VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

Magistro darbas

Lietuvos mokestinio efektyvumo tyrimas naudojantis paneliniais duomenimis

The analysis of Lithuania's tax efficiency using panel data

Lukas Kazlauskas

VILNIUS 2020

Taikomosios Matematikos Institutas

Statistinės Analizės Katedra

Darbo vadovas _____

Darbo recenzentas _____

Darbas apgintas_____

Darbas įvertintas _____

Registravimo Nr. _____

Gavimo data _____

Contents

Lietuvos mokestinio efektyvumo tyrimas naudojantis paneliniais duomenimis

Santrauka

Neseniai vykusios ekonominės suirutės parodė, jog fiskalinis sektorius gali būti naudingas įrankis valstybės stabilumui garantuoti. Siekis subalansuoti biudžetą skatina vyriausybes taikyti mokesčių didinimą arba išlaidų mažinimą. Susiklosčiusi situacija reikalauja gilesnio mokestinio potencialo suvokimo ir analizės, nes jį žinant, būtų galima spęsti, ar šaliai galima didinti mokesčius be žalos. Šiame darbe, naudojant panelinius duomenis, tiriamas Lietuvos mokestinis efektyvumas. Tikslui įgyvendinti yra naudojamas laiko eilučių klasterizavimas, impulso atsako analizė bei dinaminiai paneliniai modeliai. Tyrimo rezultatai parodė, kad Lietuvos žemės ūkio bei užsieno prekybos sektoriai yra atitinkamai mažesni ir didesni, lyginant su kitomis valstybėmis, turinčiomis panašų mokestinį lygį. Impulso atsako analizė atskleidė, jog makreokonominiai kintamieji nevienodai veikia skirtingų pajamų valstybių grupes. Galiausiai, fiksuotų efektų dynaminis panelinis modelis parodė, jog Lietuva yra labiau nutolusi nuo savo moekstinio potencialo nei didžioji dauguma valstybių.

Raktiniai žodžiai: Dinaminiai paneliniai modeliai, impulso atsako analizė, mokesčių potencialas.

The analysis of Lithuania's tax efficiency using panel data

Abstract

Recent economic turmoil has showed that the fiscal sector can be a useful tool in providing stability for a country. The need for a balanced budget has led governments imposing tax increases and various spending cuts. This calls for an inquiry into the analysis of tax potential, as having a position on the matter would shed light on how much a country can increase their taxes without causing harm. This thesis aims to analyze Lithuania's tax efficiency by using cross sectional data. To achieve this task, time series clustering, impulse response analysis and dynamic panel modeling is carried out. The results show that Lithuania's agricultural and foreign trade sectors are smaller and larger respectively, compared to other countries, that have a similar taxation level as Lithuania. Impulse response analysis showed that the macroeconomic determinants have different effects in different income country groups. Lastly, the fixed effects dynamic panel data model revealed that Lithuania is further away from it's tax potential than most of the countries.

Key words : Dynamic panel data model, impulse response analysis, tax potential.

1 Introduction

One of the key factors in robust economic growth is an effective tax system. Fiscal policy itself plays a major role in wealth redistribution, economic stability and welfare. It also acts a prominent determinant in investment, both internally and externally because a well balanced tax system can be a directional guide to a countries residential investors as well as attract foreign investments by acting as a tax haven. Besides these and other areas affected, the fiscal sector plays a role in determining the perception of the government in the position of its tax payer.

In recent years, the global economy has endured several shocks, which caused great turmoil throughout major countries. One of those shocks - the Eurozone crisis of 2012 - showed that an unbalanced fiscal system can cause severe damage to a country's financial structure and, due to deep interconnections between nations, become a heavy toll on its partners. The aftermath of these events led countries to accumulate high debts which in turn now forces them to increase their taxes. Now one can see, that multilateral institutions often are offering their help to developing or developed countries to ensure a coherent and well harmonized tax system.

Given all these reasons, a natural need for an evaluation of a country's tax system comes into demand. Having a position on the matter would allow a country to impose fiscal changes more efficiently. It is common sense, that countries who want implement new taxes or increase the rates of existing ones, should first analyze how much their revenue or rates are far from their potential. So if a country would have the need to increase their budget while being very close to their taxable limit, increasing taxes could possibly work in a reverse way, as the Laffer curve dictates. This would in turn lead to the search for other sources of income for the state. Alternately, if, after a tax rise, budget revenues would not increase, and the country would be far from its taxable potential, this would indicate that, possibly, the state's administration is doing a poor job in tax collection or the taxpayers are unwilling to pay for some reason.

In order to quantitatively estimate a country's tax potential, two concepts have to be introduced: tax effort and tax capacity. Tax capacity is the maximum limit of the tax revenues that can be collected in a country, which is usually calculated by using econometric models. Tax effort is the ratio between actual tax revenues and tax capacity. This metric allows to see how distant a country is from its taxable potential, or in other words - how efficiently it collects its revenue.

Of course, for a better understanding of tax effort, merely calculating this ratio for a country is not enough. One could evaluate tax efforts of a broad range of countries and then compare the results together. If a country's tax effort would be different from similar countries, this could indicate that, perhaps, there are some issues with the state's tax administration system. Furthermore, having this information would allow to grade the tax effort by knowing what is a relatively small or high value.

Having all this said, this thesis aims to analyze Lithuania's tax efficiency as well as estimate and compare its tax effort by using cross sectional data and econometric modeling. To do this, several tasks are set. First, time series clustering is employed in order to assess Lithuania's position among other countries. Second, panel vector autoregressive modeling is used and impulse response functions are analyzed on different income country groups, which has not been done before. Lastly, a new type of dynamic fixed effects panel data model is fitted in order to estimate tax capacity and effort.

The first part of the thesis provides theoretical background for the standard and dynamic fixed effects panel data models. Then, the panel vector autoregression is introduced alongside with panel Granger causality and unit root tests. Next, empirical works in the determination of tax revenue factors and tax effort estimates are provided. The theoretical section ends with discrete time series and functional principal component score clustering techniques.

The second part of the thesis gives the results from time series clustering. Then, impulse response functions from panel vector autoregression models are presented. Lastly, dynamic fixed effects panel data modeling is carried out which gives tax effort and tax capacity for all analyzed countries.

The methods used in this thesis expand previously done researches which estimated tax capacity using panel data. Usually static panel data models are employed for this task. By using impulse response analysis and the dynamic fixed effects panel data approach, we add more information to the model and account for heterogeneous variable effects in different income groups.

2 Theoretical background

2.1 Panel data model framework

2.1.1 Panel data model

Before beginning the analysis of tax capacity and tax effort, panel data modeling has to be introduced. The standard panel data model has the following form (Wooldridge 2010):

$$y_{it} = \beta' x_{it} + \alpha_i^* + \epsilon_{it} \tag{1}$$

where x_{it} is an independent variable, ϵ_{it} is the error term, α_i^* is the unobserved effect, $i = 1, 2, ..., N$ is the number of objects analyzed and $t = 1, 2, ..., T$ is the time periods. Fixed effects is included because it is possible that by modeling some process, which requires a panel data approach, some variables will be omitted, thus the unobserved effect deals with this problem.

Usually, α_i^* can be treated as a random or fixed effect. The main difference between these two terms is that if the unobserved effect is understood as random, then $\mathbb{E}(\alpha_i^*|x_i) = 0$ holds, where in a fixed effects model, it does not. This simply means that in a fixed effects framework, the unobserved effect is allowed to be correlated with other independent variables, whereas in a random effects approach it does not.

The latter assumption is very strict and in real life situations, the unobserved effect is likely to be correlated with other regressors, thus fixed effects is usually used. While testing for a fixed or random model specification can be done by tests (e.g. the Hausman test), Clark and Linzer (2012) state that in most applications, the true correlation between the covariates and the unobserved effects is not exactly zero. Thus failing to reject the null (if rejected - fixed effects are used) is most likely not because the true correlation is zero, but because the test does not have sufficient statistical power to identify departures from te null.

In order for the estimates of (1) to be valid, the following assumption must hold:

$$\mathbb{E}(x_{it}' \epsilon_{it}) = 0 \ \forall \ t \tag{2}$$

This means that the error term and the explanatory variables are not contemporaneously correlated. This assumption does not impose strict exogeneity. We now move on the an expansion of the standard panel model - the dynamic panel fixed effects data model.

2.1.2 Dynamic fixed effects panel data model

A lot of economic relationships have a dynamic nature. One of the advantages of panel data is that they allow the user to understand the dynamics of adjustment. These dynamic relationships are usually modeled by including a lagged dependent variable among the regressors (Baltagi, 2002):

$$y_{it} = \gamma y_{i,t-1} + \beta' x_{it} + \alpha_i^* + \epsilon_{it} \tag{3}$$

where $i = 1, ..., N$, $t = 1, ..., T$, α_i^* are fixed effect, ϵ_{it} is the error term with $\mathbb{E}(\epsilon_{it}) = 0$, and $\mathbb{E}(\epsilon_{it}\epsilon_{js}) = \sigma_\epsilon^2$ if $i = j$, $t = s$ and $\mathbb{E}(\epsilon_{it}\epsilon_{js}) = 0$ otherwise.

Since it is usually assumed that there will always be omitted information, dynamic panel data models include a fixed effect. If the lagged dependent variable in (3) appears as explanatory, exogeneity no longer holds. The fixed effects and the first differences estimators rely on this condition, thus the standard estimation methods (OLS or its alternatives) can fail. This problem is also know as the Nickell Bias. For example, demeaning the dynamic process (3) with only the lagged dependent variable in order to account for fixed effects leads to:

$$y_{it} - \bar{y}_{i.} = \gamma(y_{it-1} - \bar{y}_{i.-1}) + (\epsilon_{it} - \epsilon_{i.}) \tag{4}$$

Nickell (1981) shows that (4) creates a correlation between the regressor and the error which results in a bias. This bias arises when the data is quite short, meaning a relatively small T. Nickell in 1981 provided analytical expressions of the bias, that have been previously documented in Nerlove (1967,1971). Due to this bias, an alternative method of estimation must be used.

One of these alternatives is the use of instruments. Let Z denote a $N \times H$ matrix. An instrument, or instrumental variables Z, must satisfy these two properties:

$$\begin{aligned} \mathbb{E}(\epsilon|Z) &= 0 \\ \mathbb{E}(x_{jk}z_{jh}) &\neq 0 \end{aligned} \tag{5}$$

for $h \in 1, ..., H$ and $k \in 1, ..., K$. This means that the instruments are correlated with the independent variables X and exogenous to the error term.

Using instrumental variables, Arellano and Bond (1991) proposed a method, called the Generalized method of moments (GMM) in order to estimate β and γ from (3).

This method takes on the following assumptions from (3) about the error and fixed effects

terms:

$$\begin{aligned}
&\mathbb{E}(\epsilon_{it}) = 0,\ \mathbb{E}(\alpha_i^*) = 0 \\
&\mathbb{E}(\epsilon_{it}\epsilon_{js}) = \sigma_\epsilon^2 \text{ if } j = i \text{ and } t = s, 0 \text{ otherwise} \\
&\mathbb{E}(\alpha_i^*\alpha_j^*) = \sigma_\alpha^2 \text{ if } j = i\ , 0 \text{ otherwise} \\
&\mathbb{E}(\alpha_i^* x_{it}) = 0
\end{aligned} \tag{6}$$

The GMM approach is based on a model in first differences in order remove the individual effects of α_i^*:

$$(y_{it} - y_{i,t-1}) = \gamma(y_{i,t-1} - y_{i,t-2}) + \beta'(x_{it} - x_{i,t-1}) + \epsilon_{it} - \epsilon_{i,t-1} \tag{7}$$

It can be noticed, that all the lagged variables $y_{i,t-2-j}\ \forall\ j\ \leq\ 0$, satisfy both exogeneity and relevance properties:

$$\begin{aligned}
&\mathbb{E}(y_{i,t-2-j}(y_{i,t-1} - y_{i,t-2})) = 0 \\
&\mathbb{E}(y_{i,t-2-j}(\epsilon_{i,t} - \epsilon_{i,t-1})) = 0
\end{aligned} \tag{8}$$

This means that they are all legitimate instruments for $y_{i,t-1} - y_{i,t-2}$, meaning that lags of the dependent variable can be used as instruments.

Next, the $m + 1$ conditions $\mathbb{E}(y_{i,t-2-j}(\epsilon_{i,t} - \epsilon_{i,t-1})) = 0\ for\ j = 0, ..., m$ can be used to estimate β, γ from (3).

Assuming, that $\mathbb{E}(x_{it}'\epsilon_{is}) = 0\ \forall\ (t, s)$, for each period, the moment conditions can be expressed as:

$$\begin{aligned}
&\mathbb{E}(q_{i,t}\Delta\epsilon_{it}) = 0,\ t = 2, ..., T \\
&q_{i,t} = (y_{i0}, y_{i0}, ..., y_{i0}, x_i')'
\end{aligned} \tag{9}$$

where $x_i' = (x_{i1}', ..., x_{iT}')$, $\Delta = (1 - L)$ and L denotes the lag operator.

A problem arises with the instrumental variable approach. If the correlation between the explanatory and instrumental variables is small, then these instruments are called weak. To solve this problem, Blundell and Bond (2000) proposed an estimation method called system GMM. It uses moment conditions in both level and in first differences:

$$\begin{aligned}
&\mathbb{E}(y_{i,t-s}\Delta\epsilon_{it}) = 0,\quad \mathbb{E}(x_{i,t-s}\Delta\epsilon_{it}) = 0 \\
&\mathbb{E}(\Delta y_{i,t-s}(\alpha_i^* + \epsilon_{it})) = 0,\quad \mathbb{E}(\Delta x_{i,t-s}(\alpha_i^* + \epsilon_{it})) = 0
\end{aligned} \tag{10}$$

GMM and system GMM estimates use instruments which are usually lags of the dependent variable. The Sargan's test for over-identifying restrictions was suggested by Arellano and Bond (1991) in order to check the validity of the instruments. The test statistic is given by:

$$m = \Delta\hat{\epsilon}' Z \left[\sum_{i=1}^{N} Z_i'(\Delta\hat{\epsilon}_i)(\Delta\hat{\epsilon}_i)' Z_i \right]^{-1} Z_i(\Delta\hat{\epsilon}) \sim \chi^2_{p-K-1} \tag{11}$$

Where Z is the instrument matrix, ϵ is the error term, p is the number of columns of Z. Under the null hypothesis, the instruments are valid.

