Mathware & Soft Computing 12 (2005) 169-184

Systems of Possibilistic Regressions: A Case Study in Ecological Inference

S. Donoso[1], N. Marín[2] and M.A. Vila[2]
[1]UTEM - Chile
[2] University of Granada - Spain
sdonosos@vtr.net and {nicm,vila}@decsai.ugr.es

Abstract

This work introduces how possibilistic regression can be used in the case of non symmetrical triangular membership functions, building a system of regressions, so that suitable restrictions for each particular problem can be incorporated. We apply this methodology to the problem of ecological inference, in particular to the estimation of the electoral transition matrix. An experimentation with several examples shows the benefits of the new approach.

1 Introduction

Fuzzy regression techniques date back to 1982[11]. In this particular kind of regression, the coefficients of the regression's model are fuzzy numbers with symmetrical membership functions and the objective is the minimization of a linear function. A lot of works have been devoted to the improvement of this methodology, incorporating the fuzzy regression of least squares. In this work, we propose to use a set of regressions, which we call *system of possibilistic regressions*, instead of a unique regression. Special emphasis is paid to the use of coefficients with non symmetrical membership functions in order to ease the practice of fuzzy regression.

We apply these techniques to the ecological inference problem. Ecological inference [9] takes as its objective the estimation of individual behaviors from aggregated information. The political preferences of electors in between two electoral processes are a typical example of ecological inference. Due to the secret character of the suffrage, the particular behavior of an individual voter is not known. However, there is no doubt about the utility of the estimation of this behavior for the political science in general and for the electoral studies in particular. The *ecological inference* constitutes a statistical problem that has been studied for more than half a century, but conclusive results have not been obtained up to now. Provided that this problem is very complex from the statistical point of view, the works that can be found in the literature have focused on the estimation of 2x2 tables [6, 9, 10]. In this work we introduce a new methodology by applying the concept of possibilistic regression to the estimation of *electoral transition's matrixes* without this last restriction (we will work on pxq tables).

The paper is organized as follows: section 2 contains a brief summary of the possibilistic regression and its use with non symmetrical membership functions. Section 3 analyzes the problem of *ecological inference* and the estimation of *electoral transition's matrixes*. Section 4 introduces the concept of *system of possibilistic regressions* and its application to the estimation of the *electoral transition's matrixes*. Sections 5 and 6 present some examples of the methodology. Finally, section 7 ends the paper with some concluding remarks.

2 Possibilistic regression

The equation 1 shows the linear multiple model of the probabilistic regression for n observations, where coefficients β are unknown:

$$Y_i = \beta_1 X_{i1} + ... + \beta_m X_{im} + \epsilon_i \tag{1}$$

This model was re-formulated by Tanaka [11] so that values of X are crisp and values of coefficients β are fuzzy sets with a symmetrical triangular membership function $A_j = (a_j, c_j)$, and the values of Y can be also fuzzy with membership function $Y_i = (y_i, e_i)$. In this approach, the random error ϵ_i is replaced by the imprecision in the coefficients, turning out to be the following problem of linear programming:

$$Min \sum_{i=1}^{n} (\sum_{j=1}^{m} c_j |X_{ij}|) \tag{2}$$

subject to the possibilistic restriction that the membership function of output Y_i is contained in the membership function of the estimation $\widetilde{Y}_i$:

$$Y_i \subseteq \widetilde{Y}_i \tag{3}$$

which can be formally expressed by means of the following set of inequalities:

$$\sum_{j=1}^{m} a_j X_{ij} - (1-h) \sum_{j=1}^{m} c_j |X_{ij}| \leq y_i - (1-h) e_i;\ i = 1..n \tag{4}$$

$$\sum_{j=1}^{m} a_j X_{ij} + (1-h) \sum_{j=1}^{m} c_j |X_{ij}| \geq y_i + (1-h) e_i;\ i = 1..n \tag{5}$$

$$c_j \geq 0;\ j = 1, ..., m \tag{6}$$

with a degree of belief h for the estimated Y_i

$$\mu(Y_i) \geq h;\ 0 \leq h \leq 1 \tag{7}$$

2.1 Use of asymmetrical membership functions

The previous formulation, obtained using the Zadeh's Extension Principle[1], can be extended to estimate fuzzy regression coefficients with non symmetrical membership functions $A_i = (a_i, c_i^L, c_i^R)$ by means of the following problem of linear programming:

