International Journal of Economics, Business and Management Research

Vol. 4, No. 04; 2020

ISSN: 2456-7760

USING SOFT REGRESSION FOR DETERMINING DEMOGRAPHIC POLICY PRIORITIES FOR UNDERDEVELOPED COUNTRIES

Arthur Yosef[1] **and Eli Shnaider**[2]

[1]Tel Aviv-Yaffo Academic College, Israel, ✉yusupoa@yahoo.com

[2]Israel, ✉eli-sh@012.net.il

Abstract

Since their introduction, soft computing methods have found very wide range of practical applications, often displaying superior performance in comparison to the traditional methods. However, the application of soft computing tools has not been uniform, and it did not reach yet some domains where traditional methods still prevail despite their ineffectiveness. In this study we demonstrate advantage of utilizing a soft computing tool: "Soft Regression" for designing effective demographic policy. We conducted extensive literature survey and did not find any case of soft computing applications utilized to design demographic policy; therefore we consider this study as an initial introduction of soft computing to demographic research.

Soft Regression (SR), is a modeling tool based on Soft Computing methods: Fuzzy information processing and Heuristic approach. In contrast to traditional statistical regression methods, it does not require restrictive conditions (which often contradict the "real world" conditions), and thus avoids computational distortions when such conditions are violated. It allows us to include in the model all the relevant explanatory variables without losing some variables due to multi-collinearity problem. Moreover, SR method performs reliable computation of relative importance of the explanatory variables and hence constitutes an effective tool for policy-makers to determine policy priorities. There are additional advantages that will be explained later in the article.

Keywords: Soft computing, soft regression, demographic policies for LDCs, modeling method.

Introduction

While there have been numerous applications of Soft Computing methods in various fields such as Engineering, Computer Science, etc., there have been no applications in demographic research (to the best of our knowledge). The purpose of this study is to demonstrate the superior performance of Soft Regression (a soft computing method) in determining effective demographic policy priorities for less developed countries (LDCs). In order to demonstrate superior performance of Soft Regression (SR)the following steps were undertaken:

Vol. 4, No. 04; 2020

ISSN: 2456-7760

1. Extensive survey of demographic literature dealing with factors affecting the rate of population growth in order to determine explanatory variables to be included in the model.
2. Attempt was made to build demographic model using conventional multiple regression modeling tool, thus demonstrating difficulties involved such as inconsistency, confusion and ambiguity of numerous results.
3. SR was applied to generate one set of results using the same explanatory variables in one regression run. The results are clear, logically reasonable and allow to reach unambiguous conclusions.
4. We repeated the same process twice (for two different years: 1985 and 2000), in order to demonstrate consistency of the conclusions.

Utilizing SR helps to overcome some technical difficulties associated with quantitative modeling using conventional modeling techniques such as Multivariate Regression (MVR). For example, MVR can indicate which variables are significant and which are not, but there is no effective technique to find out the relative importance of various variables (Yosef and Shnaider, 2017), which is needed in order to set policy priorities. There are also additional limitations when using MVR (see details below). The method of SR does not require restrictive and often unrealistic conditions in order to generate reliable results. Its computation of relative importance of the explanatory variables is reliable, and provides valuable information for policy makers regarding what variables are more important (for setting policy priorities) and to what extent.

The method of soft regression is explained and compared to the traditional MVR. Various advantages of SR are presented and evaluated, both- theoretically as well as based on the practical application.

Demographic problem of underdeveloped countries

It is desirable for the underdeveloped economies to attempt to lower the pace of the natural growth rate of their population (in particular for countries where such a growth is excessively high). The reason is: the only way to increase standard of living is by increasing aggregate income faster than the rate of increase of population. Therefore, it is necessary to attain a sustainable rate of increase in aggregate income that is sufficiently larger in comparison to the rate of population growth, in order to gradually diminish the standard of living gap versus developed countries. Hence, when the rate of population growth is high it requires to attain and to maintain very rapid economic growth rate on a sustainable basis for decades. However, such an exceptional economic performance has been extremely unusual historically, and is difficult to achieve by most underdeveloped economies for reasons explained below.

1. Large majority of underdeveloped economies (characterized by a very rapid rate of natural population increase) rely to a very large extend on traditional economic activities, associated with utilization of natural resources (including traditional agriculture, forestry,

Vol. 4, No. 04; 2020

ISSN: 2456-7760

fishery, etc.), and critically depend on these natural resources. Such traditional economic activities when combined with rapid growth of population face the Law of Diminishing Marginal Returns and thus achieving a rapid long-term growth of aggregate economic activities becomes unfeasible.

2. LDC susually rely on primary commodities for exports. There have been many cases of underdeveloped economies displaying impressive rates of growth over limited (in general) number of years due to favorable conditions in the relevant for them commodity markets. However, primary commodity markets are characterized by wide fluctuations of prices over the years and this fact complicates sustainable and rapid long-term economic growth. Decline of commodity prices in a major export market or markets generally leads to economic crisis in the affected LDCs, and when combined with a rapid population growth over an extended time period, further exacerbates the situation.

During the period of economic difficulties due to the Law of Diminishing Marginal Returns and/or unfavorable commodity markets, we expect the persisting rapid population growth to exert substantial stress on political system and social stability, and if such conditions persist long enough, they can lead to very severe consequences.

