VAR model with current-optimal leading indicators

Miroslav Kľúčik[1]

Abstract. Mutual dependence of business cycles is not easy tractable in the tangle of interconnected economic relationships. The small and open Slovak economy is dependent on the performance of its main trading partners and thus faces the fluctuations of the global economy. For managers and economic policy makers it is crucial to interpret different signals from various segments of the economy for their decisions about future activities. That information is currently available in immense quantity. An artificial intelligence tool – genetic programming – can exploit the huge amount of available data and find patterns of associations between the Slovak economy and foreign economies. Symbolic regression via genetic programming is used. The individual time series and their combinations are optimized for best fit and optimal lead against the Slovak economy. Such leading indicators are used for forecasting of the Slovak economy in the structure of a VAR model. The forecasting ability is tested and compared to a proxy AR model. The forecasts of the VAR model using current-optimal leading indicators show advanced quality compared to the proxy model.

Keywords: business cycle, leading indicators, VAR model, genetic programming, symbolic regression.

JEL Classification: C32, C53, C63, E32
AMS Classification: 60G50, 68T05

1 Introduction

Fluctuations of a small and open economy are mainly driven by the foreign markets, specifically, by its main trading partners. The potential information useful for economic decisions of market participants is therefore available in large quantities considered the macroeconomic data of each trading partner. The main aim is to exploit the immense amount of information and to extract patterns of associations between the business cycle of a small and open economy of Slovakia and other economies.

Business cycle analysis and forecasts can be based on model and non-model approach. The simple non-model approach, as e.g. the OECD methodology for leading indicators construction [14], fulfils mainly the task of qualitative forecast of economic activity. On the other side, model approaches enable to make quantitative forecasts using the leading indicators. The observed non-linearity in business cycles, e.g. [6], [15], gives particular attention to nonlinear models such as regime switching models, models using neural networks or evolutionary computations.

The aim of this paper is to transform the original leading indicator of the Slovak economy as in [7], intended only for qualitative forecasts, to a model approach based on leading indicators and also to improve the adaptability of the leading indicator in relation to the non-linear character of business cycles.

Basic modelling approaches for business cycles are the unobserved variables based models (dynamic factor models), regime switching models or the classic VAR model. A VAR model using leading indicators was constructed by Mendez et al. [12], Cubbada, Hecq [3], Fichtner et al. [5] or Savin and Winker [16].

For the composite leading indicator construction, i.e. for improving its ability to adapt to nonlinear characteristics of business cycle (e.g. different development of cycle in recessions and expansions), the most suitable is the application of heuristics. This is because the space of potential solutions is very large given the more than 20 thousand of time series available in the database and beforehand unknown character of the relationship between the cycles. The group of evolution computation techniques can be regarded as a very effective data mining method for large databases, i.e. genetic algorithms, genetic programming (GP) and other evolutionary computation methods. GP allows the most abstract handling of problems while using small computer programs as individuals in the evolution process in search for the global optimum.

Although GP is predominantly applied in areas of financial markets (high frequency data) some attention was paid also to macroeconomic time series modelling, e.g. in forecasting of Gross Domestic Product (GDP) [11],

[1] VSB – Technical University Ostrava, Faculty of Economics, Department of Economics, Sokolská třída 33, 701 21 Ostrava, Czech Republic, miroslav.klucik@vsb.cz.

[17] or other time series [10], country risk early warning system as in [2]. Genetic algorithms have been used in optimizing regression coefficients for private housing demand leading indicators models [13] or business cycles indicators selection based on genetic algorithm based clustering method [18]. Kotanchek et al. [8] uses genetic programming for detecting models and outliers in large public data sets and Kronberger et al. [9] identifies variable interactions using GP symbolic regression and macroeconomic time series.

