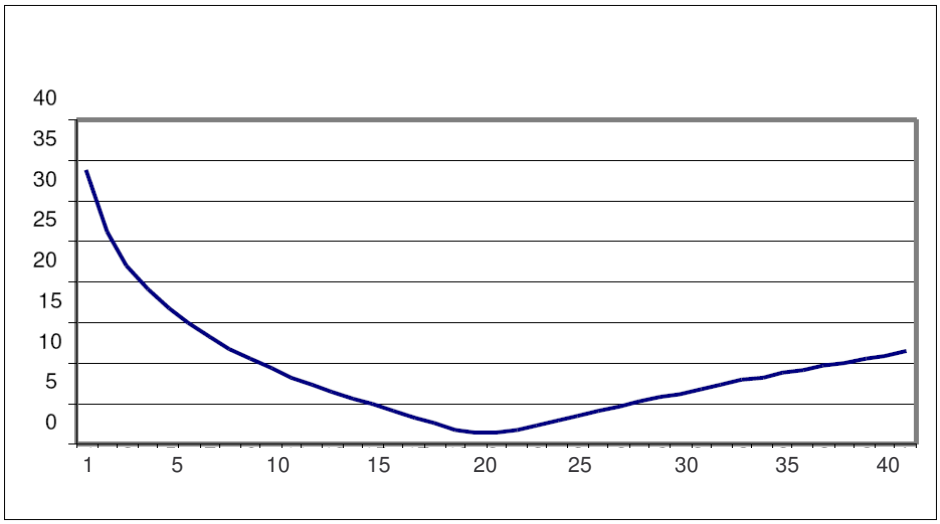

Figura 5.6: Desviaciones relativas de a_i en función de λ. No posibilístico. (Ej. simulación)

A partir de este gráfico, se concluye que el valor de $\lambda = 0{,}19$ es el valor más adecuado para los modelos de regresión Ridge, que es el valor mínimo de la suma de las desviaciones relativas de los cinco coeficientes centrales. De una desviación de 30 para $\lambda = 0$, se baja a una desviación del orden de 2 para $\lambda = 0{,}19$.

Este resultado nos indica que el método de validaciones cruzadas, con K=n podría estar subestimando el efecto del parámetro λ.

Empleando las restricciones posibilísticas en la estimación, para los modelos RRA(1,0), RRA(1,1) y RRA(0,1), el comportamiento de los coeficientes estimados es muy similar al reflejado en la tabla 5.2. Sin embargo, se nota que para el modelo RRA(0,1), que sólo supone el efecto de multicolinealidad en las extensiones, presenta una distribución más rígida en función de λ.

Sin tener información adicional sobre las extensiones de la variable de salida, y con la evaluación de los ejemplos realizados, nuestra conclusión es que el modelo

RRA(1,1), que aplica el efecto contra la multicolinealidad con igual peso tanto a los valores centrales como a las extensiones, es el más adecuado para estimar regresiones difusas con multicolinealidad en sus datos de entrada.

5.2. Regresión Ecológica

A la hora de hacer estimaciones, hay situaciones en las que no se dispone de información a nivel de individuos, sino que esta información está agregada en base a un cierto nivel geográfico o de otro tipo. Por ejemplo, se tienen resultados electorales a nivel de comunas, barrios, distritos, etc., pero no se conoce cómo votó una persona en particular. Otro ejemplo se tiene cuando se dispone de datos censales a través de las distribuciones marginales a nivel de municipio o barrio.

En el caso electoral, si se quiere conocer cómo cambió el electorado de un partido determinado en una elección respecto a la siguiente, en cuanto a si se mantuvo la adhesión al partido en un porcentaje determinado o se captaron electores que eran de otros partidos, hay que hacer inferencia a nivel de individuos en el electorado. Y este es el objetivo de la *Inferencia Ecológica*: lograr información a nivel de personas partiendo de datos agregados. Se conoce como *falacia ecológica* al hecho de afirmar que las relaciones obtenidas a nivel de grupos se sostienen a nivel de individuos.

Goodman en 1953 [Goo53], planteó la *Regresión Ecológica* como solución para este problema, con el requisito denominado *supuesto de homogeneidad*, es decir, que todas las probabilidades condicionales son iguales en las diversas unidades de información. Esto se ha aplicado hasta ahora en muchos estudios, por ejemplo, en los trabajos de Engel [Eng90, EH03].

Duncan [DD53] estimó un intervalo como solución, lo que se denomina la *solución de bordes*. King en 1997 [Kin97], planteó un nuevo enfoque en que los coeficientes de regresión son aleatorios, aplicado por Wellhofer[Wel01] para diversas elecciones italianas y Burden [BK98] para elecciones simultáneas en Estados Unidos. Cho[CG04, Cho01] ha criticado el modelo de King, ya que teniendo supuestos menos estrictos que los de Goodman, también son frecuentemente sobrepasados, y plantea algunas alternativas para mejorar los dos modelos anteriores.

Los estudios sobre la Regresión Ecológica han estado envueltos de polémica. Por un lado, los investigadores que niegan su validez general, y por otro, los que discuten las propuestas específicas enunciadas. Por ejemplo, un autor español ha escrito que

"nunca las cualidades propias de un conjunto pueden transferirse a las distintas individualidades que lo componen o, lo que es lo mismo, de la comparación entre comportamiento político de una colectividad y su específica personalidad a otros niveles, nunca puede deducirse la decantación electoral de alguno de sus integrantes".

Como ya hemos anticipado, el caso de inferencia ecológica más analizado se relaciona con los análisis de resultados electorales. Aunque se trata de un problema con cifras precisas [2], el sistema esta construido en base a un ambigüedad esencial: al tratarse de datos agregados, no existe una única distribución de las dos variables que genere la distribución agregada, puesto que hay diversas distribuciones desagregadas que producen los mismos datos agregados y, por lo tanto, esta ambigüedad hace interesante intentar modelar este problema, en base a los principios de los subconjuntos difusos.

La agregación de datos desde el nivel individual en que se generan, al nivel que se dispone para el análisis, significa una pérdida de información, que podría cuestionar los resultados. Nuestra opinión es que dicha pérdida de información se suple, en general, con las múltiples observaciones disponibles en las diversas unidades de agregación.

Para ejemplarizar, se precisará la situación más típica de Regresión Ecológica. Se considera el siguiente esquema para nuestro problema de estimación de matrices de transición electoral: se toman dos elecciones, en general consecutivas, a saber, la elección 1 y la elección 2. La elección 2 puede ser una elección simultánea para elegir otro cargo político (Senado y Cámara de Diputados, Presidente de la República, Parlamento Regional, etc.) o puede ser la segunda vuelta de la primera elección. También puede ser la siguiente elección política como, por ejemplo, las elecciones generales españolas del 2000 y 2004.

- Para la elección 1 se presentan p partidos políticos y hay n unidades de recogida

[2] En muchas elecciones el número de votos nulos es muy alto, como en la reciente elección presidencial peruana de 2006, en la que hubo 620.000 votos nulos y 1.737.000 votos blancos, frente a una diferencia de 62.000 votos entre el segundo y tercer candidato para pasar a la segunda vuelta. Por este motivo, que creemos que tales votos nulos reflejan, en un buen porcentaje, una votación mal contabilizada, o mal expresada formalmente, cuya incertidumbre podría modelarse con conjuntos difusos. En Chile, donde existe una larga tradición de tener que marcar la preferencia por un candidato, y no por un partido, en una sola mesa de votación, durante la última elección parlamentaria de 2005, hubo dos votos nulos, debido a que las dos personas votaron, en cada voto, por dos candidatos de la misma lista, debiendo sólo votar por un candidato para que el voto sea válido. La elección de Bush en el 2000 en Estados Unidos demostró que los procesos electorales no son, necesariamente, determinísticos.

de sufragios. Con x_{ik} se representa el número de votos para el partido i en cada unidad k de recepción de sufragios.

- Para la elección 2 se presentan q partidos políticos. Cada partido j obtiene y_{jk} votos en la unidad k.

Teniendo estas *distribuciones marginales* en cuenta, para cada unidad k, existe la probabilidad condicional:

$$jik = Prob(\frac{\text{un elector votó por partido j en la elección 2}}{\text{el elector votó por partido i en la elección 1}})$$

Debido al carácter secreto del sufragio, no es posible conocer estas probabilidades. Para estimarlas, se pueden modelizar con las relaciones

$$y_{jk} = \sum_i \beta_{jik} x_{ik} \tag{5.45}$$

donde se tienen pxqxn valores a estimar y sólo qxn ecuaciones, con lo cual no se tiene una solución única.