$$Min \sum_{i=1}^{n} (\sum_{j=1}^{m} ((c_j^L + c_j^R)|X_{ij}|) \tag{8}$$

subject to the restrictions:

$$\sum_{j=1}^{m} a_j X_{ij} + (1-h) \sum_{j=1}^{m} c_j^R |X_{ij}| \geq y_i + (1-h)e_i;\ i = 1..n \tag{9}$$

$$\sum_{j=1}^{m} a_j X_{ij} - (1-h) \sum_{j=1}^{m} c_j^L |X_{ij}| \leq y_i - (1-h)e_i;\ i = 1..n \tag{10}$$

$$c_j^R,\ c_j^L \geq 0;\ j = 1, ..., m \tag{11}$$

where for the m X variables, $3m$ parameters have to be estimated. If the output variable is a fuzzy set with non symmetrical membership function LR (has left function L and right function R), denoted $Y_i = (y_i, p_i, q_i)$, the restrictions of the problem have the following formulation (for a given degree of belief h):

$$\sum_{j=1}^{m} a_j X_{ij} + R^{-1}(h) \sum_{j=1}^{m} c_j^R |X_{ij}| \geq y_i + R^{-1}(h) q_i;\ i = 1..n \tag{12}$$

$$\sum_{j=1}^{m} a_j X_{ij} - L^{-1}(h) \sum_{j=1}^{m} c_j^L |X_{ij}| \leq y_i - L^{-1}(h) p_i;\ i = 1..n \tag{13}$$

$$c_j^R, c_j^L \geq 0;\ j = 1..m \tag{14}$$

The use of asymmetrical membership functions in the possibilistic regression produces more precision in the estimation than in the case of symmetrical ones. As we will see, this benefit is particularly significant in the problem of the *ecological inference*, where the range of this estimation is truncated within the interval [0,1]). We use, in our examples, asymmetrical triangular membership functions with h=0.

The main criticism that the possibilistic regression has received is that it is very sensitive to the effect of extreme points (outliers). In fact, an analysis of extreme points is highly recommended in order to obtain an appropriate diagnosis[12]. The use of asymmetrical membership functions attenuates this effect on the estimation of the central or modal value of the fuzzy coefficients.

3 Ecological regression

Usually, we do not have information about the individuals at the moment of doing estimations. The information is aggregated according to some grouping criterium. For example, electoral results are often offered by neighborhoods, districts, etc., but the particular vote of a given person is not known. Information of census has a similar problem.

In the electoral case, if we want to guess if a given party's electorate has changed with respect to the past election, it is necessary to do inference on the level of individuals. And this is the objective of what is called *ecological inference*: to obtain information about the individuals from aggregated information. In this context, to affirm that the relations obtained in the level of groups are directly supported in the level of individuals is known as *ecological deceit* .

In 1953, Goodman [9] proposed the *ecological regression* as a solution for this problem: an ordinary least square regression, with a requirement called *supposition of homogeneity*: all the conditional probabilities are equal in the diverse units of information. This method has been applied in many studies, for example, in Engel's works [7, 8]. Duncan [6] estimated an interval as solution, which is called *the solution of edges*. In 1997, King [10] introduced a new approach in which the regression's coefficients are random; this approach has been applied by Wellhofer [13] for several Italian elections and by Burden [3] for simultaneous elections in the United States. Cho [5, 4] has criticized King's model, since his suppositions are not frequently fulfilled (though they are less strict than Goodman's ones). He also proposed some alternatives to improve the previous two models.

3.1 Estimation of electoral transition's matrixes

Our problem of estimation of *electoral transition's matrixes* complies with the following scheme:

- We take two elections (in general consecutive), namely, the election 1 and the election 2.
- Election 2 can be either a simultaneous election to choose another political position (Senate and Deputies' Chamber; President of the Republic, Regional Parliament, etc.) or it can be the run-off of the first election. It can also be the following political election like, for example, the general Spain elections of 2000 and 2004.
- In the first election, p political parties appear, each of which obtains X_{ik} votes in every precinct k (there are n precincts).
- In the second election, q political parties appear, each of which obtains Y_{jk} votes per precinct.

Taking this into account, the conditional probability for precinct k:

$$\beta_{jik} = Prob(V_{j,2}|V_{i,1}) \tag{15}$$

where $V_{s,t}$ stands for "An elector voted for party s in the election t".