Both methods (GMM and system GMM) are considered to be a far better alternative than the standard OLS estimation when evaluating dynamic models. Nevertheless, this approach has met some criticism. According to Behr (2003) GMM type estimators are inefficient. Furthermore this approach imposes assumptions about the appropriateness of past values of the dependent variable which are used as instruments for estimation. These assumptions may or may be not valid.

Lancaster (2002) proposes a new type of dynamic fixed effects panel data modeling approach by suggesting a conditional likelihood estimator that can analytically compute the conditional probability distributions of the variables at hand and does not require instrumental variables. Hsiao (2014) and Hsiao et al. (2002) has shown that this approach performs better than GMM estimators.

Suppose that the fixed effects can be re-parameterized so that the likelihood function for the data for a single case factors looks like this:

$$l_i(\alpha_i^*, \gamma, \beta, \sigma^2) = l_{i1}(\alpha_i^*) l_{i2}(\gamma, \beta, \sigma^2) \tag{12}$$

Where l_{i1} and l_{i2} are likelihood functions. If the parameters α_i^* and $(\gamma, \beta, \sigma^2)$ are also variation independent, they are orthogonal. If it can be said that $\prod l_{i2}$ is the product of l_{i2} for all observations, in Lancaster (2002) it can be shown that the application of maximum likelihood to $\prod l_{i2}$ gives consistent estimates of $(\gamma, \beta, \sigma^2)$ as $N \to \infty$ for any $T \geq 2$.

However, not all likelihoods can be transformed in such a way that the parameters become orthogonal. It is possible to reparametrize the fixed effects so that they are information orthogonal. Denoting the log likelihood for the data for observation i as L_i, then the fixed effects are information orthogonal to, for example, β if the following condition holds true:

$$\mathbb{E}\left(\frac{\delta L_i}{\delta \alpha_i^* \delta \beta} \right) = 0 \tag{13}$$

From (13) it can be understood, that if the slope of the log likelihood with respect to α_i^* is independent of the slope of the log likelihood with respect to β, then α_i^* is information orthogonal to β. If it is possible to perform such a transformation that this condition is met, then it may be possible to place priors on the parameters and integrate out the fixed

effects. Flat priors are used for the α_i^* and the remaining parameters. This is essentially a Bayesian estimation technique. With such a framework, marginal posteriors for the remaining parameters are obtained. Then Monte Carlo methods can be used to sample values from the marginal posterior to produce estimates and credible intervals for the parameters.

The posterior densities are obtained as follows. Denote $X_{i,t}, y_{i,t}$ and $y_{i,t-1}$ in vector terms X_i, y_i and y_{i-}. The appropriate reparametrization of the fixed effects, forming uniform priors on γ, β, σ^2 and integrating out the fixed effects results in the following posterior density function:

$$p(\gamma, \beta, \sigma^2|data) \propto \sigma^B exp\left\{ \frac{N}{T} \sum_{t=1}^{T-1} \left(\frac{T-t}{t} \gamma^t \right) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (A - \beta X_i)' H (A - \beta X_i) \right\} \tag{14}$$

where $A = y_i - \gamma y_{i-}$, $B = -(N(T-1)-2)$ and H is defined as an operator that subtracts the mean. For example, if

$$\omega_i = y_i - \rho y_i - \beta_1 X_i \tag{15}$$

then $H(\omega_i) \equiv \omega_i - \bar{\omega_i}$.

Sampling from this posterior, gives the distributions of estimates for the parameters γ, β, σ^2. First, β must be integrated out of (14). After doing so, the following density is achieved:

$$\begin{aligned} &p(\gamma, \sigma^2|data) \propto \sigma^B exp\left\{ \frac{N}{T} \sum_{t=1}^{T-1} \left(\frac{T-t}{t} \gamma^t \right) \right\} \\ &exp\left\{ -\frac{1}{2\sigma^2} \left(\Big(\sum_{i=1}^{N} (X_i)' H(A) \Big)' \Big(\sum_{i=1}^{N} (X_i)' H(X_i) \Big) \Big(\sum_{i=1}^{N} (X_i)' H(A) \Big) \right) \right\} \end{aligned} \tag{16}$$

Next, σ^2 is integrated out of (16) giving the marginal posterior density $p(\gamma|data) \propto$:

$$\frac{exp\left\{ \frac{N}{T} \sum_{i=1}^{T-1} \left(\frac{T-t}{t} \gamma^t \right) \right\}}{\left((A)' H(A) - \Big(\sum_{i=1}^{N} (X_i)' H(A) \Big)' \Big(\sum_{i=1}^{N} (X_i)' H(X_i) \Big)^{-1} \Big(\sum_{i=1}^{N} (X_i)' H(A) \Big) \right)^C} \tag{17}$$

where $C = \left(\frac{N(T-1)-K}{2} \right)$.

Now, sampling from (17) gives γ, then given γ sample $1/\sigma^2$ from (16). Lastly, given γ and $1/\sigma^2$ sample β from (14). The medians of these samples are the estimated parameters.

Pickup et al. (2017) using data simulations showed that the GMM estimator is a large improvement compared to OLS, but the orthogonal reparametrization approach perform as

well as the GMM or even better. This model evaluation approach is also attractive because it provides the distributions of each β and γ parameters which are then used for significance hypothesis testing. Also, no instruments are necessary for the evaluation of the model.

Next, the panel vector autoregression model is introduced.

2.1.3 Panel vector auto regression model

In order to have a better understanding in the relationships between variables in a dynamic system, a panel vector auto regression (PVAR) can be used, which is an expansion of the standard dynamic panel data model. The very first PVAR was introduced by Holtz-Eakin et al. (1988). In time, it has been extended to use p lags of m endogenous, k predetermined and n strictly exogenous variables. With this in regard, the PVAR model takes on the following form:

$$\mathbf{y}_{i,t} = \sum_{l=1}^{p} \mathbf{A}_l \mathbf{y}_{i,t-l} + \mathbf{B}\mathbf{x}_{i,t} + \mathbf{C}\mathbf{s}_{i,t} + \epsilon_{i,t} \tag{18}$$

where $\mathbf{y}_{i,t-l}$ are lagged endogenous variables, $\mathbf{x}_{i,t}$ are predetermined variables, $\mathbf{s}_{i,t}$ are exogenous variables.

The disturbances $\epsilon_{i,t}$ are independently and identically distributed (i.i.d.) for all i and t with $\mathbb{E}[\epsilon_{i,t}] = 0$ and $Var[\epsilon_{i,t}] = \Sigma_\epsilon$. Σ_ϵ is a positive semi-definite matrix. It is also assumed, that all unit roots of $\mathbf{A}$ in (18) fall inside the unit circle to assure co-variance stationarity (stability condition), which is done by inspecting the eigenvalues of an estimated model.

It becomes clear, that a PVAR model is hence a combination of a single equation dynamic panel model and a PVAR model.

Before evaluating the model, a transformation of first difference or the forward orthogonal transformation is to be used:

$$\Delta^* \mathbf{y}_{i,t} = \sum_{l=1}^{p} \mathbf{A}_l \Delta^* \mathbf{y}_{i,t-l} + \mathbf{B}\Delta^* \mathbf{x}_{i,t} + \mathbf{C}\Delta^* \mathbf{s}_{i,t} + \Delta^* \epsilon_{i,t} \tag{19}$$

Here Δ^* is the first difference or the forward orthogonal transformation. The former exists for $t \in \{p+2, ..., T\}$ and the latter - for $t \in \{p+1, ..., T-1\}$. The set of indexes, for which the transformation exists is denoted by T_{Δ^*}. These two transformations can be summed up in the following way: the first difference transformation subtracts the previous value from the current value and the forward orthogonal deviation transformation subtracts the average of all available future observations from the current value. While first difference drops the first observation on each individual in the panel, forward orthogonal deviations drops the last observation for each individual.

Below are the moment conditions are provided, which are necessary for the model to be estimated via GMM, which was discussed previously. These are slightly different, due to the multi-equation model structure.

$$\begin{aligned} &E[\Delta^* \epsilon_{i,t} \mathbf{y}_{i,j}^T] = 0 \quad j \in \{1, ..., T-2\} \quad \text{and} \quad t \in T_{\Delta^*}, \\ &E[\Delta^* \epsilon_{i,t} \mathbf{x}_{i,j}^T] = 0 \quad j \in \{1, ..., T-1\} \quad \text{and} \quad t \in T_{\Delta^*}, \\ &E[\Delta^* \epsilon_{i,t} \Delta^* \mathbf{s}_{i,j}^T] = 0 \quad t \in T_{\Delta^*} \end{aligned} \tag{20}$$

Additional moment conditions can be constructed when imposing the following assumption on the structure of the process. These assumptions are sufficient enough for the system GMM estimator.

$$\begin{aligned} &\mathbb{E}[\epsilon_{i,t}(\mathbf{y}_{i,t-1} - \mathbf{y}_{i,t-2})^T] = 0 \quad t \in \{3, 4..., T\} \\ &\mathbb{E}[\epsilon_{i,t}(\mathbf{x}_{i,t-1} - \mathbf{x}_{i,t-1})^T] = 0 \quad t \in \{2, 3..., T\} \\ &\mathbb{E}[\epsilon_{i,t} \mathbf{s}_{i,t}^T] = 0 \quad t \in \{2, 3..., T\} \end{aligned} \tag{21}$$

According to Blundell and Bond (1998), this assumption is clearly satisfied in a stationary PVAR model.

Like the single equation system GMM dynamic panel model, the PVAR also uses lags of the dependent (in this case the endogenous) variable for estimation. Blundell and Bond (1998) also argue that the system GMM estimator performs better than the GMM estimator because the additional instruments remain good predictors for the endogenous variables in this model even when the series are very persistent.

Having evaluated the model, one can use impulse response functions (IRF) in order to have a better understanding of the process at hand.

IRF analysis is used in order to understand, how one endogenous variable responds to an impulse of another endogenous variable. To achieve this, the PVMA-X (panel vector moving average representation with exogenous variables) of a PVAR-X(1) process is used:

$$\mathbf{y}_{i,t} = \left(\sum_{j=0}^{\infty} \mathbf{A}^{j-1} [\mathbf{BC}] \right) \begin{bmatrix} \mathbf{x}_{i,t-j} \\ \mathbf{s}_{i,t-j} \end{bmatrix} + \left(\sum_{j=0}^{\infty} \mathbf{A}^j \right) [\epsilon_{i,t-j}] \tag{22}$$

Both predetermined and strictly exogenous variables are treated all the same. From here, we can derive the impulse response function:

$$IRF(k, r) = \frac{\delta \mathbf{y}_{i,t+k}}{\delta (\epsilon_{i,t})_r} = \mathbf{A}^k \mathbf{e}_r \tag{23}$$

where k is the number of periods after the shock to the rth component of $\epsilon_{i,t}$ with $\mathbf{e}_r$ being a vector with a 1 in the rth column and 0 otherwise.

Let $\mathbf{\Sigma}_\epsilon$ be the covariance matrix of ϵ_t. It is quite common for the off diagonal elements of this matrix to be different from 0, which leads to the fact that shocks across the m equations are not independent of each other. This in turn means that the parameters of the PVAR model have to be altered in such that the responses to the "independent" shocks are transferred through the PVAR system accordingly.

Because of the assumption that $\mathbf{\Sigma}_\epsilon$ is a symmetric positive semi-definite matrix, there exists a unique Cholesky decomposition such that $\mathbf{\Sigma}_\epsilon = \mathbf{P}\mathbf{P}^T$, where $\mathbf{P}$ is a lower triangular matrix. If one defines $\mathbf{\Theta}_k = \mathbf{A}^K\mathbf{P}$ and $\mathbf{u}_{i,t} = \mathbf{P}^{-1}\epsilon_{i,t}$ we obtain the orthogonal impulse response function:

$$OIRF(k,r) = \frac{\delta \mathbf{y}_{i,t+k}}{\delta(\mathbf{u}_{i,t})_r} = \mathbf{\Theta}_k \mathbf{e}_r \tag{24}$$

As stated in Lutkepohl (2007), the Cholesky decomposition depends on the ordering of the variables and thus has been criticized. Pesaran and Shin (1998) proposed an alternative to OIRF. Instead of shocking all elements of $\epsilon_{i,t}$ they chose to shock only one element, say the rth and integrate out the effects of other shocks using the historically observed distribution of errors, so we have:

$$GIRF(k,r,\mathbf{\Sigma}_\epsilon) = \mathbb{E}[\mathbf{y}_{i,t+k}|\epsilon_{i,t,r} = \delta_r, \mathbf{\Sigma}_\epsilon] - \mathbb{E}[\mathbf{y}_{i,t+k}|\mathbf{\Sigma}_\epsilon] \tag{25}$$

By setting $\delta_r = \sqrt{\mathbf{\Sigma}_{\epsilon,r,r}}$ the generalized impulse response function is obtained:

$$GIRF(k,r,\mathbf{\Sigma}_\epsilon) = \mathbf{A}^k\mathbf{\Sigma}_\epsilon(\sigma_{r,r})^{-1/2} \tag{26}$$

where $\sigma_{r,r}$ is the rth diagonal element of $\mathbf{\Sigma}_\epsilon$.

Using IRF can give valuable information into the dynamic relationships of the variables at hand. Usually, before implementing PVAR modeling, the Granger causality test are performed in order to see, if a time series causes another time series.