The policy aiming at reduction of such excessive pace of population growth requires understanding the significant factors affecting population growth and their relative importance. We present a quantitative model based on variables that have been discussed in literature, and for which quantitative data are available. Our purpose is to provide policy makers with information regarding the most important factors associated with decreasing rapid natural population growth. To test the consistency of the results, we perform modeling for two years, 1985 and 2000.

Literature Survey

Most publications dealing with negative effect of rapid natural population growth on long-term economic performance are based on Malthus (1798). The supporters are claiming that despite over 200 years passed since he published his theory about the principle of population, it is still relevant, while others reject it on the basis that developed countries have escaped his pessimistic predictions.

Cincotta and Engelman (1997) claimed that despite lack of clear evidence in previous decades, the more recent data make it clear that during the 1980s, population growth in LDC saffected the growth of per capita Gross Domestic Product, which is considered the primary measuring unit of economic growth. The negative effects of rapid population growth appear to have impacted mostly the poorest group of LDC sduring the 1980s and the 1990s. In contrast, slowing down of the population growth enhanced rapid economic growth in South Korea, Taiwan, Singapore, and Hong Kong. Shnaider and Haruvy (2008) conducted the study of background factors affecting the long-term economic performance. Their study utilized cross-national data for the year 1997. Among other findings, thestudy indicated significant negative

Vol. 4, No. 04; 2020

ISSN: 2456-7760

relationship between the level of economic activity(representing standard of living) and the rate of natural population growth.

Barlow (1994) named Lagged Fertility as an important explanatory variable often omitted in analyzing rate of population growth and its relation versus standard of living. Soubbotina and Sheram (2000) refer to "demographic momentum" as a phenomenon when population continues to increase rapidly for some years after fertility rate drops. They postulate that demographic momentum is in particular significant in LDCs that had the highest fertility rates 20-30 years ago.

Bongaarts and Watkins (1996) described the general worldwide demographic trends for LDCs, and addressed the uneven decline in the rate of natural population growth of various regions. They mentioned factors such as industrialization, urbanization and increased education as possible explanations. Wilson(2001) analyzed the convergence of demographic factors around the world, and pointed to the variables such as life expectancy and total fertility.

The method of SR which we use in our analysis is based on Fuzzy and Heuristic Information Processing (for more details see Kandel et. al. (2000), Maimon and Rokach (2005)).Comparison of SRto MVR appears in Yosef et. al (2015). The comparison of computing relative importance of explanatory variables (RELIMP) by utilizing SR versus traditional regression methods is presented inYosef and Shnaider (2017).The detailed explanation and evaluation of reliability of RELIMP (based on SR) are presented in Shnaider and Yosef (2018)

The model

Our objective in constructing the model is to find out which variables are insignificant, and therefore ineffective as policy targets, and which are significant and should be addressed by policy makers. In addition, in order to design affective policy approach, it is necessary to have reliable evaluation of the relative importance among the significant variables.

The dependent variable is Natural Population Growth, calculated by Birth Rate minus Death Rate.

Explanatory Variables

Based on the literature presented above, factors that affect Natural Population Growth are: standard of living, social progress, investment in human capital(education)and lagged fertility rates. We use quantitative variables (including some proxy variables) according to availability of data by utilizing all the relevant demographic variables we could find in the data bases of the World Bank. These variables include: economic activity per capita(represents standard of living), education enrollments at various levels(represent investment in human capital: education), lagged fertility rate, adolescent fertility rate, life expectancy and urbanization (last three variables represent social progress).

Vol. 4, No. 04; 2020

ISSN: 2456-7760

1. **Value of Economic Activity per capita.** It is represented by GDP (Gross Domestic Product) per capita. It is considered common and legitimate measurement of the standard of living. Some of GDP data series are in current U.S. dollars (USD), while others are in constant 1995 USD, in constant 2000 USD, and in constant 2005 USD. There are data series based on regular currency conversion method vs. PPP (purchasing power parity) conversion method. We decided to select GDP per capita, PPP (current international $).This variable is expected to be inversely related to natural population growth.
2. **Education:** This is a proxy for investment in human capital and the model addresses the question: how different levels of education are related to the natural population growth? It is represented by the percentage of relevant population groups enrolled in primary, secondary and tertiary education. Education variables are expected to be inversely related to natural population growth.
    a. **Tertiary Education(Tertiary Enrollment)**:Percentage of the relevant population group that is enrolled in tertiary education institutions.
    b. **Secondary Education(Secondary Enrollment)**: Percentage of the relevant population group that is enrolled in secondary education institutions.
    c. **Primary Education(Primary Enrollment)**: Percentage of the relevant population group that is enrolled in primary education institutions. It is also a proxy for literacy level.

    Each one of the three variables representing Education factor, represents different degree of investment in human capital, which justifies including all of them in the model.
3. **Lagged Fertility Rate:** We expect lagged fertility rates to be directly related to the present natural population growth rate (Soubbotina and Sheram ,2000). We selected lag period of 20 years.
4. **Life Expectancy**: This variable is a proxy for social progress and represents standard of living, quality of life and welfare of the population. Itis expected to be inversely related to the dependent variable.
5. **Urbanization**: This variable representing the degree of urbanization is a proxy for social progress. Itis expected to be inversely related to natural population growth. It is measured as percentage of urban population vs total population of that country.
6. **Adolescent Fertility**: This variable is also a proxy (negative) for social progress. It is expected to be directly related to natural population growth.
7.

Data

We used cross-national data for the years 1985 and 2000, obtained from the World Bank data bases and hard copy reports. We excluded from the study all the countries with small populations (of half a million or less). Additional countries were excluded due to missing data. The total of 109 countries were included in our study for year 1985, and 129 countries were included in year 2000.

