Comparative Study of the Wages in Spain by using ADRI and TFC

Juan Moreno-Garcia
Universidad de Castilla-La Mancha
E. U. de Ingeniería Técnica Industrial
Toledo, Spain
Juan.Moreno@uclm.es

Luis Jimenez
Universidad de Castilla-La Mancha
Escuela Superior de Informatica
Ciudad Real, Spain
Luis.Jimenez@uclm.es

Abstract

The aim of this work is to present a comparative study of the wages in Spain from the year 1967 to the year 1988. Leaving from a classic model, as it is the MOISSES model (econometric model of the representative series of the Spanish macroeconomics), we obtain two fuzzy models by using these sources of data. Considering the domain of the MOISSES model, and by using an inductive mechanism of hierarchical partition of the domain named ADRI, we obtain a classification and regression fuzzy tree. In the same way, we obtain a temporal fuzzy chain basing us in the temporal changes. The comparative study of the two models shows that the characteristics of one complements to the other one. So, the two models present a more enriching vision of the system.

Keywords: Temporal modelling, linguistic modelling, economic modellings.

1 Introduction

The demand of resources on the part of the society is limitless. The resources that can be generated by the society are limited. This produces a dissatisfaction in the individuals of the society. To try to palliate this inadequacy, the society is forced to choose the production of some certain resources in decrement of others. This selection is carried out by evaluating the grade of satisfaction that they will take place in the society. The derived problems of this selection are treated for the economy. The economic theories explain and model the behavior of the society. The behavior of people that forms the society is extremely complex and unforeseeable. To define the conditions that affect to the decisions that a person can make in a concrete circumstance is a very complex task. The necessity to carry out quantitative treatments and the impossibility of carrying out controlled experiments has forced to choose a modeling mathematician, allowing that these can be contrasted by means of the use of the great quantity of data that the economic activity generates. The use of mathematical models based on statistical techniques determines the birth of a branch of the economy denominated *econometrics*.

As alternative to this classic form of construction and validation of econometric theories, and to palliate their inconveniences, it seems reasonable to create another methods. We will choose fuzzy models to obtain these characteristic. In this paper, we apply two induction methods named Fuzzy Tree of Regression and Identification (ADRI) [2] and Temporal Fuzzy Chains (TFCs). These two methods of modelling are complementary. The use of these methods is more intuitive than the mathematical expressions, and the fuzzy models reflect the treatment of the uncertainty in the structure of the own rules. Also, the selection of the variables that ADRI carries out to model the behavior of the series is obtained of the data, extracting of the total group of variables those that better describe the global behavior of the series, without necessity of determining them a priori. And like complement is the TFCs that reflect the temporary evolution of the system, reflecting how the variables change in each instant. The combined use of ADRI and TFCs allow us to select the most outstanding variables

and the TFCs allow to show the temporary evolution of these variables. The two next sections recall the definition of ADRI and TFCs. The wages modelling is shown in section 4. Finally, the conclusions are exposed in section 5.

2 ADRI: an Induction Method

ADRI [2] is a generalization of the regression technique [1]. Let $S = \{s_1, s_2 \dots s_n\}$ be a set of data defined for the set of values that takes a set of variables $X = \{X^1, X^2 \dots X^d\}$, thus, $s_i = \{x_i^1, x_i^2 \dots x_i^d, y_i\}$. Let F be a function that only is known in the set S, such that, $F(s_i) = y_i$. The aim of the regression methods, expressed by means of a parameterized function F', is to minimize the distance between the real output value $y_i = F(s_i)$ and its estimated value using $F'(s_i)$.

The regression methods are different to the classification methods. The regression methods consider continuous output values instead of a set of categories in the classification methods. From this point of view, the classification methods could be considered a particularization of the regression methods. Methods based on the successive partitioning of the problem domain (for example, the decision trees $ID3$ [7]) are used like technique of regression for restrictions in Classification and regression trees (CART) [1].

CART is based on a sequence of questions and its possible answers (structured in tree form) over the variable values that define the problem. CART obtains a separate divisions $SR = \{r_1 \dots r_p\}$ of the domain of each variable of the problem. It splits the domains by using the answers to the questions. ADRI generalizes the divisions obtained in CART by means of fuzzy logic [8].

Now, we expose how the set of rules is obtained. Let A_T be a fuzzy set of the tree node T defined over the set of data S. The root node contains all data of the set S, and its membership function is $A_T : S \rightarrow 1$. The output associated to the node T is defined using the membership grade $A_T(s_i)$ of each data s_i and the output y_i (equation 1).

$$F''(T) = \frac{\sum_{i=1}^{n} A_T(s_i)^m * y_i}{\sum_{i=1}^{n} A_T(s_i)^m} \qquad (1)$$

where $A_T(x)$ is calculated as $\min_{j=1}^{d} A_T^j(x^j)$ and $A_T^j(x^j)$ is the membership function of a fuzzy set defined over the j-th variable.

