Estimates of Foreign Trade Using Genetic Programming

Miroslav Kľúčik

Abstract Genetic programming, a methodology based on evolutionary algorithm, has proven to be very useful in global search for solutions of user specified problems. Its dynamic properties make it suitable for applications in the area of data mining, specifically, for analyzing data of official statistics. This work tries to investigate usefulness of a tree-based genetic programming in estimation and analysis of foreign trade via symbolic regression. The genetic programming is introduced in its basic properties. Genetic programming run parameters as crossover and mutation probabilities, number of constants, tournament size and initial population size are the subject of sensitivity analysis. With help of symbolic regression the possibilities for model estimates of total imports of Slovakia are demonstrated based on business statistics data on industry and administrative data (Extrastat).

Key words: Intrastat, genetic programming, symbolic regression, official statistics

1 Outline of the Problem

Foreign trade official statistics consists of huge amount of data collected in the system of Intrastat and Extrastat. The European Central Statistical Offices deal with problems of late- or non-response, resulting in consequent revisions of official data on total exports and imports of goods. This paper investigates the usefulness of an heuristic approach for estimating the total imports of goods of Slovakia using data collected in the Intrastat system, administrative data (Customs declarations) and business statistics (Industry exhausting survey of big companies).

The main purpose is to use data from various official resources and to capture all new information that cannot be obtained by using classical methods. Artificial intelligence tools can make use of the vast amount of micro-data collected at NSIs. The

[1] Miroslav Kľúčik, INFOSTAT, e-mail: klucik@infostat.sk

European Plan of Research in Official Statistics concludes (2007): "Also the use of such techniques as neural networks/artificial intelligence in data mining and comparisons with classical statistical approaches need to be further researched."This is an attempt to produce an artificial intelligence tool for National Statistical Institutes in EU.

Contemporary there is little evidence in literature on studies oriented at symbolic regression via genetic programming in official statistics. In similar areas, Kronberger et al. (2011) aims at identification of variable interactions using macroeconomic time series and symbolical regression via genetic programming. Kotanchek et al. (2011) aims at detecting models and outliers in large public data sets.

2 The Approach

In practice the estimation of total imports and exports is a complex task, covering processing of questionnaires, requesting additional and specific information from the companies, providing feedback to companies and estimation and calculation of aggregates in different foreign trade classifications. Estimation of the main aggregates (1st official release) deals with high level of uncertainty due to incomplete information from reporting units and also due to estimation of units under the reporting threshold.

At the time of estimation huge amount of data is disposable, covering Intrastat data (CN, SITC, BEC and CPA classifications), Extrastat data from Customs' questionnaires (extra EU trade), but also data of business statistics, e.g. monthly surveys among businesses covering the production side of the economy.

Taking the time series of trade between Slovakia and individual countries of the CN, SITC and BEC classifications (only lagged), together with a 3-digit data of business statistics, the database will consist of more than 10 thousand time series. In a regression with total imports as dependent variable and other time series as explanatory variables, there exist approximately 10^8 combinations with 2 explanatory variables or 10^{12} combinations of 3 explanatory variables. Including the various possibilities of the character of the relationships between the variables and problems of regression technique constraints (spurious regression etc.), there is a space for introducing more powerful estimation methods from the area of heuristics.

One of numerous computation methods dealing with large data sets is genetic programming (GP). In computer science it is enlisted among the artificial intelligence tools (evolutionary computation method). It uses the knowledge from other science branches as genetics, biology and ecology. In general, it is a very simple method. In the computation process it works with mathematical symbols that are grouped together according to chosen representation. The most used method is a tree representation; one tree is representing one individual, i.e. one solution. In the case of classical symbolic regression (with no prior assumption about the linearity or nonlinearity of the solution) a tree consists of branches and leafs, where leafs are representing various symbols (names of variables, constants). The branch is "holding" together more leafs with help of a function (terminal), the same way as in classical equation the parameters and function members are being held together by symbols of '=' or '+'.

2.1 Description of Genetic Programming Parameters

The genetic programming "run" is comprised of number of parameters during the execution of the program, i.e. from the point of initiating the population to the termination of the execution. According to Poli et al. (2008) the parameters include number of initial population, number of constants used in the symbolic regression, number of generations (termination criterion), selection method (tournament method), tournament size, number of elites, probability of crossover, probability of mutation and some restrictions (maximum depth of individual, constant population) etc.