Si se supone que los valores β_{jik} son constantes en las diversas unidades i, es decir,

$$\beta_{ji} = \beta_{jik}\ k = 1, ..., n \tag{5.46}$$

entonces es posible realizar algún tipo de estimación estadística para los coeficientes β_{ji}. El paso desde el modelo indicado por la ecuación 5.45 al modelo basado en la ecuación 5.46 se conoce como *supuesto de homogeneidad* en la Regresión Ecológica. Si este supuesto no se cumple, entonces la regresión de Goodman anteriormente citada entrega resultados incorrectos.

Si los valores *jik* varían de unidad en unidad y el supuesto de homogeneidad no se cumple, entonces el objetivo de las matrices de transición electoral es estimar el promedio ponderado de los *jik* para las n unidades de información.

En la práctica, el problema de la inferencia ecológica ha sido planteado como una tabla 2x2, en términos de proporciones de personas en cada unidad de información, como se aprecia en la Tabla 5.3.

Tabla 5.3: Notación clásica de la regresión ecológica

		Elección 2	
Elección 1	Partido B	Otros partidos	Total
Partido A	β^B	$1-\beta^B$	X
Otros partidos	β^W	$1-\beta^W$	1 - X
Total	T	1 - T	1

donde X es la proporción de votos por el partido A en la primera elección, T es la proporción de votos por el partido B en la segunda elección y las probabilidades condicionales β^B y β^W son los valores a estimar, como valor promedio entre todas las unidades de información. β^B representa la proporción de los votantes por el partido A en la primera elección que votan por el partido B en la segunda elección.

El desarrollo teórico más profundo de los realizados hasta el momento [Kin97], basado en la teoría probabilística, asume una distribución binormal para las dos variables, con importantes dificultades en el cálculo de las estimaciones. Además no puede ser extendido a tablas de mayor tamaño.

De la Tabla 5.3 se tiene la relación

$$T = \beta^B X + \beta^W (1 - X) \tag{5.47}$$

que es estimada efectivamente como

$$T = \beta^W + (\beta^B - \beta^W) X \tag{5.48}$$

Goodman[Goo53] propone la estimación de esta regresión (por lo que es conocida con el nombre de *Regresión de Goodman*) basándose en el supuesto de que todos los β^W y β^B son constantes en las diversas unidades de información.

Duncan y Davis[DD53] proponen un método para acotar los coeficientes β^B y β^W. Para ello, transforman la ecuación 5.45 en una recta en función de los coeficientes para cada dato disponible:

$$\beta_i^W = (T_i/(1 - X_i)) - (X_i/(1 - X_i))\beta_i^B \tag{5.49}$$

De esta ecuación se obtienen las cotas para los coeficientes a estimar:

$$Max(0, (T_i - (1 - X_i))/X_i) \leq \beta_i^B \leq Min(T_i/X_i, 1) \tag{5.50}$$

$$Max(0, (T_i - X_i)/(1 - X_i)) \leq \beta_i^W \leq Min(T_i/(1 - X_i), 1) \tag{5.51}$$

El promedio ponderado de estas cotas sobre los n datos parece ser una buena medida de cota inferior y cota superior para el valor central a estimar en la Regresión Ecológica.

King [Kin97] propone otro método de estimación, que se llamará EI (Ecological Inference, que es también el nombre del programa computacional que creó el mismo King para realizar la estimación), considerando los coeficientes como aleatorios y suponiendo

1 una distribución bivariante normal truncada en el intervalo [0,1] para los dos coeficientes.
2 β_i^W y β_i^B son independientes *en promedio* de X_i.
3 Los valores T_i son independientes en las diversas mesas receptoras de votación, para un X_i dado.

El esfuerzo de King es un avance considerable en el campo de la Regresión Ecológica, aunque, en cualquier caso, descansa en tres supuestos que pueden no cumplirse, a pesar de que King fundamenta que la violación de estos supuestos no invalida el modelo.

Cho[CG04, Cho01] ha destacado las limitaciones de los enfoques de Goodman y King y ha propuesto un test para medir el efecto de otras variables en el comportamiento de los electores. También ha planteado, junto con Anselin [AC02], incorporar el efecto espacial de las diversas unidades receptoras de sufragios, en el sentido de que unidades territoriales vecinas pueden generar covariantes entre β^B y β^W.

La estimación de mínimos cuadrados ordinarios de β^B y β^W puede producir valores fuera del intervalo [0,1] [AC02]. Es una muestra de lo que se ha llamado *sesgo de agregación*. Menos frecuente es que la estimación de la regresión difusa entregue valores fuera de rango. Sin embargo, con la flexibilidad del modelo posibilístico, se mostrará a continuación que este sesgo puede ser corregido.

Cuando los coeficientes de mínimos cuadrados están fuera de rango, es conveniente hacer un análisis de los datos, considerando variables tales como el tamaño de las unidades de información, tamaño de X y T, o recurrir a otras variables [Cho01] tales como nivel socio-económico de cada unidad geográfica de información, para

una posible división de la estimación en dos grupos. De esta manera se puede hacer la estimación independientemente en cada grupo y luego calcular los coeficientes del conjunto mediante promedios ponderados. Cho en su trabajo presenta dos tests para determinar si la variable efectivamente discrimina entre grupos de observaciones. La forma en que opera esta discriminación consiste en ordenar los datos de acuerdo a la variable que influiría en la variabilidad de los coeficientes y determinar los puntos de corte o de cambio de tendencia. En el ámbito de la Regresión Difusa, también se encuentra una proposición para determinar los puntos de cambio de tendencia [YTL01].

El hecho de estimar datos sólo de tablas 2x2, como en el modelo EI, convierte en tediosa la estimación de muchas matrices de transición electoral. En el estudio de las elecciones italianas [Wel01], se afirma que "el numero total de posibles combinaciones basado sólo en elecciones consecutivas es un número mayor que 400; sin embargo pueden ser excluidas como improbables, por ejemplo, MSI intercambió votantes con el PDS, o que votantes DC puedan cambiarse a la RC. Sin embargo, existen alrededor de 200 combinaciones binarias de interés". No parece razonable excluir de las estimaciones ciertas combinaciones partidistas, por el solo hecho de que su coeficiente pueda ser pequeño o el número de estimaciones que hay que realizar sea alto. Estimar tablas 3x2, 4x2, 3x3, etc. debiera ser una opción disponible de la metodología.

5.2.1. Sistema de Regresiones Posibilísticas

Si se quiere estimar la matriz de transición electoral 5.45 con la condición de que los coeficientes sean conjuntos difusos con función de pertenencia no simétrica, se tiene que

$$Y_j = \sum_i \beta_{ji} x_i \tag{5.52}$$

donde los coeficientes a estimar $\beta = (a, c_L, c_R)$ son conjuntos difusos.

Si se expresa cada una de las ecuaciones 5.52, como un sistema de ecuaciones integrado para cada registro de información k, se tiene

$$Y_{kj} = \sum_k \beta_{ji} x_{ki} \tag{5.53}$$

Como los coeficientes representan probabilidades condicionales, entonces su máxi-

mo valor de pertenencia a_{ji} debe estar entre 0 y 1. Para lo cual cada una de las q regresiones debe agregar las condiciones

$$a_{ji} \leq 1;\ a_{ji} \geq 0;\ j = 1, ..., qi = 1, ..., p \tag{5.54}$$

Además, las q probabilidades condicionales de los votantes de un partido de la primera elección que votaron por los diversos partidos en la segunda elección deben sumar 1, lo que se expresa en la relación

$$\sum_i \beta_{ji} = 1 \tag{5.55}$$

Esto involucra no sólo a una de las regresiones sino al conjunto de ellas (q), lo que hace necesario que su estimación sea realizada como un sistema.

Por esta razón, se debe formar con los q problemas mencionados un solo problema de optimización, que para las p variables X tendrá que estimar pxq número difusos. Si se utilizan funciones de pertenencia no simétrica, se tienen un total de 3xpxq valores a estimar.

Como, en general, los coeficientes con subíndices jik, no son iguales para las n unidades de recepción de votos, lo que se estima realmente es los coeficientes ji promedio. Por lo tanto, es posible colocar la condición adicional

$$\sum (\sum_i a_{ji} X_{ik}) = \sum Y_{jk} \tag{5.56}$$

para cada uno de los q partidos de la segunda elección.

