European Conference on Quality in Official Statistics (Q2016)
Madrid, 31 May–3 June 2016

GDP flash estimates: sophistication through simplicity

Kristina Kiriliauskaitė

Statistic Lithuania, Vilnius, Lithuania; Kristina.Kiriliauskaite@stat.gov.lt

Abstract
The sustainable development of GDP measurement is very important for the evaluation of the economy. Thus, huge attention is given to producing quarterly GDP estimates, their accuracy, timeliness and accessibility. Qualitative and relevant GDP estimates lead to accurate economic, political and business decision-making. Lithuania is one of the European countries which produce GDP estimates at a very early stage – within 30 days of the end of the corresponding quarter. When calculating actual GDP at such an early stage, a considerable amount of statistical data is still missing. Therefore, mathematical and econometric methods have to be used to derive GDP flash estimates.
Lithuania's GDP flash estimate based on the production approach is obtained using an indirect method by estimating the value added of each economic activity and intermediate consumption separately. The paper presents Lithuania's experience and good practice of making GDP flash estimates in the context of limited data availability using classical econometric methods, such as linear regression, time series analysis (ARIMA, ARIMAX) and combined forecasting techniques. It describes the sophisticated modelling procedure for deriving hundreds of models and their incorporation into one GDP estimation system – to produce high quality, timely and accurate estimates through simplicity. The paper explains how the models were built, how the explanatory variables were chosen, and how estimate modelling was performed through the combination of linear regression and time series models. A chain-linking method, used to obtain real GDP volumes, and ways of estimating quality analysis are also covered by this topic. Moreover, the paper discusses estimation system stability and adaptation to the global methodological changes (ESA 2010).

Keywords: GDP flash estimate, regression and time series analysis, estimate quality analysis.

1. Introduction

Gross domestic product (GDP) is a very important measure in the macroeconomic analysis, monitoring and forecasting, as it is the first signal of the current economy situation, its trends.

Lithuania's Gross Domestic Product (GDP) flash estimation was performed for the first time in late 1998, when the first quarterly estimate of GDP at current prices and at 1995 year's constant prices was evaluated. From that time methodological analysis and experiments were carried out to obtain the most reliable and qualitative estimates.

Statistics Lithuania (SL) is among the European countries (United Kingdom, Belgium, Spain, Latvia) which produce GDP estimates at a very early stage – within 30 days of the end of the corresponding quarter. Such early estimate requires strong methodological background and reliable statistical information. Usually, necessary statistics are available much later, when GDP estimates are particularly required to accurate evaluation of countries economy, political and business decision-making. This causes a problem of appropriate real data deficiency.

Despite of many structural methodological changes (NACE classification revisions, changes in the European System of National and Regional Accounts (ESA 2010)), GDP estimation methodology is always being updated in order to bring timeliness and accuracy information for users, but foundation/base of methodology remains constant and time confirmed. This article will present the latest SL GDP flash estimate methodology, implemented mathematical tools, difficulties and their solutions as well as quality assurance issues.

2. Methodology background

SL GDP flash estimate is based on production approach method. This choice was done because of sufficient level of data sources and easier perception by the users. GDP flash values are estimated at current and previous year's price using indirect method. This means that GDP is estimated as the sum of gross value added by each kind of activity produced by all resident producer units in the economy, plus taxes on products, minus subsidies on products:

$$\widehat{GDP} = \sum_{i=1}^{n} \widehat{VA}_i + \widehat{D.21} - \widehat{D.31}, \quad (1)$$

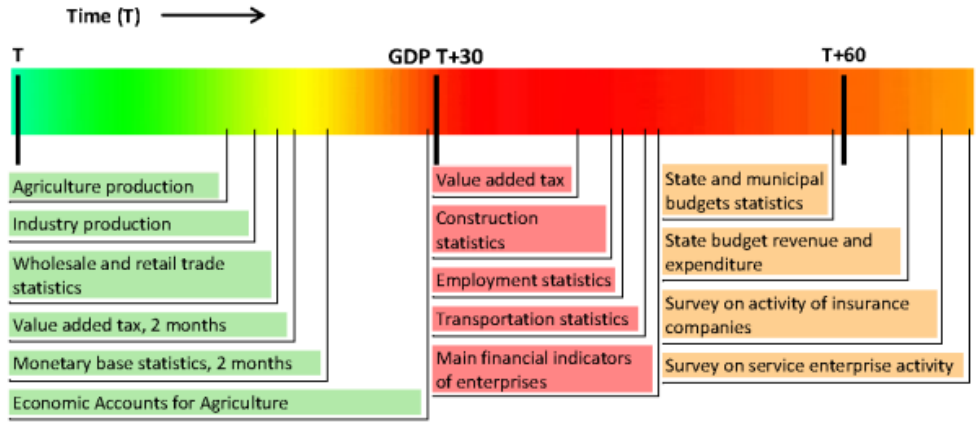
here $\widehat{VA}_i$ – value added at activity i, $\widehat{D.21}$ – taxes on products, $\widehat{D.31}$ – subsides on products.