2.1.4 Panel Granger causality

In order to test whether a time series has an effect on the forecasts of another time series, one can perform the Granger causality test. In 1969 Granger proposed this test for univariate time series. It is said, that y causes x if the inclusion of y as a regressor in x improves the forecast of x.

In 2012, Dumitrescu and Hurlin expanded Grangers framework and modified the test to be able to detect causal relationships in panel data. Let x and y be stationary processes. The underlying regression writes as follows:

$$y_{it} = \delta_i + \sum_{k=1}^{K} \gamma_{ik} y_{i,t-k} + \sum_{k=1}^{K} \beta_{ik} x_{i,t-k} + \epsilon_{it} \tag{27}$$

Note that the coefficients are allowed to differ across individuals. The test's null hypothesis is defined as:

$$H_0 : \gamma_{i1} = ... = \gamma_{iK} = 0 \ \ \forall \ i = 1, ..., N \tag{28}$$

(28) corresponds to the absence of causality of all individuals in the panel.

The alternative hypothesis of the test states that there can be causality for some individuals but not necessarily for all:

$$\begin{aligned} H_1 : &\gamma_{i1} = ... = \gamma_{iK} = 0 \ \ \forall \ i = 1, ..., N_1 \\ &\gamma_{i1} \neq 0 \ \ \text{or} \ \ ... \ \ \text{or} \ \ \gamma_{iK} \neq 0 \ \ \forall \ i = N_1 + 1, ..., N \end{aligned} \tag{29}$$

Granger causality is usually done before constructing VAR models, as a complementary analysis.

2.1.5 Panel unit root test

It is well known, that some time series models, usually in forms of autoregression, require the process to be stationary. A time series is said to be stationary if its mean and variance does not change over time. To test if a process is stationary or not, unit root test can be performed. A time series is said to have a unit root if it is not stationary. The first well known unit root test was constructed for univariate time series by Dickey and Fuller in 1984, but a few decades later, other academics introduced unit root test for panel data.

Kleiber and Lupi (2011) denote a series having a unit root with $I(1)$. Thus a series without unit roots is $I(0)$. Naturally, the null hypothesis of a panel unit root test states that all of the series are $I(1)$. The alternatives are split into two: H_1^A : *all series are* $I(0)$ and H_1^B : *at least one series is* $I(0)$. In practice, test that consider H_1^A are less flexible due to a more strict alternative.

The test can be conducted in two approaches - through t ratios or p values. If the former is chosen, then consider the following process:

$$\Delta y_{it} = \mu_i + \rho_i y_{i,t-1} + \sum_{j=1}^{k_i} \phi_{i,j} \Delta y_{i,t-j} + \epsilon_{it} \tag{30}$$

Here ϵ_{it} is independent and identically distributed (i.i.d) with $\mathbb{E}(\epsilon_{it}) = 0, \ \mathbb{E}(\epsilon_{it}^2) = \sigma_i^2 < \infty, \ \mathbb{E}(\epsilon_{it}^4) < \infty$. Under the null, $H_0 : \rho_i = 0 \ \forall \ i$.

(30) can be rewritten:

$$\mathbf{\Delta y_i} = \rho_\mathbf{i} \mathbf{y_{i,-1}} + \mathbf{\Upsilon_i} \gamma_\mathbf{i} + \epsilon_\mathbf{i} \tag{31}$$

where $\mathbf{\Delta y_i} = (\Delta y_{i,k_i+2}, ..., \Delta y_{i,T})'$, $\mathbf{y_{i,-1}} = (y_{i,k_i+1}, ..., y_{i,T-1})'$, $\mathbf{\Upsilon_i} = (\mathbf{i}, \mathbf{\Delta y_{i,-1}}, ..., \mathbf{\Delta y_{i,-k_i}})'$, $\mathbf{i} = (1, ..., 1)'$, $\gamma_\mathbf{i} = (\mu_i, \phi_{i,1}, ..., \phi_{i,k_i})$ and $\epsilon_\mathbf{i} = (\epsilon_{i,k_i+2}, ..., \epsilon_{i,T})$.

The tests based on the t ratios are panel extensions of the standard Augmented Dickey-Fuller test either pooling the units (equations) before computing a pooled test statistic, or averaging the individual test statistics in order to obtain a group-mean test. In the latter case we implicitly refer to the alternative hypothesis H_1^A, in the former to H_1^B.

For example, Im et al. (2003) developed a mean-group test based on (30) and assuming that $\epsilon_{i,t} \sim N(0, \sigma_i^2)$:

$$\hat{t}_i = \frac{\hat{\rho}_i}{\left[\hat{\sigma_i}^2 \left(\mathbf{y'}_{i,-1} \mathbf{M}_{\boldsymbol{\Upsilon}_\mathbf{i}} \mathbf{y}_{i,-1} \right)^{-1} \right]^{1/2}} \tag{32}$$

where $\hat{\rho}_i$ is the OLS estimator of ρ_i in (30), $\mathbf{M}_{\boldsymbol{\Upsilon}_\mathbf{i}} = \mathbf{I_T} - \boldsymbol{\Upsilon}_\mathbf{i}(\boldsymbol{\Upsilon}_\mathbf{i}'\boldsymbol{\Upsilon}_\mathbf{i})^{-1}\boldsymbol{\Upsilon}_\mathbf{i}'$ and $\hat{\sigma_i}^2 = \frac{1}{T-k_i-1}(\mathbf{M}_{\boldsymbol{\Upsilon}_\mathbf{i}}\boldsymbol{\Delta}\mathbf{y_i} - \hat{\rho}_i \mathbf{M}_{\boldsymbol{\Upsilon}_\mathbf{i}}\boldsymbol{\Delta}\mathbf{y_{i-1}})'(\mathbf{M}_{\boldsymbol{\Upsilon}_\mathbf{i}}\boldsymbol{\Delta}\mathbf{y_i} - \hat{\rho}_i \mathbf{M}_{\boldsymbol{\Upsilon}_\mathbf{i}}\boldsymbol{\Delta}\mathbf{y_{i-1}})$.

Other researchers like Levin et al. (2002), have also provided t ratios which take on a similar form.

Tests conducted on p values and p values combinations have been proposed by Maddala and Wu (1999) and Choi (2001). The tests have an alternative hypothesis of H_1^B. They are based on the idea that the p values from N independent Augmented Dickey-Fuller tests can easily be combined to obtain a test on the joint hypothesis concerning all the N units. Both papers highlight that under the null the p values p_i are independent $U_{(0,1)}$ variables so that $-2\log p_i \sim \chi^2(2)$. This leads to the facts that for fixed N, as $T \to \infty$, under the null hypothesis:

$$P = -2 \sum_{i=1}^{N} \log p_i \xrightarrow{d} \chi^2(2N) \tag{33}$$

There are other forms of both t and p type test, but they follow the same ideas and thus are not presented. Panel unit root tests are a necessary tool in order to ensure that the data used in models that required stationary meet this condition.

2.1.6 Data imputation

The data used in this thesis contains missing values thus in order not to lose information, time series imputing methods are employed.

Mainly, time series imputation methods are divided into two groups: those who deal with multivariate data, and those who deal with univariate data. The former employs inter-attribute correlations, meaning that missing values are imputed by taking into account information from other variables. Univariate time series imputation takes into account inter-time correlations, which means that each process is treated individually, and imputation is done taking into account the structure and inertia (autocorrelation) of the series. Moritz et al.

(2015) conducted an analysis which measured different imputation algorithms (multivariate and univariate). Their work showed that a univariate time series approach performed at least as good or even better then the multivariate framework.

This thesis uses univiariate time series imputation. This approach is chosen in accordance with Mortizz et al. (2015) and due to the fact that using multivariate imputation provided questionable results.

Elissavet (2017) conducted an analysis of a wide range of methods which are used in dealing with missing data in univariate time series. The analysis showed that structural models using Kalman smoothing provided the best results. Thus this thesis employs structural modeling with Kalman smoothing.

Structural time series models are based on a decomposition of the process into a number of components. The simplest model is the local level model, which has and underlying level μ_t, which evolves by:

$$\mu_{t+1} = \mu_t + \xi_t, \quad \xi_t \sim N(0, \sigma_\xi^2) \tag{34}$$

Then the process of such model is defined as:

$$x_t = \mu_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_\epsilon^2) \tag{35}$$

The local linear trend model has the same specification as before, with an added time varying slope in the dynamics of μ_t, defined by:

$$\begin{aligned} \mu_{t+1} &= \mu_t + \nu_t + \xi_t, \quad \xi_t \sim N(0, \sigma_\xi^2) \\ \nu_{t+1} &= \nu_t + \zeta_t, \quad \zeta_t \sim N(0, \sigma_\zeta^2) \end{aligned} \tag{36}$$

Lastly, the basic structural model is defined as local linear trend model with a seasonal component γ_t:

$$x_t = \mu_t + \gamma_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_\epsilon^2) \tag{37}$$

where the seasonal component has dynamics:

$$\gamma_{t+1} = -\gamma_t + ... + \gamma_{t-s+2} + \omega_t, \quad \omega_t \sim N(0, \sigma_\omega^2) \tag{38}$$

After fitting a selected model, Kalman smoothing is carried out. The objective of this is to estimate the vector a_t, given the entire sample of X_n. Here, a_t is called the state vector and is defined:

$$a_t = (\mu_t, \nu_t, \gamma_t)^T \tag{39}$$

a_t varies depending on the model at hand. The conditional distribution of $a_t|X_n$ is considered to be normal $N(\hat{a}_t, V_t)$, where $\hat{a}_t = \mathbb{E}(a_t|X_n)$ and $V_t = Var(a_t|X_n)$. $\hat{a}_t$ is called

the smoothed state, V_t is the smoothed state variance, and the process of estimating $\hat{a}_1, ..., \hat{a}_n$ is state smoothing.

We now move on to empirical research on the determinants of tax to GDP and tax effort.

2.2 Empirical findings

2.2.1 Macroeconomic factors determining tax to GDP

Gupta (2007) analyzed the determinants of tax revenue in developing countries. The author used a panel of 105 countries over the past 25 years and various panel data models in order to find the most significant variables in determining the tax to GDP ratio. By employing fixed and random effects models, Gupta found out that agricultural share in GDP has a strong negative relationship with tax to GDP. The author suggested that this is possibly due to the fact that agricultural goods are usually quite hard to tax. Log per capita GDP was significantly positive in both models which is logical, because as income increases, tax collection also tends to increase. Next, trade openness (import share of GDP) showed a mild and positive impact on revenue. This could be due to the fact that trade-related taxes are easier to impose because of higher control at posts at which the goods enter of exit the country. Foreign aid appeared to have a weak and positive effect, probably because of if foreign aid comes primarily in the form of loans then the burden of future loan repayments may induce policymakers to mobilize higher revenues. Institutional factors capturing government stability, corruption, law and order showed quite mixed effects, the most important being corruption with a negative effect. Gupta also used a dynamic panel model by including lagged tax to GDP ratio and found out that agriculture and import share and log GDP per capita proved to be significant although had smaller effects.

Castro and Ramirez (2014) also used both static and dynamic panel data models in an effort to estimate the effects of economic, structural, institutional and social factors on tax revenue. They used a panel of 34 OECD countries over a period of 11 years (2001-2011). According to the results, GDP per capita had a positive effect, trade volume (sum of imports and exports) was not significant, FDI (foreign direct investments) relative to gross fixed capital formation and agricultural share had a negative effect. Institutional variables like civil liberties and political rights had a positive effect, but only the former was significant. Child mortality rates and education were not statistically significant while life expectancy was. Castro, Ramirez (2014) explained that the insignificance of trade volume can be explained by fact that OECD economies are open and have reduced import taxes

gradually. The sign of FDI variable is explained by that the creation of government incentives to attract foreign investment reduces the potential to collect taxes. The insignificance of child mortality and education states that social factors are not robust determinants of tax revenue. Dynamic panel data models showed that tax revenue in middle income countries depend less on its lagged values than high income countries, which indicates that the roll of economic, institutional, social and structural factors are more important to determine current values of tax revenue in the middle income countries.

Dioda (2012) analyzed the determinants of tax revenue to GDP in Latin America and The Caribbean (a total of 32 countries) from 1990 to 2009. The author used a random effects panel regression and found out that GDP per capita and trade openness (sum of imports and exports) have a positive effect on tax revenue. The share of agriculture over GDP and size of shadow economy were both statistically significant but had a negative effect on tax to GDP ratio. GDP per capita growth rate and lagged fiscal deficit were almost always not significant. The level of education, female labor force and population density have a positive and significant impact on tax ratio. Such political variables as higher degree of civil liberties and more political stability are associated with higher tax to GDP ratio. The author found that the structure of taxation does not diverge significantly among the regions analyzed.

Basheer, Ahmad, Hassan (2018) examined the impact of economic and financial factors on tax revenue of Bahrain and Oman from 1990 to 2010 by using a static panel data model. The authors concluded that GDP growth, bank capital to asset ratio, risk premium on lending, foreign direct investment, net inflow and Cash surplus have the largest impact on tax revenues. Financial variables seemed to have a lower effect, than economic variables. Bank capital asset ratio and risk premium on lending had a negative, and all other mentioned variables – a positive effect on tax revenue.

Morrisey et al (2016) used variuos static panel data models and analyzed the tax revenue to GDP of 152 countries from 1980 to 2010. They showed that mineral and fuel exports have a positive effect on revenue and have a quite consistent outcome on taxes in various panel models. Manufacturing exports and agricultural share of GDP have negative signs but are not consistently significant. The contribution of imports is very small and together with agricultural share of GDP not consistent throughout. In all models, GDP per capita had a significant and positive outcome to tax to GDP. By splitting the sample into two groups - high/low income, democracy/non-democracy, resource-rich/non-resource-rich the authors concluded that mineral exports as well as imports tend to act different (have different sign) among these groups. Manufacturing exports tend to have a more negative effect in low

income countries and democracies, which the authors found quite unusual.

