Available online at http://scik.org
Math. Finance Lett. 2016, 2016:4
ISSN 2051-2929

INVESTIGATION OF DETERMINANTS OF NET MIGRATION USING VARIABLE SELECTION ALGORITHMS

HASAN BULUT, EMRE DÜNDER, ERDİNÇ YÜCESOY*
Department of Statistics, Ondokuz Mayıs University, Samsun, Turkey

Abstract: Net migration is an important issue concerning all the world. Especially developed countries consider about migration movement and struggle to overcome this problem. There are also some social and economic factors that trigger net migration. In this article we investigate the determinants of net migration for the world by using regression modeling and variable selection techniques. We used differential evolution algorithm and four well-known information criteria to select related factors with the net migration. We consider 26 global factors and reduced the variable set by performing variable selection. We evaluated the common factors that affect the net migration which obtained from selected regression models. Finally we concluded the remarks for the global determinant of net migration flow.

Keywords: net migration; regression modeling; differential evolution algorithm.
2010 AMS Subject Classification: 91D10.

Introduction

Sprenger (2013) investigated the determinants of migration between 21 developed countries which are members of the EU and the OECD. In the study, least squares, poisson and negative binom regression models are used [1]. Poveda (2007) take an interest in the current migration in rural population in the south of Veracruz State (Mexico). In the paper, three different spaces of migration are identified, traditional markets, the northern border and the United States. The multinomial logistic regression is used as method [2]. Kim and Cohen (2010) evaluated determinants of international migratory inflows to 17 Western countries of Europe and outflows from 13 of these countries between 1950 and 2007 in 77,658 observations from multiple sources using panel-data analysis techniques [3]. Miller (2012) explored the economic determinants of

*Corresponding author
Received July 13, 2016

attitudes toward immigration across forty seven countries from North America, South America, Western Europe, Eastern Europe, Sub-Saharan Africa, the Middle East and North Africa and Asia [4]. Mayda (2005) investigated inflows into fourteen OECD countries by country of origin between 1980 and 1995 using panel data analysis in terms of both economic and non-economic determinants of migration [5]. In the literature some studies are interested in only immigration flow of a county. The studies of Isserman et al. (1985) [6], Greenwood and McDowell (1999) [7] and Clark et al. (2007) [8] can be cited for United States. Similarly; for United Kingdom the studies of Hatton (2005) [9] and Mitchell and Pain (2003) [10] and for Germany the study of Vogler and Rotte (2000) [11] can be given example. Dunn and Dyck (2000) investigated differences in health status and health care utilization between immigrants and non-immigrants, immigrants of European and non- European origin, and immigrants of <10 years and >10 years' residence in Canada [12].

Differential Evolution Algorithm

Various methods are developed for solving non-linear problems such as genetic algorithm, fuzzy logic and ant colony algorithms. The genetic algorithm is the most popular optimization method. This algorithm based on the alternative solutions set called as chromosome. In this algorithm generally the binary coding is used. In solution with binary coding genetic algorithms there are some problems with real parameters [13],hence GA has been developed. One of newly developed algorithm to solve these problems is the differential evolution (DE). DE is proposed by Storn and Price (1997) [14]. DE is used for optimization of some non-linear, non-differentiable and non-convex functions. The general strategy of DE can be defined as follows for minimization problem [14,15,16]:

i. Generation of initial population:

The DE algorithm begins with the initial population $X = (x_{ij})_{mxn}$ and this population is generated as follows:

$$x_{ij}(0) = x_j^l + rand(0,1)(x_j^u - x_j^l)$$

where $i = 1,2, \dots, m; j = 1,2, \dots, n$, x_j^u denotes the upper boundary constraints and x_j^l denotes the lower boundary constraints of the decision variables.

ii. Mutation

For each target vector x_i $(i = 1,2, \dots, m)$, a mutant vector $k_i(t + 1)$ is generated according to:

$$k_i(t+1) = x_{r1} + F(x_{r2} - x_{r3})$$

where $i, r1 \; and \; r2 \in (1,2,\ldots,m)$ are randomly selected and cannot be equal. $F\epsilon \; [0,1]$ is the scaling factor and it has an effect on the difference between x_{r2} and x_{r3}.