Además, si se quieren incorporar las limitaciones de borde para la estimación, en forma más estricta que la indicada en la ecuación 5.54, se pueden tener las condiciones adicionales

$$a_{ji} + c_{R_{ji}} \leq 1; a_{ji} - c_{L_{ji}} \geq 0; j = 1, ..., q\ i = 1, ..., p \tag{5.57}$$

De igual manera, si de un problema particular de estimación de matrices de transición electoral surgen nuevas condiciones, ya sea entre los coeficientes o a través de la incorporación de nuevas variables que pueden complementar el comportamiento electoral de los ciudadanos, estas nuevas condiciones pueden ser agregadas a la formulación del problema de programación matemática.

En general, un sistema de regresiones posibilísticas permite incorporar condiciones particulares de cada modelo, por ejemplo, información obtenida de encuestas independientes de los datos, lo que conlleva obtener estimaciones de regresión más precisas.

Para evaluar la calidad del ajuste de la estimación difusa, se considera la suma del valor absoluto de las diferencias entre el valor a estimar y el valor estimado:

$$Dif = \sum |T_i - (\beta^B X_i + \beta^W (1 - X_i))| \tag{5.58}$$

El hecho de que *Dif* aumente de una estimación a otra, lo convierte en un indicador de alejamiento de la tendencia central en la estimación de los coeficientes, y deberá considerarse cuando se disponga de diversos modelos disponibles.

5.2.2. Ejemplos de Estimación de Matrices de Transición Electoral

Veamos ahora la aplicación de lo anteriormente propuesto a un caso práctico concreto. Se considerará primero el caso de tablas 2x2, construido expresamente para conocer los coeficientes a estimar, y para comparar con los resultados de la regresión de Goodman y la de King, se tomarán las siguientes alternativas de regresión difusa para el sistema de ecuaciones:

- Difusa SI 1: modelo que sólo incluye las restricciones de que los coeficientes estén en el intervalo [0,1].
- Difusa SI 2: modelo difuso SI 1, más la restricción de que la suma de los coeficientes de cada partido de la primera elección debe ser 1.
- Difusa SI 3: modelo difuso SI 2, más la condición 5.56.
- Difusa SI 4: agregando a las condiciones anteriores el límite de cotas de las relaciones 5.57.

El ejemplo 1 considera una situación relativamente corriente en que el coeficiente β^B es alto y el otro coeficiente es bajo. Se generaron 40 datos con una desviación estándar para la variable dependiente de 0.179.

En este primer ejemplo se aprecia que el método de Goodman es el que entrega una máxima diferencia entre el valor estimado para los coeficientes y el valor verdadero. Además, hay una superioridad de los métodos difusos frente al método EI. El

Tabla 5.4: Comparación de métodos: Regresión Ecologica (ejemplo 1)

Método	β^B	β^W	error
Verdadero	0,9390	0,1231	
Goodman	0,9905	0,0756	0,0990
EI	0,9703	0,0934	0,0610
Difuso SI 1	0,9799	0,1364	0,0542
Difuso SI 2	0,9669	0,1695	0,0743
Difuso SI 3	0,9742	0,0899	0,0684
Difuso SI 4	0,9693	0,1080	0,0454

método difuso EI 2 entrega los resultados menos correctos, tal vez por forzar la *unicidad* de los coeficientes, unicidad que se cumple, prácticamente automáticamente, en el caso de que la segunda elección tenga sólo dos opciones.

El segundo ejemplo es real, forma una tabla 4x2, y no se conocen los verdaderos valores de la matriz de transición electoral. Se trata de uno de los miles de ejemplos que pueden obtenerse en Internet, que corresponde al condado de Amoostook, en le Estado de Maine, Estados Unidos. Se desarrollaron elecciones simultaneas de Gobernador, Senador, Representantes, y otros cargos, en Noviembre del 2002 [3].

En este caso, para Senadores gana el candidato republicano y para el caso de Gobernador gana el candidato demócrata, ambos por amplio margen. Por este motivo resulta muy interesante conocer la matriz de transición electoral. Los datos constan de 71 mesas receptoras de sufragios en el condado y la principal limitación que tienen es que el número de electores en cada mesa no es uniforme: de un mínimo de 16 electores se llega a un máximo de 3060 electores, con una mediana de 153 electores, una media de 370 y una desviación estándar de 585.

Para la estimación de Goodman, realizando estimaciones 2x2, se obtiene el resultado representado en la tabla 5.5.

Tabla 5.5: Estimación de Goodman para ejemplo 2

	Sen rep.	Sen dem.	Total
Gob rep.	1.099	-0.099	8401
Gob. dem.	0.439	0.561	15783
Gob. verde	0,600	0,4004	1.279
Gob. indep.	-0.780	1.780	812
Total	17867	8408	26275

Se puede apreciar la presencia de valores fuera del rango [0,1] lo que invalida abiertamente los resultados. Aún truncando a 0 o 1 los valores fuera de rango, se

[3] http://www.state.me.us/sos/cec/elec/prior.htm

dispondría de una estimación poco confiable.

Para la estimación de King, realizando también 4 estimaciones 2x2 (King sólo recomienda extender su modelo a tablas 3x2, pero tomamos como una prueba de robustez del modelo el forzarlo en tablas 4x2), se tienen los resultados de la tabla 5.6.

Tabla 5.6: Estimación de King (EI) para ejemplo 2

	Sen rep.	Sen dem.	Total
Gob rep.	0.9495	0.05059	8401
Gob. dem.	0.5096	0.4904	15783
Gob. verde	0.9942	0.0058	1.279
Gob. indep.	0.0265	0.9735	812
Total	17867	8408	26275

Esta estimación tiene coeficientes razonables para la relación entre los candidatos demócratas y republicanos. Sin embargo, llama la atención lo extremos que son los coeficientes para el partido verde y el candidato a gobernador independiente, lo que sugiere que ambos estimaciones son erróneas. ¿Cómo va a ser real que el coeficiente correspondiente al candidato a gobernador republicano sea menor que el coeficiente del candidato a gobernador verde? ¿Cómo va a ser posible que el candidato ganador para senador con una amplísima ventaja, prácticamente no obtenga ningún voto de los que votaron por el candidato independiente a gobernador? La estimación para el total de la votación del candidato a senador republicano es de 17013, lo que significa una subestimación de unos 500 votos.

La estimación de cualquiera de los sistemas difusos, a diferencia de las dos estimaciones anteriores, se ha realizado de forma simultánea, como tabla 4x2. Los resultados para el modelo SI 1, con h=0 y utilizando el modelo MCP con $k_1 = 1$ y $k_2 = 1$, se muestran en la tabla 5.7.

Tabla 5.7: Estimación con el modelo difuso SI 1, ejemplo 2

	Sen rep.	Sen dem.	Total
Gob rep.	0.9586	0.0415	8401
Gob. dem.	0.4595	0.5402	15783
Gob. verde	0.1953	0.8047	1.279
Gob. indep.	0.6175	0.3825	812
Total	17867	8408	26275

Esta estimación tiene una diferencia absoluta de 2366, produce valores para los que votaron por el candidato a gobernador verde e independiente muchísimo más razonables, pero tiene como dificultad que la estimación para el total de la votación del candidato a senador republicano es de sólo 16057, subestimando su votación en

Tabla 5.8: Estimación con el modelo difuso SI 3, ejemplo 2

	Sen rep.	Sen dem.	Total
Gob rep.	0.9851	0.0149	8401
Gob. dem.	0.5248	0.4752	15783
Gob. verde	0.6300	0.3700	1.279
Gob. indep.	0.6175	0.3825	812
Total	17867	8408	26275

Tabla 5.9: Estimación de parámetros para el modelo difuso SI 3

	a_i	c_{L_i}	c_{D_i}
Gob rep.	0.9851	0.0650	0.2793
Gob. dem.	0.5248	0.1734	0.0279
Gob. verde	0.6300	1.2492	0.3847
Gob. indep.	0.6175	0.0000	0.0000

cerca de 1800 votos.

La estimación con el sistema SI 2 produce una diferencia absoluta de 2703, que es más alta que la anterior, por lo que no se considera interesante detenerse en ella. La estimación de modelo SI 3 se muestra en la tabla 5.8.

Esta estimación hace bajar la diferencia absoluta a 1410 y el valor estimado para el total de la votación del candidato a senador republicano es de 17866, por condición de la estimación, pero mucho más exacta que la entregada por EI.

La estimación con el modelo SI 4, que trunca las extensiones de la función de pertenencia para que la estimación completa esté en el intervalo [0,1], eleva la diferencia 5.57 a 1670, por lo que no se considerará.