Andreoni (2019) examined the determinants of environmental tax revenue in 25 European union member states from 2004 to 2016. The author used index decomposition techniques. The main results show just 5 of the 25 Member States have moved toward a more sustainable system. In particular, Italy, Greece, Slovenia, Estonia and Latvia have been the only countries to increase the role of taxation rates and regulations and to reduce the relative contribution that economic factors have played in the generation of the revenue collected. For all the other Member States, economic growth and structural change effect have been the main drivers of environmental tax revenue variations.

This analysis shows that the main variables that are significant to tax revenue are GDP per capita, which acts as a revenue proxy, agricultural share to GDP and trade openness (sum of imports and exports). Authors also include some variables to represent institutional sector efficiency, like bureaucracy indexes and etc. Finally, population (workforce share in total population or its growth rate) is also included. It can also be noted, that panel data modeling is the usual tool for the analysis.

2.2.2 Implemented models for tax effort

There exists a quite a few papers, concerning the evaluation of tax effort and tax capacity. Usually, a very similar methodology is used like in the determination of factors that influence tax revenue. In essence, the same panel data modeling is used for the task of tax effort and capacity evaluation.

Minh Le et al. (2012) used a static panel data model of 110 countries in order to estimate the aforementioned criterion. OLS was used to find the model parameters, meaning that a simple, pooled regression was implemented. The authors used GDP per capita, demography (population growth rate or age dependency rate), trade openness (sum of exports and imports as a percentage of GDP), agricultural value added, governance quality (bureaucracy quality or corruption index) as the independent variables to model tax capacity. They also included regional and time dummies. The variables were chosen by reviewing literature that discusses the determinants of of tax collection. After performing some robustness checks, the authors used coefficients of the evaluated model to forecast the tax to GDP ratio and thus calculated tax effort. They then grouped countries by the results and gave insights into what actions should be taken by each group to benefit.

Eltony (2002) analyzed tax effort in a panel of 16 Arab countries in a period of 1994-

2000. The author used share of agriculture, share of mining, share of manufacturing, share of exports, share of imports, share of foreign debt (all in share of GDP) and per capital income. A static fixed panel effects model was estimated, because the author was interested in making inferences conditional on the effects that are in the sample. Eltony (2002) concluded, that the most important variables are income, share of manufacturing and share of agriculture. Also, by calculating the tax effort for each country in every year, the author noticed that some of the countries significantly increased their tax effort, while others decreased it.

Piancastelli (2001) also used a static panel of 75 countries with a yearly period from 1985 to 1995. The author employed GDP per capita and the shares of trade, agricultural, industrial and services sectors in GDP as variables for the panel data model. The static panel model was estimated with fixed, random and no effects. The results showed that including random or fixed effects significantly increases the quality fo the model. Tax effort was calculated by taking the coefficients and predicting the tax to GDP ratio. By grouping countries into 3 groups in terms of income, Piancastelli showed that as a country's income level increases, its tax effort also tends to be higher.

The approach by estimating random and fixed effects static panel data regression has also been used by Davoodi and Grigorian (2007). The authors used a panel of 141 countries through a period of 1990-2004 in an attempt to estimate Armenia's tax effort. Amongst the more exotic variables, used in the regression, were a dummy variable for fuel exporters and the share of urban population in a country's total population. The oil variable was included because countries that export oil are more likely to generate higher tax revenues. Urban variable represents the demand for public services, as the authors postulated that residents, living in the city, create a need for public services. The results showed that random effects regression was a more solid choice, because it explained more of the tax-to-GDP ratio variance. The authors concluded, that the country's tax effort falls short of its potential by about 6.5 percent of GDP and that the improvements in institutions as well as policy measures designed to reduce the size of the shadow economy are prominent factors in increasing tax performance.

Pessino and Fenochietto (2010) expanded the panel data approach by mixing it with a stochastic frontier framework. The data was composed of 96 countries and a yearly period from 1991 to 2006. The variables used were: GDP per capital, sum of imports and exports as percent of GDP, consumer price index, public expenditure on education, agricultural value added, GINI and corruption indexes. The authors concluded, that most European countries with high GDP per capital and education, open economies and low inflation are near their

tax capacity.

Stochastic frontier analysis integrated with panel data was also used by Garg, Goayl, Pal (2017) and Valles-Gimenez, Zarate-Marco (2017) in order to estimate tax effort in Indian states and Spanish municipalities respectively. The former found out that higher income is correlated with larger tax capacity. They also noticed that the disparity in tax effort between states is widening in time. The latter was able to conclude that Spanish municipalities are nearing their full tax potential and maybe some intervention is needed.

In a quite new approach, Dalamagas et al. (2019) used the assumptions of the Arrow-Debreu to analytically find tax capacity. By using a Lagrangian framework, the authors concluded that it is equal to GDP minus consumption. The tax effort is then the ration between actual tax revenue and the tax capacity. After calculating the tax effort, the authors then performed a correlation analysis between approaches in panel data models and showed that their results highly correlated with those obtained from econometric modeling.

A natural question of whether there exists an optimal value for tax effort arises, which would give further insight into a countries tax system. Mahdavi and Westerlund (2018) analyzed 48 states of the US over a period of 1981-2013 in order to find convergence in tax capacity or tax effort. The authors used a Bootstrap sequential quantile test (BSQT) of unit root. They found that neither tax capacity nor tax effort of the state-local (and state) government units showed any sign of convergence in terms of narrowing of the gaps. However, there was evidence of partial effort convergence only when own-source revenue was more broadly defined to include non-tax revenue items beyond charges and fees.

Finally, one can understand, that the estimation of a countries tax effort mainly falls on two keystones: variable selection and model structure. First, it is obvious that including or not a single variable can lead to severe changes in the models' coefficients. Although, as was seen before, there exists some literature on the fundamental factors that determine tax revenue it is not clearly determined, which variables should one choose. Second, primarily static panel models are used. This asks for the implementation of dynamic panel models for tax effort evaluation as it is possible that these models would give more information to the model. It should be noted that these models do not aim to best fit the model to the data, but to capture the general tendencies of tax revenue by using macroeconomic and institutional determinants. Lastly, it should be noted, that tax capacity is calculated by taking the fitted values of the model for tax revenue.

A table is provided in appendix E which is taken from Atsan (2017) which summarizes a large number of research done in the analysis of tax effort. The table shows the dependent

(second row) and independent variables used in panel models. It can be seen, that the most frequent are trade openness, agriculture, GDP per capita and population growth.

Next, time series clustering is introduced as this approach can shed more light about a country's tax dynamic.

2.3 Time series clustering

In the case where there exists multiple time series objects, one can implement time series clustering to have a better understanding about the analyzing systems in hand. As the size of data generating processes increase, this method becomes more involved in data analysis techniques. The area of time series clustering is quite broad: bio medicine, computational biology, electronic manufacturing, physics, seismology and even speech recognition. Also, the field of econometrics has also benefited from this method. For example, according to Focardi and Fabozzi (2004), clustering of economic and financial time series includes the following areas of application:

- identifying ares of sectors for policy-making purposes;
- identifying structural similarities in economic processes for economic forecasting;
- identifying stable dependencies for risk and investment management.

Augustynski and Laskos-Grabowski (2018) suggest that one of the most valuable advantages of time series clustering is the identification of structural similarities at different points in time and space.

Since this thesis deals with a large number of countries over a period of time, time series analysis can be implemented. This is done with an incentive to have a better understanding on how tax to GDP can be grouped and what characteristics these groups have. Two approaches are implemented for time series clustering: by working with discrete values and functions.

It should be noted, that time series clustering acts as a descriptive statistics approach to the data. Its main goal is to see where Lithuania stands among other countries.

2.3.1 Discrete time series clustering

One of the main parameters in discrete time series clustering is the dissimilarity measure. Various analysis software programs offer a broad range of distance functions which may often lead to different results.

According to Lin and Li (2009) and Corduas (2010) all dissimilarity measures can be

grouped to two categories: shape and structure based. The former is aimed for the comparison the geometric profiles of the series, or alternatively, representations of them designed to reduce the dimension of the problem. This leads to fact that that shape-based dissimilarities are mainly dominated by local comparisons. Structure-based dissimilarity is focused at comparing the underlying dependence structures of a process. Higher level dynamic structures describing the global performance of the series must be captured and compared in this case.

As the interest leans towards shape-based dissimilarity, standard distances (for example L_p type) or complexity-based measures (for example CID dissimilarity) can give acceptable outcomes, although sometimes measures invariant to specific distortions of the data could be required. For example, time series that have different scales will require previous normalization to cancel differences in amplitude and then match well similar shapes. In fact, conventional metrics like Minkowski, DTW or Frechet distances can lead to common misunderstandings unless the processes at hand are recorded in the same units (Rakthanmanon et al. 2012).

In this thesis shape-based distances, provided below in table 1, will be used because the data in hand is quite short in length and without any anomalous values (outliers).

Distance	Formula	Summary
Euclidean	$d(X_T, Y_T) = \left(\sum_{t=1}^{T} (X_t - Y_t)^2 \right)^{1/2}$	"Ordinary" (i.e.straight-line) distance between two points
Frechet	$d(X_T, Y_T) = \min_{r \in M} \left(\max_{i=1,,,m} \lvert X_{ai} - Y_{bi} \rvert \right)$	Takes into account the location and ordering of the points
DTW	$d(X_T, Y_T) = \min_{r \in M} \left(\sum_{i=1,...,m} \lvert X_{ai} - Y_{bi} \rvert \right)$	Similar to Frechet's distance
CID	$d_{CID}(X_T, Y_T) = CF(X_T, Y_T) \cdot d(X_T, Y_T)$	Calculates a correction of the Euclidean distance based on the complexity estimation of the series
COR	$d_{COR,1}(X_T, Y_T) = \sqrt{2(1 - COR(X_T, Y_T))}$	Uses Pearson's correlation coefficient as a distance between two points
CORT	$d_{CORT}(X_T, Y_T) = \phi[CORT(X_T, Y_T)] \cdot d(X_T, Y_T)$	Uses temporal correlation coefficient as a distance between two points

Table 1: Metrics used in discrete time series clustering

Summing up Montero and Vilar (2014) conclude that shaped-based distances work well with short time series but they can fail by working with long sequences. This can be especially probable when a high amount of noise or anomalous records are present. In these situations,

a structure-based dissimilarity aimed to compare global underlying structures can be more appropriate.

A more thorough inquiry into each distance is provided in appendix A. Next, we introduce the concept of functional data.

2.3.2 Functional data and data smoothing

Usually data analysis is performed on data which contains a set of observations, meaning that the information is discrete. If functional data analysis is carried out, the observed data are depicted as a function, which means that the data is continuous. If, for example, in a discrete framework one deals with time series, then in a functional data framework, this time series becomes a function, or curve, which is then treated as a single functional entity. The continuum of a function is often time.

By "transforming" the data from discrete observations to a function, one tries to capture the general tendency of the data, rather than fitting a curve that completely fits the process at hand. For example, the tax to GDP data for each country can and possibly is with some level of measurement error. While countries try to be precise in gathering their tax data, it is quite common to make mistakes, miscalculations ir the process of information aggregation, approximation and etc. To account for this measurement error, one can try to capture the general tendency of the data, which can be done by functional data. The notion of continuum in tax data also has a meaning, as taxes are payed every day or even hour, they are just aggregated yearly, which also holds true for GDP.

The first step to perform functional data analysis is to smooth discrete data points. Let t be a one-dimensional argument which can be referred as time. Functions of t are observed over a discrete grid $t_1, .., t_j$ at sampling values t_j, which may or may not be equally spaced. For a functional datum to be created, a basis needs to be specified. This basis is a linear combination of functions which defines the functional object. A functional observation X_i is defined by:

$$X_I(t) \approx \sum_{k=1}^{K} c_{ik}\phi_k(t), \ \forall t \in \mathcal{T} \tag{40}$$

where $\phi_k(t)$ is the k^{th} basis function of the expansion and c_{ik} is the corresponding coefficient. As mentioned before, the data usually contains observational errors that are superimposed on the underlying process. In reality, one often meets a scenario which involves N processes being observed at the same time. Let $\mathbf{y}$ be a vector of N functional $\mathbf{y} = [\mathbf{y_1}, ..., \mathbf{y_N}]^T$,

where each functional data are written as:

$$Y_{ij} = X_i(t_j) + \epsilon_{ij} \tag{41}$$

where $1 \leq j \leq J,\ 1 \leq i \leq N$ and ϵ_{ij} is the error term with 0 mean and variance σ_i^2.

Although there exits several bases for functional data, this thesis employs the Fourier basis, which takes on the following form:

$$\phi_0(t) = \frac{1}{\sqrt{|\mathcal{T}|}},\ \phi_{2r-1}(t) = \frac{sin(r\omega t)}{\sqrt{|\mathcal{T}|/2}},\ \phi_{2r}(t) = \frac{cos(r\omega t)}{\sqrt{|\mathcal{T}|/2}} \tag{42}$$

where $r = 1, ..., \frac{K-1}{2}$ with K being the number of basis functions. Note that this number must be odd. The frequency ω determines the period and the length of the interval $|\mathcal{T}| = 2\pi/\omega$.

(42) can also have a penalty parameter λ which acts as a smoothing instrument, meaning that as λ increases, the smoothness of the functional datum also increases.

The Fourier basis was chosen because during practical applications in the thesis, other bases gave unusable results. This can be due to the fact that the data used in this thesis is quite short, as each series has only 16 observations.