Vol. 4, No. 04; 2020

ISSN: 2456-7760

Method

The above description of the explanatory variables points to a possibility that there is a mathematical correlation among some of the variables described above. This means that it becomes impossible to include all of them together in the model when utilizing traditional modeling tools such as MVR. Due to multicollinearity, some of the explanatory variables become insignificant not because they are not related enough to the dependent variable, but because of technical limitations of the MVR. We avoid this problem by utilizing SR modeling tool, where explanatory variables are not required to be independent of each other. Detailed mathematical description ofSR as well as mathematical comparison ofSR versus MVR appear in Yosef and Shnaider (2017) and Yosef et. al (2015).

Weaknesses of the traditional modeling tools such as MVR

In this section we present more detailed evaluation of some weaknesses of the traditional modeling tools such as MVR. The purpose of this analysis is to demonstrate why it is essential to utilize SR as our modeling tool instead of MVR.

MVR is a modeling tool, and in the process of using it we distinguish between the important factors correlated with the variable we model and the unimportant factors. Modeling by definition is a process of simplification, where we attempt to simplify a complex reality and try to understand it by focusing only on the most important factors, while leaving unimportant factors out of the model, so that they will not obscure our ability to analyze and understand the most important things. Therefore, by definition, modeling involves a certain degree of imprecision, caused by the factors (supposedly unimportant) left out of the model.

The factors that are left out of the model are in reality still interacting with the dependent variable causing some variation in its behavior that the included explanatory variables cannot explain, and represent randomness. Randomness is supposed to cause minor deviations in the behavior of the depended variable versus its expected behavior based on the behavior of explanatory variables. This, of course, expected to be the case if the factors left out of the model are truly of minor importance and tend to cancel each other out over a large enough number of measurements. However, if for whatever reason one or more of the important factors influencing the dependent variable is/are left out, and is/are causing deviation in the expected behavior of the dependent variable, this is already not a randomness error (normal and expected statistical imprecision) but a modeling error causing mistaking results of large magnitude. It is termed “misspecification of model” and leads to biased, distorted results. Regular statistical tests cannot detect misspecification of the model. In some cases, model misspecification can be detected because the coefficients of explanatory variables appear with illogical signs (plus instead of minus or vice versa). However, in many cases models appear to be logical, signs of their coefficients appear to be correct and all the statistical tests look satisfactory; however, the model might still be mis specified. In this case we will discover the problem only when the model fails(leading to wrong decisions, incorrect forecasts, etc.).

Vol. 4, No. 04; 2020

ISSN: 2456-7760

Model misspecification may occur due to incorrect set of the explanatory variables because we are not aware of some important factors influencing the behavior of dependent variable or if we cannot measure them. For example, in the survey of literature above, we encountered factors such as modernization, industrialization etc., that can supposedly influence natural population growth. However, we did not find quantitative data for these variables.

In addition, model misspecification may occur because of incorrect functional form of the equation. In general we apply linear function because of convenience (assuming it is a close enough approximation of real behavior), and not because we have definite theoretical proof that the function is linear. If we decide to use non-linear specification, there is an infinite amount of possibilities and we do not know which is the correct one.

Additional factor for the model misspecification arises from purely technical reasons, since it is assumed that explanatory variables are independent of each other; However, in reality very often explanatory variables are highly correlated mathematically among themselves (even if logically they are unquestionably separate factors). This often causes either one or both of the correlated explanatory variables to appear as statistically insignificant, and therefore redundant (even though based on common sense, they should definitely be a part of the model in order to have correct model specification).

Hence, the modeling process (using MVR) raises many questions that are very difficult to answer positively. Do we know with certainty all the important factors that affect natural population growth? Are all of them measurable quantitatively and appear in statistical publications and data bases? Do we have any idea regarding the correct functional form of the equation? Are all the explanatory variables independent of each other?

In the following section we present a modeling process using conventional regression method(MVR)in order to illustrate the problems described above.

Results based on conventional regression method

Due to expected multicollinearity among the explanatory variables, when all explanatory variables were included in the same regression run, most of them came out insignificant. This required additional regression runs consisting of smaller amounts of explanatory variables in each regression run. In such cases, the general practice of the traditional regression users is to run different combinations of explanatory variables and attempt to find a combination that contains as many as possible significant explanatory variables (provided all the coefficients have logical signs, and the given combination of variables makes sense). Of course, it is a common knowledge (but not common practice) that eventually it is necessary to justify why a final selected group of explanatory variables (which is apart of the initial model) constitutes a correct and complete model specification. It is also necessary to justify why variables which were part of the initial theoretical model were eventually excluded from the final version (of the model). Legitimate reason to exclude explanatory variable is not just because it is insignificant in agiven

Vol. 4, No. 04; 2020

ISSN: 2456-7760

regression run, but because there is a good reason to believe that it represents a truly unimportant factor and its lack of importance was confirmed by the regression runs. As can be seen, it is impossible to justify any of the specifications presented below as correct vs. other specifications. Under similar circumstances, some justifications provided by the modeling professionals are visibly a patch-work that is very vulnerable to serious scrutiny.

Table1 and Table 2 represent the results of various regression runs: Table1 for 1985, and Table2 for the year 2000. In both Tables the first regression run (Run1) included all the explanatory variables.