A la vista de los resultados, en función de la calidad de ajuste, se recomienda la estimación del modelo SI 3 como la más adecuada para calcular la matriz de transición electoral. Como se dijo, la estimación SI 1, aunque muy razonable, subestima la estimación global del candidato a senador republicano, por lo que tiene plena justificación el aumento de valor en tres coeficientes, entre las estimaciones SI 1 y SI 3, para compensar la subestimación indicada.

La estimación completa para los coeficientes del candidato a senador republicano con SI 3, se muestra en la tabla 5.9.

Las extensiones para los coeficientes correspondientes al gobernador republicano y al demócrata son bastante acotadas, como la del candidato independiente. Sin embargo, la extensión izquierda del candidato a gobernador verde sobrepasa largamente la cota de 1, por lo que habría que pensar en una función de pertenencia triangular

truncada a los límites que impone el problema.

La función de pertenencia *truncada* para el coeficiente difuso de gobernador verde para la elección del senador republicano se define como:

$$\begin{aligned} 0\ para\ \beta > 1 y \beta < 0 \\ (\beta + 0{,}6192)/1{,}2492\ para\ 0 \leq \beta \leq 0{,}63 \\ (1{,}0147 - \beta)/0{,}3847\ para\ 0{,}63 < \beta \leq 1 \end{aligned} \tag{5.59}$$

En Diciembre de 1999 se realizó la última elección presidencial en Chile. En Enero del 2000 se realizó, por primera vez en la historia del país, la segunda vuelta de esta elección, en que salió elegido el actual Presidente de la República Ricardo Lagos. Una estimación con esta metodología se encuentra en nuestra referencia [DMV05].

5.2.3. Ejemplos con Datos del Censo de Población

Otro caso de Regresión Ecológica se presenta cuando se quiere obtener cierta información detallada de un censo de población y sólo se dispone de la información a un nivel anterior de desagregación.

Por ejemplo, si se cuenta con la estadística (distribuciones marginales), para cada departamento del Perú, de la lengua materna de los censados (quechua, castellano, etc.), y también de la estadística del nivel educacional por departamento (educación primaria, secundaria, etc.) y queremos estimar cuánta población, para cada lengua materna, tiene cada nivel educacional (es decir, el cruce de información entre lengua materna y nivel educacional), la regresión ecológica podría ser una alternativa de estimación.

En nuestra experimentación, para el Censo de Población del Perú del año 1993[4], se tomarán cuatro categorías de niveles educacionales: sin instrucción, educación primaria, educación secundaria y educación superior (universitaria y no universitaria) cuyos totales nacionales y totales por departamento se muestran en la tabla 5.10.

También se dispone de la información de lengua materna en las siguientes categorías: quechua, aymara, otras lenguas maternas nativas, castellano, idiomas extranjeros y no responde. Como la categoría idiomas extranjeros es muy poco frecuente

[4] http://www.inei.gob.pe/

Tabla 5.10: Nivel educacional, por departamento y total del país (Perú 1993)

Sin educacion	*Primaria*	*Secundaria*	*Superior*
64505	162526	43203	12928
183617	369473	203500	78471
108122	145140	49428	18467
89468	286743	271093	172533
132827	190031	66300	32691
298703	577207	144131	55010
41903	169500	237884	127086
225601	391638	185885	82704
102895	161684	46496	13789
147686	274473	95771	37857
47940	186552	179596	86987
140635	393013	258780	107456
187267	481740	292854	150528
115780	345386	252565	93628
460985	1734890	2201581	1345055
99618	296467	134766	42166
7350	26238	18327	4863
12718	43844	38567	20689
32746	90921	50659	20700
237459	574130	280413	109929
202492	452015	206845	76923
80269	255805	103503	30153
20026	69774	68940	35960
13532	62180	43405	16188
36171	130998	77093	21474
3090000	7872000	5551000	2891000

(35 mil en 19 millones), se considerara junto a la categoría Castellano. Los totales por departamento están indicados en la tabla 5.11, incluyendo los totales nacionales.

Tabla 5.11: Lengua materna, por departamento y total del país (Perú 1993)

Quechua	*Aymara*	*Otras nativas*	*Castellano*	*No responde*
1050	234	32572	247194	2112
299185	3287	777	525826	5086
245953	927	250	72145	1882
140535	16692	552	658371	3687
297737	651	411	120210	2850
6038	6284	1447	1051568	9714
34916	3246	475	535680	2056
560101	2836	7257	308779	6855
215972	733	274	105560	2325
171052	2161	1680	375596	5298
33004	1097	389	464287	2298
113189	3221	19664	757408	6402
4716	812	1122	1099091	6648
18173	5059	810	779006	4311
547397	26018	5106	5141736	22254
9603	236	21107	536308	5763
13687	786	1761	40186	358
12580	14741	87	87907	503
21208	1055	5157	165874	1732
2717	420	1568	1186287	10939
405596	305951	1216	219350	6162
10138	2070	2365	450772	4385
6847	40411	196	146333	913
563	60	137	134018	1527
5991	1392	25794	230640	1919
3177000	440000	132000	15405000	117000

Un primer modelo para mostrar el esquema de regresión ecológica descrito anteriormente, se tiene agrupando las tres categorías de lenguas nativas más las categoría *no responde*, quedando sólo dos agrupaciones de lengua materna. Para cuatro estimaciones independientes de mínimos cuadrados (una para cada nivel educacional), resultan los estimadores de tendencia central que se muestran en le tabla 5.12.

Tabla 5.12: Estimación de mínimos cuadrados con dos categorías de lengua materna (Perú 1993)

Categoria	*Sineduc*	*Primaria*	*Secundaria*	*Superior*
Nativas	0,286	0,478	0,177	0,060
Castellano	0,081	0,317	0,376	0,225

Una propiedad de la estimación de mínimos cuadrados es que si se suma cada fila de la tabla 5.12 se tiene exactamente la unidad.

En nuestro análisis, se obtuvo un resultado muy interesante: para las categorías de menor instrucción educacional, las personas con lenguas maternas nativas pertenecen a ellas en una 78.4 %. En cambio, para las categorías de mayor instrucción educacional, las personas con lengua materna Castellano pertenecen en un 60 % (sólo el 6 por ciento de las personas con lengua materna nativa tiene educación superior, en cambio este porcentaje aumenta al 22.5 para las personas con lengua materna Castellano).

Como se observa, todas las celdas deben estar entre 0 y 1, puesto que son probabilidades condicionales.

Si se repite la estimación, pero desagregando la categoría lengua materna quechua, que es la más numerosa del resto, se tienen los resultados de la tabla 5.13

Tabla 5.13: Estimación de mínimos cuadrados con tres categorías de lengua materna (Perú 1993)

Categoria	*Sineduc*	*Primaria*	*Secundaria*	*Superior*
Quechua	0,286	0,390	0,211	0,110
Otras lenguas Nativas	0,272	0,810	0,047	**-0,129**
Castellano	0,081	0,322	0,374	0,223

En esta tabla, se aprecia que aparece un valor fuera del rango [0, 1], lo que resta validez a toda la estimación de *otras lenguas nativas*. Este resultado es de tan común ocurrencia que ha llevado a cuestionar toda la metodología de la Regresión Ecológica.

Si se repite la estimación, pero ahora con cuatro categorías de lengua materna, los resultados con mínimos cuadrados ordinarios se muestran en la tabla 5.14.

Tabla 5.14: Estimación de mínimos cuadrados con cuatro categorías de lengua materna (Perú 1993)

Categoria	*Sineduc*	*Primaria*	*Secundaria*	*Superior*
Quechua	0,280	0,374	0,224	0,122
Aymara	0,173	0,641	0,180	0,003
Otras Nativas y NR	**2,545**	**4,633**	**-3,020**	**-3,159**
Castellano	0,066	0,296	0,395	0,243

Finalmente, con las cinco categorías de lengua materna y con regresión de mínimos cuadrados, se tienen los resultados de la tabla 5.15.

Se aprecia, de inmediato, que hay valores fuera de rango y *valores mucho mayores que 1*. Y esto ocurre, no sólo en las categorías con menor cantidad de personas, sino en categorías tan significativas como *lengua castellana*. Esto muestra que la re-

Tabla 5.15: Estimación de mínimos cuadrados para Perú (1993) con cinco categorías de lengua materna

Categoria	*Sineduc*	*Primaria*	*Secundaria*	*Superior*
Quechua	0,112	0,160	0,431	0,296
Aymara	**-0,052**	0,357	0,457	0,238
Otras nativas	**-0,637**	0,573	0,901	0,163
Castellano	**-0,049**	0,150	0,537	0,363
No responde	**30,06**	**39,74**	**-36,93**	**-31.87**

gresión de mínimos cuadrados no entrega resultados razonables para nuestro modelo y algunos de sus supuestos, como independencia de las variables de entrada o que los coeficientes se regresión sean constantes, no se cumple.