2.3.3 Functional principal components and score clustering

Functional principal components (FPCA) is a useful tool for functional data analysis. In essence, FPCA does not deviate much from regular principal component anaysis. Ramsay and Silverman (2002) state that in functional context, every principal component is specified by a principal component weight function $\xi(t)$. These functions are defined over the same range of t as the functional object. The principal component scores of an individual process are then defined:

$$z_i = \int \xi(t) X_i(t) dt \tag{43}$$

Next, the first principal component $\xi_1(t)$ is found by maximizing the variance of the first principal component scores $\sum_{i=1}^{N} z_{i1}^2$ subject to the constraint:

$$\int \xi(t)^2 dt = 1 \tag{44}$$

The second, third and higher order principal components are defined in the same approach as the first component, but with additional constraints. The second principal function $\xi_2(t)$ are subject to the constraint (44) and:

$$\int \xi_2(t)\xi_1(t) dt = 0 \tag{45}$$

In general, for the j^{th} component, additional constraints take the following form:

$$\int \xi_j(t)\xi_1(t)dt = \int \xi_j(t)\xi_2(t)dt = ... \int \xi_j(t)\xi_{j-1}(t)dt = 0 \tag{46}$$

Note, that (46) ensures orthogonality between the components. In tax time series context, the first principal component, which explains the largest amount of variance, can capture the underlying trend of all the analyzed countries. This gives insight into the dynamics of global taxation variation.

The scores calculated in FPCA can also be used for clustering. Since every process or individual has a score, they (the individuals) can be represented by the score of each component. If the first 2 or 3 components explain a sufficient amount of variation, these scores can be plotted on a 2D or 3D domain in order detect outliers or to have a better understanding of the data at hand. Naturally, these scores can be clustered with classical clustering algorithms like hierarchical or k-means. This practice has been used by Shang (2014), Illian et al. (2009), Suyundykov et al. (2010) and others.

It is noteworthy, that there exist other functional data object clustering methods. Usually, these approaches are quite computationally expensive. Jacques et. al (2014) conducted a survey of clustering methods for functional data and showed that working with FPCA scores provided at least as good or better results than other frameworks.

3 Application results

3.1 Used data

The data used in this thesis is gathered from the World Bank database from years 2002 to 2017. The taxation variable is represented by the tax revenue to GDP ratio. Tax revenue refers to compulsory transfers to the central government for public purposes. Certain compulsory transfers such as fines, penalties, and most social security contributions are excluded. Refunds and corrections of erroneously collected tax revenue are treated as negative revenue.

The independent variables are trade openness (sum of imports and exports ratio with GDP), agriculture (agriculture, forestry, and fishing, value added to GDP), final consumption expenditure (private and general government consumption ratio to GDP), GDP per capita, age dependency ratio (people younger than 15 or older than 64 ratio to people aged between 15-64), corruption index (calculated by the World Bank, spans from -2.5 and 2.5, higher value indicates less corruption). The corruption index is "lifted" by 2.5 (add 2.5 to every observation) in order to avoid negative values.

These variables will be referenced as TAX, TRADE, AGR, CONS, GDPPC, POP and CORRU henceforth. The independent variables were chosen in accordance with empirical findings in subsection 2.2 and appendix E. Among the most popular independent variables (TRADE, AGR, GDPPC and POP) we add CORRU as a institutional proxy and CONS because it is known that a lot of tax income comes from indirect taxes (for example the value added tax).

As mentioned before, almost every variable contains missing values:

Percent of missing values						
TAX	AGR	TRADE	GDPPC	CONS	POP	CORR
5%	1.51%	0.96%	0.41%	1.47%	0%	0.41%

Table 2: Percentage of missing values

As can be seen, the taxation variable has the largest percent of missing values. A structural model is fitted for every variable with Kalman smoothing. Since the data is yearly and quite short, the issue of seasonality becomes irrelevant, thus a local linear trend model is chosen. Only time series with no more than 5 missing values (per country) are imputed. This is done so that the series as a whole would have around 30% of its values missing and not more, so as not to include a large amount of error. After the imputation, the data contains 99 countries, which makes the total data set 1584 observations long.

Since the data covers the period of the 2008-2009 financial crisis, in further modeling this information must be accounted for. The following table shows how much negative values (from logged differences) appear in each year in the TAX variable:

Year	Count of negative values	Year	Count of negative values	Year	Count of negative values
2003	39	2008	50	2013	36
2004	27	2009	68	2014	40
2005	22	2010	44	2015	35
2006	26	2011	32	2016	41
2007	36	2012	36	2017	31

Table 3: Negative values in TAX by year

It can be seen, that the year 2009 contains the largest amount of negatives values. Therefore the inclusion of a time dummy variable in future modeling will be necessary.

Next, the results from discrete and functional clustering are provided. It should also be noted, that the results in every section emphasize not only Lithuania, but its neighbors

(Latvia, Estonia and Poland) also. This is done in order to have a reference point.

All calculations further in this thesis are done with the statistical software program R and can be provided upon request.

3.2 Clustering

In order to have a better understanding about Lithuania's position among other countries, we perform time series clustering. This information is useful in order to have a "bigger picture" perspective about the data at hand. Since we are dealing with a large number of variables and countries, visualization becomes quite hard and thus clustering is a more useful tool.

The clustering for discrete time series is done in the following principle:

1. Use the variable TAX and calculate the distance matrices with each of the 6 shape-based dissimilarity measures;
2. Use hierarchical clustering (4 clusters) on each metric with Ward's linkage;
3. Calculate the mean values of all 7 variables for each cluster. We can think of the results as cluster centers;
4. Divide Lithuania's, Latvia's, Estonia's and Poland's mean value of the variables (from the vector containing years 2002-2017) from the cluster mean value to which each country belongs. The results give following information: how much each country's mean constitutes of its cluster's mean (for every variable). Lithuania, Latvia, Estonia and Poland are chosen because of their economic similarities;
5. Repeat step 4 for each metric and then calculate the mean for each country from all the metrics.

Hierarchical clustering was chosen because it provides reproducible results and requires fewer assumptions about the data. The main idea behind these calculations is to see, how much each country differs (if it does) from its cluster. If the value in step 5 is close to 1 for a particular variable, then this means that a country is in the center of its cluster in that particular variable. With this information, one can have an understanding, how similar countries (in the context of tax revenue) look with respect the macroeconomic forces that govern them.

The dendogram from discrete time series clustering can be shown with the Euclidean distance metric in order to see the number of clusters possible (figure 1). It should be noted, that for both FPCA and discrete time series approach these dendograms looked very similar,

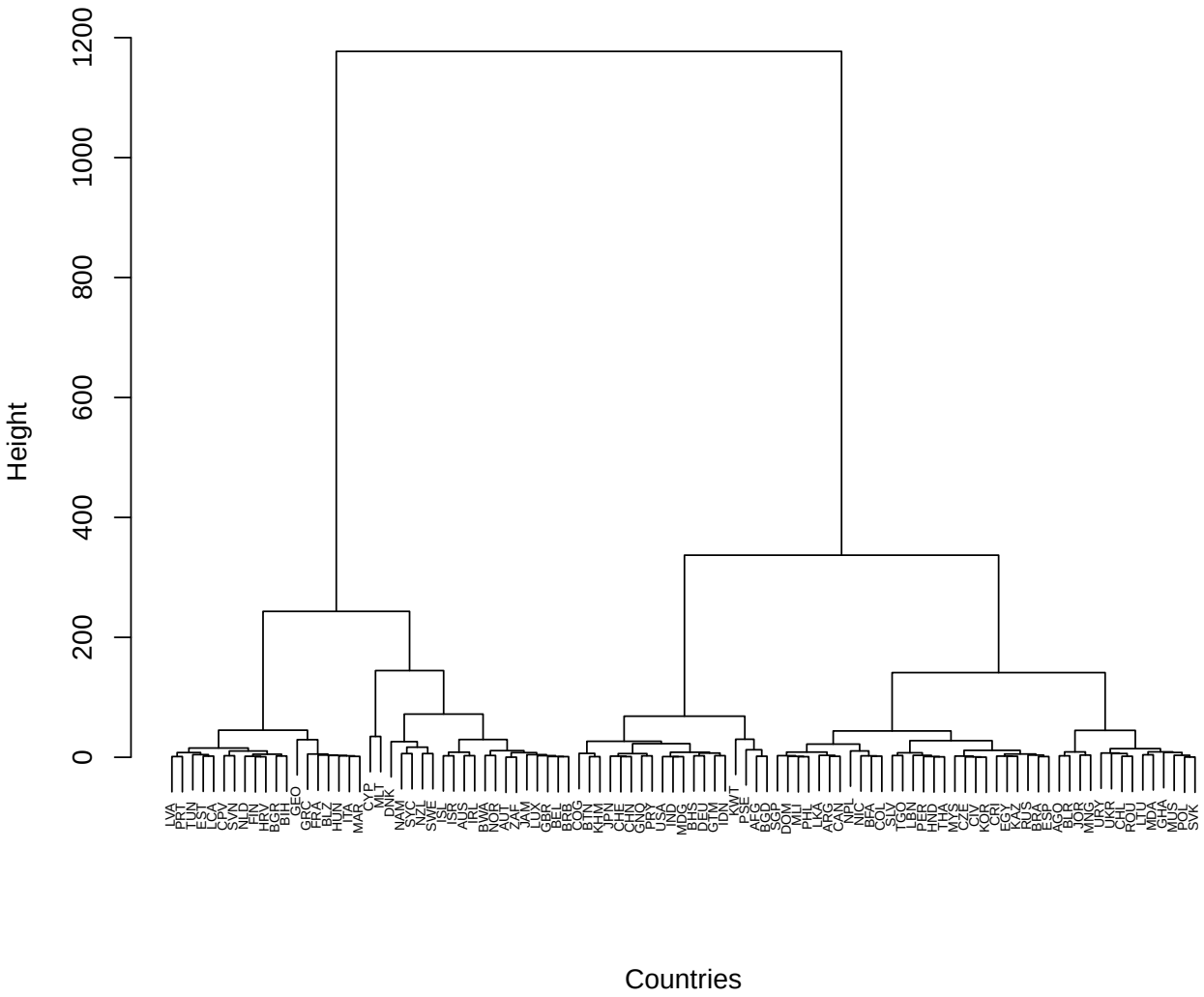

Figure 1: Dendogram from discrete time series clustering with Euclidean metric

thus they are not shown in this thesis.

As seen from figure 1, cutting the tree at the height of about 170, provides 4 clusters. 4 groups of countries seems logical as usually various development indexes (OECD or World Bank) divide countries into 3 or 4 groups. Choosing less than that would defeat the purpose of clustering. This is why 4 clusters were chosen.

Functional data clustering is also performed. As mentioned before, smoothing the data can help avoid measurement errors which in turn can lead to better quality results.

The algorithm is quite similar for functional data but not the same:

1. Smooth the data using the Fourier basis;
2. Perform FPCA on TAX for all of the 99 countries;
3. Cluster (hierarchical, Ward's method, 4 clusters) the scores from a number of principal

components, that explain at least 90% of variation. Here, the standard distances are chosen (Euclidean, Minkowski, Maximum, Canberra, Manhattan);

4. Repeat steps 3, 4, 5 from the algorithm for discrete data.

Before providing the results, the "side" results from functional smoothing are shown as they give quite interesting observations. Each TAX time series is smoothed with the Fourier basis using 11 basis functions and imposing a penalty of $\lambda = 0.001$. The picture below shows Lithuania's TAX fact, it's smoothed version and the first component of all 99 countries.

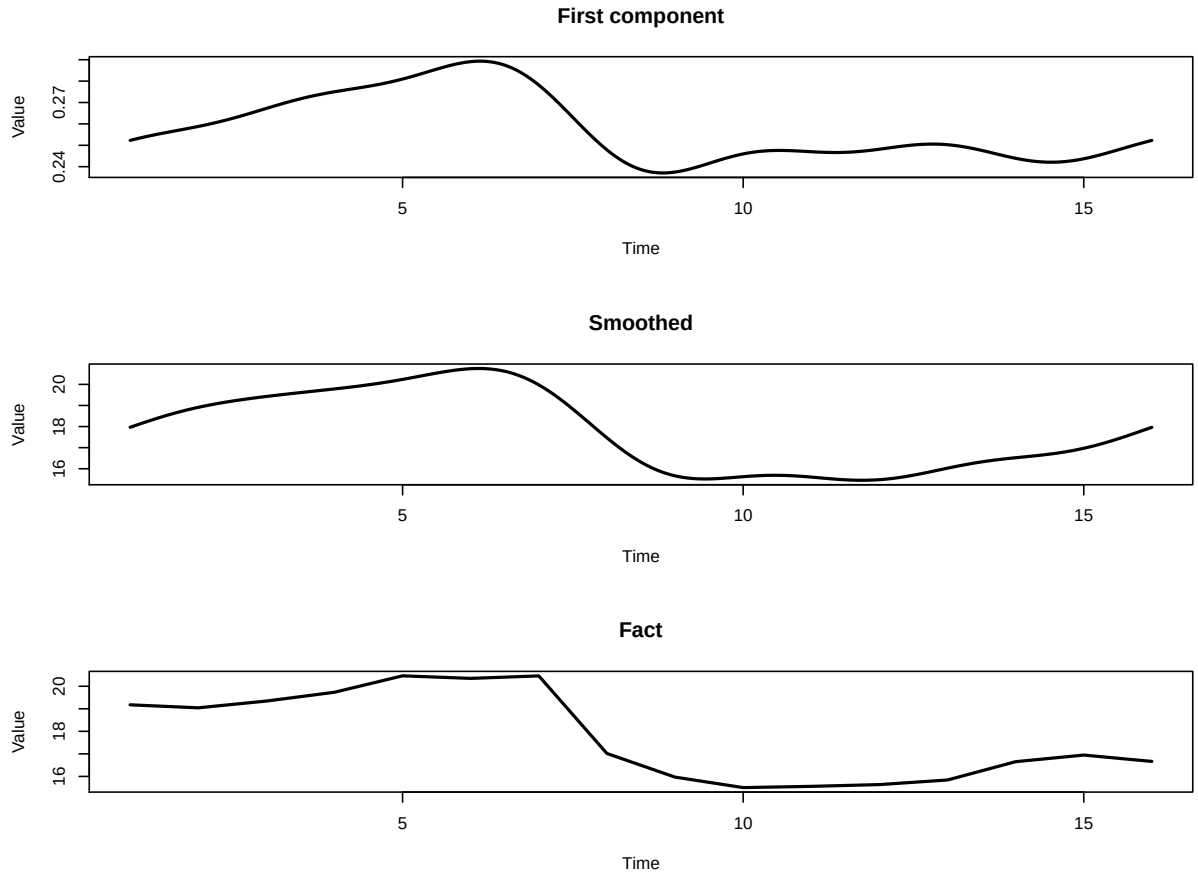

Figure 2: Results from smoothing

As can be seen, from figure 2, the smoothed version of Lithuania's TAX captures the general tendency of it's fact quite good. In the end one can observe that the smoothed version has a little spike. These "curls" or "turns" are usually seen in data, that are smoothed with the Fourier basis, as this basis uses sines and cosines. What is more interesting, the first component, which accounts for 89% of total variation of all countries, is quite similar to Lithuania's smoothed TAX curve. This means that the general tendency of TAX in all 99 countries exhibits similar movement as Lithuania.