Results of Table 1:As expected, due to multicollinearity, most of the explanatory variables came out insignificant (in the Run 1).As stated above, the following up, additional steps in the modeling process consist of attempts to combine as many as possible explanatory variables such that all of them will be significant. Obviously it is not feasible to show here all the results of hundreds of unsuccessful regression runs(where some variables are significant and others are not). We did not even try all the possible combinations(the amount of possible regression runs would be unreasonable), but rather applied common sense in combining explanatory variables in order to find meaningful equation. We did not find any combination above two explanatory variables, where all of them are significant. In addition, we tried to find out whether everyone of our initial explanatory variables can appear in at least one two-variable equation, such that both variables are significant(regression runs 2 through 8).The variable "Primary Enrollment" was the only explanatory variable that did not appear significant in any of the regression runs, and thus very likely it is an irrelevant variable.

Next step in building the model(for the purpose of designing effective policy) is to decide which of the regression runs represents correct model specification. Initially Run 2 looks as the best candidate. Both of its variables are also significant in the Run 1. Run 2 maintains the same high value of the Adjusted R square: 0.904. However, selecting Run 2 is inappropriate and unhelpful for the objectives of this study: to design an effective demographic policy. The reason is: the variable "Lagged fertility rate" represents "demographic momentum (see literature survey above). Governments cannot address or change fertility rate that took place 20 years previously. They need to focus on variables they can affect at the present time. Hence, the most important explanatory variable of the model is not a policy variable. It must be part of the model in order to have correct model specification, but it is useless as a policy variable. Which regression run is then the most appropriate? It looks as GDP per capita is a good explanatory variable to guide us in selecting the appropriate regression run. However, GDP per capita isnot a policy variable but another policy "target", similarly to the rate of population growth. In fact, as stated above, the main purpose to lower the rapid population growth in the underdeveloped countries is raise their standard of living (which is measured as GDP per capita).However, this variable, similarly to the variable "Lagged Fertility Rate" must be included in the model in order to maintain correct model specification.

All the explanatory variables (except "Primary Enrollment") appear significant in at least one of the regression runs. This means that most of them are possibly relevant variables for a correct model specification. High degree of mathematical correlation among the explanatory variables is

Vol. 4, No. 04; 2020

ISSN: 2456-7760

causing their insignificance in many regression runs (even-though logically each one of these variables is relevant and distinct factor). Obviously, any regression run including only two explanatory variables out of 8 (or 7 without "Primary Enrollment"), represents model misspecification, which cannot be justified on logical and theoretical ground. Therefore, based on Table 1, except of rejecting "Primary Enrollment" variable, it is impossible to reach meaningful and reliable conclusion regarding relative importance of the explanatory variablesand policy priorities based on the results of traditional modeling method.

Table 1: Conventional regression results for year 1985

	Run 1	Run2	Run3	Run4	Run5	Run6	Run7	Run8
GDP/Cap	0.158		-6.087*	-4.398*	-5.491*			
Tertiary Enrollment	-0.589		-2.623*				-8.078*	
Secondary Enrollment	-1.445			-4.817*				-5.848*
Adolescent Fertility	2.276*	3.197*						2.367*
Lagged Fertility Rate	13.94*	20.24*						
Life Expectancy	**1.546**				-3.262*			
Urban Population	-0.179					-6.263*		
Primary Enrollment	-0.045					-1.119	-1.611	
Adjusted R square	0.903	0.904	0.618	0.671	0.631	0.382	0.484	0.626

* mean that t-value is significant at the 0.05 level.

Notes: 1. All the values appearing in the Table (except in the last row) are t-values.

2. Value marked in bold has wrong sign in comparison to what is logically expected.

Vol. 4, No. 04; 2020

ISSN: 2456-7760

Results of Table 2demonstrate that the modeling difficulties presented in Table 1 are not an exception, but rather expected (on theoretical grounds) outcome. Regression run 1 generates four significant variables, however in two of them signs are illogical and four other variables are insignificant. These are indications of model misspecification. Regression run 2, performed after the deletion of the insignificant variables from the Run 1, generated results consisting of 5 highly significant explanatory variables (also Adjusted R square went up from 0.906 to 0.915). However, two explanatory variables (GDP per capita and Life expectancy) have illogical signs and are definitely absurd from policy stand point:

a. Long term policy of lowering GDP per capita, in order to achieve lower rate of natural population growth for the purpose of raising back the GDP per capita Therefore, there is a contradiction here: on the one hand policy of lowering GDP per capita and on the other hand policy objective of increasing GDP per capita. In addition, as was stated above, the GDP per capita is not relevant (directly) as a policy variable because it is a policy target variable.
b. To reduce "Life Expectancy" means a policy to shorten life span of the population in order to increase standard of living. Can anyone imagine government announcing such policy?

Hence, regression Run 2 is an excellent example to demonstrate incorrect and absurd policy conclusions that could be derived from mis specified models displaying excellent statistical results, based on traditional regression methods.

Rest of the results of Table 2end up with similar outcome as in Table 1 (trying to find various combinations of explanatory variables so that all of them will be significant - see Regression runs 3 through 8). As in Table 1, all the regression runs that had all their explanatory variables significant - did not exceed two explanatory variables per regression run (regression runs 3 through 8). The only different outcome when comparing to Table 1: the variable "Primary enrollment" appears significant in regression runs 4 and 8.