Sin embargo, no sólo se debe a la desagregación la aparición de valores fuera de rango. En las tablas 2x2, generalmente las categorías están bien equilibradas en cantidad de observaciones, e igualmente se presentan valores fuera de rango.

A continuación, veremos cómo este problema de números precisos, se estima empleando el modelo de regresión difusa de ecuaciones simultáneas definido anteriormente.

En el caso de la tabla 5.12 se tienen dos regresiones independientes, y en el caso de la tabla 5.15, se tienen 5 regresiones independientes. Pero existe una condición que agrupa a las regresiones, dado que, para cada lengua materna, la suma de las probabilidades condicionales de los diversos niveles de educación debe sumar 1. Esta condición puede plantear un sistema conjunto de regresiones, en el que para cada Departamento k del Perú, para cada lengua materna j, y para cada nivel de instrucción i, se postula el modelo

$$Nivel_{jk} = \sum_{i} \beta_{ji} * Lengua_{ik} \;\; para\; j = 1, ..., q \; y \; k = 1, ..., n \tag{5.60}$$

donde la condición de probabilidades condicionales de los estimadores β_{ji} está dada por

$$\sum_{i=1}^{i=p} \beta_{ji} = 1 \quad para\; j = 1, ..., q \tag{5.61}$$

Esta condición puede ser incorporada a nuestro modelo de regresión difusa en la forma de q restricciones adicionales.

Vamos ahora a repetir las estimaciones que se hicieron con mínimos cuadrados, ahora con nuestra propuesta de sistema de regresiones. La estimación con el modelo MCP(1,1) produce los coeficientes para el valor central (es decir, con valor de pertenencia 1), que se muestran en la tabla 5.16.

Tabla 5.16: Estimación del valor central de nivel educacional con 2 categorías de lengua materna con MCP(1,1)(Perú 1993)

Categoria	Sin educ	Primaria	*Secundaria*	*Superior*
Lenguas nativas	0,277	0,536	0,149	0,038
Castellano	0,126	0,406	0,310	0,159

Si se estima con tres variables de entrada, se tiene la tabla 5.17, cuyos resultados son bastante consistentes comparados con la tabla 5.16.

Tabla 5.17: Estimación de valor central de nivel educacional con 3 categorías de lengua materna con MCP(1,1)(Perú 1993)

Categoria	Sin educ	Primaria	*Secundaria*	*Superior*
Quechua	0,327	0,470	0,149	0,053
Otras nativas y NR	0,221	0,760	0,018	0,000
Castellano	0,134	0,374	0,315	0,175

Si se desagrega la lengua nativa en quechua, aymara y otras lenguas, se tiene una estimación de un sistema de 4 ecuaciones simultáneas, con los resultados dados en la tabla 5.18 para la tendencia central y el modelo de estimación MCP(1,1).

Tabla 5.18: Estimación de valor central de nivel educacional con 4 categorías de lengua materna con MCP(1,1)(Perú 1993)

Categoria	Sin educ	Primaria	*Secundaria*	*Superior*
Quechua	0,304	0,434	0,192	0,070
Aymara	0,199	0,588	0,195	0,017
Otras nativas y NR	0,000	1,000	0,000	0,000
Castellano	0,133	0,368	0,318	0,180

Si se desagregan los datos en todas las categorías disponibles, quedan dos categorías con muy poca cantidad de individuos: *otras lenguas maternas* y *no responde*. Por este motivo, la estimación tiende a resentirse, al aparecer más valores cero y uno, como se aprecia en la tabla 5.19.

Esta estimación que, a diferencia de los mínimos cuadrados ordinarios, cumple con la restricción que los coeficientes estimados se encuentren entre 0 y 1, adolece, sin embargo, de la presencia de algunos coeficientes exactamente 0 o 1, lo que indica

Tabla 5.19: Estimación de valor central del nivel educacional con 5 categorías de lengua materna con MCP(1,1)(Perú 1993)

Categoria	Sin educ	Primaria	*Secundaria*	*Superior*
Quechua	0,418	0,581	0	0
Aymara	0,234	0,744	0,021	0
Otras nativas	0	1	0	0
Castellano	0,107	0,353	0,359	0,181
No responde	0	1	0	0

la presencia de ciertas rigideces en los datos a los que se aplicó el modelo. A nuestro entender, la limitación de estos resultados se debe:

1 Los datos están agregados a nivel de Departamentos del Perú, que son 25, con un claro predominio de Lima, con casi 6 millones sobre un total levemente superior a 19 millones.
2 Las categorías de lengua materna son muy desequilibradas en el número total de población, ya que Castellano representa más de 15 millones. En cambio, la categoría *otras lenguas nativas* suman 132.000 y la categoría *No responde* 117.000.

A pesar de estas dos fuertes limitaciones de la información inicial, si se comparan, por ejemplo, las cuatro estimaciones para los coeficientes correspondientes a *Lengua nativa Castellano*, se aprecia una gran consistencia en los resultados, que confirman y cuantifican, la hipótesis de que a un mayor nivel de educación corresponde una mayor proporción de personas con lengua nativa castellana de la población del Perú.

En resumen, disponiendo de dos fuentes de datos independientes, para las mismas unidades geográfico-administrativas, es posible hacer una estimación que relaciona ambas distribuciones. Este es un problema de minería de datos, en el que definiendo un modelo de regresión en un sentido difuso, es posible relacionar datos precisos con buenos resultados y con una interpretación respaldada por la teoría.

Capítulo 6

Conclusiones

En nuestro trabajo de investigación hemos intentado desarrollar una metodología para el análisis de regresión difuso, a partir de datos de salida con funciones de pertenencia triangulares no simétricas, y se ha establecido un conjunto de resultados que potencian lo conocido hasta ahora en los fundamentos y la aplicabilidad del Análisis de Regresión Difusa.

Como conclusiones específicas, en este resumen final, se pueden destacar diversos aportes originales al tema de estudio:

- Se ha desarrollado un nuevo criterio para modelar la incertidumbre en los problemas de Regresión Difusa, que consiste en implementar un enfoque de programación cuadrática, incorporando el nuevo concepto del cuadrado de las diferencias de las extensiones de las funciones de pertenencia entre el dato observado y el dato estimado, en la función objetivo. Este enfoque permite, además, superar las limitaciones que presentaba la programación lineal como medio de estimación de la Regresión Difusa.

- Esta nueva propuesta también ha integrado el criterio de minimización de mínimos cuadrados con el criterio de minimización de la incertidumbre en un solo modelo que permite ponderar ambos criterios, con lo que se funden en un solo modelo las dos grandes líneas de desarrollo que había tenido hasta ahora la Regresión Difusa.

- Se han flexibilizado las opciones de restricción en el modelo de programación cuadrática. El criterio tradicional de restricciones posibilísticas de la Regre-

sión Difusa, se ha ampliado con la capacidad de la sola restricción por la naturaleza de las extensiones de las funciones de pertenencia. Se han incorporado, además, nuevos criterios de restricción basados en la teoría de la posibilidad, como el índice de posibilidad de la igualdad entre el dato estimado y el dato original, o los índices de necesidad de inclusión de un número difuso en otro número difuso.

- Se ha desarrollado una interpretación del numero difuso estimado por la regresión, diferenciando su sentido cuando se trata de estimaciones posibilísticas y cuando se refiere a estimaciones no posibilísticas.

- Frente a la mínima atención que se le ha prestado hasta ahora a la medición de la calidad de la estimación difusa, en esta memoria se ha definido un conjunto de seis indicadores, propios de la inferencia difusa, para medir la bondad de ajuste de las estimaciones. Se ha generalizado el clásico R^2 probabilístico a un R^2 difuso, también con el rango en el intervalo [0, 1], y con una interpretación similar, para medir el ajuste de la tendencia central de la regresión. Se han definido cinco índices construidos a partir de las funciones de pertenencia de los datos estimados y los datos originales, que enriquecen la interpretación y el análisis de los resultados, y se ha escogido el indicador SIM como aquel que representa en forma más completa la comparación de dichas funciones de pertenencia.

