

**THE QUALITY OF INSTITUTIONS: A GENETIC PROGRAMMING
APPROACH***

Marcos Álvarez-Díaz (Columbia University)

Gonzalo Caballero (University of Vigo)**

ABSTRACT

The new institutional economics has studied the determinants of the quality of the institutions. Traditionally, the majority of the empirical literature has adopted a parametric and linear approach. These forms impose ad hoc functional structures, sometimes introducing relationships between variables that are forced and misleading. This paper analyses the determinants of the quality of institutions using a non-parametric and non-linear approach. Specifically, we employ a Genetic Program (GP) to study the functional relation between the quality of institutions and a set of historical, economical, geographical, religious and social variables. Besides this, we compare the obtained results with those employing a parametric perspective (Ordinary Least Square Regression). We conclude that, at least for our application, the parametric perspective adopted in previous papers about institutional quality could be accurate.

Keywords: Quality of Institutions, Institutional Determinants, Non-Parametric Perspective, Genetic Programming.

JEL: O10, O50, C14.

Category: Institutional Analysis, Positive Political Economy.

* A previous version of this paper was presented at the Annual Conference of the International Society for New Institutional Economics (Barcelona, September 2005), and later it was presented as a FUNCAS Working Paper (FUNCAS, 2006).

** Email: gcaballero@uvigo.es

1-.INTRODUCTION

In recent decades, the new institutional economics (NIE) has constituted a program of research that has propelled the return of institutions into the agenda of mainstream economics. The *coasean* notion of transaction costs (Coase, 1937, 1960) and the *northian* notion of institutions (North, 1990) established the foundations for the theoretical framework of the NIE. Political rules, informal norms and enforcement mechanisms constitute the “rules of the game” of a society and these rules establish an incentives structure that affects the level of transaction costs and the efficiency in the economy.

The NIE is a young program that is in a stage of development and it includes some academic debates and controversies, but has already allowed significant advances in different areas such as economic history, economics of organization, law and economics, policy analysis and development economics (Williamson, 2000; Ménard and Shirley, 2005). The progress of the NIE is generated via a “guerrilla action” (Coase, 1999) that stems from several social sciences, and that was propelled by the award of the Nobel Prize to Ronald Coase in 1991 and to Douglass North in 1993. Since then, the NIE has experienced a growing process in which its analytical abilities have been recognized (Caballero, 2001, 2002). Nevertheless, although North (2005) already propose an extension of the NIE, this program continues requiring efforts in the theoretical and applied work. In fact, problems of definition (for example, Greif’s notion of institutions as an equilibrium versus North’s view as rules), methodology and measurement are present, and we need small pieces of work that expand the stock of knowledge on institutions and economy. In this sense, empirical work is the best way to improve this knowledge.

The contribution of institutions in determining income levels around the world has been one of the main programs of empirical research that has been developed in the last decade (Knack and Keefer, 1997; Hall and Jones, 1999; Acemoglu, Johnson and Robinson, 2001; Rodrik, Subramanian and Trebbi, 2004). There is now widespread agreement among economists studying economic growth that institutional quality holds the key to prevailing patterns of prosperity around the world (Rodrik, 2004). In this way, economics understands the relevance of analysing the quality of institutions and its determinants.

The study of the quality of institutions include works such as La Porta *et al* (1999) or Islam and Montenegro (2002). Traditionally, this program of research, which analyzes the effect of a set of variables on the quality of institutions, has adopted a parametric perspective; therefore a specific functional form (usually linear) is assumed and the unknown parameters are later estimated using some optimization procedure as ordinary least square (OLS). The theoretical validity of the model is easily analysed considering the signs of the coefficients, the statistical significance of the parameters estimated and some fit criterion such as the R-Square. However, assuming a parametric point of view might cause misspecification problems and, in consequence, originate a bias in the results, a loss of predictive ability and an absence of generalization of the model in the face of new observations.