Due to the secret character of the suffrage, it is not possible to know these probabilities. To estimate them, we can write the relations

$$Y_{jk} = \sum_{i=1}^{p} \beta_{jik} X_{ik} \tag{16}$$

where there are pxqxn values to estimate and only qxn equations, and more than one solution will be obtained. If we suppose that the values are constant in the different units i, that is

$$\beta_{ji} = \beta_{jik};\ k = 1..n \tag{17}$$

then it is possible to do some type of statistical estimation for the coefficients. This supposition is known as *supposition of homogeneity* in the ecological regression. If this supposition is not fulfilled, then Goodman's regression, previously mentioned, produces incorrect results. In this case, the objective of the *electoral transition's matrix* is to estimate the weighted average of the β_{jik}, instead of the particular β_{jik}. In [10], King does not demand that the coefficient is constant, but it imposes that it is the central value of certain probabilistic distribution, having a certain variance and covariance.

3.2 Conventional ecological estimation

The problem of *ecological inference* has been considered as a 2x2 table, where X is the proportion of votes for the party A in the first election, T is the proportion of votes for the party B in the second election, and the conditional probabilities β^B and β^W are the values to estimate, as average value of all the units of information. β^B represents the proportion of voters that chose party A in the first election but chose party B in the second election while β^W represents the proportion of voters that did not chose party A in the first election and chose party B in the second election.

Table 1: Problem of ecological inference

Election 1	Election2		
	Party B	Others	Total
Party A	β^B	1-β^B	X
Others	β^W	1-β^W	1-X
Total	T	1-T	1

King[10] made considerable advances in the field of the ecological regression, but his model is based on three suppositions that cannot be fulfilled. Cho [5, 4] has emphasized the limitations of Goodman and King's approaches and has proposed a test to measure the effects of other variables in the behavior of the electors [4]. He has also proposed (together with Anselin) to incorporate the spatial effect of the different precincts [2], in the sense that territorial neighboring units can generate covariance between β^B and β^W. The estimation of ordinary least square of β^B and β^W can produce values out of the interval [0,1] (see [2] and table 4 of this paper). It is a sample of what is called "bias of aggregation". It is less frequent that the estimation of the fuzzy regression (8) - (11) gives values out of range. Nevertheless, with the flexibility of the possibilistic model for adding new restrictions, we will show that this bias can be corrected.

When the coefficients of least square are out of range, it is convenient to do an analysis of the data so that we can divide the estimation in two groups. To do it, we can consider variables such as the size of the units of information, the size of X and T, or other variables[4] as, for example, the socio-economic level of every geographical unit of information. This way it is possible to do the estimation independently in every group, and then to calculate the coefficients of the set by means of weighted averages. Cho in his work presents two tests to determine wether the variable really discriminates between groups of observations or not [14].

The fact of considering only 2x2 tables makes the estimation of many electoral transition's matrixes tedious [13].

4 A fuzzy approach for ecological regression with a system of possibilistic regressions

If we want to estimate the *electoral transition's matrix* (16), the homogeneity restriction is too strict, given the natural variability that takes place between districts. This fact makes the model of Goodman non suitable. In order to face this problem, though our Xs and Ys are crisp values, we assume that the coefficients can be fuzzy sets (see Figure 1). Thus, we have

$$\beta_{ji} = (a_{ji}, c^L_{ji}, c^R_{ji}) = A_{ji} \tag{18}$$

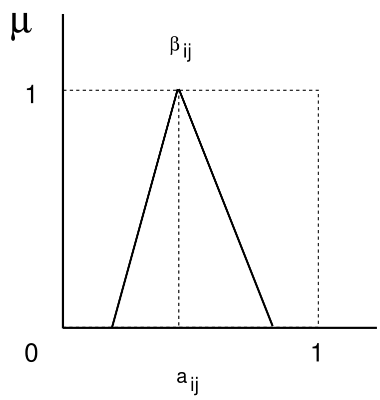

Figure 1: Coefficient β_{ij} as a fuzzy set

For each one of the q options in the second election, we have k observations in the electoral districts, forming the following model with conditional probabilities β_{ji}:

$$Y_{jk} = \sum_{i=1}^{p} \beta_{ji} X_{ik} \; for \; j = 1, ..., q \; and \; k = 1, ..., n \tag{19}$$