Equation 2 calculates the estimated error $E(T)$ of a node T.

$$E(T) = \frac{\sum_{i=1}^{n} (F''(T) - y)^2 * A_T(s_i)^m}{\sum_{i=1}^{n} A_T(s_i)^m} \qquad (2)$$

Now, our problem is how to establish a set of questions to divide the node T. These questions are carried out for each variable, thus, a binary division of the node T is obtained for each one of the variables. We suppose binary division for the fuzzy set associated to the node T (by means of the fuzzy set A_T^j of the variable j), that is, $p_T^j = \{B(x), C(x)\}$ where $A_T^j = B(x) + C(x)$.

This partition creates two new nodes and two new fuzzy sets associated to them (Equations 3 and 4). The obtained rules with these fuzzy sets are: **if** $variable^j$ is A_{T_1} **then** ... **else if** $variable^j$ is A_{T_2} **then** ...

$$A_{T_1}(s_i) = \min(A_T(s_i), B(x_i^j)) \qquad (3)$$

$$A_{T_2}(s_i) = \min(A_T(s_i), C(x_i^j)) \qquad (4)$$

Following the ADRI algorithm, we obtain the proportion of the fuzzy sets $B(x)$ and $C(x)$ in relation to the fuzzy set A_T^j (Equation 5).

$$P(T_1) = \frac{\sum_{i=1}^{n} B(x_i^j)}{\sum_{i=1}^{n} A_T^j(x_i^j)} \quad P(T_2) = \frac{\sum_{i=1}^{n} C(x_i^j)}{\sum_{i=1}^{n} A_T^j(x_i^j)} \qquad (5)$$

The quality of this partition is estimated with the equation 6.

$$C(T, p^j) = (E(T_1)P(T_1)) + (E(T_2)P(T_2)) \qquad (6)$$

where p^j is the fuzzy partition of the j-th variable.

The selected partition is the one that has the minimum value $C(T, p_j)$. This technique generates a hierarchical fuzzy partition in each one of the variables to obtain the questions. This process of division obtains the relevant variables to define the model. The mechanism of division is stopped when some condition is verified. In our case, the condition is that the error is less than a constant:

$$ERROR = \max_{T \in \overline{T}} \{E(T)\} <= c, \qquad (7)$$

where $\overline{T}$ is the set of tree leaves, thus, each leaves is a fuzzy region of the model.

Equation 8 calculates the final output value for each input value s_i.

$$F''(s) = \frac{\sum_{t\in\overline{T}} A_T(s)^m * F''(T)}{\sum_{t\in\overline{T}} A_T(s)^m} \quad (8)$$

The interested reader is referred to [2].

3 Temporal Fuzzy Chains

We suggest to represent the temporal side of a DS making use of the TFCs, unplublished in [6]. A TFC is formed by linguistic states and transitions. A linguistic state is defined to represent the system at a time. Between two consecutive linguistic states is established a linguistic transition that indicates the conditions necessary to enters into the next linguistic state. The change of state is described in a *linguistic way* (using linguistic labels).

Let Ξ be a DS MISO with a set of m real input variables $X_1, X_2 \dots X_m$ and an output real variable S. The behavior of the system is given by means of a set of examples $E = \{e_1, e_2 \dots e_n\}$ with $e_i = (x_1^i \dots x_m^i, s^i, t_i)$ where $x_j^i \in X_j$, $s^i \in S$ and t_i is the time in which occurs the example i.

TFCs work with linguistic variables [9]. These variables have defined an ordered set of linguistic labels over its domain named *continuous linguistic variables*, from now on *variables*. The linguistic labels (from now on labels) associated to these variables are defined before the TFC will be obtained. Thus, an *ordered set of labels* SA_j is defined for each input variable X_j. Its structure is $SA_j = \{SA_j^1, SA_j^2 \dots SA_j^{i_j}\}$, where i is the position of SA_j^i in SA_j and i_j is the number of linguistic labels in SA_j, that is $i_j = |SA_j|$. An ordered set of labels SC is defined for the output variable S. Its structure is $SC = \{SC^1, SC^2 \dots SC^{i_y}\}$ where i is the position of SC^i in SC and i_y is the number of linguistic labels in SC ($i_y = |SC|$).