The optimal parameters for the given data set are chosen according to results of sensitivity analysis following the criteria of best fitness of individuals (lowest root mean square error – RMSE), best elites and diversity of population (number of leafs in the population). The most optimal parameters are found to be the following: initial population – 1000, number of constants in the population – 1000, number of generations – 70, tournament size – 22, number of elites in tournament – 1, probability of crossover – 0.8, probability of mutation – 0.05[1]. The data sample covers the period from 2004 to 2010.

The search for optimal fitted model of total imports has been conducted on 50 runs with maximum 70 generations for each run using the mentioned parameters. Analysis of this result can be approached in two different ways. Firstly, the best individual is evaluated, secondly, the structure of the results is analysed in detail as in Kotanchek et al. (2010).

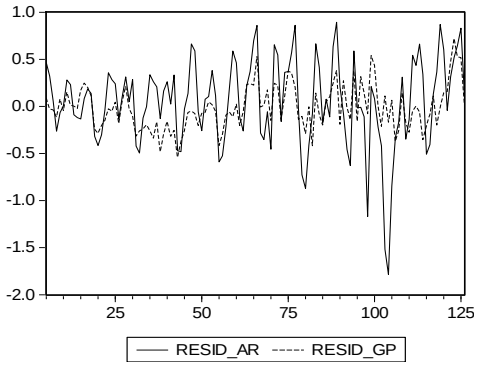
3 Comparison and results

The best individual from all runs on imports is a simple linear model consisting of turnover in industry (for export into Euro area), turnover in automobile industry and turnover in chemicals. As a proxy model an AR model is used. Comparing the RMSE of genetic programming' best individual (best individual model versus dependent variable) to a simply AR model, the model of genetic programming (RMSE = 0.025) is superior to the AR model with 12 lags (RMSE = 0.028). Also other 16 individuals had lower RMSE than the AR model. The residuals of both models are depicted in Figure 1.

The heuristic approach is able to imitate the traditional methods. Sometimes the explanatory variables can be considered a coincidence and so can be the results. In this case the examination of the structure of the genetic programming runs results is adequate. A detailed look into the structure of the results allows us a look into the best individuals' structure. For the most repeated leafs (i.e. explanatory variables), we can assume the highest association with the dependent variable. Surprisingly, in most occurring explanatory variables (first 50) there is no evidence of time series of CN, BEC or SITC. The mostly repeated variables are time series of turnover and orders in Industry. From the total of over 8 thousand leafs in the last generation of genetic programming evolution, the best time series is the total orders for export, which reports over 150 counts from all the leafs of individuals.

[1] Calculations were carried out by the author and documentation of these calculations is available by request from the author.

Figure 1: *Comparison of residuals between AR and genetic programming (GP) model*



4 Conclusions

The consideration that the data of Intrastat are generally "full of holes" forces one to use data from other resources as business statistics or Extrastat. These data can be used for reliable and qualified first estimates. Heuristic method is tested for the estimation of total imports aggregate, and generally, more than 10 thousand time series are at disposal for such estimation as possible explanatory variables. The resulting models can compete with a proxy AR model.

Acknowledgements

The research reported herein was funded by the European Commission through the 7th Framework Programme (FP7/2007-2013) under grant agreement n°244767.
This work was supported by the Slovak Research and Development Agency under the contract No. DO7RP-0024-10.

References

1. European Plan of Research in Official Statistics: Main conclusions from the activities in the 5th Framework Programme, Eurostat/European Commission (2007) http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-07-003/EN/KS-RA-07-003-EN.PDF, Accessed 27. February 2012:
2. Kotanchek, M. E, Vladislavleva, E. Y., Smits, G. F.: Symbolic Regression Via Genetic Programming as a Discovery Engine: Insights on Outliers and Prototypes. In: R. Riolo et al., eds. 2010. Genetic Programming Theory and Practice VII, Genetic and Evolutionary Computation Vol. 8., 55-72 Springer Science+Business Media, LLC (2010)
3. Kronberger, G., Fink, S., Kommenda, M., Affenzeller, M.: Macro-economic Time Series Modeling and Interaction Networks. In: C. Di Chio et al. (Eds.): EvoApplications, Part II, 101–110, LNCS 6625 (2011)
4. Poli, R., Langdon, W. B., McPhee, N. F.: A Field Guide to Genetic Programming, available free at www.lulu.com (2008)