Therefore, the model of fuzzy regression, as a problem of linear programming, can be expressed as follows:

$$Min \sum_{j=1}^{q} \sum_{k=1}^{n} (\sum_{i=1}^{p} (c_{ji}^{L} + c_{ji}^{R}) |X_{ik}|) \tag{20}$$

subject to the possibilistic conditions

$$\sum_{i=1}^{p} a_{ji} X_{ik} + (1-h) \sum_{i=1}^{p} c_{ji}^{R} |X_{ij}| \geq y_{ik}; \; k = 1..n \; and \; j = 1, ...q \tag{21}$$

$$\sum_{i=1}^{p} a_{ji} X_{ik} - (1-h) \sum_{i=1}^{p} c_{ji}^{L} |X_{ij}| \leq y_{ik}; \; k = 1..n \; and \; j = 1, ...q \tag{22}$$

Taking into account that the coefficients represent conditional probabilities, its maximum value of membership (center value of the fuzzy coefficients) must be between 0 and 1. A new restriction is:

$$a_{ji} \leq 1 \text{ and } a_{ji} \geq 0; \; j = 1...q, i = 1...p \tag{23}$$

Additionally, if we want to incorporate the limitations of edge for the estimation, in form more strict than (23), we need the following additional restrictions

$$a_{ji} + c_{ji}^{R} \leq 1 \text{ and } a_{ji} - c_{ji}^{L} \geq 0; \; j = 1..q, i = 1..p \tag{24}$$

Nevertheless, when the estimation of the central coefficient is close to 0 or to 1, the membership function is forced to be 0, which does not seem reasonable. For example, see figure 2 for a coefficient close to 1. This is the reason why we prefer to use (23) (see figure 3).

To deal with this fact, specially for predictive aims, truncated membership functions are used in order to be able to keep the membership function within the interval [0, 1] without imposing membership value 0 in the extremes (see figure 4).

The q conditional probabilities of the voters of a party in the first election who voted for the diverse parties in the second election must add 1:

$$\sum_{j=1}^{q} a_{ij} = 1 \; i = 1, ..., p \tag{25}$$

which involves not only one of the regressions but the whole set of regressions(q).

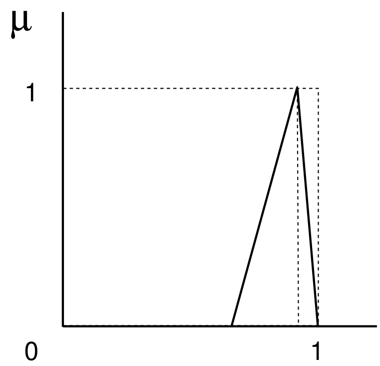

Figure 2: Coefficient subject to restriction 24

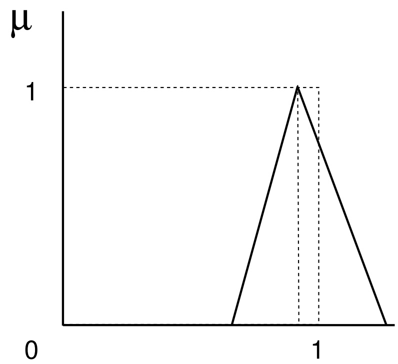

Figure 3: Coefficient subject to restriction 23

For this reason, it is necessary to form with the q problems of the form (8)-(11) a unique problem of linear programming (LP) (20)-(22), which will have a whole of 3pq values to estimate[1].

Therefore, it is suitable to set the additional condition

$$\sum_{i=1}^{p} a_{ij} (\sum_{k=1}^{n} X_{ki}) = \sum_{k=1}^{n} Y_{kj} \quad for\ j = 1, ..., q \tag{26}$$

or the equivalent relation

$$\sum_{i=1}^{p} a_{ij} (avg(X_i)) = avg(Y_j) \quad for\ j = 1, ..., q \tag{27}$$

[1]The use of a software as MatLab easies the implementation of a problem of mathematic programming of major size.