The outline of this paper is the following: the 2nd chapter consists of introducing the database of time series and a basic genetic programming model allowing the search for the best matching composite leading indicator for Slovak economy (identification of leading indicators, comparison with a classical average composite indicator). In the 3rd chapter the composite leading indicators are used for constructing a simple VAR model. This is used for forecasting of the Slovak GDP retrospectively for the years 2008-2011. A simple autoregressive model is taken as a proxy for comparison of forecasts. The main findings are concluded in the 4th chapter.

2 Search for leading indicators

The database of time series gathered by the author consists of basic macroeconomic indicators of the European economies and other trade partner of Slovakia (USA, China, South Korea etc.). The analysis is based on quarterly data from the sources of Eurostat, OECD and IMF. The data cover the areas of the whole economy – national accounts data, different branches of real economy (industry, construction, retail trade, services), consumer and business tendency surveys, financial markets data (equity and commodity indexes, exchange rates, short-term and long-term interest), employment, consumer and producer prices, foreign trade data etc.

Together the database comprises over 21 thousand time series. This amount is reduced to over 9 thousand time series by requiring full sample from 1998 to 2011. To avoid unnecessary inaccuracies connected with seasonal adjustment, all the data are transformed to year-over-year changes. Exceptions are the data from business and consumer surveys and time series of balances. These are understood as a deviation from long-term growth, i.e. they are comparable to the year-over-year changes of other time series. However, these time series need seasonal adjustment, therefore both versions of the data are used in the analysis (seasonally adjusted and not adjusted).

2.1 Basic GP model

Genetic programming is a natural selection based algorithm successfully tested mainly for searching large spaces of potential solutions. The aim is to apply symbolic regression of GP for finding the best leading indicators for Slovakia.

The procedure of GP is the following: consider a population where each single individual represents a solution of a problem. In a process of natural selection the best individuals (best solutions) are more likely to be successful in passing the good genes onto the next generation. With the continuing selection procedure new generations are arising with better solutions. Generally, each individual tries to adapt to the current environment through crossover with other individuals, i.e. each solution attempts to get closer to the global solution of the modelled problem. Individuals are encoded in trees (tree representation of individuals). The trees exchange their branches or leafs (crossover) and can mutate.

The search for the best model is based on symbolic regression, where the GDP of Slovakia is the dependent variable and other time series from the database are the potential explanatory variables. The explanatory variables are leafs of the tree, together with constants. The roots of the branches are the terminals – functions (+, -, *, /). Best individuals are chosen according to fitness of individual solutions, which is in the case of symbolic regression a difference measure between the estimated model and true values. Symbolic regression contrary to the classical regression does not assume a relationship between the dependent and explanatory variables in advance. The resulting relationship is mostly nonlinear.

This procedure is written in the EViews language by the author of this paper. The basic structure of the program is the following:

```
Random population
    For (number of generations)
        Evaluation of individuals (fitness measure)
        Crossover (crossover probability)
        Mutation (mutation probability)
                ⇨ New individuals – new generation
    Next
```

The GP parameters for controlling the process of evolution are the following: number of generations, number of initial population, number of constants, selection method, tournament size, number of elites, probability of crossover, probability of mutation and other restrictions (maximum depth of individuals, constant population) etc. The optimal parameters are set according to a sensitivity analysis.

The choice of optimal parameters is based on 10 preliminary runs of sensitivity analysis for each variation of the parameter. The optimal parameter is chosen according to the best average fitness measure of the individuals in last generation. In this case the RMSE measure is used (root mean square error). The final parameters are the following: initial population number – 1000 individuals, number of constants – 1000, tournament size – 29, number of elites in tournament – 1, probability of crossover – 0.9, probability of mutation – 0.01[2].

No considerable progress in fitness has been recorded after the 50th generation (Figure 1); therefore this value is set as the termination criterion.