The value added is estimated at the 2-digit level of NACE Rev. 2: value added of 88 economical activities at current prices and 88 at previous prices. This composes the set of 176 endogenous variables to be observed. Such GDP flash preparation creates a complex system of mathematical, econometric methods (regression analysis, time series theories) and implementation stages of logical algorithms, where a huge amount of observable variables and additional available statistical information is included. Moreover, there are a lot of connections in this system, needed to be considered: i.e. current prices and previous prices connection through GDP deflator, real GDP growths.

3. Data sources: appropriate data challenge

GDP by production approach is considered as main estimation method in SL. However, monthly or quarterly statistical data from surveys, registers or administrative sources are provided almost at the last days or later, when GDP t+30 should be announced (See Picture 1).

Picture 1. Data availability for GDP t+30 estimates.



As it could be seen above in *Picture 1*, the timeline of GDP t+30 is intensive; missing appropriate data challenges are faced. The necessary data are available between the 20–26th days after the end of the corresponding quarter; however, some monthly data are usually late, only two months of quarter are received (value added tax, wholesale and retail trade statistics). It was evaluated that only 40% of all necessary data for GDP preparation is accessible in 30

days after the end of the corresponding quarter, and remaining 60 % of missing data must be extrapolated or nowcasted using mathematical, econometrical techniques. All this leads to situation that GDP t+30 estimation must be done on the limited real data conditions.

4. Modelling techniques:

Econometrics theory suggests a large variety of models for economic analysis and forecasting: one-dimensional, structural or non-linear models. Classical linear regression (Multiple, multivariate regression analysis) or time series models (models of ARIMA class family) are the most popular in economic forecasting (Račkauskas, 2003).

GDP t+30 modelling techniques of Statistics Lithuania, such as linear regression (Green, 2003), and ARIMA family models (Leipus, 2013), will be discussed in this chapter.

We indicate dependent (regressant) variable as y_i (the value added at each i economic activity at 2-digit level of NACE Rev. 2 classification) and independent (regressor) variable as x_j. We assume that y_i can be the best described by available short term business and price statistics. The set of the additional different regressors x_j – is concluded. Correlations are calculated to identify potential regressors. If the regressor has a strong correlation with regressant, there is high probability that it will bring significant information into regression equation (Čekanavičius, 2002). Regressors of the same activity (such as activity production, value added tax statistics (VAT), employees, etc.) are applied.

During analysis of each y_i it was noticed that most of these times series have seasonality, trend and outliers. Thus, classical multivariate regression models were extended: multivariate regression models with trend and seasonal dummies, outliers are applied:

$$y_{i,t} = \mu_{i,t} + \alpha_1 x_{1,t-l} + \alpha_2 x_{2,t-l} + \cdots + \alpha_n x_{n,t-l} + \beta_k S_{k,t} + \tau_i O_{i,t} + \varepsilon_{i,t}, \qquad (2)$$

here, $y_{i,t}$ is dependent variable of the value added of i activity at time – t, $t = (1, \cdots, T)$, T is time series length, $\mu_{i,t}$ – trend of the value added, $x_{j,t-l}$ – regressor j, $j = \overline{1,n}$, with a time lag $l = (0, \cdots, 4)$, $S_{k,t}$ – seasonal variables, which are equal to 1 if quarter depends to first,

second, third or fourth season, otherwise – 0, $k = (1, \cdots 4)$, $O_i(t)$ represents outliers: additive outlier, level shift or transitory change; α_j, β_k, τ_i are model parameters, coefficients of regressors, $\varepsilon_{i,t} \sim N(0, \sigma^2)$ is random disturbance. Model's (2) parameters (coefficientsα_j,β_k, τ_i) are estimated by Least Squares method, we check if errors ε_t follows Gauss-Markov assumptions $(E(y_t) = \mu_y = const, E\left(y_t - \mu_y\right)^2 = \sigma_y^2 = const, E\left(y_t - \mu_y\right)\left(y_{t-k} - \mu_y\right) = const)$ (Račkauskas, 2003).