Nowadays, the great advances made in the field of Computer Science allow us to develop, improve and apply powerful and sophisticated techniques for the estimation and prediction of different phenomena. One of these techniques, called Genetic Programming (GP), is inspired by Genetics and by the darwinian theories of natural selection and survival (Holland, 1975; Koza, 1992; Mitchell 2001). The method has already been used satisfactorily in different scientific areas, including economics

(Beenstock and Szpiro, 2002), finance (Álvarez-Díaz and Álvarez, 2003, 2005) and environmental economics (Álvarez-Díaz and Domínguez-Torreiro, 2005). This increasing and intense spread of GP is mainly due to its advantages. Firstly, they do not have any initial restriction on the functional form underlying in the data. Moreover, unlike other methods based on Computer Science, the GP also offers explicitly a mathematical equation which allows a simple ad hoc interpretation of the results. However, as opposed to these advantages, these techniques usually have the difficulty of being computationally intensive and the construction of confidence intervals and hypothesis contrasts is not trivial.

In this work we intend to verify the existence of a bias motivated by employing a parametric perspective; therefore, we try to detect possible misspecifications problems associated to the traditional parametric models. In our opinion, this verification is crucial in an empirical application and it should be always done in order to verify and corroborate the adequacy of the parametric results. In our specific application we use a Genetic Programming called DARWIN (Álvarez et al., 2001) to realize this verification and, additionally, to model what factors explain the institutional quality in different countries. To this purpose, we compare the GP results with those obtained from the traditional parametric point of view and analyse their differences and similitude.

The article is presented as follows. After this introduction, Section 2 presents a brief explanation of the methods used in our study. In Section 3, the data are described and the results obtained for each method are presented. Finally, in Section 4, we draw our conclusions.

2-. GENETIC PROGRAMMING

Genetic Algorithms, originally developed by Holland (1975), enclose a whole series of procedures inspired in biology and, to be more precise, in genetics and in the theory of evolution of species. From the evolution of a random set of possible solutions and by means of applying operators based on natural selection concepts such as *survival of the fittest individuals* and *genetic heritage*, these computing procedures allow finding an optimal approximation to the solution of a certain problem.

In the specialized literature there is no a commonly accepted definition of genetic algorithms which allows distinguish them from other computational evolutionary methods. However, there exist many programs considered as genetic algorithms which present the following common elements: initial population of possible solutions to the problem, selection process using some fit criterion, and use of crossover and random mutation to generate new solutions (Mitchell, 2001). In this paper we have used a kind of genetic algorithm, called genetic programming (Koza, 1992; Álvarez et al., 2001), as a tool to model the relationship between the quality of institutions and a set of historical, economical, geographical, religious and social variables. The evolution process developed by the genetic program can be explained by means of a series of stages. At a first stage, the genetic programming creates a random initial population of N mathematical equations susceptible of representing accurately the relationship between the dependent variable IG_i (institutional quality index) and historical, economical, geographical, religious and social variables $X = \{X_{1i}, X_{2i}, \dots, X_{ki}\}$. These mathematical equations are created by means of a random combination of operators and arguments in the following way:

$$S_j : ((A \otimes B) \otimes (C \otimes D)) \quad \forall 1 \leq j \leq N$$

where A, B, C, and D are the arguments (operand genes), the symbol \otimes represents the mathematical operators (operator genes) and the subscript j refers to each one of the N equations belonging to the initial population. These arguments can be real numbers included in a certain interval (the equation coefficients) or independent variables (delays of the variable). Besides, the mathematical operators (\otimes) used will be sum (+), subtraction (-), multiplication (\cdot) and division (/), being the latter ‘protected’ to prevent zero divisors. It is also possible to include other mathematical operators (such as logarithm or the trigonometric ones) but at the expense of increasing the complexity in the functional optimisation process. Moreover, previous studies on genetic programming have demonstrated that it is possible to describe complex dynamics with mathematical expressions that are built simply with these four arithmetical operators (Szpiro, 1997; Yadavalli et al., 1999; Álvarez et al., 2001).

At a second stage, after determining the initial population of candidates, the evolution process starts selecting those equations that fit best to the problem. For this purpose, the R-Square has been adopted as fitness criterion. This performance measure is defined as:

$$R^2_j = 1 - \frac{\sum_{i=1}^M (IG_i - \hat{IG}_i)^2}{\sum_{i=1}^M (IG_i - \text{mean}(IG_i))^2} \quad \forall 1 \leq j \leq N$$

where R^2_j is the R-Square obtained by equation j , IG_i is the observed value, \hat{IG}_i is the predicted value, and M is the total number of observations in the sub-sample employed to train the genetic program. Later on, all equations of the initial population are classified in decreasing order according to the value of R^2_j . Those equations whose

value of R_j^2 is very low are rejected, while those with a high value are more likely to survive, being the base for the next generation of equations.