Our variable takes *linguistic interval* as value. A linguistic interval (from now on interval) $LI_{j,p}^c = \{SA_j^p \dots SA_j^{p+(c-1)}\}$ for a variable X_j is defined as a subset of the ordered set of labels SA_j that

Figure 1: Linguistic interval

begins in the label p and has c labels (Figure 1). Its membership function is the sum of the membership grade of a value a_j to each label belonging to the interval (Equation 9, where $z \in [p..c-1]$).

$$\mu_{LI_{j,p}^c}(a_j) = \sum_{SA_j^z \epsilon LI_{j,p}^c} \mu_{SA_j^z}(a_j) \quad (9)$$

A set of m intervals defined on m variables is an *ordered set of m intervals* for each one of the input variable, and is represented as $SLI_m = \{LI_{1,p_1}^{c_1}, LI_{2,p_2}^{c_2} \dots LI_{m,p_m}^{c_m}\}$. The membership function of a SLI_m is calculated applying a t-norm to the membership grade of the intervals in the SLI_m (Equation 10). SLI_m is used to represent linguistically the range of values of the m input variables.

$$\mu_{SLI_m}(e_i) = *(\mu_{LI_{j,p_j}^{c_j}}(x_j)) \quad (10)$$

where $e_i = (x_1^i \dots x_m^i, s^i, t_i)$ is an example belonging to the set E, $j\epsilon[1..m]$ and * is a t-norm.

Finally, a *linguistic state* i (from now on state) is defined as a tuple $est_i =< A_m^i, SE_i >$ where A_m^i is an ordered set of m intervals of the state i corresponding to the m input variables of the DS, and SE_i is the output label of the state i corresponding to the output variable of the DS. A *linguistic transition* i (from now on transition) is a tuple $trans_i =< T_m^i, ST_i >$ where T_m^i is an ordered set of m intervals of the transition i corresponding to the m input variables of the DS, and ST_i is the output label of the transition i corresponding to the output variable of the DS.

A *TFC* is a tuple $CHAIN =< EST, TRANS >$ where $EST = \{est_1 \dots est_{ns}\}$ is an ordered set of ns states, and $TRANS = \{trans_1 \dots trans_{ns-1}\}$ is an ordered set of $ns-1$ transitions. Transition i reflects the conditions to change from est_i to est_{i+1}.

Algorithm 1 Inference Method

```
cur ← 1
for i = 1 to |E| do
  if μ_{A_m^{cur}}(e_i) > μ_{T_m^{cur}}(e_i) then
    s ← SE_cur
  else
    s ← ST_cur
    cur ← cur + 1
  end if
end for
```

In order to reproduce the DSs with TFCs, the algorithm 1 offers an inference method [5]. The inference algorithm needs a set of examples E as input and is based on the definition of a state est_{cur} named *current state*. est_{cur} indicates the state in which the DS is, and allows to calculate the output at this time. A defuzzification process is needed to obtain a discrete output. To do that, it is possible to use the gravity center method, middle of the maxima criterium, or another classical one [3].

4 Wages Modelling

This section shows the model of the wages by using economic model (Section 4.1), ADRI (Section 4.2) and TFCs (Section 4.3).

4.1 Wages in MOISESS.

The nominal wage of the economy is the result of a negotiating process among companies and workers previous to the realization work. MOISESS [4] proposes the equation 11 as wage model.

$$log(W) = a_0 + a_1 log(P) + a_2(\frac{K_{-1}}{L}) + a_3 log(1 + TWCE) - a_4 U + a_5 Z \qquad (11)$$

where the nominal labor cost (W) depends on the price (P), the productivity of the work-employment (K_{-1}/L), the unemployment rate, the taxes on the work in charge of the employers (TWCE), and a vector of influences Z that can affect to the capacity of union pressure on the wages or to the decision of participating in the work market. Possible components of Z are the indirect taxes and other variables that impact on the real wage of consumption, the benefits to the unemployed, the minimum wage, the legal protection of the employment, etc.

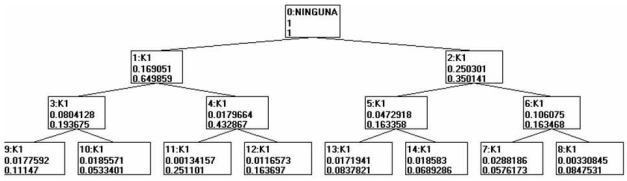

Figure 2: Space partition

4.2 Modelling of the Wages by using ADRI

ADRI can build the fuzzy model Y of the function of wages. We use the following variables: *CL* is the *nominal labor cost*, *PCF* is the *deflator of the PIB (gross interior product) to the cost of the factors*, *T3* is the *half effective type of the net indirect taxes of grants*, *K1* is the *capital stock in the previous year*, *K* is the *capital stock*, *LD* is the *Employment*, *U* is the *rate of unemployment*, *D* is the equation 12 and *SAL* is $CL/PCF(1 + TWCE)$.