- Como consecuencia de la definición de R^2_{difuso}, se ha definido un indicador de ajuste parcial, RP^2_{difuso} que permite medir el aporte que una nueva variable hace a la explicación de la tendencia central de un modelo de regresión difusa. Este indicador sirve para implementar un criterio de selección de variables de entrada.

- Se ha desarrollado un procedimiento de selección de variables (regresión paso a paso), que es general en cuanto permite que el analista de los datos elija flexiblemente entre una diversidad de criterios, y que combina las visiones de selección de variables *hacia adelante* y *hacia atrás*.

- También se ha aplicado el enfoque de optimización cuadrática para definir un modelo de estimación con datos de entrada que presenten el problema de multicolinealidad, en el enfoque llamado Ridge. Creemos que esta propuesta significa un importante paso para avanzar hacia la formalización de una econometría de la Regresión Difusa.

Dentro de este aporte, se ha propuesto un criterio de normalización de datos para el modelo difuso Ridge, para compatibilizar resultados con distinta escala de medición de los datos. Además, se ha demostrado que existe una relación directa, para la estimación con datos normalizados, entre los modelos con el parámetro Ridge λ constante, y los modelos generalizados con este parámetro dependiendo de cada variable.

- Dentro también de las aplicaciones presentadas en la memoria, se ha implementado el modelo de regresión difusa para estimar un sistema de regresiones, en el que es posible incorporar restricciones entre las diversas ecuaciones. Este resultado constituye un segundo aporte para una econometría de la Regresión Difusa.

  Se ha aplicado esta nueva modalidad de múltiples ecuaciones al caso de la Regresión Ecológica (estimación de caracteres individuales a partir de datos agregados), problema que no ha sido resuelto teóricamente tras más de 50 años de estudio y grandes debates entre estadísticos y sociólogos. Nuestra propuesta no pretende ser una solución analítica sofisticada, sino una solución realista consistente y con la apreciable ventaja de ser generalizada al caso de múltiples variables.

6.1. Futuras líneas de investigación

El trabajo descrito en esta memoria está empezando a ser publicado para su difusión en congresos y revistas de prestigio en el área [DMV04, DMV05, DMV06a, DMV06b]. Sin embargo, vamos a terminar esta memoria señalando algunas líneas de investigación que han surgido del estudio realizado y que no han sido abordadas, debido a que sobrepasan los objetivos de este proyecto de investigación, o que han aparecido como consecuencia de sus resultados. Estas líneas se abordarán en nuestra investigación futura:

1 Incorporar dentro del modelo de regresión difusa de optimización cuadrática, variables difusas de entrada. Se sugiere modelar estas variables con coeficientes precisos, dado que la multiplicación de dos números difusos, por el principio de extensión, entrega extensiones demasiadas amplias.
2 En relación a avanzar en la teoría econométrica de la Regresión Difusa, estudiar la relación entre el concepto econométrico de heterocedasticidad (cam-

bios de magnitud en la varianza de la incertidumbre), con la propuesta de la Regresión Difusa de restricciones de tipo posibilístico. En esta línea, también consideramos interesante estudiar la relación entre los modelos econométricos de múltiples ecuaciones, con nuestra propuesta de sistema de ecuaciones de regresión difusa. Además, aplicar el modelo propuesto a un procedimiento de regresión por *trazos* puede ser interesante para estimar fenómenos como los ciclos económicos y cambios de tendencia en el empleo.

3 En relación a la Minería de Datos, aplicar el modelo de optimización cuadrática propuesto en esta memoria a técnicas conocidas dentro del campo de los árboles de clasificación y regresión. También resulta atractivo estudiar el efecto de la agregación/desagregación de los datos en los modelos de Regresión Difusa, y comprobar la hipótesis de que a mayor desagregación de los datos, más precisa es la estimación de los modelos. Analizar la relación entre los indicadores de bondad de ajuste propuestos en la memoria con los índices de correlación entre números difusos que se han propuesto en la literatura de conjuntos difusos es un trabajo que queda pendiente para el futuro.

4 En relación a la multicolinealidad en la Regresión Difusa, pretendemos estudiar con mayor detalle la elección óptima del parámetro λ, incorporar la corrección Ridge para datos multicolineales a la Regresión Ecológica, y desarrollar modelos alternativos de Regresión Difusa para corregir el problema de la multicolinealidad de los datos de entrada, como serían la *Regresión Lasso*, la *Regresión de Componentes Principales* y la *Regresión de Mínimos Cuadrados Parciales*.

5 Finalmente, en relación a la interpretación de los coeficientes difusos estimados, queda pendiente para el futuro determinar la relación de cada variable de entrada con el número difuso estimado con restricciones posibilísticas.

Bibliografía

[AC02] Anselin L. y Cho W. K. T. (2002) Spatial effects and ecological inference. *Political Analysis* 10(3): 276–297.

[Alu01] Aluja T. (2001) La minería de datos, entre la estadística y la inteligencia artificial. *Questto* 25(3).

[Bar90] Bardossy A. (1990) Note on fuzzy regression. *Fuzzy Sets and Systems* 37: 65–75.

[BBD91] Bardossy A., Bogardi I. y Duckstein L. (1991) Fuzzy set and probabilistic techniques for health-risk analysis. *Applied Mathematics and Computation* 45: 241–268.

[BBD93] Bardossy A., Bogardi I. y Duckstein L. (1993) Fuzzy nonlinear regression analysis of dose-response relationships. *European Journal of Operational Research* 66: 36–51.

[BBK88] Bardossy A., Bogardi I. y Kelly W. E. (1988) Fuzzy regression for electrical resistivity-hydraulic conductivity relationships. *International Journal of Approximate Reasoning* 2: 98.

[BK98] Burden B. C. y Kimball D. C. (1998) A new approach to the study of ticket splitting. *American political Science Review* 92(3): 533–544.

[BPDN97] Boreux J.-J., Pesti G., Duckstein L. y Nicolas J. (1997) Age model estimation in paleoclimatic research: fuzzy regression and radiocarbon uncertainties. *Palaeogeography, Palaeoclimatology, Palacoccology* 128: 29–37.

[Buc04] Buckley J. J. (2004) *Fuzzy Statistics*. Springer.

[BW97] Bell P. M. y Wang H. (1997) Fuzzy linear regression models for assesing risks of cumulative trauma disorders. *Fuzzy Sets and Systems* 92: 317–340.

[CA01] Chang Y.-H. O. y Ayyub B. M. (2001) Fuzzy regression methods - a compartive assessment. *Fuzzy Sets and Systems* 119: 187–203.

[Cel87] Celmins A. (1987) Least squares model fitting to vector data. *Fuzzy Sets and Systems* 22: 245–269.

[CG04] Cho W. K. T. y Gaines B. J. (2004) The limits of ecological inference: The case of split ticket voting. *American Journal of Political Science* 48(1): 152–171.

[Cha01a] Chang Y.-H. O. (2001) Hybrid fuzzy least-squares regression analysis and its reliability measures. *Fuzzy Sets and Systems* 119: 225–246.

[Cha01b] Chang Y.-H. O. (2001) Hybrid regression analysis with reliability and uncertainty measures. *Ph.D. Dissertation, University of Maryland* .

[Che01] Chen Y.-S. (2001) Outliers detection and confidence interval modification in fuzzy regression. *Fuzzy Sets and Systems* 119: 259–272.

[Cho01] Cho W. K. T. (2001) Latent groups and cross-level inferences. *Electoral Studies* 20: 243–263.

[Cho03] Chou C.-C. (2003) The canonical representation of multiplication operation on triangular fuzzy numbers. *Computers and Mathematics with applications* 45: 1601–1610.

[CL94a] Chang P. T. y Lee E. S. (1994) Fuzzy least absolute deviations regression based on the ranking of fuzzy numbers. *IEEE World Congress Fuzzy Systems, IEEE Poc.* páginas 1365–1369.

[CL94b] Chang P. T. y Lee E. S. (1994) Fuzzy linear regression with spreads unrestricted in sign. *Computer Math. Applc.* (4): 61–70.

[CL99] Cheng C. y Lee E. (1999) Nonparametric fuzzy regression: k-nn and kernel smoothing techniques. *Computers and Mathematics with Applications* 38: 239–251.

[CLK96] Chang P. T., Lee E. S. y Konz S. A. (1996) Applying fuzzy linear regression to vdt legibility. *Fuzzy Sets and Systems* páginas 197–204.

[dAT04] de Andres J. y Terceño A. (2004) Estimating a fuzzy term structure of interes rates using fuzzy regression techniques. *European Journal of Operational Research* 154(3): 804–818.