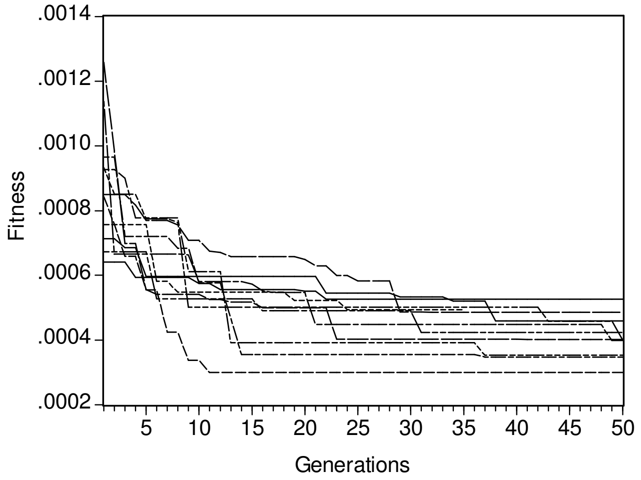


Figure 1 Fitness progress

2.2 Results and comparison

Total of 50 runs have been executed on the GP symbolic regression with maximum of 50 generations in each run. It means there are 50 best individuals/solutions/models available. Except these competing models we can evaluate best individuals in each tournament (29 in each run). The results are mostly nonlinear solutions, but relatively simple, which is acceptable in view of the problems of over-fitting. Evaluating the structure of the best individuals the following variables are mostly repeated in the trees structure as explanatory variables for the Slovak GDP: final consumption expenditure of general government in Euro area (leading 3 quarters), deflated turnover index of Euro area (+ 1 quarter), final consumption expenditure of Austria (+ 1 quarter), GDP of Czech Republic (+ 2 quarters), final consumption expenditure of Czech Republic (+ 1 quarter), gross capital formation of the European Union (EU - + 1 quarter), external balance of goods of Germany (+ 1 quarter), competitive position over the past 3 months EU companies (+ 1 quarter) etc. Some of the variables were not taken into account due to their potential spuriousness, e.g. the GDP of Norway, Denmark etc. The relationship between Slovakia and Norway or Denmark is not direct, but could be regarded as indirect (e.g. through the relationships with Germany). This assumption needs rather further analysis and therefore it is left out from this work.

As an example the formula of the best individual is the following[3]:

$$GDP_{SK} = \left((\_NAB111DE(+2))(\_NAB112EU27(+2))(\_NAB111DE(+1))(\_NA0B1GMCZ(+1))\right) + E_t \qquad (1)$$

To compare the performance of the above GP best individual, a simple composite leading indicator can be used as a proxy. The most associated time series from the database of over 9 thousand time series are chosen according to cross correlation with the GDP of Slovakia. Omitted are again time series with potential threat of spuriousness. Finally, the composite leading indicator used as a proxy is computed as a simple average of the

[2] Calculations performed by the author are provided in the documentation available by request.

[3] GDP_{SK} – GDP of Slovakia in constant prices, _NAB111DE(+2) – external balance of goods of Germany (leading 2 quarters), _NAB112EU27(+2) – external balance of services of EU27 (+ 2 quarters), _NAB111DE(+1) – external balance of goods of Germany (+ 1 quarter), _NA0B1GMCZ(+1) – GDP of Czech Republic (+ 1 quarter).

following five time series: retail confidence indicator of EU27 (leading 1 quarter), Dow Jones Euro Stoxx Basic Materials index Euro area (changing composition, + 1 quarter), trend of activity compared with preceding months in UK (+ 1 quarter), expected business situation in EU27 (+ 1 quarter), balance of goods of Germany (+ 1 quarter). The comparison is depicted in Figure 2 (equalized phases – time shifted series).