The equations that survived after the selection process are used to create the equations of a new solutions generation (i.e., reproduction process). In order to do that the so-called genetic operators will be applied: cloning, crossover and mutation. With the cloning operator, the fittest equations are replicated in the next generation. With the crossover operator pairs of equations with high values of R_j^2 are selected and they exchange part of their arguments and of their mathematical operators. Finally, mutation means that any operator or argument is randomly replaced in a small number of equations. The first top ranked individuals are exempted from mutation, so that their information is not lost. Let us consider, for example, that the following equations belong to the initial population:

$$S_1 : (A + B) / C$$

$$S_2 : (D \cdot E) - G$$

where A, B, C, D, E and G are the equation arguments (coefficients and independent variables). Let us suppose that both expressions will survive the selection process and so they become the base equations for the next generation. The crossover operator means the random selection of a block of operators and arguments in each equation and their later exchange. For instance, let us suppose that the block (A+B) in expression S_1 and the argument G in expression S_2 have been selected. By means of an exchange of blocks two new equations appear as follows:

$$S_3 : G / C$$

$$S_4 : (D \cdot E) - (A + B)$$

As one can observe, the new equations inherit certain features from their parents. Now let us suppose that the expression S_1 is selected again and the mutation operator is applied. So, the following equation can be obtained from S_1 :

$$S_5 : (A \cdot B) / C$$

where the mutation was the random alteration of a mathematical operator.

In short, the new population created from the initial population of equations is composed of cloned equations (such as S_2), mutated expressions (such as S_5), or crossed (such as S_3 and S_4). From this moment, the process will repeat the selection and reproduction stages in an iterative way. After a given number of generations, determined by the user, the iteration procedure ceases and an optimal mapping $IG = F(X_1, X_2, \dots, X_k)$ is given by the strongest mathematical equation in the population.

3-. DATA AND RESULTS

This paper analyses the functional relation between the quality of institutions and a set of historical, geographical, economical, religious and social variables. In Table 1 a brief description of the employed variables is showed. Regarding to the dependent variable, we construct a general index of institutional quality (IG) adding the six particular indicators of governance that were elaborated by Kaufmann, Kraay and Mastruzzi (2003); in this sense, we follow the index aggregation process proposed by Easterly and Levine (2003). On the other hand, the explanatory variables include the ethnolinguistic fractionalization, the legal tradition (English Common Law, Socialist/Communist Law, French Commercial Code), the religion (Roman Catholic,

Protestant, Others), the geographical (latitude) and economical (GNP) condition (La Porta *et al*, 1999). In this way, the database constructed for our study contains complete information about 117 countries.

In order to detect the possible existence of overfitting, the total sample was divided in two sub-samples: In-Sample and Out-of-Sample. The In-Sample is composed of 90 observations randomly chosen and it was reserved exclusively for obtaining the models. On the other hand, the Out-of-Sample contains the rest of observations. Its main function is to verify the validity and consistence of the obtained models and, in consequence, detect possible overfitting problems. Therefore, it will be necessary that the R-Square obtained in In-Sample and Out-of-Sample are similar and relatively high. If this condition was verified, it would be proved the ability of the constructed models to generalize new observations and, therefore, the no-existence of overfitting problems.

In order to study the institutional determinants and their temporal dynamics, we estimate our model for different years of the dependent variable. Table 2 depicts the results obtained using OLS regression and GP. At a first glance, we should highlight the temporal consistency in the results for both methods. In spite of considering different years, the results in terms of R-Square, the explanatory variables finally chosen and their effects on quality index do not show temporal divergences considering both methods.