Multicollinearity and explanatory variable overfitting issue often stays as a problem. It is important to test multicollinearity and optimal explanatory variables number. Otherwise, multicollinearity and overfitting of explanatory variables, can distort results. Model acceptance statistics: the regression coefficients, p-values, and R-squared, can be weak and this can cause poor predictive performance (Čekanavičius, 2002). These modelling problems are very important for the short time series. Most of the economical time series produced by SL takes only 20–25 years quarterly or monthly periodical data. Therefore, multicollinearity and optimal regressors are tested during model specification.

If a significant regressor is not found, or the necessary data are missing for the component of the value added, or regression model brings poor results, ARIMA family models (AR(p), MA(q), ARMA(p,q), ARIMA(p,d,q), ARIMAX(p,d,q)) are used. The general model with the addition regressors X_t included into model is used:

$$\phi(L)(1 - L)^d Y_t = \Theta(L) X_t + \theta(L)\varepsilon_t, \quad (3)$$

here $\phi(L)$ is an autoregressive polynomial, $\theta(L)$ – moving average polynomial, $\Theta(L)$ – explanatory variables polynomial, L is a lag operator and d is d^{th} difference operator.

This model is extended ARIMA model, which includes more realistic dynamics and additional information from the explanatory variables. ARIMA(X) model selection process is done by the Box-Jenkins procedure. It is checked whether errors (ε_t) are independently and normally distributed and whether coefficients are statistical significant (Enders, 2014).

For example, according (2) extended multivariate regression model using indirect method, we will estimate wholesale and retail trade (G) value added $y_{G,t} = \sum_{i=45}^{47} y_{i,t},\ i$ is value added of G section activities (45, 46, 47) at 2-digit level of NACE Rev. 2.

$$y_{45,t} = \underset{(0.000)}{86} + \underset{(0.000)}{0.0002} \cdot x_{1,t} - \underset{(0.004)}{0.0001} \cdot x_{2,t} - \underset{(0.000)}{30.9} \cdot S_{1,t} - \underset{(0.049)}{5.7} \cdot S_{4,t} + \underset{(0.049)}{10} \cdot O_t + \varepsilon_{45,t}, R^2 = 0.89,$$

here, $x_{1,t}$ – turnover of G45, $x_{2,t}$ – VAT of G45, $S_{1,t}, S_{4,t}$ – 1st and 4th seasonal dummies, O_t – level shift outlier.

$$y_{46,t} = \underset{(0.024)}{163} - \underset{(0.049)}{0.0001} \cdot x_{1,t} + \underset{(0.006)}{0.0001} \cdot x_{2,t} - \underset{(0.049)}{29.5} \cdot S_{1,t} - \underset{(0.049)}{32.6} \cdot S_{4,t} + \underset{(0.049)}{6.7} \cdot O_t + \varepsilon_{i,t}, R^2 = 0.92,$$

here, $x_{1,t}$ – VAT of G46, $x_{2,t}$ – wholesales trade results, $S_{1,t}, S_{4,t}$ – 1st and 4th seasonal dummies, O_t – level shift outlier.

$$y_{47,t} = \underset{(0.049)}{81} + \underset{(0.000)}{0.0002} \cdot x_{1,t} + \underset{(0.049)}{0.0003} \cdot x_{2,t} - \underset{(0.004)}{35.7} \cdot S_{1,t} - \underset{(0.049)}{4.4} \cdot S_{4,t} + \varepsilon_{i,t}, R^2 = 0.93,$$

here, $x_{1,t}$ – retail trade results, $x_{2,t}$ – VAT of G47, $S_{1,t}, S_{4,t}$ – 1st and 4th seasonal dummies.

All model's parameters above are significant, R^2 is acceptable high and models errors $\varepsilon_{i,t}$ satisfies Gauss-Markov assumptions. G section value added estimate will be obtained from these models fitted values.

ARIMAX model is used for telecommunications (J61) value added estimation:

$$y_{61,t} = 8.1 + 0.0015 \cdot y_{61,t-1} + 0.0001 \cdot x_{1,t} + 0{,}52 \cdot x_{2,t} - 2{,}34.4 \cdot S_{4,t} + \varepsilon_{i,t}, AIC = 207,$$

here explanatory variables and seasonal dummy are included: $x_{1,t}$ – VAT of J61, $x_{2,t}$ – J61 services incomes.