Table 2:Conventional regression results for year 2000

	Run 1	Run2	Run3	Run4	Run5	Run6	Run7	Run8
GDP/Cap	0.881	**2.144***				0.547	-2.255*	
Tertiary Enrollment	0.237			-11.20*		-5.455*		
Secondary Enrollment	-2.226*	-2.256*			-6.468*			
Adolescent Fertility	8.024*	8.318*	4.593*		4.270*			

Vol. 4, No. 04; 2020

ISSN: 2456-7760

Lagged Fertility Rate	12.09*	17.31*	15.84*					
Life Expectancy	**6.140***	**7.512***				-3.391*	-6.538*	
Urban Population	0.590							-6.836*
Primary Enrollment	-0.162			-2.433*				-2.791*
Adjusted R square	0.906	0.915	0.867	0.629	0.689	0.618	0.522	0.422

* mean that t-value is significant at the 0.05 level.

Note: 1. All the values appearing in the Table (except in the last row) are t-values.

2. Values marked in bold are significant but have wrong sign.

Hence, there is a consistency of the results appearing in the two Tables: both fail to identify highest priority policy variables. In addition, out of 8 initial explanatory variables we are unable to include more than two significant variables (having logical coefficient sign) per regression run, and have different significant explanatory variables in each regression run. All these symptoms point to a conclusion that the regression runs included in both Tables represent model misspecification and therefore useless for policy decisions. The results of the SR presented below definitely confirm this conclusion.

In order to mitigate the uncertainty regarding the reliability of our results, we utilize SR which is a soft computing modeling tool that is not affected by the problems discussed above.

Soft Regression

SR is a modeling tool based on soft computing concepts such as Fuzzy Logic (Zadeh, 1965) and Heuristic information processing.. The technical details of the SR method are described in Yosef et al. (2015),Yosef and Shnaider (2017) and Shnaider and Yosef (2018). Previous works leading to the development of Soft Regression are: Shnaider et. al (1997), Kandel et. al (2000) and Shnaider et. al. (2001).

We will briefly describe several of the important characteristics of the SR that are different from those of traditional MVR, and thus justify the need to include this method in the set of modeling tools. These characteristics are:

Vol. 4, No. 04; 2020

ISSN: 2456-7760

1. Soft regression does not require precise model specification. This regression tool is based on Fuzzy Logic, which is designed in the first place to handle information under severe conditions of uncertainty and imprecision (Zadeh, 1965). The idea here is to give up on the possibility of building a precise model and satisfying ourselves with the opportunity to work with whatever data are available. We generate a partial/less-precise model that could still be very reliable in a general direction of its conclusions because it avoids the problem of misspecification bias. In other words, it allows the user to utilize whatever partial and not very reliable data are available to generate some general conclusions that are expected to be more reliable in comparison to mis specified (MVR) model based on the same data. Of course, in the cases where some potentially important variables are excluded due to lack of dataor because of appearing insignificant due to multicollinearity (MVR method), the models are mis specified by definition
2. The relative importance of the explanatory variables among themselves is not affected by adding or removing variables. When a partial model is constructed, the significance of the explanatory variables and the relative importance of those variables among themselves are not affected by adding additional variables to the model, or removing some variables from it. This is in contrast to the behavior of MVR, where addition or removal of an explanatory variable can change drastically the significance and even coefficient sign of other explanatory variables of the model. This characteristic of the SR adds an important feature of stability into the research/decision making.
3. We introduce the heuristically determined maximum and minimum thresholds (for maximum and minimum values in membership function – see explanation below).This helps to handle the distortions due to outlying values in a user-based logical approach (in contrast to strictly mathematical method utilized in sophisticated traditional methods such as Robust Regression).

In SR we have a dependent variable (single numerical one-columnvector) and m numerical vectors (m columns) of explanatory variables.

Let $Y=(y_1,y_2,...,y_n)$ be the n-dimensional vector of dependent variable to be explained, and let $\{X_j\}_{j=1}^{m}$ be the corresponding n-dimensional vectors of explanatory variables when $X_j=(x_{j,1},x_{j,2},...,x_{j,n})$.

We denote $V=(v_0,v_1,...,v_m)$ when $v_0=Y$ and $v_j=X_j$ for all $j=1,2,...,m$ (In other words, $v_0(i)=y_i, v_1(i)=x_{1,i},...,v_m(i)=x_{m,i}$). (1)

The conversion of numerical vectors into fuzzy sets requires their projection into equivalent vectors of the corresponding grades of membership (between zero and one, where 1 represents full membership, and 0 represents no membership at all), based on predefined membership function which is expected logically to reflect the membership of each element in the fuzzy set.

Vol. 4, No. 04; 2020

ISSN: 2456-7760

Based on Kandel et. al. (2000), Shnaider et. al. (1997) and Shnaider et. al. (2001) we define the membership function as follows: Let's define $Max(v_j)$ as the value in a given vector such that all elements equal to or greater than $Max(v_j)$ have full membership in the fuzzy set. We assign all elements that are above or equal $Max(v_j)$ value of one. Let's define $Min(v_j)$ as the value in that vector such that all elements equal to or smaller than $Min(v_j)$ have zero membership in the fuzzy set (do not belong to the fuzzy set at all). We assign all elements that are below or equal $Min(v_j)$ value of zero. $Max(v_j)$ and $Min(v_j)$ must be determined based on logic and common sense for each domain (for details and example see Shnaider and Haruvy,2008). Thus $Max(v_j)$ and $Min(v_j)$ are Maximum cut-off point and Minimum cut-off point correspondingly. In this study we selected "Average of Low-Income Economies" and "Average of High-Income Economies" as our $Max(v_j)$ and $Min(v_j)$,representing the Maximum cut-off point and Minimum cut-off point for each numerical vector. Such average values appear in the data bases and hard copy publications of the World Bank for all variables. By turning all the numbers above $Max(v_j)$ into 1, and all the numbers below $Min(v_j)$ into 0, we neutralize the negative effect of the outliers having excessive values without deleting these data points. In other words, we normalize the data in reference to average performance of "Low Income Economies" and "High Income Economies".