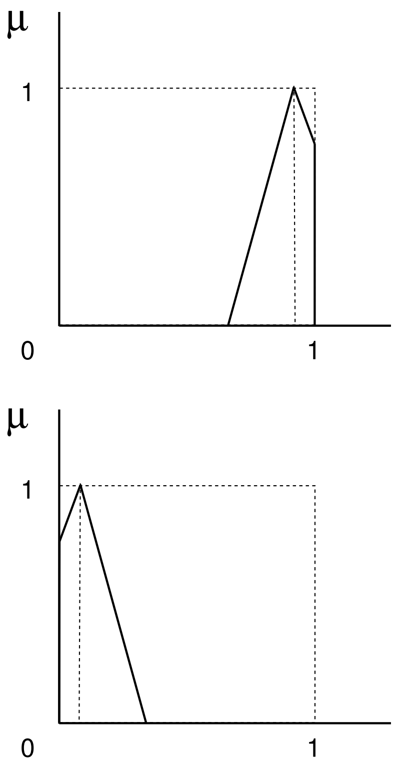

Figure 4: Truncated membership functions closed to 1 and 0

for each of the q parties of the second election.

For the enunciated mathematical programming problem, if the restrictions of each q original regressions have the form

$$AZ \leq b_i \tag{28}$$

then the restrictions for the system of fuzzy regressions have, for example, the following form in the case of $q = 4$ parties in the second election:

$$\left| \begin{array}{cccc} \text{A} & 0 & 0 & 0 \\ 0 & \text{A} & 0 & 0 \\ 0 & 0 & \text{A} & 0 \\ 0 & 0 & 0 & \text{A} \\ & C_{conj} & & \end{array} \right| \quad \left| \begin{array}{c} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_{conj} \end{array} \right|$$

where the coefficient's matrix (C_{conj}) and the vector of the right side (b_{conj}) represent the joint restrictions. Matrixes C_{conj} and b_{conj} contain the particular restrictions of the system of equations, for example (26) or (27).

Similarly, if new restrictions arise in a particular problem of estimation, either among the coefficients or through the addition of new variables that can complement the electoral behavior of the citizens, these new restrictions can be added to the formulation of the problem of mathematic programming.

In general, a system of possibilistic regressions allows to incorporate particular conditions of every model, as for example, information obtained from independent surveys of electoral data, which produces more precise estimations of regression.

To evaluate the quality of the adjustment of the fuzzy estimation, the sum of the absolute value of the differences between the value to estimate and the estimated value is considered:

$$Dif = \sum_{k=1}^{n}\sum_{j=1}^{q}|Y_{jk} - (\sum_{i=1}^{p} a_{ji}X_{ik})| \tag{29}$$

The fact that Dif increases from an estimation to another is an indicator of deviation from the central trend in the estimation of the coefficients, and will have to be considered when diverse models are available.

5 Examples of ecological regressions

Let us begin with the case of a table 2x2 in which the coefficients to estimate are known so that we can contrast the obtained results with those of Goodman's regression and the King's model. We will take the following alternatives of fuzzy regression for the system of equations:

- Fuzzy SI 1: The model only includes the restrictions of coefficients in the interval [0,1] (23).
- Fuzzy SI 2: Coefficients are in the interval [0,1] and restriction (25), namely, the sum of the coefficients of every party of the first election must add 1.
- Fuzzy SI 3: Fuzzy model SI 2 plus the condition (26).

The example 1 considers a relatively common situation in which a coefficient is high and the other coefficient is low. 40 data were generated with a standard deviation for the dependent variable of 0.179.

Table 2: Comparisons of results, example 1

Method	β^B	β^W	error
REAL	0,9390	0,1231	0
Goodman	0,9905	0,0756	0,0990
King's EI	0,9703	0,0934	0,0610
Fuzzy SI 1	0,9799	0,1364	0,0542
Fuzzy SI 2	0,9669	0,1695	0,0743
Fuzzy SI 3	0,9742	0,0899	0,0684

As can be observed in table 2, Goodman's method is the one that produces a major difference between the estimated values and the real ones. Fuzzy method SI 1 performs better than EI. The fuzzy method SI 2 produces the least correct results.

The second example is a 4x2 table. The actual values of the *electoral transition's matrix* are not known, so we will have to analyze the results in terms of interpretation. This example corresponds to Amoostook's county, in the State of Maine, United States: simultaneous elections of Governor, Senator, Representatives, and other positions, in November, 2002 (http://www.state.me.us/sos/cec/elec/prior.htm). The total values for Senator and Governor, by political party, are in table 3.