[DD53] Duncan O. D. y Davis B. (1953) An alternatiave to ecological correlation. *American Sociological Review* 18: 665–666.

[DG00] D'Urso P. y Gastaldi T. (2000) A least-squares approach to fuzzy linear regression analysis. *Computational Statistics and Data Analysis* 34: 427–440.

[DG02] D'Urso P. y Gastaldi T. (2002) An orderwise polynomial regerssion procedure for fuzzy data. *Fuzzy Sets and Systems* 130: 1–19.

[Dia88] Diamond P. (1988) Fuzzy least squares. *Information Sciences* 46: 141–157.

[DK97] Diamond P. y Korner R. (1997) Extended fuzzy linear models and least squares estimates. *Computers Math. Applic.* 33(9): 15–32.

[DMV04] Donoso S., Marín N. y Vila M. (2004) Sistemas de regresiones posibilísticas: una alternativa para la inferencia ecológica. En *Estylf 2004 - XII Congreso Español sobre Tecnologías y Lógica Fuzzy*, páginas 479–484.

[DMV05] Donoso S., Marín N. y Vila M. A. (2005) Systems of possibilistic regression: A case study in ecological inference. *Mathware and Soft computing* XII(2-3): 169–184.

[DMV06a] Donoso S., Marín N. y Vila M. (2006) Fuzzy regression with quadratic programming: An application to financial data. *Lecture Notes in Computer Science* 4224: 1304–1311.

[DMV06b] Donoso S., Marín N. y Vila M. (2006) Quadratic programming models for fuzzy regression. En *International Conference on Mathematical and Statistical Modeling*.

[DP80a] Dubois D. y Prade H. (1980) *Fuzzy Sets and Systems: Theory and Applications*. Academic Press.

[DP80b] Dubois D. y Prade H. (1980) *Fuzzy sets and systems. Theory and Applications*. Academic Press Inc.

[DP83] Dubois D. y Prade H. (1983) Ranking fuzzy numbers in setting of possibility theory. *Information Science* 30: 183–224.

[DP88] Dubois D. y Prade H. (1988) *Possibility Theory. An Approach to Computerized Processing of Uncertainty*. Plenum Press.

[D'U03] D'Urso P. (2003) Linear regression analysis for fuzzy o crisp input and fuzzy o crisp output data. *Computational Statistics and Data Analysis* 42(1-2): 47–72.

[EH03] Engel E., y Hernando A. (2003) Chile: ¿dos o más bloques? *Perspectivas en Política, Economía y Gestión* 6: 219–230.

[Eng90] Engel E. (1990) Evolución del comportamiento electoral desde el plebiscito a la elección presidencial. *Colección de Estudios CIEPLAN* 28: 73–83.

[Fel71] Feller W. (1971) *An Introduction to Probability Theory and Its Applications*. John Wiley and Sons.

[Goo53] Goodman L. A. (1953) Ecological regressions and the behavior of individuals. *American Sociological Review* 18(6): 663–666.

[GP02] Grigoriev A. y Popov A. (Sep 2002) Hypothesis testing in fuzzy regression analysis. *International Conference on Actual Problems of Electronic Instrument Engineering Proceedings* páginas 230–232.

[GT06] Guo P. y Tanaka H. (2006) Dual models for possibilistic regression analysisin press.

[HbLZH98] Huang L., bai Ling Zhang y Huang Q. (1998) Robust interval regression analysis using neural networks. *Fuzzy Sets and Systems* 97:3: 337–347.

[HBS05] Hojati M., Bector C. y Smimou K. (2005) A simple method for computation of fuzzy linear regression. *European Journal of Operational Research* 166: 172–184.

[HH04] Hong D. H. y Hwang C. (2004) Ridge regression procedure for fuzzy models using triangular fuzzy numbers. *Fuzziness and Knowledge-Based Systems* 12:2: 145–159.

[HHA04] Hong D. H., Hwang C. y Ahn C. (2004) Ridge estimation for regression models with crisp input and gaussian fuzzy output. *Fuzzy Sets and Systems* 142:2: 307–319.

[HK70] Hoerl A. E. y Kennard R. (1970) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12: 55–67.

[Hon06] Hong D. H. (2006) Fuzzy measures for a correlation coeficient of fuzzy numbers under tw (the wakest t-norm)- based fuzzy arithmetic operations. *Information Sciences* páginas 150–160.

[HSD01] Hong D. H., Song J.-K. y Do H. Y. (2001) Fuzzy least-squares linear regression analysis using shape preserving operations. *Information Sciences* 138: 185–193.

[HT90] Hayashi I. y Tanaka H. (1990) The fuzzy gmdh algorithm by possibility models and its application36.

[HTF01] Hastie T., Tibshirani R. y Friedman J. (2001) *The Elements of Statistical Learning: data mining, inference, and prediction*. Springer.

[IKW03] Ip K. W., Kwong C. y Wong Y. (2003) Fuzzy regression approach to modelling transfer moulding for microchip encapsulation. *Journal of Materials Processing Technology* 140(1-3): 147–151.

[IN96] Ishibuchi H. y Nii M. (1996) Fuzzy regression analysis by neural networks with non symmetric fuzzy number neights. *IEEE International Conference on Neural Networks* 2: 1191–1196.

[IT92] Ishibuchi H. y Tanaka H. (1992) Fuzzy regression analysis using neural networks. *Fuzzy Sets and Systems* 50: 257–265.

[KB98] Kim B. y Bishu R. R. (1998) Evaluation of fuzzy linear regression models by comparison membership function. *Fuzzy Sets and Systems* 100: 343–352.

[KC02] Kao C. y Chyu C.-L. (2002) A fuzzy linear regression model with better explanatory power. *Fuzzy Sets and Systems* 126: 401–409.

[KC03] Kao C. y Chyu C.-L. (2003) Least-squares estimates in fuzzy regression analysis. *European Journal of Operational Research* 148: 426–435.

[Kin97] King G. (1997) *A solution to the ecological inference problem: reconstructing individual behavior from aggregate data*. Princeton University Press.

[KL05] Kao C. y Lin P.-H. (Nov 2005) Entropy for fuzzy regression analysis. *International Journal of Systems Science* 36(14): 869–876.

[KMK96] Kim K. J., Moskowitz H. y Koksalan M. (1996) Fuzzy versus statistical lineal regression. *European Journal of Operational Research* 92: 417–434.

[KTKF] Kaneyoshi M., Tanaka H., Kamei M. y Furuta H.New system identification technique using fuzzy regression analysis.

[KY95a] Klir G. J. y Yuan B. (1995) *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice Hall.

[KY95b] Klir G. J. y Yuan B. (1995) *Fuzzy sets and fuzzy logic: theory and applications*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[LT72] Luca A. D. y Termini S. (1972) A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. *Information and Control* 20: 301–312.

[MK93] Moskowitz H. y Kim K. (1993) On assessing the h value in fuzzy linear regression. *Fuzzy Sets and Systems* 58: 303–327.

[MKIK02] Mori H., Kossemura N., Ishiguro K. y Kondo T. (2002) Short-term load forecasting with fuzzy regression tree in power systems. En *International Conference on Actual Problems of Electronic Instrument Engineering Proceedings*, páginas 230–232.

[MNN04] Modarres M., Nasrabadi E. y Nasrabadi M. M. (2004) Fuzzy linear regression analysis from the point of view risk. *International Journal of Uncertainty, Fuzziness and Konwledge-Based Systems* 12(5): 635–649.

[MP03] Mogilenko A. y Pavlyuchenko D. (2003) Development of fuzzy regression models using genetic algorithms. *International Journal of Uncertainty, Fuzziness and Konwledge-Based Systems* 11(4): 429–444.

[NZ99] Nazarko J. y Zalewski W. (1999) The fuzzy regression approach to peak load estimation in power distribution systems. *IEEE Transactions on Power Systems* (3): 6809–814.

[OD00] Ozelkan E. C. y Duckstein L. (2000) Multi-objetive fuzzy regression: a general framework. *Computers and Operations Research* 27: 635–652.

[OS03] Oussalah M. y Schutter J. D. (2003) Approximated fuzzy lr computation. *Information Sciences* 153: 155–175.

[Pet94] Peters G. (1994) Fuzzy linear regression with fuzzy intervals. *Fuzzy Sets and Systems* 63: 45–55.

[PGBSZ95] Pawlak Z., Grzymala-Busse J., Slowinski R. y Ziarko W. (1995) Rough sets. *Communications of the ACM* 38(11): 88–95.