For all other elements (between $Max(v_j)$ and $Min(v_j)$) we project all other i vector elements of $v_j(i)$ into the interval [0,1] proportionally for all vectors, by

$$v_j^{Norm}(i) = \begin{cases} 0 & , v_j(i) \le Min(v_j) \\ \dfrac{v_j(i) - Min(v_j)}{Max(v_j) - Min(v_j)} & , Min(v_j) < v_j(i) < Max(v_j) \\ 1 & , Max(v_j) \le v_j(i) \end{cases} \quad \text{For all } j = 0,...,m \quad (2)$$

The result is: $$V^{Norm} = (v_0^{Norm}, v_1^{Norm}, ..., v_m^{Norm}) \quad (3)$$

We compute the similarity between the dependent variable and every explanatory variable v_j $(j = 1,...,m)$ in the following way: we define distance for direct relation between variables:

$$d_{Y,X_j}^{direct}(i) = \left| v_0^{Norm}(i) - v_j^{Norm}(i) \right| \text{ For all } j = 1,...,m \quad (4)$$

and distance forinverse relation between variables:

$$d_{Y,X_j}^{inverse}(i) = \left| v_0^{Norm}(i) - (1 - v_j^{Norm}(i)) \right| \text{ For all } j = 1,...,m \quad (5)$$

Based on (4) and (5) we can define, for each j:

$$\text{If } \sum_{i=1}^{n} d_{Y,X_j}^{direct}(i) \le \sum_{i=1}^{n} d_{Y,X_j}^{inverse}(i) \text{ then } d_{Y,X_j}(i) = d_{Y,X_j}^{direct}(i) \text{ else } d_{Y,X_j}(i) = d_{Y,X_j}^{inverse}(i), \text{ for all } i = 1,...,n \quad (6)$$

The similarity or closeness (denoted by S_{Y,X_j}) of each explanatory variable X_j to Y is then computed as:

$$S_{Y,X_j} = 1 - \tfrac{1}{n} \sum_{i=1}^{n} d_{Y,X_j}(i) \quad \text{For all } j = 1,...,m. \quad (7)$$

The measure of similarity indicates the degree to which explanatory variable behaves in a similar pattern (direct or inverse) in comparison to dependent variable. Therefore, the measure of similarity S_{Y,X_j} is an equivalent to the traditional statistical measures of significance (t-tests or sig.). However, in addition to significant relation (similarity of $S_{Y,X_j} \ge 0.8$), there is an option of partial significance $0.7 < S_{Y,X_j} < 0.8$, so that as S_{Y,X_j} is approaching closer to 0.7, it is closer to insignificance. The gradual transition from being fully significant to being fully insignificant adds additional stability to modeling process when utilizing soft regression.

Once similarity measures are computed for all the explanatory variables, the next step is to calculate collective contribution of all the explanatory variables combined in explaining the behavior of dependent variable. For every observation, we select the element from one (or more) of the explanatory variables, that is the most similar (has the shortest distance) to the dependent variable, thus creating the vector of minimum distances:

$$d_{Y,X_1,...,X_m}^{Min}(i) = \min_{1 \le j \le m} d_{Y,X_j}(i) \quad (8)$$

A combined similarity (S_comb) of all the explanatory variables to the dependent variable is

$$S_{Y,X_1,...,X_m}^{Comb} = 1 - \tfrac{1}{n} \sum_{i=1}^{n} d_{Y,X_1,...,X_m}^{Min}(i) \quad (9)$$

$S_{Y,X_1,...,X_n}^{Comb}$ explains, to what degree all the explanatory variables combined – explain the behavior of the dependent variable, and in this respect, it is parallel to R^2. One important difference between the two measurements is that in $S_{Y,X_1,...,X_n}^{Comb}$ we allow for overlap of explanatory variables in their relations with the dependent variable (which is of course more reasonable and more in

Vol. 4, No. 04; 2020

ISSN: 2456-7760

line with the "real world" behavior), and therefore explanatory variables are not required to be independent of each other.

The way to compute relative importance of the explanatory variables is to find outhow much each of them contributes to the vector of minimum distances (8)(that was used to compute $S^{Comb}_{Y,X_1,\ldots,X_m}$). This is done by finding the difference between the vector of minimum distances $d^{Min}_{Y,X_1,\ldots,X_m}(i)$ (overall closeness of all the explanatory variables combined to the dependent variable) and the distance of each of the explanatory variable from the dependent variable(d_{Y,X_j})(see Shnaider et. al., 2014). Therefore, relative importance in the SR(in contrast to traditional regression methods) is not affected by correlation with other explanatory variables, and is determined solely by the contribution of a given explanatory variable to explaining the behavior of the dependent variable.

We can calculate relative weight or relative importance (denoted by $RELIMP$) of each explanatory variable in explaining the behavior of the dependent variable based on the following principles (for more details see Shnaider et. al., 2014):

$$RELIMP_j = \frac{S_j}{\sum_{k=1}^{m} S_k} \quad , j = 1,2,3\ldots,m \text{ where } S_j = 1 - \frac{1}{n}\sum_{i=1}^{n}\left| d^{Min}_{Y,X_1,\ldots,X_m}(i) - d_{Y,X_j}(i)\right|.$$

Results

Table 3below summarizes all the results of the Soft regression runs.