Before analysing the OLS results, we should mention that the standard backwards stepwise procedure with a 10% level of significance was considered to select the final variables in OLS model. In Table 2 we can observe how the R-Square does show a relatively high value (around 0.70) and how there exists a small divergence between the in-sample and the out-of-sample period. This characteristic reveals the absence of a possible lack of generalisation using OLS. It seems that the method has

discovered the general pattern existing in the data rather than memorise some specific features of the individual observations (overfitting problem). As we can observe for the different years, the relevant variables to explain the institutional quality are *GNP*, *LATIT*, *FRENCH* and *SOCI* (and marginally once *CAT*). The sign of the estimated coefficients seem to be in accordance with the a priori expectative. Specifically, the coefficients on *GNP* and *LATIT* are positive, while on *SOCI* and *FRENCH* are negative. In general terms, these results are coherent with those obtained by La Porta *et al* (1999), when they conclude that countries that are poor, close to the equator, ethnolinguistically heterogeneous, use French or Socialist laws exhibit inferior government performance.

Up to this point we have introduced the results assuming the common linear and parametric perspective. Nevertheless, as it was mentioned in the introduction, this perspective can originate some problems of misspecification, biasing the results and, therefore, misunderstanding our conclusions. For example, are the selected variables the most important to explain the institutional quality? We could question as well if the effect of the selected variables are real or spurious because of assuming a specific and rigid functional form. In order to valid and investigate the possible existence of a bias in our results, we compare them with those obtained employing a GP.

Table 2 also provides specific information about the GP results and, certainly, we can find certain similitude with OLS. First of all, among all possible arithmetic equations, the GP approach has obtained a very similar functional form to the OLS regression. The functional structures found by GP are quasi-linear; and therefore, a simple linear relation would be a valid approach to link the General Institutional Quality Index and the explanatory variables. Secondly, as in the OLS case, the R-Square is relatively high and constant when In-Sample and Out-of-Sample are considered. Lastly, some survival variables to the evolutionary process coincide with the selected variables

in OLS. For instance, *GNP* and *LATIT* appear in all equations offered by the genetic program for the different years, and their positive effects corroborate those obtained by OLS. However, there are other variables which have survived the evolutionary process but they were not selected by the OLS backwards stepwise procedure, such as *ENG*, *ETHF* and, sometimes, *SOCI*. In this case, *ENG* shows a positive sign, and *ETHF* and *SOCI* have a negative effect. These results obtained via a GP approach are coherent with the conclusions of the traditional literature on the quality of institutions in the sense of La Porta *et al.* (1999).

In summary, we can affirm that adopting a rigid functional form (parametric perspective) does not provoke a loss of out-of-sample predictive ability, compared with a flexible technique such as the GP. Moreover, there exists certain similitude in the variables considered as the most relevant to explain the regressand. Therefore, we can conclude that, at least for our application, the parametric perspective adopted in previous papers about institutional quality could be accurate.

4-. CONCLUSIONS

The general procedure to model the quality of institutions has been based almost exclusively on a linear and parametric point of view. Therefore, a-priori and rigid functional forms are discretionally imposed by the researcher rather than observed in the data. This leads to a possible misspecification problem and, in consequence, a bias in the results. In order to validate the results obtained in a parametric framework, it is relevant to open an avenue of research on the determinants of the institutional quality which adopts a non-parametric approach. Our paper has tried to initiate this avenue, and in fact, it constitutes the first case in which the new institutional economics employs a

genetic programming approach. In particular, the main focus of this paper has been to validate the parametric structure traditionally used in the literature (La Porta *et al*, 1999). In this sense, we have opened a new frontier which demands future efforts of research.

Our results have revealed that the parametric perspective, which has been commonly adopted to study the determinants of the institutional quality, could be considered an accurate analytic approach. We have to point out that, among all possible arithmetic equations, the GP approach has obtained a very similar functional form to the OLS regression for all our regressions. Moreover, our comparison seems to corroborate the results obtained by the parametric perspective (OLS) in terms of the variables that were finally selected (*GNP* and *LATIT*); besides this, their effects on regressand coincide. There exist some divergences in the variables selected as the most relevant by the different methods (for example, *FRENCH* is considered as relevant using the OLS backwards stepwise procedure, while *ETHF* and *ENG* are considered as relevant by the GP), however, in all cases the effects on regressand are in accordance with previous results obtained by the parametric approach (La Porta *et al*, 1999).

Analysing the fit criterion, the R-Square shows a similar value for both methods. Moreover, for both cases, there exists a small divergence between the R-Square in the in-sample and the out-of-sample period. Therefore, we can confirm the absence of overfitting problems using both OLS and GP.