Combining linear regressions or ARIMA(X) models all section's economical activities are estimated. Then chain – linking method is used to obtain constant price values and real GDP growth (Bloem, 2001), additionally observing 88 time series. Moreover, the GDP deflators are analysed in comparison to logical economic sense, in order to prepare qualitative results.

5. Combined forecasting

Use of different length of time series for estimation, explanatory variables choice, models specifications can bring different results in forecasting. Other models can capture some information that is not contained in the others. To choose the best result we can forecast using all of the plausible models and then combine forecasts to obtain final estimate. There are several ways to combine forecasts and choose the most accurate estimate (Enders, 2014):

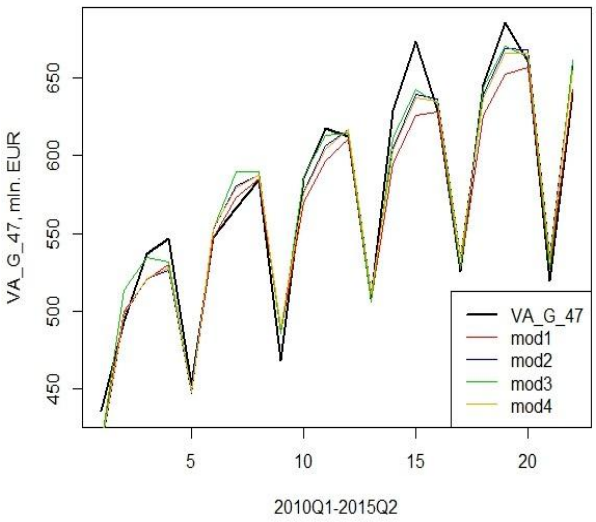
Table 1. Forecasts possible choices.

No.	Option	Formula
1.	*Average of all forecasts*	$\hat{f} = \frac{1}{n}\sum_{i=1}^{n}\hat{f}_i$, here n – number of forecasts, $i = (1, \cdots n)$ – index of the model, $\hat{f}_i$– forecast of i model.
2.	*Minimum forecast*	$\min_i \hat{f}_i$.
3.	*Maximum forecast*	$\max_i \hat{f}_i$.
4.	*Weighted average of forecasts*	$\hat{f} = w_1\hat{f}_1 + w_2\hat{f}_2 + \cdots + w_n\hat{f}_n$, here $\sum_{i=1}^{n} w_i = 1$.

Based on these main four combined forecasting methods from all plausible forecast and economic assumptions according current situation, group of experts makes final decision for GDP flash estimates in SL. To make final decision, analytical economic analysis of current situation in economy is obligatory. Obtained estimations can be compared with short term business, prices statistics available on this time.

For example, if different length of time series is chosen for economic activity G47 value added evaluation, different results will be obtained (See Picture 2).

Picture 2. Estimates comparison.



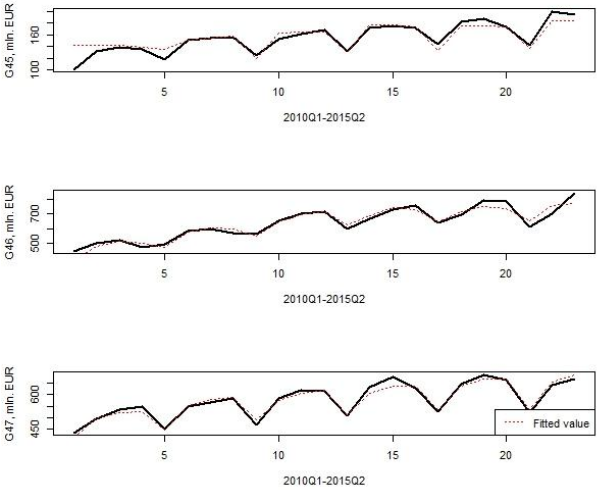
There are 4 models which with different time's series length of all variables brought different results. To pick up 1 final result, the following characteristics can be used: min, max, average or weighted average of estimates. At this situation all obtained estimates are relatively good. However, the model's No 4 estimates would be chosen, because their Mean absolute percentage error is lower than others.

6. Quality assurance

To ensure quality of the results, the analysis of error monitoring results is carried out. It consists of two parts:

1. **Graphical analysis**. True values of value added and fitted values are compared.