[RW96] Redden D. T. y Woodall W. H. (1996) Further examination of fuzzy linear regression. *Fuzzy Sets and Systems* 79: 203–211.

[Sar94] Sarle W. S. (1994) Neural networks and statistical models. *Proceeding 9th. anual SAS user group international conference* .

[SATEH02] Soliman S. A., Alammari R. A., Temraz H. K. y El-Hawary M. (2002) Fuzzy linear parameter estimation algorithms: a new formulation. *International Journal of Electric Power and Energy Systems* 24(5).

[SB75] Shortliffe E. y Buchanan B. (1975) A model of inexact reasoning in medicine. *Mathematical Biosciences* 23: 351–379.

[Sch76] Schafer G. (1976) *A Mathematical Theory of Evidence*. Princeton University Press.

[SIT04a] Sugihara K., Ishii H. y Tanaka H. (2004) Interval priorities in ahp by interval regression analysis. *Europeian Journal of Operatin Research* 158: 745–754.

[SIT04b] Sugihara K., Ishii H. y Tanaka H. (2004) Interval priorities in ahp by interval regression analysis. *European Journal of Operational Research* 158: 745–754.

[SJ02] Schjaer-Jacobsen H. (2002) Representation and calculation of economic uncertainties: intervals, fuzzy numbers and probabilities. *International Journal of Production Economics* 78: 91–98.

[SP91] Savic D. A. y Pedrycz W. (1991) Evaluation of fuzzy lineal regression models. *Fuzzy Sets and Systems* 23: 51–63.

[SP92] Savic D. A. y Pedrycz W. (1992) Fuzzy lineal regression models: construction and evaluation. En Kacprzyk J. y Fedrizzi M. (Eds.) *Fuzzy regression Analysis*, páginas 91–100.

[SREH97] Soliman S. A., Rahman M. H. A. y El-Hawary M. (1997) Applicatin of fuzzy linear regression algoritm to power system voltage measurements. *Electric Power Systems Research* 42(3).

[SY92] Sakawa M. y Yano H. (1992) Fuzzy linear regression and its applications. En Kacprzyk J. y Fedrizzi M. (Eds.) *Fuzzy regression Analysis*, páginas 61–80.

[Tan87] Tanaka H. (1987) Possibilistic regression analysis based on linear programming24.

[TAT95] Trillas E., Alsina C. y Terricabras J. M. (1995) *Introducción a la lógica borrosa*. Ariel Matemática.

[TI91] Tanaka H. y Ishibuchi H. (1991) Identification of possibilistic lienar systems by quasratic membership functions of fuzzy parameters. *Fuzzy Sets and Systems* 41: 145–160.

[TI92] Tanaka H. y Ishibuchi H. (1992) Possibilistic regression analysis based on linear programming. En Kacprzyk J. y Fedrizzi M. (Eds.) *Fuzzy regression Analysis*, páginas 47–60.

[TIY95] Tanaka H., Ishibuchi H. y Yoshikawa S. (1995) Exponential possibilistic regression analysis. *Fuzzy Sets and Systems* 69: 305–318.

[TKL96] Tanaka H., Koyama K. y Lee H. (1996) Interval regression analysis based on quadratic programming. *IEEE Trans. on Fuzzy Systems* páginas 325–329.

[TL98] Tanaka H. y Lee H. (1998) Interval regression analysis by quadratic programming approach. *IEEE Trans. on Fuzzy Systems* 6(4).

[TL99] Tanaka H. y Lee H. (1999) Exponential possibility regression analysis by identification method of possibilistic coefficients. *Fuzzy Sets and Systems* 106: 155–165.

[TLss] Tseng F.-M. y Lin L. (In press) A quadratic interval logit model for forescasting bankruptcy. *The International Journal of Management Science* .

[TSWA87] Tanaka H., Shimomura T., Watada J. y Asai K. (1987) Fuzzy lineal regression analysis of the number of staff in local boernment. En Bezdele J. C. (Ed.) *Analysis of fuzzy information, Vol. III: Application in Engineering and Science*, páginas 191–203.

[TUA80] Tanaka H., Uejima S. y Asai K. (1980) Fuzzy linear regression model. *International Congress on Applied Systems Research and Cybernetics. Acapulco. Mexico* .

[TUA82] Tanaka H., Uejima S. y Asai K. (1982) Linear regression analysis with fuzzy model. *IEEE Trans. on Systems, Man, and Cybernetics* 12(6): 903–907.

[TW88] Tanaka H. y Watada J. (1988) Possibilistic linear systems and their application to the linear regression model. *Fuzzy Sets and Systems* 27(3): 275–289.

[TW99] Tsaur R.-C. y Wang H.-F. (1999) Outliers in fuzzy regression analysis. *International Journal of Fuzzy Systems* 1(2): 113–119.

[TWY02] Tsaur R.-C., Wang H.-F. y Yang J.-C. O. (2002) Fuzzy regression for seasonal time series analysis. *International Journal of Information Technology and Decision Making* 1(1): 165–175.

[UW05] Urbanski M. K. y Wasowski J. (2005) Fuzzy arithmetic based in boundary weak t-norm. *Internacional Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 1: 27–37.

[Vap98] Vapnik V.Ñ. (1998) *Statistical Learning Theory*. John Willey & Sons, Inc.

[Wat00] Watada J. (2000) Analysis of software reliability by fuzzy regression model. *TENCON 2000. Proceedings* 3: 439–445.

[Wat01] Watada J. (2001) The thought and model of linguistic regression. *IFSA Wordld Congress* 5: 2955–2959.

[Wel01] Wellhofer E. S. (2001) Party realignment and voter transition in italy, 1987-1996. *Comparative Political Studies* 34(2): 156–186.

[WL99] Wen C.-G. y Lee C.-S. (1999) Development of a cost function for wastewater treatment systems with fuzzy regression. *Fuzzy Sets and Systems* 106(2).

[WT00a] Wang H.-F. y Tsaur R.-C. (2000) Bicriteria variable selection in a fuzzy regression equation. *Computers and Mathematics with Applications* 40: 877–883.

[WT00b] Wang H.-F. y Tsaur R.-C. (2000) Insight of a fuzzy regression model. *Fuzzy Sets and Systems* 112: 355–369.

[WT02] Wu B. y Tseng N.-F. (2002) A new approach to fuzzy regression models with application to business cycle analysis. *Fuzzy Sets and Systems* 130(1): 33–42.

[Wu03] Wu H.-C. (2003) Fuzzy estimates of regression parameters in linear regression models for imprecise input and output data. *Computational Statistics and Data Analysis* 42: 203–217.

[WY94] Watada J. y Yabuuchi Y. (1994) Fuzzy robust regression analysis. *IEEE World Congress on Fuzzy Systems* 2: 1370–1376.

[XL01] Xu R. y Li C. (2001) Multidimensional least-squares fitting with a fuzzy model. *Fuzzy Sets and Systems* 119: 215–223.

[YC04] Yang M.-S. y Chen H.-M. (2004) Fuzzy class logistic regression analysis. *International Journal of Uncertainty, Fuzziness and Konwledge-Based Systems* 12(6): 761–780.

[YK97] Yang M.-S. y Ko C.-H. (1997) On cluster-wise fuzzy regression analysis. *IEEE Transactions on systems, man and cybernetics* 27(1): 1–13.

[YTL99] Yu J.-R., Tzeng G.-H. y Li H.-L. (1999) A general piecewise necessity regression analysis based on linear programming. *Fuzzy Sets and Systems* 105: 429–436.

[YTL01] Yu J.-R., Tzeng G.-H. y Li H.-L. (2001) General fuzzy piecewise regression analysis with automatic change-point detection. *Fuzzy Sets and Systems* 119: 247–257.

[Zad65] Zadeh L. A. (1965) Fuzzy sets. *Information and Control* 8: 338–353.

[Zad75a] Zadeh L. A. (1975) The concept of linguistic variable and its application to aproximate reasoning I. *Information Sciences* 8: 199–251.

[Zad75b] Zadeh L. A. (1975) The concept of linguistic variable and its application to aproximate reasoning II. *Information Sciences* 8: 301–357.

[Zad75c] Zadeh L. A. (1975) The concept of linguistic variable and its application to aproximate reasoning III. *Information Sciences* 9: 43–80.

[ZCB87] Zwick R., Carlstein E. y Budescu D. V. (1987) Measures of similarity among fuzzy concepts: A comparative analysis. *International Journal of Aproximate Reasoning* 1: 221–242.