$$D = \begin{cases} 0.5 & 1970 \\ 1 & 1971 \\ 0 & the\ others\ years \end{cases} \qquad (12)$$

After running the ADRI algorithm, with maximum value of $RN = 0.01$, a fuzzy regression and identification tree is obtained with $RN = 0.00987695$ and an error average $SEE = \sqrt{E} = 0.00817058$ (Figure 2). Each node shows three values. the first one has a pair $n : v$, where n is the node number and v is the variable partition that originates the node; the second one is the local error; and the last one is the percentage of covered examples. A set of 10 rules is obtained to define the model Ψ of the wages (Table 1). In this case, all variables have all the domain except the $K1$ variable (capital stock). We highlight as the fuzzy model Ψ depends exclusively on one variable, the capital stock in the previous time $K1$. This unique dependence makes that Ψ projects on the group of data a fuzzy partition of 11 subsets. A inference process by using the same data used in the induction is shown in the Figure 3, where the discontinuous line is the obtained line and the continue one is the real line. This partition is originated by the fuzzy partition induced by ADRI for the only variable of the model $log(K - 1)$. The fuzzy sets induced about the values of $Log(K1)$

and the value of the wage assigned as output in each one of them is shown in Figure 4.

Table 1: Obtained model.

	node 13	node 14	node 8	node 7
K1	4.100	4.147	4.190	4.313
	4.100	4.190	4.275	4.344
	4.100	4.190	4.313	4.344
	4.147	4.275	4.344	4.450
SAL	-0.230	-0.195	-0.016	-0.0906
	node 10	**node 9**	**node 12**	**node 11**
K1	4.344	4.450	4.501	4.579
	4.450	4.478	4.555	4.597
	4.450	4.503	4.579	4.632
	4.478	4.555	4.596	4.632
SAL	-0.028	0.0250	0.076	0.0964

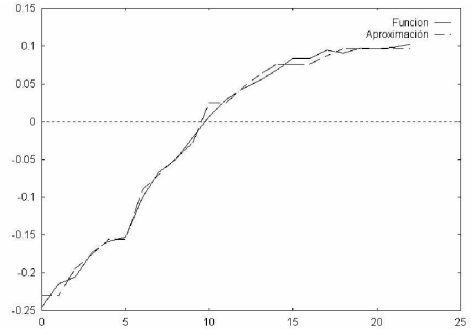

Figure 3: Inference process

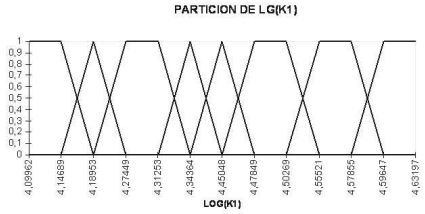

Figure 4: Fuzzy sets induced

To finish this section, observing the value of output of the wage, we see that this grows at the same time that the value of the capital stock grows, except in the central period (Node 7) that a decrement of the wages occurs. This rule would include the period from the year 71 until the year 75, and it checks that the increase of the capital stock is not reflected in the wages.

4.3 Modelling of the Wages by using TFCs

This section shows the obtained wages model. The input and output variables are the same that

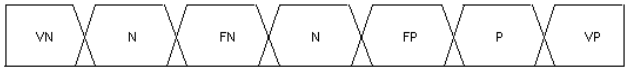

Figure 5: Ordered set of 7 labels

in the previous section. An ordered set of seven labels for each input and output variable is used (Figure 5), where VN is Very Negative, N is Negative, FN is Few Negative, NR is Norm, FP is Few Positive, P is Positive, VP is Very Positive y O is One.

Table 2: Examples.

State	examples
1	{1, 2, 3}
2	{4, 5, 6}
3	{7}
4	{8, 9}
5	{10, 11}
6	{12, 13, 14}
7	{15, 16, 17, 18, 19, 20, 21, 22, 23}

Figure 6 shows the obtained TFC.It is formed for 7 states and 6 transitions.If you see the output values, you can see that the output label of the states are the sequence $< VN, N, FN, NR, FP, P, VP >$, that is, the output labels are continuous along the time. The same output labels have the transitions. The intervals for each input variables are continuous too. Table 2 shows the times for each state. When we obtain a TFC we have information about the evolution in time of each variable (intervals), the duration of the interval (examples that the state covers) and the variable that changes between a groups of consecutive times and another group of consecutive times. Finally, an inference process is done by using the algorithm 1, the input data are the same values that in the induction process. The obtained error is 0.000231. Figure 7 shows the comparison between the real (black line) and obtained output (grey line). The obtained line is very similar to the real one.