By plotting the scores of the first two components (which account for 96.6 % of all variation) on a 2 dimensional surface, one can see how all the countries are scattered in the

context of TAX.

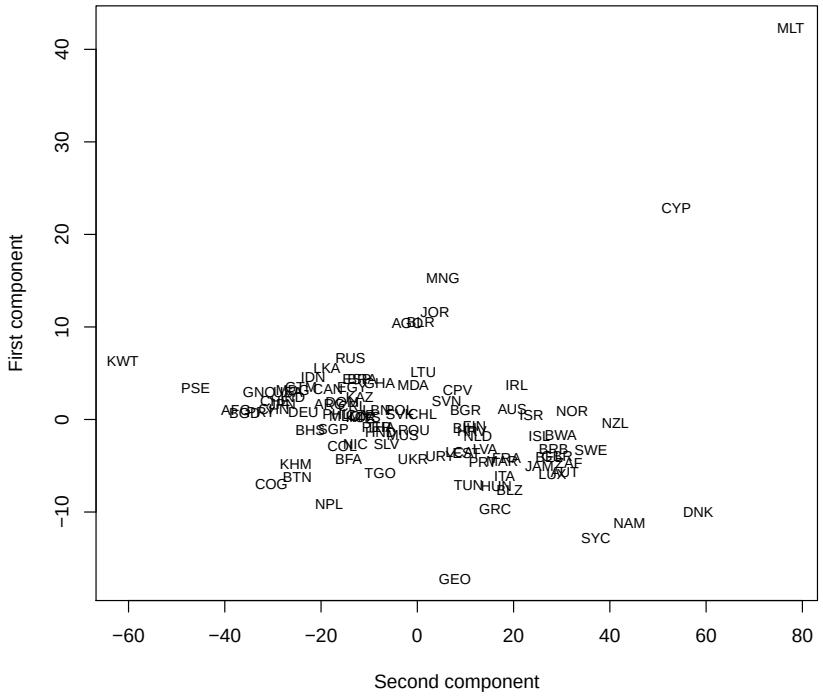

Figure 3: First two component plot

Several things can be observed from this figure 3. First,there are visible clusters, which gives further incentive to perform cluster analysis. Second, aberrations can be identified, as Cyprus, Malta, Kuwait or even Denmark appear to be further (some especially further) from the rest of the countries. It should be noted, that these cases are outliers in the context of TAX tendency, meaning that the dynamics of the variable of these countries differs from others. It is well known, that for example, Malta or Cyprus has been a tax haven for investors and businesses for some time, which might explain their departure from the rest of the world.

Next, the information to which cluster each country was assigned is provided below:

	LTU	LVA	EST	POL
EUCL	2	3	3	2
CORT	2	3	3	2
FRECHET	3	3	3	2
DTWARP	2	4	4	2
COR	2	1	1	4
CID	3	2	2	3

Table 4: Results from discrete time series clustering

As seen in table 4, in most of the distances, Lithuania and Poland are separated from Latvia and Estonia. Only the Frechet and Correlation distances provide different results.

Next, the same results for FPCA scores is shown. The results in table 5 show, that every metric separates Lithuania and Poland from Estonia and Latvia.

	LTU	LVA	EST	POL
EUCL	2	4	4	2
MAX	2	4	4	2
MANH	2	4	4	2
CANB	2	3	3	2
MINK	2	4	4	2

Table 5: Results from FPCA score clustering

In order to see, if these results are good or bad, TAX is shown for each of the 4 countries:

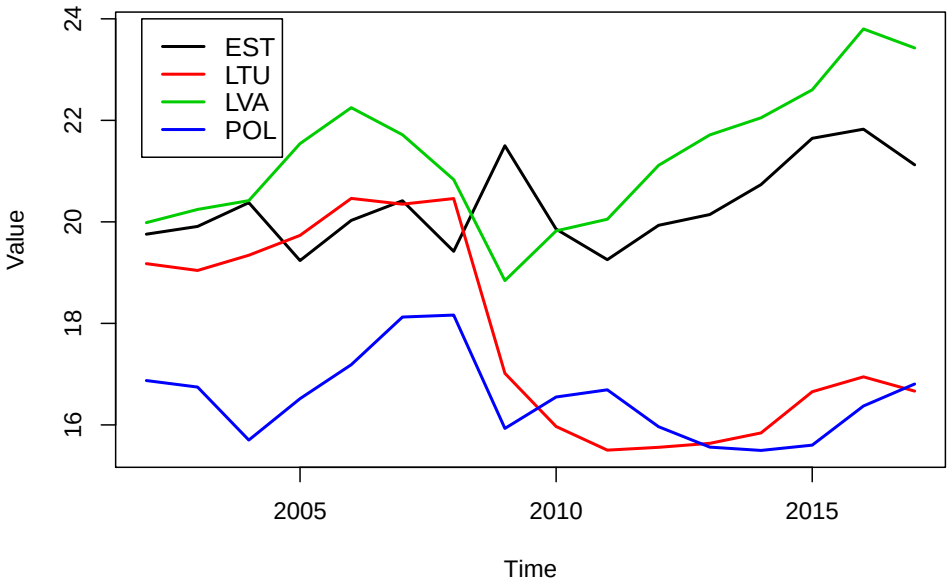

Figure 4: TAX dynamics of Lithuania, Estonia, Latvia and Poland

It is quite clear from figure 4 that indeed Poland and Lithuania have similar dynamics. After the crisis, a divergence can be seen with Estonia and Latvia rising greatly upwards. This means that cases, where these countries are separated give plausible results. Tables 4 and 5 show that both methods perform good, but the FPCA score clustering gives better results.

The results of the clustering can be visualized in figure 5. Here the clustered FPCA scores (with the Euclidean metric) are presented. Each color represents one of the 4 clusters. It

appears that the green cluster is somewhat further from the blue one, and both black and red clusters have thick cluster centers.

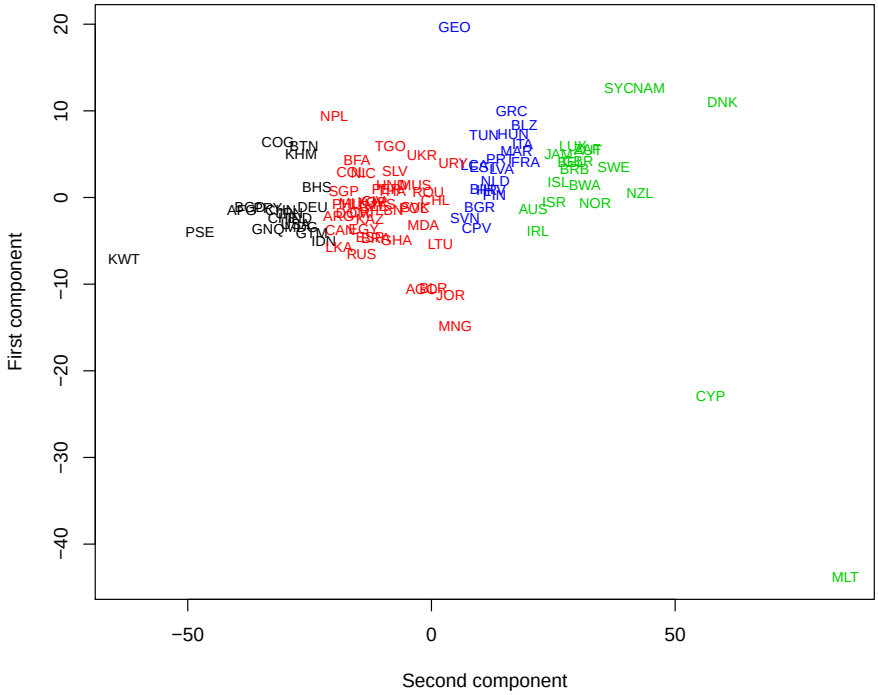

Figure 5: FPCA clustering results with Euclidean metric

Lastly, the results from both discrete and FPCA clustering are shown in table 6.

Discrete time series							
	TAX	AGR	GDPPC	TRADE	CORRU	POP	CONS
LTU	1.03	0.53	1.10	1.36	1.12	0.90	1.04
LVA	1.02	0.77	0.56	1.10	0.86	0.92	1.00
EST	0.97	0.63	0.72	1.48	1.14	0.94	0.90
POL	1.04	0.31	1.23	0.86	1.25	0.78	1.00

Table 6: Results from discrete time series

Discrete time series clustering shows that AGR and TRADE variable means are not in the centers of their clusters. This means that countries, that achieve a similar dynamic and level of taxation, usually have a higher agricultural sector. The inverse situation is seen for trading, as Lithuania is around 36% higher than other countries with similar level of TAX. What is more interesting, that the TAX value stands at 1.03 which would indicate the Lithuania is in the middle of the cluster across all metrics. CORRU in Lithuania, Poland and

Estonia are larger than 1, which means that countries in each country cluster have higher corruption.

Since it is known from empirical research that agriculture has a negative and trade openness - a positive effect on tax, it would seem that by being the lower position (AGR) and higher (TRADE) should boost Lithuania's position above 1 in the TAX variable. A possible explanation for this is that these sectors are not efficiently used taxwise. Also it is possible that after the crisis, the TAX drop is not natural in the sense of macroeconomics; perhaps Lithuania's fiscal policy agents tried to cope with the crisis and imposed new regulations and etc., which had this effect. Appendices H and I provide cluster centers for discrete and FCPA score clustering. It can be seen that clusters, who have a higher TAX tend to have a smaller AGR and higher TRADE.

Other variables of Lithuania seem (POP, GDPPC and CONS) seem to be near the center of their clusters. The large difference between Latvia's and Lithuania's GDPPC value means that Latvia reached its TAX level with a substantially lower GDPPC, which also reinforces the theory that perhaps after the crisis, TAX acts not entirely governed by macroeconomic factors.

Looking at table 7 it is clear that FPCA approach gives very similar results. Of course, to some extent the values differ, but the main observation remains essentially the same - both AGR and TRADE remain lower and higher respectively. Of course, lower AGR values are also visible for other neighbors.

Functional time series							
	TAX	AGR	GDPPC	TRADE	CORRU	POP	CONS
LTU	1.11	0.45	1.22	1.39	1.19	0.89	1.04
LVA	1.00	0.71	0.65	1.10	0.88	0.95	0.99
EST	0.95	0.58	0.84	1.48	1.17	0.96	0.88
POL	1.15	0.27	1.26	0.95	1.34	0.75	1.00

Table 7: Results from FPCA scores

The results can also be plotted in a radar chart for visual aid. The results from table 7 are used to form a radar plot which is provided below. The values are given ir percents, meaning that 100 is equal to 1 from table 7.

Looking at the figure 6, it is clear that no country reaches its center in AGR. The countries also are quite widely distributed in GDPPC and TRADE and close to one another in POP, CONS and TAX. The overall shapes of each neighbor are different indicating that between

each country, they are somewhat unique.

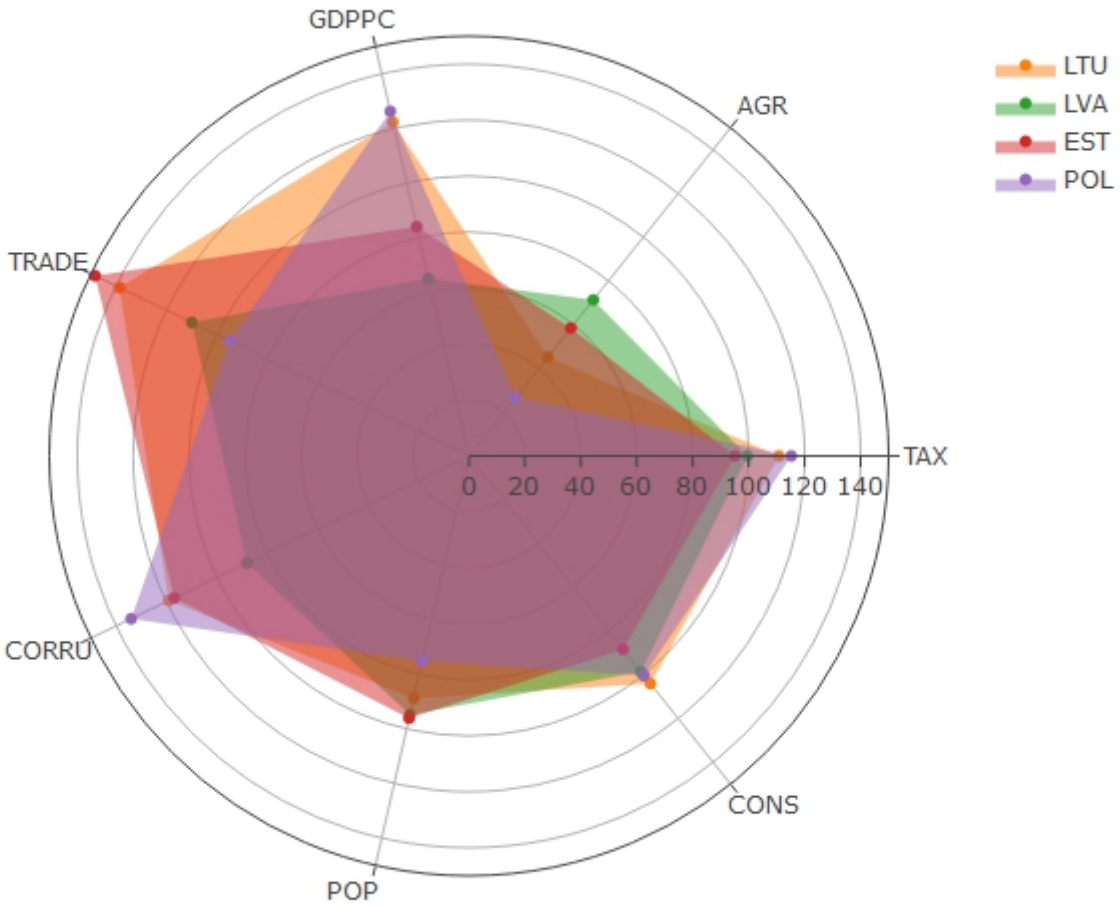

Figure 6: TAX dynamics of Lithuania, Estonia, Latvia and Poland

Next the results from a PVAR model and IRF are presented.