Table 3: Senator and Governor totals, for each party

	sen rep	sen dem	total
gover rep	?	?	8.401
gover dem	?	?	15.783
gover green	?	?	1.279
gover indep.	?	?	812
total	17.767	8.408	26.275

In the case of senators, the republican candidate wins, while in the case of governor the democratic candidate is the most voted. In both situations the margin is wide, so to know the *electoral transition's matrix* turns out to be very interesting. The information consists of 71 precinct in the county, and the principal limitation of the information is that electors' number in every table is not uniform: from a minimum of 16 electors to a maximum of 3060 electors, with a median of 153 electors, an average of 370, and a standard deviation of 585. Table 4 shows Goodman's estimation for ordinary least squares regressions.

Table 4: Goodman's estimation for example 2

	sen rep	sen dem	total
gover rep	1,0996	-0,0996	8.401
gover dem	0,439	0,561	15.783
gover green	0,6	0,4	1.279
gover indep.	-0,78	1,78	812
total	17.767	8.408	26.275

As can be observed, values out of the range [0,1] appears, what invalidates the results. Even truncating the values out of range, the estimation will be slightly reliable. For King's estimation, doing 4 estimations 2x2 (King recommends to extend his model to 3x2 tables as much, but we test the robustness of the model forcing 4x2 table) we obtain the results of the table 5.

This estimation has reasonable coefficients for the relation between the democratic and republican candidates, but coefficients for the independent and green

Table 5: King's estimation (EI) for the example 2

	sen rep	sen dem	total
gover rep	0,9495	0,0505	8.401
gover dem	0,5096	0,4904	15.783
gover green	0,9942	0,0058	1.279
gover indep.	0,0265	0,9735	812
total	17.767	8.408	26.275

candidates are extreme: it is very strange that the republican candidate coefficient is lower than the green one; similarly, it is very rare that the winning candidate for senator with the most wide advantage, does not obtain any vote of those who voted for the independent candidate for governor. The estimation for the whole voting of the republican candidate for senator is 17013, which means an underestimation of 500 votes. The estimation of all the fuzzy systems were performed only once, as a table 4x2.

The results for the model SI 1 are shown in table 6. This estimation has an absolute difference of 2366 and gives more reasonable values for those who voted for the green and independent candidates, but, in contrast, the estimation of the whole voting of the republican candidate to senator is 16057, underestimating the total voting in almost 1800 votes.

The estimation with the fuzzy system SI 2 produces an absolute difference of 2703, which too high and we do not consider the results significant enough.

Table 6: Fuzzy estimation SI 1 for example 2

	sen rep	sen dem	total
gover rep	0,9586	0,0415	8.401
gover dem	0,4595	0,5402	15.783
gover green	0,1953	0,8047	1.279
gover indep.	0,6175	0,3825	812
total	17.767	8.408	26.275

Table 7 shows the estimation of model SI 3. This estimation decreases the absolute difference to 1410 and the value estimated for the whole of the voting of the republican candidate to senator is 17866, due to the estimation condition, much more exact than the solution of King's method EI.

According to the quality of the adjustment, the estimation of the fuzzy model SI 3 is the most suitable in order to compute the *electoral transition's matrix*. As was said, the estimation SI 1, though very reasonable, underestimates the global estimation of the republican candidate to senator. This justifies the increase of value in 3 coefficients (specially the coefficient corresponding to green votes, which comes from 0,1953 to 0,63), between the estimations SI 1 and SI 3, to compensate the indicated underestimation.

The complete estimation for the coefficients of the republican candidate to senator with model SI 3, are in the table 8.

Table 7: Estimation SI 3 for the example 2

	sen rep	sen dem	total
gover rep	0,9851	0,0149	8.401
gover dem	0,5248	0,4752	15.783
gover green	0,63	0,37	1.279
gover indep.	0,6175	0,3825	812
total	17.767	8.408	26.275

Table 8: Estimation of model SI 3 for republican senator

	center a_i	c_i^L	c_i^R
gover rep	0,9851	0,0650	0,2793
gover dem	0,5248	0,1734	0,0279
gover green	0,6300	1,2492	0,3847
gover indep.	0,6175	0,0000	0,0000