1. Based on the columns of <u>similarities</u> (S_{Y,X_j}),one can see that only one explanatory variable is insignificant (less than 0.70) namely, Primary Education (Enrollment). The variable "Urban Population", which represents the degree of urbanization, expressed as percentage of urban population vs total population of that country, is only partially significant, but is very close to being insignificant**.**

   The two other variables - Life Expectancy and Adolescent Fertility, are also partially significant (similarity levels between 0.70 and 0.80), but are very close to the borderline of being fully significant variables.

   The additional four variables that are fully significant (similarity higher than 0.80) include: Value of economic activity per capita (GDP/Cap), Secondary Education Enrollment, Tertiary Education Enrollment and Lagged Fertility Rate.

   We can also notice the similarity between the results generated for 1985 vs. year 2000. This is an indicator of the stability and reliability of the model.

Vol. 4, No. 04; 2020

ISSN: 2456-7760

Table 3: Soft Regression results

Natural Population Growth	1985		2000	
	Similarity	RELIMP	Similarity	RELIMP
GDP/Cap	0.852 I	0.178	0.825 I	0.141
Secondary Enrollment	0.858 I	0.177	0.856 I	0.175
Tertiary Enrollment	0.810 I	0.119	0.820 I	0.150
Lagged Fertility Rate	0.929 D	0.245	0.903 D	0.263
Adolescent Fertility	0.781 D	0.102	0.788 D	0.112
Life Expectancy	0.798 I	0.118	0.773 I	0.099
Urban Population	0.736 I	0.058	0.729 I	0.060
Primary Enrollment	0.624 I	------	0.642 I	------
S_comb	0.976		0.963	

I-Inverse; D-Direct;

2. Based on the RELIMP columns representing Relative Importance of the variables, one can see that:
    a. There is no weight assigned for the variable Primary Education since we found this variable to be insignificant.
    b. The results point to the Lagged Fertility Rate as the most important explanatory variable. This supports the conclusions of Soubbotina and Sheram (2000) regarding the "Demographic Momentum". This is not very encouraging result and points to a difficulties involved in effectively carrying out policy to quickly reduce rapid natural population growth, because Lagged Fertility Rate is not a policy variable. It represents a status of fertility rate which occurred 20 years ago and obviously no present or future government actions can affect it. Hence, natural population growth is a stable phenomenon, strongly influenced by past behavior, and only drastic and even draconian measures like in China, can force quick changes in natural population growth in spite of this variable. Otherwise, the changes are expected to be gradual and long term.
    c. Next after the "Lagged Fertility Rate", the most important variables are Secondary Education and the "Value of Economic Activity per capita" (GDP/Cap).The variable representing "Value of Economic Activity per capita", reflecting average standard of living, is not a simple policy variable. Overwhelming majority of countries desire to

Vol. 4, No. 04; 2020

ISSN: 2456-7760

find any possible way to raise their standard of living, and vast majority of underdeveloped economies are not successful in these efforts. In fact, in this case we have a vicious cycle: high natural population growth negatively affects the ability of the underdeveloped economies to raise their standard of living over long term periods (see above),which in turn negatively affects the possibility to reduce the rapid population growth.

However, based on Shnaider and Haruvy (2008), natural population growth is only one of several factors affecting successful economic performance. Other variables affecting long term economic performance include Tertiary Education, which is also a significant variable affecting natural population growth in the present model. Hence, Tertiary Education variable affects natural population growth in two ways: It affects natural population growth directly and also indirectly by influencing economic performance.

d. The goal of reaching high(comparable to the developed countries)tertiary education enrollment levels in vast majority of underdeveloped countries is not feasible, since in most of these countries the Secondary Education enrollment is still way behind that of the developed world. Hence, the analysis above leaves Secondary Education as the most effective policy variable for underdeveloped countries to lower rapid natural growth of their population. The reasons include:
    I. The variable of Secondary Education Enrollment has more or less the same weight (Relative Importance) as the Value of Economic Activity variable, but is much easier to implement successfully.
    II. Based Shnaider and Haruvy (2008) it also affects to some extend the Value of Economic Activity, although to a much lesser degree in comparison to Tertiary Education, so it affects natural population growth also indirectly.
    III. Secondary Education is a prerequisite to achieving Tertiary Education, which significantly affects natural population growth directly and indirectly by significantly affecting Value of Economic Activity" (GDP per capita).
    IV. Of course, in the case of underdeveloped countries, where primary education enrollment is still very low, the policy goal of increasing secondary education enrollment rate could be implemented only to a limited extendand thus raising Primary Education enrollment must become the highest priority policy in those countries.

Note: Out of Four Fully significant explanatory variables, only two (Secondary Education Enrollment and Tertiary Education Enrollment) are policy variables. In addition, Secondary Education Enrollment attained higher relative importance (weight) in comparison to Tertiary Education Enrollment, thus pointing to the Secondary Education Enrollment as the preferable policy variable.

e. Other variables included in this study came out as only partially significant. Even though Life Expectancy and Adolescent fertility are fairly close to being fully

Vol. 4, No. 04; 2020

ISSN: 2456-7760

significant variables, nevertheless as policy variables they are expected to have much lower influence on natural population growth in comparison to Education variables (due to lower relative importance). In addition, it is expected that the investment in human capital by education, will also positively affect these two partially significant variables, and hence have additional indirect positive influence on reducing natural population growth.

f. The variable Urban Population, which refers to the percentage of urban population out of total population, is only partially significant, and very close to the borderline of insignificance. Our model assigned very low relative importance to this variable, and therefore it definitely should not be addressed for policy purposes.