3.3 Impulse response functions

As mentioned in the literature overview, authors usually are unified when determining the effect of independent variables to TAX. Generally, GDPPC, TRADE, CONS have a positive effect and POP, AGR have negative effects. Increasing institutional quality has also a positive outcome.

It is interesting to see, how these effects vary in countries with different income levels. Since in general higher income countries are associated with more developed economies, tax systems and etc., perhaps the effects of macroeconomic factors act different. As seen in the literature overview, usually researchers evaluate models for all countries and then analyze their model coefficients. Papers on different country groups are difficult to find, so the question remains with large amount of potential research. Since the software for a PVAR

model has just recently been developed, implementing the model on the data at hand and giving further insights into the subject becomes a possibility.

Before implementing the IRF analysis, panel unit root and panel Granger causality tests are performed. Below, various unit root tests and their p values are shown. The tests are performed on logged data.

Undifferenced					
	Levinlin	Madwu	PM	Logit	Invnormal
TAX	1	1	1	1	1
CONS	0.09	0.98	098	1	1
AGR	0	0	0	0	0
TRADE	0.99	1	1	1	1
GDPPC	1	1	1	1	1
CORRU	0	0.07	0.07	0.08	0.09
POP	1	0	0	1	1

Table 8: P values of various panel unit root test results

Various test show that most of the variables have a unit root. It would also seem that AGR and CORRU do not have a unit root. Also, two tests show that POP is $I(0)$. Next, we difference each variable and perform the tests again:

Differenced					
	Levinlin	Madwu	PM	Logit	Invnormal
TAX	0	0	0	0	0
CONS	0	0	0	0	0
AGR	0	0	0	0	0
TRADE	0	0	0	0	0
GDPPC	0	0	0	0	0
CORRU	0	0	0	0	0
POP	0	0	0	0	0

Table 9: P values of various panel unit root test results

Since all the p values are 0, the null is rejected in the favor of the alternative, meaning that all variables are $I(0)$ after the differentiation. It is noteworthy that if a model would require variables to be stationary, then AGR, CORRU and maybe POP would not require differentiating, but in order to work with the same units of measure (logged differences can be thought as growth rates) performing this transformation would not be inexpedient.

Next, Granger causality for each variable with TAX is conducted in order to see if each variable Granger causes TAX. The test is performed by including lags until the test shows that a variable causes TAX. The data is in logged differences because the procedure requires the data to be stationary. The results are provided below:

Variable	GDPPC	AGR	TRADE	CONS	POP	CORRU
First significant lag	3	1	1	3	1	1

Table 10: First significant lag in Granger causality test

All variables except GDPPC and CONS Granger cause in the first lag. This does not necessarily means that when constructing a PVAR model, these number of lags will be used; the test merely shows that the macroeconomic and institutional factors have and effect on TAX and that PVAR modeling is viable, which follows next.

All 99 countries are divided into 3 groups. This is done by calculating each country's mean GDPPC for 2002 - 2017. Then, the vector of 99 means is cut in 3 parts by the 33 and 66 percentiles (Lithuania belongs to the second group). Then, for each of the 3 groups, a PVAR model was fitted. Since the inclusion of more than 3 variables in the model gave unstable results, a pairwise analysis was done, meaning that for each group a total of 6 models were estimated: a model with TAX and AGR, with TAX and TRADE and so on.

Because the information criterion (provided in Appendix B) showed lowest values for a PVAR(1) model, only one lag was included for all models. Also, system GMM estimation provided usually unstable models, so GMM was used with forward orthogonal deviation transformation. Lags (2,3,4,5) of the dependent variable are used as instruments for estimation. Up to 5 lags are used because including more caused errors/warnings in the program. A dummy exogenous variable was also used for the crisis in the year 2009. All variables are used in logs. The eigenvalues of each model (for stability diagnostics) are provided in appendix J. Below an example PVAR(1) model for AGR is provided:

$$\begin{aligned} \Delta^* ln(TAX_{i,t}) &= \alpha_1 \Delta^* ln(TAX_{i,t-1}) + \beta_1 \Delta^* ln(AGR_{i,t-1}) + \gamma_1 CRISIS + \epsilon_{i,t} \\ \Delta^* ln(AGR_{i,t}) &= \alpha_2 \Delta^* ln(AGR_{i,t-1}) + \beta_2 \Delta^* ln(TAX_{i,t-1}) + \gamma_2 CRISIS + \epsilon_{i,t} \end{aligned} \tag{47}$$

Next, the GIRF results with 95 % confidence interval from bootstrap 100 runs are shown for TRADE. The dynamics of the shock is shown though a total of 8 years, meaning that if a positive shock in TRADE (or any other variable) happens at $t-1$, these plots show how TAX responds for the next 8 years from $t-1$.

As can be seen from figure 7, for the lowest income groups, the response from a shock in

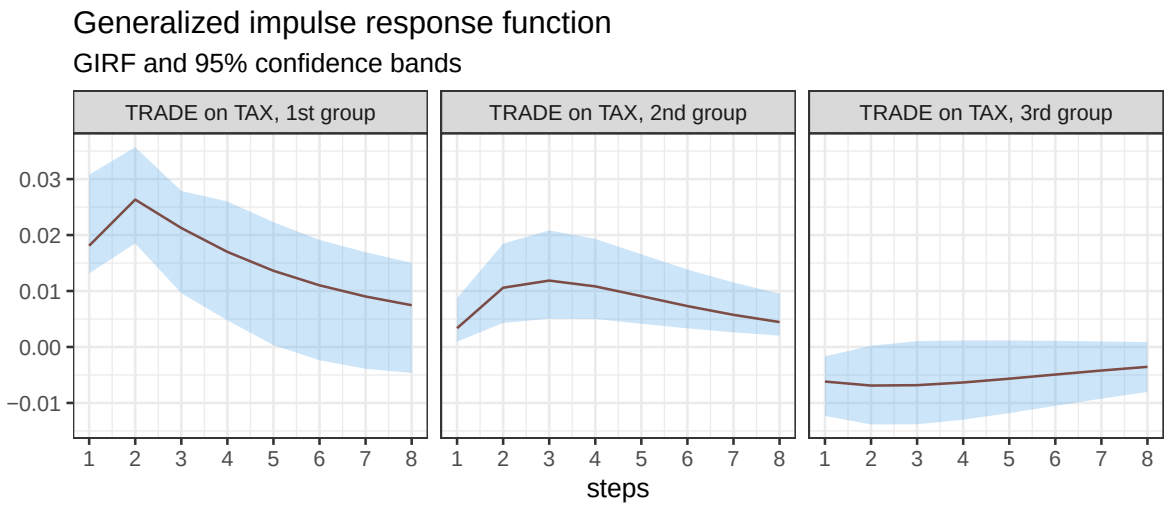

Figure 7: Response of TAX from a shock in TRADE

TRADE has a generally positive effect on TAX. Since the confidence interval bands do not cover up 0 until the 8 lag, these shocks are significant. Since forward orthogonal deviations transformation was used, the quantitative evaluation of these shocks becomes somewhat clouded. Also, it is known, that some software packages measure shocks in standard deviations of the errors. Despite this, it is clear, that as the income level of countries increases, the effect becomes insignificant, as the confidence intervals in the third group cover up 0.

Since it is known that higher developed countries become more intertwined economically (for example more trading between one another as in the EU), the decrease in TRADE as income rises has a logical interpretation. Countries that trade with each other intensively usually lower export or import tariffs for the good to be cheaper thus boosting demand. This in turn lowers tax income which explains the 0 effect of a TRADE shock in the third group.

Next, the AGR shocks are shown in figure 8:

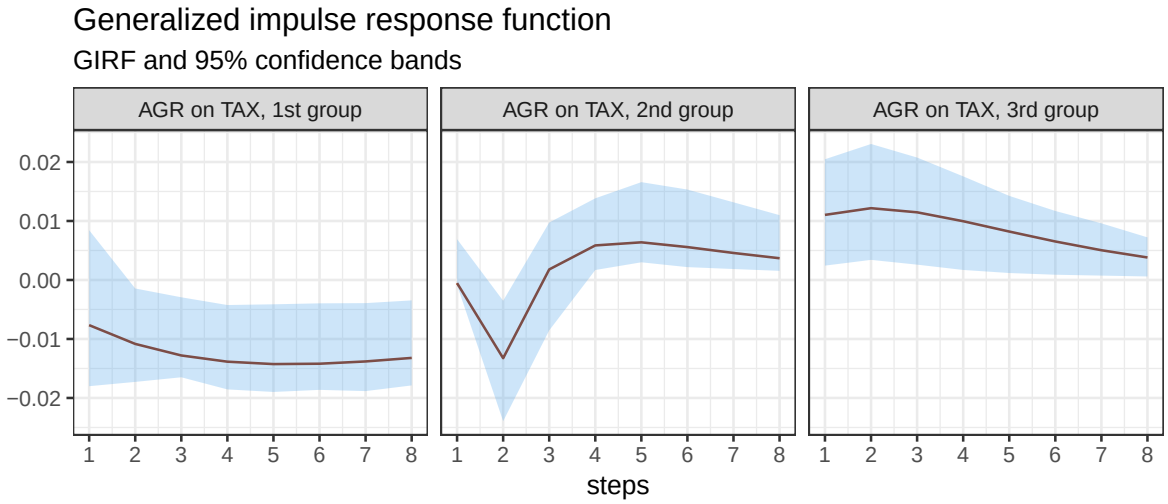

Figure 8: Response of TAX from a shock in AGR

The situation is quite similar to TRADE, because as income rises, the shocks become positive from negative. The confidence intervals show that in general all responses are significant. As Minh Le et al. (2012) states, the negative effect of AGR can be explained due to the fact that agricultural goods are harder to tax. It is possible, that as a country develops, the effectiveness of its tax system becomes better, meaning improved tax collection, more effective tariffs and etc. This could transpire to the taxation of agricultural goods, which would lead to a situation where the sector no longer poses as a burden to a country's tax system. It is possible, that higher income countries actually have benefits from the increase of agriculture.

The GDPPC shocks on TAX can be seen below:

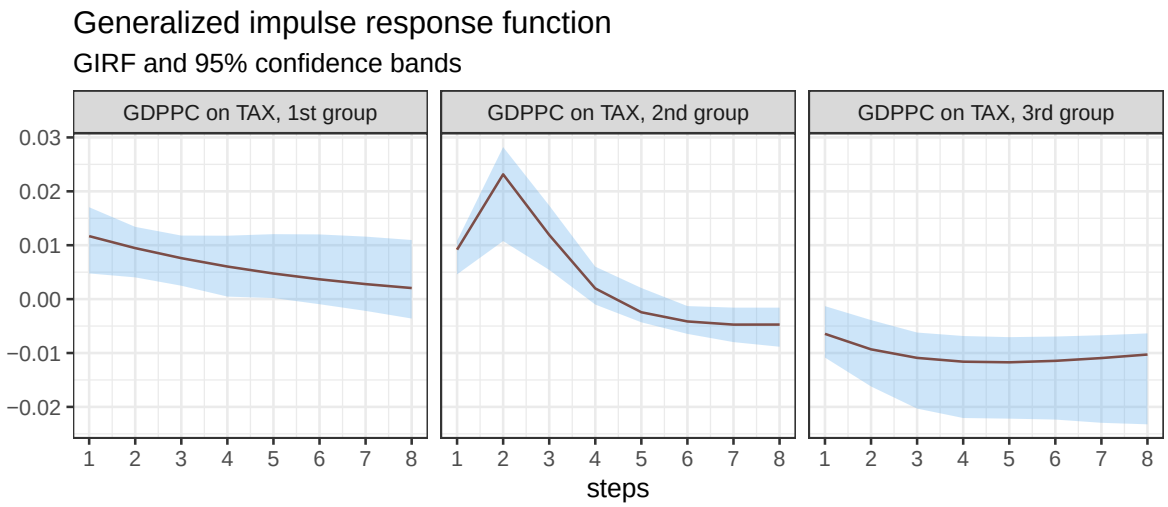

Figure 9: Response of TAX from a shock in GDPPC

From figure 9 it can be seen that a positive shock in GDPPC leads to a decrease in TAX at the highest income group. This is in fact the opposite of what can be expected, as GDPPC can in interpreted as income, thus their increase should rise tax revenue. In other groups, the shock is positive at first, but in later periods it converges towards 0. A negative response in the third group can be due to the fact in high income countries, possibly a large amount of income is concentrated in a small number of companies. When their income increases, due to ineffective progressive taxes, the budget might not increase by a proportional amount.