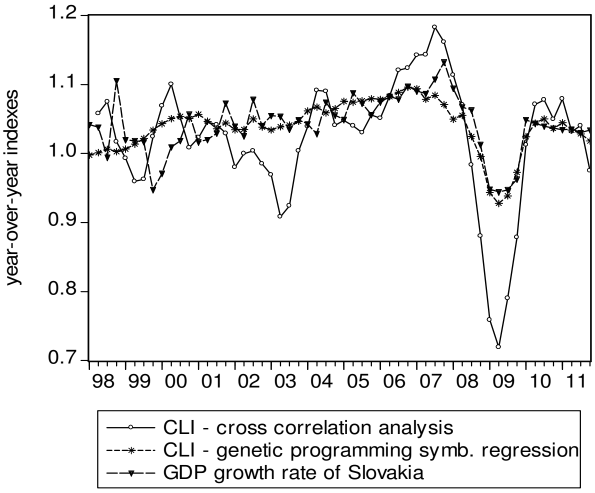


Figure 2 Comparison of composite leading indicators

Clearly, the composite leading indicator from the GP symbolic regression performs better than the leading indicator from cross-correlation analysis (Figure 2), regarding the size of fluctuations and generally, the fitness between the time series.

3 VAR model forecast

Enabling the use of leading indicator for forecasts, the competing 50 composite leading indicators from GP runs are taken preliminary for the VAR model construction. A simple VAR model of the following form is regarded:

$$Y_t = a_0 + A_1 Y_{t-1} + ... + A_p Y_{t-p} + B_1 X_{t-1} + ... + B_q X_{t-q} + E_t \qquad (2)$$

The matrices of parameters of endogenous variables (of vector Y_p) are denoted A_p and of exogenous variables (of vector X_p) as B_q, where p, q are lags. E_t is a vector of random disturbances and a_0 the constant.This form of VAR model is called VARX model, while it includes also exogenous variables. Initially, the composite leading indicator is regarded as an exogenous variable considering only one way dependence relationship between the small Slovak economy and outside economies.

The primary condition of endogenous and exogenous variables for entering the VAR model is the stationarity of the time series. This condition is tested for the 50 leading indicators via the ADF test. All time series are stationary at 10% significance level. This was to be expected due to the year-over-year transformation of the time series (removed trend).

3.1 Model selection

The VAR model is constructed gradually according to the basic criteria deciding about the usability and quality of the model: stationarity of the VAR model, exogeneity test of endogenous variables, test of lag exclusions and residual test.

In the first step the optimum number of lags is determined using the Akaike information criterion. Maximum 6 lags are taken for the test arbitrary. The model of the lag order with lowest AIC value is chosen and consequently, non-significant lags are excluded according to the Wald statistics (5% significance level). All 50 models are estimated using the above mentioned properties. The stationarity of each model is judged following the value of roots of autoregressive polynomials, which must lie inside the unit circle. The test indicates stationarity of all 50 models. Initially the composite leading indicator was regarded as exogenous variable; this assumption is confirmed by the two-way Granger causality test. As expected the GP composite leading indicator indicates only one-way relationship according to the test – from the composite indicator to the Slovak GDP, this is confirmative for all 50 GP competing indicators. Lastly, the presence of autocorrelation in the residuals is tested via the Breusch-Godfrey LM test and the normality of residuals. All of the VARX models do not violate any of these conditions (non-presence of autocorrelation and normality of residuals) at a 5 % significance level. The model with the best individual is chosen for the forecasts. The model is estimated with 5 lags, with 2^{nd} a 3^{rd} lag of the endogenous variable excluded. The exogenous variable (composite leading indicator) is presented in the same form as in (1).

3.2 Forecasts

The predictive ability of the VARX model is tested on the sample 2008-2011, covering the beginning of the current financial/economic crisis in Slovakia (3^{rd} quarter of 2008) and the slow economic recovery since 2010. The robustness of the model is questionable taking into account the short quarterly time series and properties of the model in adjusted sample, but nevertheless the forecasting ability is tested in two steps for each quarter - estimation of the VARX model, one-quarter forecast, two-quarter forecast, prolongation of the sample by one quarter and anew estimation of the model with forecasts for one and two quarters ahead, and so on. The insignificant parameters of the model have been excluded stepwise (due to over-fitting) from the estimated GDP equation of the model following [1] and [4].