The extensions for the coefficient of republican and democratic governor are enough bounded, as well as the independent candidate coefficient, but the left extension of green candidate to governor far exceeds the bound 1. Thus, it would be necessary to think about a *triangular membership function* truncated to the limits that imposes the problem (in the ecological regression, the interval [0,1] for each coefficient). The *truncated triangular membership function* for the fuzzy coefficient of Green candidate, from Table 8, is defined like:

$$\begin{array}{cc} 0 & \text{if } \beta > 1 \text{ or } \beta < 0 \\ (\beta + 0.6192)/1.2492 & \text{if } 0 \leq \beta \leq 0.63 \\ (1.0147 - \beta)/0.3847 & \text{if } 0.63 < \beta \leq 1 \end{array} \tag{30}$$

6 Example of a estimation with system of possibilistic regressions

The last presidential election in Chile happened in December 1999. In January 2000, for the first time in the history of the country, the runoff of this election was done, resulting elected the current President of the Republic Ricardo Lagos. To compute the transition's matrix of these 2 elections, a table 4x3 was constructed, considering the three candidates with the best results. The categories for the first round are:

- Voting for the candidate Lavin
- Voting for the candidate Lagos
- Voting for the candidate Marin
- The remain (voting for other 3 candidates, null votes, votes in target, and abstention)

For the runoff, 3 categories were considered:

- Voting for the candidate Lavin
- Voting for the candidate Lagos
- The remain (null votes, white votes, and abstention).

The considered model corresponds to equation (20), subject to the possibilistic restrictions. We use 96 major counties of the 341 counties in which the country is divided, which corresponds to towns and cities with more than 20.000 citizens. An independent estimation was performed for every county, separated by gender.

We present here the transition's matrix of one county and the sum of votes of each candidate in the county (table 9).

Table 9: Presidential election in Chile 1999-2000. County: Lota

	Lavin	Lagos	Other	Sum of Votes
Lavin	0,917	0,000	0,084	3.019
Lagos	0,029	0,971	0,000	9.076
Marin	0,204	0,768	0,029	1.337
Other	0,092	0,133	0,775	2.826
Sum of Votes	3.304	10.622	2.332	16.258

Each of these estimated fuzzy coefficients, 4x3 in total, has a truncated membership function. The truncated triangular membership function for the coefficient LAGOS - LAGOS, for the estimated counties, is shown at table 10.

7 Conclusions

The possibilistic regression for fuzzy or crisp output data, with coefficients with non symmetrical membership function, can be organized as a system of regressions if the model is multiple in the number of equations. This way, the combined estimation of all the equations allows to incorporate additional conditions which can be independently added to every regression.

In this work, we have proved that the systems of fuzzy regressions allow the computation of the *electoral transition's matrix*, a typical example in *ecological inference*. Diverse groups of restrictions have been defined to be incorporated into the estimation of the system of possibilistic regressions(models SI 1 to SI 3). While the ecological regression has been mainly used in the case of 2x2 tables, our formulation to estimate the *electoral transition's matrix* allows the estimation of pxq tables.

The numerical examples show that the fuzzy methods defined with linear minimization perform reasonably compared to Goodman's method and King's method EI. In particular, the fuzzy model SI 3 seems to give better estimations for the real example of an election in Maine's State. The table shown for the Chilean's election gives a robust estimation in the interval [0,1] for a 4x3 table.