iii. Crossover

In order to increase the diversity of the population crossover is introduced. For this aim, the trial vector is defined as follows:

$$U_i = (u_{i1}(t+1), u_{i2}(t+1), \ldots, u_{im}(t+1)), i = 1,2,\ldots,m$$

where

$$u_{ij}(t+1) = \begin{cases} k_{ij}(t+1), if \; rand \leq CR \; or \; j = z_i \\ x_{ij}, \quad otherwise \end{cases}$$

where $i = 1,2,\ldots,m; j = 1,2,\ldots,n; CR\epsilon[0,1]$ is crossover constant and $z_i = (1,2,\ldots,n)$ is the randomly selected index.

iv. Selection

In order to decide whether or not the trial individual $U_i(t+1)$ should be a member of the next generation, it is compared to its parent $x_i(t)$.tle selection operation is done by looking this function:

$$x_i(t+1) = \begin{cases} u_i(t+1), if \; f(u_i(t+1)) < f(x_i(t)) \\ x_i(t), \quad otherwise \end{cases}$$

Information Criteria for Regression Models

Information criteria are fitness functions of the regression models that show the quality of the models. Information criteria are very important during variable selection. Selected models highly related to chosen information criteria in regression modeling. According to selection approach, information criteria produce different solutions. The right choice of the information criteria leads the determination of true independent factors which effects the response variable. The main objective is to select a predictor set that minimizes information criteria value. All information criteria are based on penalizing the regression models. There are three components to penalize the models: number of variables, observations and complexity of the covariance matrix of the regression coefficients. These components can vary according to structure of the information criteria.

We considered four well-known information criteria for variable selection process. These are Akaike information criteria (AIC), corrected Akaike informaiton criteria (AICc), Bayesian information criteria (BIC) and consistent information complexity criteria (CICOMP) [17,18,19,20,21]. The formulations of the information criteria are given below.

$$\mathrm{AIC} = -2\log L(\widehat{M}) + 2\mathrm{k} \tag{1}$$

$$\mathrm{AICc} = -2\log L(\widehat{M}) + 2\mathrm{k}(\mathrm{k}+1)/(\mathrm{n}-\mathrm{k}-1) \tag{2}$$

$$\mathrm{BIC} = -2\log L(\widehat{M}) + \mathrm{k}\log(\mathrm{n}) \tag{3}$$

$$\mathrm{CICOMP} = -2\log L(\widehat{M}) + \mathrm{k} + \mathrm{k}\log(\mathrm{n}) + 2\mathrm{C}(\hat{\Sigma}_{model}) \tag{4}$$

AIC is based on only penalizing the number of variables in the model. AICc is a modified version of classical AIC and it take the sample size into consideration. BIC also considers number of variables and observations similar to AICc but the penalization differs. Unlike other criteria, CICOMP penalizes the complexity of the regression coefficients via covariance matrix. Also CICOMP penalizes the sample size and variable as the other criteria. The penalization of the covariance matrix of the coefficient is accomplished using a complexity function $\mathrm{C}(.)$ defined as the following:

$$C(.) = \frac{1}{2}\log\big(tr(.)\big) - \frac{1}{2}log|.| \tag{5}$$

We used the mentioned four information criteria to select relevant determinants for net migration. Differential evolution algorithm employed to minimize information criteria value with assigning $0-1$ values to each predictors and selected the optimal variable set.

Application

In application part we collected a dataset for 30 countries all over the world for 2012. Because of the missing observations, we could only access the full dataset for 2012. Dataset consists of 26 socio- economic variables as explanatory variables and net migration counts as response variable. All the variables are available in http://data.worldbank.org/. To specify the optimal determinants of migration, four information criteria employed within DE algorithm to select optimal variable set within linear regression analysis. We get the alpha level for interpretation of significance as $\alpha = 0.05$.

Table1: The variables used in analysis.

y	Net Migration
x1	Energy use (kg of oil equivalent per capita)
x2	Exports of goods and services (% of GDP)
x3	Fertility rate, total (births per woman)
x4	Foreign direct investment, net inflows (BoP, current US$)
x5	Forest area (sq. km)
x6	GDP (current US$)
x7	GDP growth (annual %)
x8	GNI per capita, Atlas method (current US$)
x9	GNI, Atlas method (current US$)
x10	GNI, PPP (current international $)
x11	Gross capital formation (% of GDP)
x12	Imports of goods and services (% of GDP)
x13	Improved water source (% of population with access)
x14	Industry, value added (% of GDP)
x15	Inflation, GDP deflator (annual %)
x16	Life expectancy at birth, total (years)
x17	Merchandise trade (% of GDP)
x18	Military expenditure (% of GDP)
x19	Mobile cellular subscriptions (per 100 people)
x20	Mortality rate, under-5 (per 1,000 live births)
x21	Population density (people per sq. km of land area)
x22	Population growth (annual %)
x23	Population, total
x24	Services, etc., value added (% of GDP)
x25	Surface area (sq. km)
x26	Urban population growth (annual %)