A final comment must be mentioned about GP. A genetic program can be very useful to model and validate parametric results (analysing the survival variables and their sign, for example). However, we should not forget that the field of Genetic Algorithms, and of evolutionary computing in general, is relatively new and many of its problems are still under study Mitchell (2001). More research needs to be done in order

to improve and perfect the procedure (for example, our genetic program requires new technical developments and improvements to incorporate a higher number of variables and reduce the computational time).

REFERENCES

- Acemoglu, D., S. Johnson, and J. A. Robinson (2001). "The colonial origins of comparative development: An empirical investigation", *American Economic Review*, 91, pp. 1369-1401.
- Álvarez A., A. Orfila and J. Tintoré (2001). "DARWIN- an evolutionary program for nonlinear modeling of chaotic time series", *Computer Physics Communications*, 136, pp. 334-349.
- Álvarez-Díaz M. and A. Álvarez (2003). "Forecasting exchange rates using genetic algorithms", *Applied Economic Letters*, 10, pp. 319-322.
- Álvarez-Díaz M. and A. Álvarez (2005). "Genetic multi-model composite forecast for non-linear forecasting of exchange rates", *Empirical Economics*, Vol. 30.
- Álvarez-Díaz M. and M. Domínguez-Torreiro (2005). "Using Genetic Algorithms to Estimate and Validate Bioeconomic Models: The Case of the Ibero-atlantic Sardine Fishery", *Journal of Bioeconomics*, in press.
- Beenstock, M. and G. Szpiro (2002). "Specification search in non-linear time-series models using genetic algorithms", *Journal of Economic Dynamic & Control*, 26, pp. 811-835.
- Caballero, G. (2001). "La Nueva Economía Institucional", *Sistema*, N. 156, pp. 59-86.

- Caballero, G. (2002). “El programa de la Nueva Economía Institucional: lo macro, lo micro y lo político”, *Ekonomiaz*, N. 50, pp. 230-261.
- Coase, R. H. (1937). “The Nature of the Firm”, *Economica*, N. 4, pp. 386-405.
- Coase, R. H. (1960). “The Problem of Social Cost”, *Journal of Law and Economics*, V. 3, N. 1, pp. 1-44.
- Coase (1999). “The task of the society”, *ISNIE Newsletter*, V. 2, N. 2, pp. 1-6.
- Easterly, W. and R. Levine (2003). “Tropics, Germs, and Crops: How endowments influence economic development”, *Journal of Monetary Economics*, N. 50, pp. 3-40.
- Hall, R. and C. I. Jones (1999). “Why do some countries produce so much more output per worker than others?”, *Quarterly Journal of Economics*, 114, pp. 83-116.
- Holland J. H. (1975). *Adaptation in natural and artificial systems*, Ann Arbor, The University of Michigan Press.
- Islam, R. and C. E. Montenegro (2002). “What determines the quality of institutions?”, Policy Research Working Paper, N. 2764. The World Bank.
- Kaufmann, D., A. Kraay and M. Mastruzzi (2003). “Governance Matters III: Governance Indicators for 1996-2002”. The World Bank.
- Knack, S. y P. Keefer (1997). “Does Social Capital Have an Economic Payoff?. A Cross-Country Investigation”, *Quarterly Journal of Economics*, Vol. 112, N. 4, pp. 1251-1288.
- Koza J. R. (1992). *Genetic programming: On the programming of computers by means of natural selection*, The MIT Press, Cambridge.
- La Porta, R.; Lopez de Silanes, F.; Shleifer, A. y R. Vishny (1999). “The quality of government”, *Journal of Law, Economics and Organization*, Vol. 15, N. 1, pp. 222-279.