5 Conclusions

This paper presents the use of fuzzy logic to model the wages evolution in Spain. For these pur-

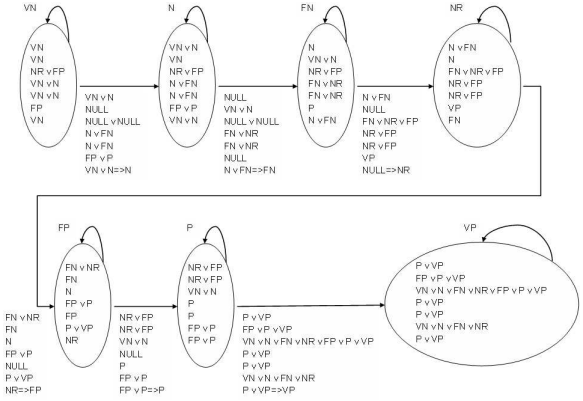

Figure 6: TFC obtained

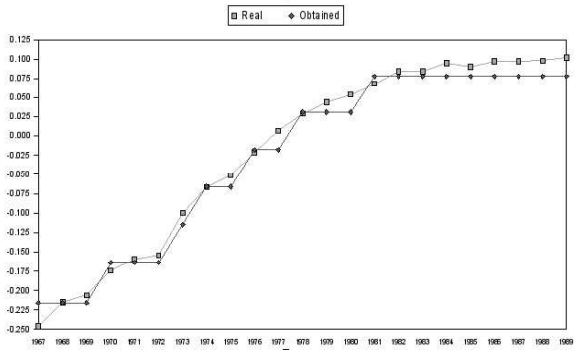

Figure 7: Inference process with TFCs

pose, we use two different methods of modelling systems, ADRI and TFCs. ADRI is a general method that allows us to center in the interpretation of the results. The selection of the variables that ADRI carries out is appropriate, extracting the variables that better describe the global behavior of the series without necessity of determining them a priori. TFCs represents to the dynamic systems by means of a similar methodology to the one used traditionally (states). TFCs has a high qualitative component that facilitates the understanding of the system that represents. This component is very important in our method, and it allows us to define a priori the labels that are wanted to represent each variable of the system. TFCs offer a good approach of the system that represents. Finally, and like complement to ADRI, TFCs reflect the temporal evolution of the system, reflecting how the variables change in each instant. The combined use of ADRI and TFCs allow us to select the most outstanding variables and to show its temporal evolution.

Acknowledgments

This work has been funded by the Spanish Ministry of Science and Technology and Junta de Comunidades de Castilla-La Mancha under Research Projects "DIMOCLUST" TIC2003-08807-C02-02 and PREDACOM PBC-03-004.

References

[1] Breiman J., Friedman J., Olshen R., Stone, "Classification and regression tree", Monterey, Ca:Wadsworth, 1984.

[2] Delgado M., Skarmeta A F., Jimenez L., "A regression methodology to induce a fuzzy Model", International Journal of Intelligent Systems, 16, 2, 169-190, 2001.

[3] Leekwijck W. V. , Kerre E. E., "Defuzzification: criteria and classification",' Fuzzy Sets and Systems, 108 n.2, 159-178, 1999.

[4] Molinas C., Ballabriga F.C., Canadell E., Escribano A., López E., Manzanedo L., Mestre R., Sebastián M., Taguas D., "MOISEES. Un modelo de investigación y simulación de la economía española", Instituto de estudios fiscales. Antonio Bosch.1991.

[5] Moreno J., Jimenez L., Castro-Schez J.J., Rodriguez L.. "Definition of Temporal Fuzzy Chains for Dynamic Systems", Proceedings of Eurofuse 2002, pages 99-105, Varenna, Italy, 2002.

[6] Moreno-Garcia J., Castro-Schez J.J., Jimenez L., "A fuzzy inductive algorithm to model dynamical systems in a comprehensive way". Submitted to IEEE Transactions on Fuzzy Systems, 2005.

[7] Quilan J.R, "Induction of decision tree", Machine Learning, vol 1, 81-106, 1986.

[8] Zadeh L., "Fuzzy sets", Inform Control, 338-353, 1965.

[9] L. Zadeh", The concept of linguistic variable and its applications to approximate reasoning part I,II and III, Inform Sci vol 8 and 9, 199-249 301-357 43-80, 1975.