Table2: Parameter estimation results for each selected models

Coefficient	AIC		AICc		BIC		CICOMP	
	Estimate	p	Estimate	p	Estimate	p	Estimate	p
(Intercept)	1.0E+07	0.004	-2.2E+06	0.000	-	-	6.4E+04	0.651
x1	8.5E+01	0.307	-	-	-	-	-	-
x2	1.3E+05	0.000	8.3E+04	0.000	1.4E+05	0.000	-	-
x3	-8.5E+05	0.002	-	-	-8.6E+05	0.001	-	-
x4	-7.0E-06	0.033	-6.1E-06	0.010	-9.3E-06	0.001	-	-
x5	-1.1E-01	0.015	-	-	-7.7E-02	0.062	-	-
x6	-1.8E-06	0.023	-	-	-1.3E-06	0.072	-	-
x7	-	-	-1.1E+05	0.002	-3.6E+04	0.251	-	-
x8	1.1E+01	0.335	-	-	1.6E+01	0.054	-	-
x9	1.5E-06	**0.042**	-	**-**	9.2E-07	**0.124**	3.3E-07	**0.000**
x10	6.7E-07	0.007	4.8E-07	0.000	7.5E-07	0.001	-	-
x11	-1.0E+05	0.000			-9.8E+04	0.000	-	-
x12	1.8E+04	0.149	6.1E+04	0.000	2.3E+04	0.069	-	-
x13	-6.2E+04	0.002	-	-	-5.5E+04	0.005	-	-
x14	-	-	-	-	-	-	-	-
x15	6.2E+04	0.000	4.7E+04	0.001	6.5E+04	0.000	-	-
x16	-3.2E+04	0.090	-	-	-4.8E+04	0.003	-	-
x17	-8.5E+04	0.000	-7.7E+04	0.000	-8.9E+04	0.000	-	-
x18	2.9E+05	0.000	1.0E+05	0.073	2.8E+05	0.000	-	-
x19	4.2E+03	0.189	9.9E+03	0.004	-	-	-	-
x20	-2.2E+04	0.002	-	-	-2.5E+04	0.000	-	-
x21	-1.1E+03	**0.000**	-5.9E+02	**0.037**	-1.1E+03	**0.001**	-1.3E+03	**0.005**
x22	-	-	-	-	-	-	-	-
x23	-2.8E-03	**0.005**	-3.9E-03	**0.000**	-3.3E-03	**0.000**	-2.6E-03	**0.000**
x24	-	-	-	-	-	-	-	-
x25	-	-	-	-	-	-	-	-
x26	5.0E+05	0.000	2.2E+05	0.002	5.2E+05	0.000	-	-

Table2 shows the regression coefficients and significances for selected variables with AIC, AICc, BIC and CICOMP. The " - " sign indicates the exclusion of related variable from the regression model. It is possible to see the influence of each explanatory variables from regression coefficients for each model. Only x21 and x23 are common factors for the migration for each selected model.

Table3: Characteristic results for each information criteria

Model characteristic	AIC	AICc	BIC	CICOMP
Adjusted R-Squared	0.9807	0.9503	0.9806	0.8318
Residual Standard Error	175700	281700	176200	518100
Number of variables	21	12	20	3
Number non-significant coefficents	4	1	6	0

Table3 shows the descriptive and predictive characteristic results about selected models. Adjusted r-squared (ADJRSQ) and residual standard error (RSE) values represent the performance of the regression models. We can also see the number of variables and significance situations from Table 3. AIC seems has the highest ADJRSQ and lowest RSE value and it seems to the best one. Also BIC is so close to AIC for the performance measures. This result can be misleading because many variables are not significant in these models. Multitude of the number of variables can cause these results for AIC and BIC. AICc selects less variables when comparing with AIC and BIC and there is only one non-significant variable. There is only % 3 difference among AICC and AIC, BIC. In other respects, CICOMP selects the simplest model and all the variables are statistically significant. ADJRSQ values is lowest for CICOMP' s model but the explanation power is above % 80. This value is very satisfactory because CICOMP can explain the variability pretty much with only three variables.