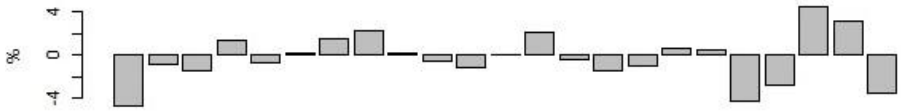
Picture 3. G45-47 estimates plots.



Plots in the picture 3 presents' good estimated results of the each economic activity in section G, fitted values are close to the real values.

2. **Error analysis**. In this part Absolute Percentage Error $APE(t) = \frac{|Y(t) - \hat{Y}_i(t)|}{Y(t)} \cdot 100$ and Mean Absolute Percentage Error $MAPE = \frac{1}{k}\sum_{t=N-k}^{N} \left|\frac{y(t) - \hat{y}_i(t)}{y(t)}\right| \cdot 100$ are compared (Enders, 2014). Here, $k = 1, \cdots, N$ is number of observations, $\hat{y}_i(t)$ – fitted value of the i economic activity. We assume that fitted values are good and acceptable to publish if the $\mathrm{APE(t)} < 5$ %, $MAPE < 5$ %.

Picture 4. Errors of G activity models.



$MAPE_G = \frac{1}{k}\sum_{t=N-k}^{N}\left|\frac{y(t)-\hat{y}_i(t)}{y(t)}\right| \cdot 100 = 1.68\ \%$. Errors are small (see Picture 4), APE and MAPE are less then 5 %. It fulfils quality criteria.

2014-2015 was the period of methodological changes in ESA2010. A huge revision was done for GDP and value added components. Revision of linear regressions and ARIMA(X) models were also done. Non-significant regressors were eliminated from models, new regressions were added, and seasonal dummies were revised. Models revision let to minimize MAPEs. Despite, ESA2010 changes and revision of the models, modelling system remains stable.

Table 2. Mean absolute percentage errors of real GDP components.

Activity	A	B- E	C	F	G- I	J	K	L	M- N	O- Q	R- T	**GDP**
MAPEs, before ESA2010	5.6	1.9	1.9	5.5	3.5	3.0	5.5	38	3.4	3.2	5.5	**2.1**
MAPEs, after ESA2010	4.6	1.6	1.6	5.4	3.0	2.9	4.8	3.4	2.9	2.7	5.0	**0.3**

The calculated aggregation A*10 MAPEs after the ESA2010 revision shows that only 1–2 aggregates are lower than quality acceptations criteria limits, which, of course, require additional work and analysis. Despites this, the main indicator GDP is of a very good quality, its MAPE does not overcome 0.3 %; moreover, models revision allows reaching better results.

7. Conclusions

GDP t+30 estimation methodology of SL is a complex system with many classical extended econometrics' equations, mathematical tools. Large data set and connections, which cannot be broken, are integrated. The GDP t+30 estimation process insists data analysis and adjustments, modelling, estimates graphical and statistical analysis, economic critical evaluation and final

estimate preparation. Although, estimation process is very intensive (6 SL's 10 days), SL performs high quality, timeline GDP t+30 estimates, where MAPEs do not reach 1 %. This allows satisfying user needs and preparing first signal of current economies' situation. Extended multiple regressions and ARIMA family models compose a stable system to obtain GDP t+30 estimates. It allows adapting to the global methodological changes (like ESA 2010) and carrying out estimates of the same high quality as before. Nevertheless, constant models and quality monitoring is necessary. SL intends to supervise models to according accurate, timelines and quality of estimates.

Reference

Bloem, A. M., Dippelsman, R. J., Maehle, N. O., 2001, *Quarterly National Accounts Manual*, International Monetary Fund, Washington, USA.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C. 1994, *Time Series Analysis. Forecasting and Control*, New Jersey, USA.

Charemza, W. W., Deadman, D., 1997, *New Directions in Econometric practice*, Edgard Elgar Cheltenham, UK.

Cuthbertson K., Hall S., Taylor M., 1992, *Applied Econometric Techniques*. New York, Harvester Wheatsheaf, USA.

Čekanavičius, V., Murauskas, G., 2002, *Statistika ir jos taikymai II*, Vilnius, Lithuania.

Enders, W., 2014, *Applied Econometric Time Series*, 4th ed., University of Alabama, USA.

Leipus, R., 2013, Ekonometrija II, Vilnius, Lithuania.

Greene, W. H., 2003, *Econometric analysis*, 5th ed., New Jersey, USA.

Račkauskas, A., 2003, *Ekonometrijos įvadas*, Vilnius, Lithuania.