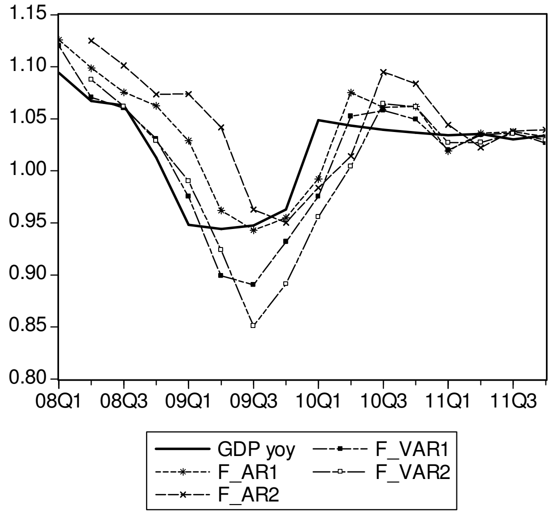
An AR model is used as a proxy model with 5 lags (same number of lags as the VARX model). We can also compare the VARX model with GP indicator with a VARX model containing the composite indicator from correlation analysis (with evaluated properties as the previous VARX model). For evaluation of the forecast the Theil's U is used (TU), which is the share of RMSFE (root mean squared forecast error) of the VARX model (e_t) on the forecast error of the AR proxy model (u_t):

$$TU = \sqrt{\sum_{t=T_1}^{T_2} e_t^2} \,/\, \sqrt{\sum_{t=T_1}^{T_2} u_t^2} \qquad (3)$$

$$RMSFE = \sqrt{\frac{1}{T_2 - T_1 + 1} \sum_{t=T_1}^{T_2} e_t^2} \qquad (4)$$

The T_1 and T_2 is the first and the last forecasting period, e_t is the difference between the true value and forecasted value of the GDP growth rate. The forecast with the lowest RMSFE is considered the best.

In Figure 3 the graphical comparison of VARX model forecast and AR forecast is shown.



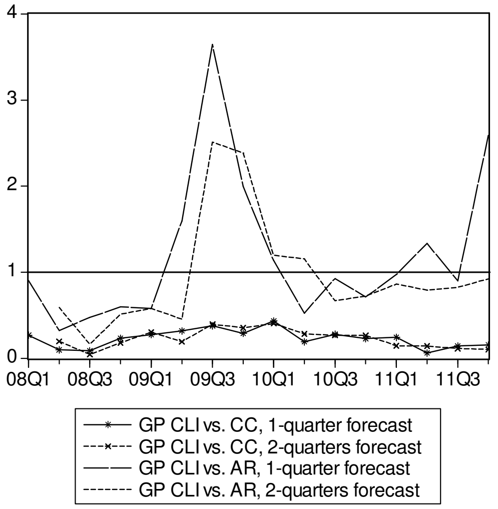


Figure 3 Forecasts comparison

Figure 4 Theil's U – evaluation of forecasts

The VARX model based on the GP composite leading indicator is outdoing the AR model forecasts in 10 out of 16 cases for the one-quarter forecasts and in 11 out of 15 cases for the two-quarter forecast. Figure 3 shows the performance of both forecasts. Also, surprisingly, the GP based VARX model has outperformed all the forecast from the VARX model with composite leading indicator based on the correlation analysis – CLI CC (from Figure 2). The results of the Theil's U are given in Figure 4 above. Theil's U below 1 denotes better forecast of the GP based VARX model. The VARX model shows lower forecast ability during the period of GDP growth after the business cycle turning point in 2009 and in the period of slow GDP growth in 2011.

4 Conclusions

This work attempts to quantify the forecasts of composite leading indicator using a VAR model. Additionally, the forecast is improved by using genetic programming symbolic regression for the composite leading indicator construction. VAR model with exogenous variable representing the leading indicator based on genetic programming clearly outperforms a VAR model with simple average composite indicator and in most cases also an AR proxy model. This proves the possibility of improving the forecasts based on linear relationships between leading indicators by simple nonlinear models. These models have better potential to adapt to the fluctuations of business cycles.