The last row of the Table 3 indicate to what extend all the explanatory variables combined explain the behavior of the dependent variable ($S^{Comb}_{Y,X_1,\ldots,X_m}$). We can see, that all the results are above 0.96, which is an important indicator of a successful model.

Summary and Conclusions

In this study we presented quantitative model of natural population growth based on the data of the World Bank from years 1985 and 2000. We included all the variables mentioned in demographic literature that were available to us. We conducted comparative modeling process utilizing traditional regression tool(MVR)as well as SR.

1. Modeling process based on traditional regression method required large number of regression runs. All the regression runs having above two explanatory variables had at least one or more variables that were insignificant. On the other hand, all of the explanatory variables (except "Primary Education Enrollment") appeared significant in some model specifications, which points to a possible relevancy of these variables in line with the demographic literature. All this points to a conclusion that the various regression runs represent model misspecification. However, it was impossible to keep all the explanatory variables in the same model due to Multicollinearity (which is a very common problem in numerous modeling projects);and regression models consisting of only two significant variables (out of much larger amount of variables in the original model) are mis specified by definition(unless convincing arguments can be presented that the eventually excluded variables are indeed irrelevant). Obviously, mis specified models are unreliable for determining recommended policy variables. In fact, some of the policy recommendations based on the results presented in Tables 1 and 2 would be dismissed as illogical. Also, inability to have all of the explanatory variables in the same regression run (and being significant) means that it is impossible to determine relative importance of policy variables, thus making the whole study useless for policy makers. Similar modeling difficulties for years 1985 and 2000 indicate that the failures described above were not a one-time co-incidence, but a theoretically expected outcome (see section "Weaknesses of traditional modeling tools" above).

Vol. 4, No. 04; 2020

ISSN: 2456-7760

2. SR required only one regression run for each of the years under study (1985, 2000). Model generated similar and consistent results for both years. The study clearly identified priority policy variables (secondary education enrollment and tertiary education enrollment). Based on Shnaider and Yosef (2018), soft regression is a reliable tool to determine relative importance of explanatory variables. In addition, the results do not contradict common sense and are in line with theoretical studies in the field of Human Capital.

Hence we conclude that soft computing tool "Soft Regression" is a superior modeling tool (vs. traditional tools) for determining demographic policy priorities.

References

1. Barlow R. (1994). "Population Growth and Economic Growth: Some More Correlations". Population and Development Review. Vol. 20, No. 1, pp. 153-165.
2. Bongaarts J. and Watkins S. (1996). "Social Interactions and Contemporary Fertility Transitions". Population and Development Review. Vol. 22, No. 4, pp. 639-682.
3. Cincotta R. and Engelman R. (1997). "Economics and rapid change: the influence of population growth". Population Action International.
4. Kandel A. (editor) et. al. (2000)."Data Mining and Computational Intelligence". Physica-Verlag Publishing.
5. Malthus, T. (1798). "An Essay on the Principle of Population", London.
6. Maimon O. and Rokach L. Eds. (2005). "Data Mining and Knowledge Discovery Handbook", Springer Science and Business Media, pages 522-525.
7. Shnaider E. and Haruvy N. (2008). "Background Factors Facilitating Economic Growth Using Linear Regression and Soft Regression". Fuzzy Economic Review, vol. XIII, No 1, pages 41-55.
8. Shnaider E., Haruvy N. and Yosef A. (2014). "The Soft Regression method - suggested improvements", Fuzzy Economic Review, International Association for Fuzzy-set Management and Economy (SIGEF), vol. XIX, issue 2, pages 21-33.
9. Shnaider E., Schneider M. and Kandel A. (1997). "A fuzzy measure for similarity of numerical vectors", Fuzzy Economic Review 2(1), 17-38.
10. Shnaider E. and Schneider M. (2001). "Heuristic significance test for economic modeling", Fuzzy Economic Review 5(2), 49-59.
11. Shnaider E. and Yosef A., (2018) "Relative Importance of explanatory variable: Traditional method vs Soft Regression" ,International Journal of Intelligent System, vol. 33, issue 6, pages 1180-1196.
12. Soubbotina T. and Sheram K. (2000). "Beyond economic growth: meeting the challenges of global development". The International Bank for Reconstruction and Development, the World Bank.
13. Wilson C. (2001). "On the Scale of Global Demographic Convergence 1950–2000". Population and Development Review. Volume 27, Issue 1, pages 155–171.

Vol. 4, No. 04; 2020

ISSN: 2456-7760

14. Yosef A., Shnaider E. and Haruvy N. (2015). "Soft Regression vs Linear Regression". Pioneer Journal of Theoretical and Applied Statistics. Volume 10, Numbers 1-2, 2015, Pages 31-46.
15. Yosef A. and Shnaider E. (2017)."On Measuring the Relative Importance of Explanatory Variables in a Soft Regression Method". Advances and Applications in Statistics. Vol. 50, No. 3, p. 201 – 228.
16. Zadeh L. A.(1965)."Fuzzy sets", Information and Control 8 (3), 338-353.