- Ménard, C. and M. Shirley (2005). *Handbook of New Institutional Economics*, Springer Ed.
- Mitchell, M. (2001). *An introduction to genetic algorithms*, Cambridge, Mass.: MIT Press
- North, D. C. (1990). *Institutions, Institutional Change and Economic Performance*. Cambridge University Press. Cambridge
- North, D. C. (2005). *Understanding the process of economic change*, Princeton University Press.
- Rodrik, D. (2004). “Getting institutions right”, Harvard University. Mimeo.
- Rodrik, D., A. Subramanian and F. Trebbi (2004). “Institutions Rule: The primacy of institutions over geography and integration in economic development”, *Journal of Economic Growth*, 9, 2, pp. 131-165.
- Szpiro, G. G. (1997). “Forecasting chaotic time series with genetic algorithm”, *Physical Review E*, 55, 3, pp. 2557-2568.
- Williamson, O. E. (2000). “The New Institutional Economics: Taking Stock, Looking Ahead”, *Journal of Economic Literature*, Vol.38, pp. 595-613.
- Yadavalli, V. K, R. K. Dahule, S. S. Tambe and B. D. Kulkarni (1999). “Obtaining functional form for chaotic time series evolution using genetic algorithm”, *American Institute of Physics*, 9, 3, pp. 789-794.

TABLE 1: Variables included in the Analysis

IG	General Institutional Quality Index: Sum of the six quality indicators by Kaufmann <i>et al</i> (2003): Voice and Accountability, Political Stability, Government Effectiveness, Regulatory Quality, Rule Of Law and Control of Corruption.
ETHF	Ethnolinguistic Fractionalization: Average value of indices of ethnolinguistic fractionalization.
ENG	English Common Law: Identifies the Legal Origin of the English Common Law.
SOCI	Socialist/Comunist Law: Identifies the Legal Origin of the Socialist/Communist Law.
FRENCH	French Commercial Code: Identifies the Legal Origin of the French Commercial Law.
PROT	Protestant Religion: Identifies the percentage of the population of each country that is protestant.
CAT	Catholic Religion: Identifies the percentage of the population of each country that is catholic.
OTHERS	Other Religion: Identifies the percentage of the population of each country that belongs to other religions (non-catholic and non-protestant).
LATIT	Latitude: Absolute value of the latitude of the country.
GNP	Logaritm of GNP per capita (expressed in current US dollars for the period 1970-1995).

Tabla 2: OLS and GP Results

	MODEL		R-SQUARE			SURVIVAL VARIABLES	
			In-Sample	Out-of-Sample	Total	Variables	Sign
2002	OLS	$IG = -15.862 - 2.884 \cdot SOCI - 1.805 \cdot FRENCH + 10.084 \cdot LATIT + 2.045 \cdot GNP$ (0.00) (0.00) (0.01) (0.00) (0.00)	0.7236	0.7314	0.7276	SOCI FRENCH LATIT GNP	- - + +
	GP	$IG = -16.15 + (2 + LATIT) \cdot GNP - ETHF + ENG$	0.7175	0.764	0.7325	ETHF ENG LATIT GNP	- + + +
2000	OLS	$IG = -14.301 - 3.472 \cdot SOCI - 1.815 \cdot FRENCH + 10.496 \cdot LATIT + 1.877 \cdot GNP$ (0.00) (0.006) (0.012) (0.00) (0.00)	0.7086	0.6978	0.7086	SOCI FRENCH LATIT GNP	- - + +
	GP	$IG = -14.88 + (2 + LATIT) \cdot GNP + ENG + \frac{3.78}{SOCI - 3.3}$	0.7015	0.7356	0.7125	SOCI ENG LATIT GNP	- + + +
1998	OLS	$IG = -13.426 - 3.618 \cdot SOCI - 2.739 \cdot FRENCH + 0.025 \cdot CAT + 10.912 \cdot LATIT + 1.69 \cdot GNP$ (0.00) (0.006) (0.002) (0.02) (0.00) (0.00)	0.7015	0.7572	0.7165	SOCI FRENCH CAT LATIT GNP	- - + + +
	GP	$IG = -15.72 + (2 + LATIT) \cdot GNP - ETHF - SOCI + ENG$	0.6824	0.7675	0.7048	SOCI ENG LATIT GNP ETHF	- + + + -
1996	OLS	$IG = -14.040 - 2.558 \cdot SOCI - 1.739 \cdot FRENCH + 7.806 \cdot LATIT + 1.871 \cdot GNP$	0.6935	0.7517	0.7104	SOCI FRENCH LATIT GNP	- - + +
	GP	$IG = -14.62 + (1.74 + LATIT) \cdot GNP - ETHF + 2 \cdot ENG$	0.6944	0.7449	0.7093	ENG LATIT GNP ETHF	+ + + -