Acknowledgements

This work was financially supported by the Student Grant Competition of Technical University of Ostrava, Faculty of Economics, within the project SP2012/62 “Modelling the impact of cyclical development on a small open economy (VAR approach)”.

References

[1] Alles, M. G., Kogan, A., Vasarhelyi, M. A. and Wu, J.: *Continuous data level auditing: Business process based analytic procedures in an unconstrained date environment*. Department of accounting & information systems, Rutgers Business School, 2006.

[2] Apoteker, T. and Barthelemy, S.: *Genetic Algorithms and Financial Crises in Emerging Markets*. TAC Financial, Saint Hilaire des Landes, 2000.

[3] Cubadda, G. and Hecq, A.: The Role of Common Cyclical Features for Coincident and Leading Indexes Building. *Economics & Statistics Discussion Papers*. University of Molise, Dept. SEGeS, 2003.

[4] Di Fonzo, T. and Marini, M.: *Benchmarking a system of time series: Denton’s movement preservation principle vs. data based procedure.* University of Padova, 2005.

[5] Fichtner, F., Rüffer, R., and Schnatz, B.: Leading indicators in a globalised world. *Working Paper Series 1125*, European Central Bank, 2009.

[6] Hamilton, J. D.: A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica* **57** (1989) 2, 357–384.

[7] Kľúčik, M.: *Composite Indicators for Small Open Economy: The Case of Slovakia*. 30th CIRET conference, New York, United States, 2010. Retrieved from: https://www.ciret.org/conferences/newyork_2010/papers/upload/p_54-548193.pdf

[8] Kotanchek, M. E, Vladislavleva, E. Y., and Smits, G. F.: Symbolic Regression Via Genetic Programming as a Discovery Engine: Insights on Outliers and Prototypes. In: *Genetic Programming Theory and Practice VII, Genetic and Evolutionary Computation* (Riolo, R. et al., eds.), Vol. 8, 55-72 Springer Science+Business Media, LLC, 2010.

[9] Kronberger, G., Fink, S., Kommenda, M., and Affenzeller, M.: Macro-economic Time Series Modeling and Interaction Networks. In: *EvoApplications* (Di Chio, C. et al., eds.), Part II, LNCS 6625, 2011, 101–110.

[10] Lee ,Y. S. and Tong, L. I.: Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming. *Knowledge-Based Systems* **24** (2011) 1, 66-72.

[11] Li, M., Liu, G., and Zhao, Y.: Forecasting GDP growth using genetic programming. *Proceedings of the Third International Conference on Natural Computation* **4**, IEEE Computer Society Washington, DC (2007), 393-397.

[12] Mendez, G. C., Kapetanios, G., and Weale, M. R.: The Forecasting Performance of the OECD Composite Leading Indicators for France, Germany, Italy and the UK. *NIESR Discussion Papers 155*, National Institute of Economic and Social Research, 1999.

[13] Ng, S. T., Skitmore, M., and Wong, K. F.: Using genetic algorithms and linear regression analysis for private housing demand forecast. *Building and Environment* **43** (2008) 6, 1171-1184.

[14] OECD: *OECD System of Composite Leading Indicators,* 2008. Retrieved from: http://www.oecd.org/dataoecd/26/39/41629509.pdf

[15] Paap, R., Segers, and R. Dijk, D.J.C. van: Do Leading Indicators lead Peaks more than Troughs? *Journal of Business & Economics Statistics* **27** (2009), 528–543.

[16] Savin, I. and Winker, P.: Heuristic model selection for leading indicators in Russia and Germany. *Working Papers 046*, COMISEF, 2011.

[17] Wagner, N. and Brauer, J.: Using dynamic forecasting genetic programming (DFGP) to forecast United States gross domestic product (US GDP) with military expenditure as an explanatory. *Defence and Peace Economics*, **18** (2007) 5, 451-466.

[18] Zhang, D., Yu, L., Wang, S., and Song, Y.: A novel PPGA-based clustering analysis method for business cycle indicator selection. *Frontiers of Computer Science in China* **3** (2009) 2, 217-225.