Table 10: Example of a truncated non symmetrical triangular membership function

COUNTY	$trun(a_i - c^L_i)$	a_i	$trun(a_i + c^D_i)$	$fp(trun(a_i + c^D_i))$ para 1	c^D_i
Ancud	0,8660	0,9659	1,0000	0,3685	0,0540
Angol	0,9525	0,9773	1,0000	0,8185	0,1251
Antofagasta	0,9751	0,9751	0,9751	0,0000	0,0000
Arica	0,9650	0,9850	0,9850	0,0000	0,0000
Buin	1,0000	1,0000	1,0000	1,0000	0,0770
Calama	0,9481	0,9481	0,9847	0,0000	0,0366
Calera	0,8416	0,9179	0,9774	0,0000	0,0595
Cauquenes	0,8644	0,9202	0,9202	0,0000	0,0000
Cerrillos	0,9886	0,9886	0,9886	0,0000	0,0000
Cerro Navia	0,8900	0,9603	1,0000	0,5458	0,0874
Chiguayante	0,9865	0,9865	0,9906	0,0000	0,0041
Chillán	0,9252	0,9810	0,9810	0,0000	0,0000
Coihaique	0,9497	0,9497	1,0000	0,0804	0,0547
Colina	0,8998	0,9449	1,0000	0,4320	0,0970
Concepción	0,9289	0,9824	0,9853	0,0000	0,0029
Conchalí	0,8546	0,8546	0,9063	0,0000	0,0517
Constitución	0,3664	0,9341	1,0000	0,4757	0,1257
Copiapó	0,9410	0,9648	0,9913	0,0000	0,0265
Coquimbo	0,9761	0,9761	0,9761	0,0000	0,0000
Coronel	0,9690	1,0000	1,0000	1,0000	0,0584
Curicó	0,9003	0,9859	1,0000	0,7241	0,0511
El Bosque	0,9405	0,9793	1,0000	0,7466	0,0817
Estación Central	0,9893	0,9893	1,0000	0,9054	0,1131
Huechuraba	0,8902	0,9195	0,9368	0,0000	0,0173
Independencia	0,8756	0,9595	0,9595	0,0000	0,0000
Iquique	0,9374	0,9642	1,0000	0,5320	0,0765
La Cisterna	0,9500	0,9732	1,0000	0,6869	0,0856
La Florida	0,8779	0,9654	1,0000	0,6099	0,0887
La Granja	0,9808	0,9808	1,0000	0,6678	0,0578
La Pintana	0,9088	0,9630	0,9630	0,0000	0,0000
La Reina	0,9418	0,9418	0,9418	0,0000	0,0000
La Serena	0,9856	0,9889	0,9932	0,0000	0,0043

We have obtained promising results with this first study of our approach, but, in order to know if it performs significantly better than previous models in the case of ecological inference, we have to further develop our experimentation using a set of representative benchmark examples. We are currently working in preparing this set because, nowadays, it is not available in the field of ecological inference.

Possibilistic regression, though useful in this kind of problems, usually produces crisp estimated coefficients. This is the reason why in our future work we intend to study a quadratic programming approach for fuzzy linear regression analysis.

References

[1] Zadeh L. A. The concept of a linguistic variable and its application to aproxímate reasoning i, ii, iii. *Information Sciences*, 8-9:199–251, 301–357, 43–80, 1999.

[2] Luc Anselin and Wendy K. Tam. Cho. Spatial effects and ecological inference. *Political Analysis*, 10(3):276–297, 2002.

[3] Barry C. Burden and David C. Kimball. A new approach to the study of ticket splitting. *American political Science Review*, 92(3):533–544, 1998.

[4] Wendy K. Tam. Cho. Latent groups and cross-level inferences. *Electoral Studies*, 20:243–263, 2001.

[5] Wendy K. Tam Cho and Brian J. Gaines. The limits of ecological inference: The case of split ticket voting. *American Journal of Political Science*, 48(1):152–171, 2004.

[6] Duncan O D and Davis B. An alternatiave to ecological correlation. *American Sociological Review*, 18:665–666, 1953.

[7] E. Engel. Evolución del comportamiento electoral desde el plebiscito a la elección presidencial. *Colección de Estudios CIEPLAN*, 28:73–83, 1990.

[8] E. Engel and A. Hernando. Chile: ¿dos o más bloques? *Perspectivas en Política, Economía y Gestión*, 6:219–230, 2003.

[9] Leo A. Goodman. Ecological regressions and the behavior of individuals. *American Sociological Review*, 18(6):663–666, 1953.

[10] G. King. *A solution to the ecological inference problem: reconstructing individual behavior from aggregate data*. Princeton University Press, 1997.

[11] Vejima S. Asai-K. Tanaka, H. Linear regression analysis with fuzzy model. *IEEE Trans. on Systems, Man, and Cybernetics*, 12(6):903–907, 1982.

[12] Hsiao-Fan. Tsaur, Ruey-Chyn; Wang. Outliers in fuzzy regression analysis. *International Journal of Fuzzy Systems*, 1(2):113–119, 1999.

[13] E. Spencer Wellhofer. Party realignment and voter transition in italy, 1987-1996. *Comparative Political Studies*, 34(2):156–186, 2001.

[14] Tzeng Gwo-Hshiung Yu Jing-Rung and Li Han-Lin. General fuzzy piecewise regression analysis with automatic change-point detection. *Fuzzy Sets and Systems*, 119:247–257, 2001.