Conclusion

It is an attractive subject to determine the related factor with migration counts in the world. This paper investigated the determinants of migration using variable selection methods. Variable selection was implemented by benefiting the power of heuristic optimization and information criteria. The relevant factors of migration were decided in consequence of selected variable sets. AIC and BIC selected intensive models but AICc is less intensive. CICOMP selects a sparse model that containing only three factors. Although some predictive measures were well in AIC and BIC, there are non-significant factors for the migration. CICOMP's explanation power was very competitive with a simple model and this result is very attractive. From the statistical perspective, all the selected models have high performance but some of them include redundant factors. Because of that, non-significance occurred in the models for certain factors.

When looking at overall selected variables, population density and population number were commonly selected by each information criteria. Due to this result, the mentioned factors have a strong impact on migration. The increment of population and population density decreases the migration counts for all the selected models. According to this result, the migration movement perform to the crowded countries.

Conflict of Interests

The author declares that there is no conflict of interests.

REFERENCES

[1] Sprenger E., The Determinants of International Migration in the European Union: An Empirical Analysis, Institut für Ost-und Südosteuropaforschung. (2013).

[2] Poveda A. R., Determinants and consequences of internal and international migration: The case of rural populations in the south of Veracruz, Mexico, Demographic Research: 16 (2007), 287-314.

[3] Kim K., Cohen J. E., Determinants of International Migration Flows to and from Industrialized Countries: A Panel Data Approach Beyond Gravity, International Migration Review, 44 (2010), 899–932.

[4] Miller B., Exploring the Economic Determinants of Immigration Attitudes, Poverty & Public Policy, 4 (2012).

[5] Mayda A. M., International Migration: A Panel Data Analysis of Economic and Non-Economic Determinants, IZA Discussion, Paper No:1590 (2015).

[6] Isserman, A. M., D. A. Plane, P. A. Rogerson, and P. M. Beaumont, Forecasting Interstate Migration with Limited Data: A Demographic-Economic Approach, Journal of the American Statistical Association, 80 (330) (1985), 277–285.

[7] Greenwood, M. J., and J. M. McDowell, The Supply of Immigrants to the United States. In The Gateway: U.S. Immigration Policies and Issues. Ed. B. R. Chiswick. Washington, DC: American Enterprise Institute, Pp. 54–85, (1982).

[8] Clark, X., T. J. Hatton, and J. G. Williamson, Explaining U.S. immigration, 1971– 1998, The Review of Economics and Statistics,89 (2) (2007),359–373.

[9] Hatton, T. J., Explaining Trends in UK Immigration, Journal of Population Economics, 18 (2005), 719–740.

[10] Mitchell, J., and N. Pain, The Determinants of International Migration into the UK: A Panel based Modelling Approach. London: National Institute of Economic and Social Research. (2003).

[11] Vogler, M., and R. Rotte, The Effects of Development on Migration: Theoretical Issues and New Empirical Evidence, Journal of Population Economics 13 (2000): 485–508.

[12] Dunn J. R., Dyck I., Social determinants of health in Canada's immigrant population: results from the National Population Health Survey, Social Science & Medicine, 51 (2000), 1573-1593.

[13] Keskintürk T., Diferansiyel Gelişim Algoritması, İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 9 (2006), 85-99.

[14] [14] Storn, R., Price, K.: Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. J. Glob. Optim. 11(1997), 341–359.

[15] [15] Kashan M. H., Kashan A. H., Nahavandi N., A novel differential evolution algorithm for binary optimization, Computational Optimization and Applications, 55 (2013), 481-513.

[16] [16] Deng C., Weise T., Zhao B., Pseudo Binary Differential Evolution Algorithm, Journal of Computational Information Systems, 6(2012), 2425-2436.

[17] [17] Akaike, H. Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika, 60(2)((1973)), 255-265.

[18] [18] Schwarz, G. Estimating the dimension of a model The Annals of Statistics 6 (2) (1978), 461–464.

[19] [19] Sugiura, N. Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's. Communications in Statistics-Theory and Methods, 7(1) (1978), 13-26.

[20] [20] Bozdogan, H. Akaike's information criterion and recent developments in information complexity. Journal of mathematical psychology, 44(1) (2000), 62-91.

[21] [21] Pamukçu, E., Bozdogan, H., & Çalık, S. A Novel Hybrid Dimension Reduction Technique for Undersized High Dimensional Gene Expression Data Sets Using Information Complexity Criterion for Cancer Classification. Computational and mathematical methods in medicine, 2015.