Below in figure 10 the results from a positive shock in CONS are displayed. Generally, consumption should have a positive effect as usually goods are taxed with a value added tax or something similar, thus consuming more should lead to higher TAX. In the first group, however, the response is 0 throughout all periods. A possible interpretation for this is that lower income countries might have an effective tax system which could cause such a reaction. The shocks in the third group causes a positive response. The main conundrum appears in

the second group, where in the first lag we can see a positive reaction but it becomes negative in later moments. Perhaps the second group has a sizable shadow economy in which goods or services are bought but not taxed. The last group provides logical results, as a positive shock in CONS gives a positive response in TAX.

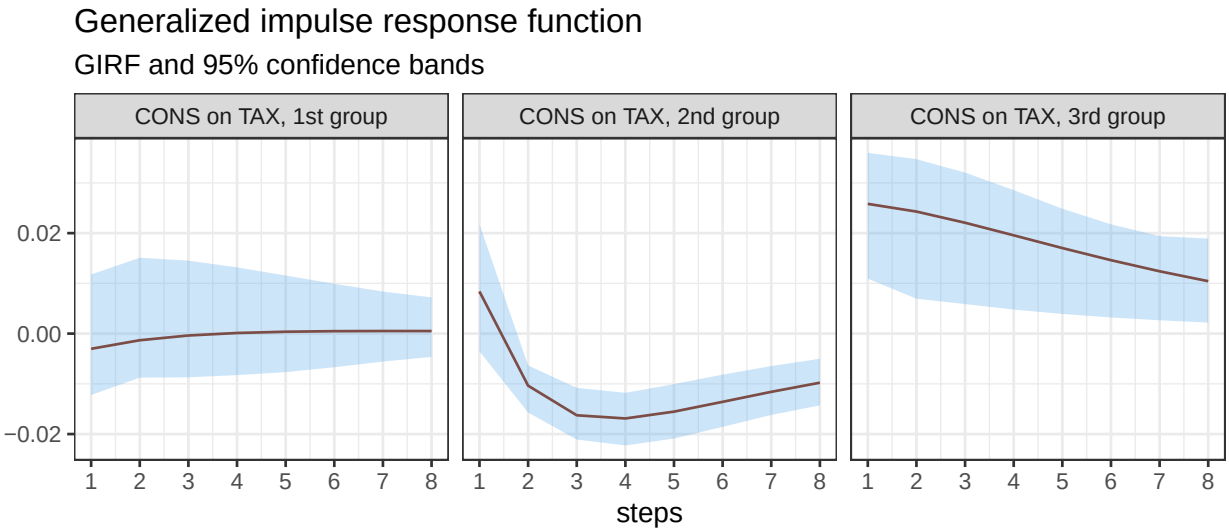

Figure 10: Response of TAX from a shock in CONS

Looking at TAX responses to POP shocks in figure 11, very large confidence intervals can be seen (probably due to the eigenvalues being close to 1). The shocks tend to become less negative going from the first group to the last, which coincides with AGR and TRADE results. This would mean that as the tax system improves with income and general development, the problem of administrating taxes of a large workforce no longer is an issue.

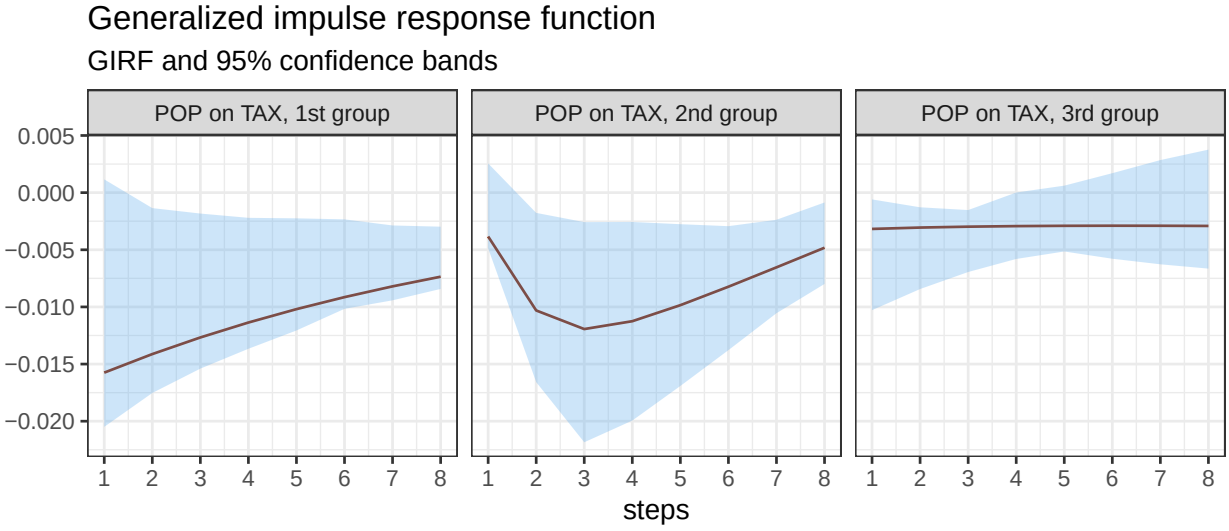

Figure 11: Response of TAX from a shock in POP

Lastly, CORRU shocks are presented below in figure 12. Shocks in the first two groups can be regarded as not significant due to confidence intervals covering 0. The third group

exhibits a negative reaction from the second period. Since two groups do not respond to CORRU shocks, it is a possibility that the index might not have an effect on TAX.

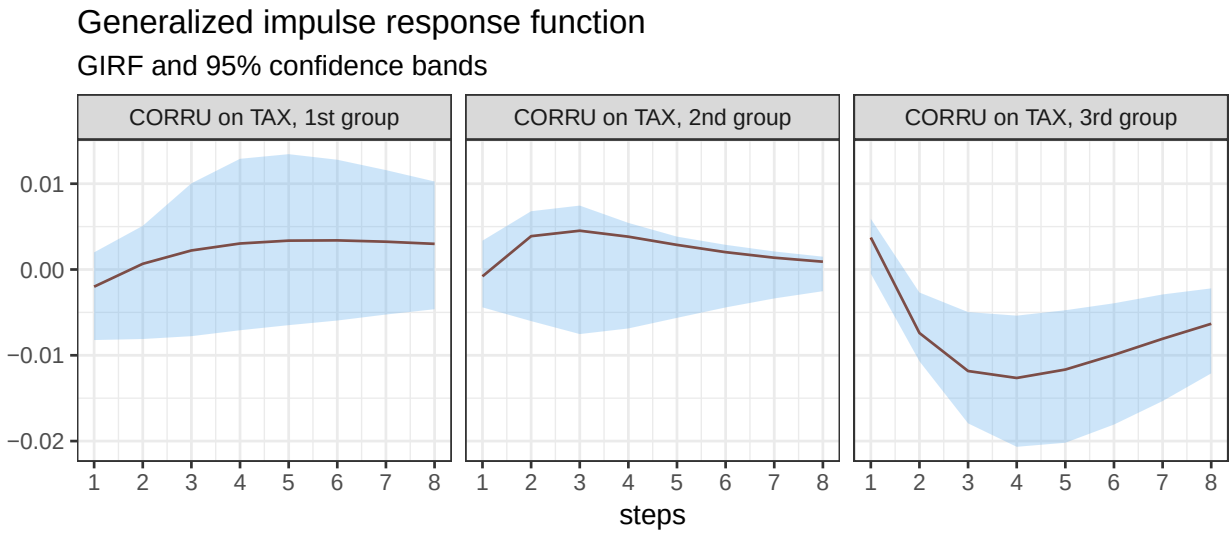

Figure 12: Response of TAX from a shock in CORRU

The results of IRF analysis show that the variables have a different effect in each income group. This information has to be accounted for in further modeling. Since Lithuania belongs to the second group, it would appear that the country is in a transition phase. It is possible that as Lithuania develops to a high income country, the effects of the factors at hand will have a different outcome and create new challenges.

3.4 Dynamic panel data modeling

As mentioned before, a dynamic panel data model is one that includes a lagged dependent (or independent) variable. In inclusion of this information enriches the model by capturing the dynamics of the variables at hand. The lagged dependent variable can also encompass information that affects the dependent variable but has has been excluded.

In this section dynamic panel data models are implemented. These models are estimated in two approaches: the system GMM and the orthogonal reparametrization, proposed by Lancaster (2002). The latter will be called MLE (maximum likelihood) henceforth. The regular GMM is not implemented because it gave poorer results than system GMM, therefore it was excluded. Since IRF analysis showed that variables effect the dependent variable differently in each income group, the dynamic models have a fixed effect included in them. This will account for group effects as well as other not included variables in the analysis. Also, the data is in logged differences, because when estimation was done on only logged data, the lagged dependent variable coefficient reached nearly 1, indicating non stationarity.

For each estimation method, the lag selection is a bit different. Since the system GMM approach does not have a formal method for lag selection, the initial model consists of the lag of the dependent variable with independent variables not lagged. A crisis dummy for 2009 is also included. In this model, AGR and CORRU were not significant, so their first lags were included, but were insignificant also. Second lags were not tested, because this would mean that the present TAX is determined by AGR and CORRU from 2 years behind which is very unlikely. Thus the initial model is chosen. Like in the PVAR model, lags from second to fifth of the dependent variable are used as instruments for estimation. Appendix L provides the estimated models with Sargan's test p values.

For the MLE estimation, the deviance information criterion (DIC) is used of lag selection. Once again, the initial model is the same as for the system GMM. AGR and CORRU were also insignificant, so their first lags were included and the DIC was calculated in order to see, which model is the best suited for the data. Once again, the initial model was best suited for the data. Appendix K provides the estimated models with DIC values.

Thus, in both the MLE and system GMM approach, the following model is estimated:

$$\begin{aligned}\Delta ln(TAX_{i,t}) =& \gamma \Delta ln(TAX_{i,t-1}) + \beta_1 \Delta ln(AGR_{i,t}) + \beta_2 \Delta ln(TRADE_{i,t}) + \\ & \beta_3 \Delta ln(GDPPC_{i,t}) + \beta_4 \Delta ln(CONS_{i,t}) + \beta_5 \Delta ln(POP_{i,t}) + \\ & \beta_6 \Delta ln(CORRU_{i,t}) + \phi CRISIS + \alpha_i^* + \epsilon_{i,t}\end{aligned} \tag{48}$$

Of course, the fixed effect is "removed" in the estimation process. The estimated coefficients for both models are presented below:

	MLE	System GMM	Significant?
TAX lag	-0.095	-0.094	YES_YES
CONS	0.193	0.168	YES_YES
AGR	0.013	0.011	NO_NO
TRADE	0.082	0.104	YES_YES
GDPPC	0.329	0.271	YES_YES
CORRU	0.012	0.015	NO_NO
POP	-0.394	-0.441	YES_YES
CRISIS	-0.04	-0.04	YES_YES

Table 11: Model coefficient estimates and significance at $\alpha = 0.95$

From table 11 it can be seen that the estimates of each variable are somewhat similar. POP, lag of TAX and CRISIS have negative values. The results obtained do not differ

from previous researches. Although insignificant, AGR and CORRU are kept in the models because they carry some sort of information about the corruption and agricultural sector in countries, which makes the model more realistic.

Having evaluated both models, the TAX potential can be calculated for each country. As mentioned before, the tax potential is essentially the fitted values of the model. Below the results from both models are presented for Lithuania. Since the data was differenced, 1 observation is lost, so the year starts from 2003.

Looking at the results below, the MLE approach shows that before the crisis, Lithuania's TAX fact was very close to its potential.

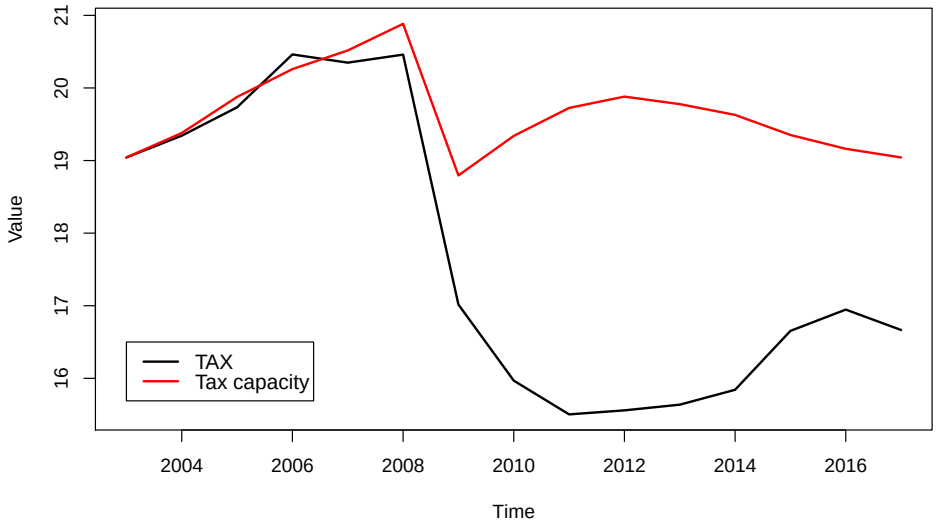

Figure 13: Lithuania's TAX fact and potential from the MLE approach

After the crisis, both the fact and potential experience a drop, although the latter begins to rise in future periods while the former continues to drop. The mean ration of these two curves (or tax effort) from 2003 to 2017 is 0.9, which means that Lithuania is 10% away from its TAX potential. According to the results, this also means that tax policy imposers can increase taxes or the tax base because in order to boost revenue, without causing too much harm to the economy.

Looking at the results in figure 14 from system GMM, they appear to be quite similar, although Lithuania never reaches its potential during the period. The ratio now reaches 0.83, which is a bit lower than the one from MLE estimation. Despite these little differences, both models show that Lithuania is not "living up" to its TAX potential and that a tax increase can be a justifiable action by law enforcers.

Although both models provide very similar results, only one must be chosen for further

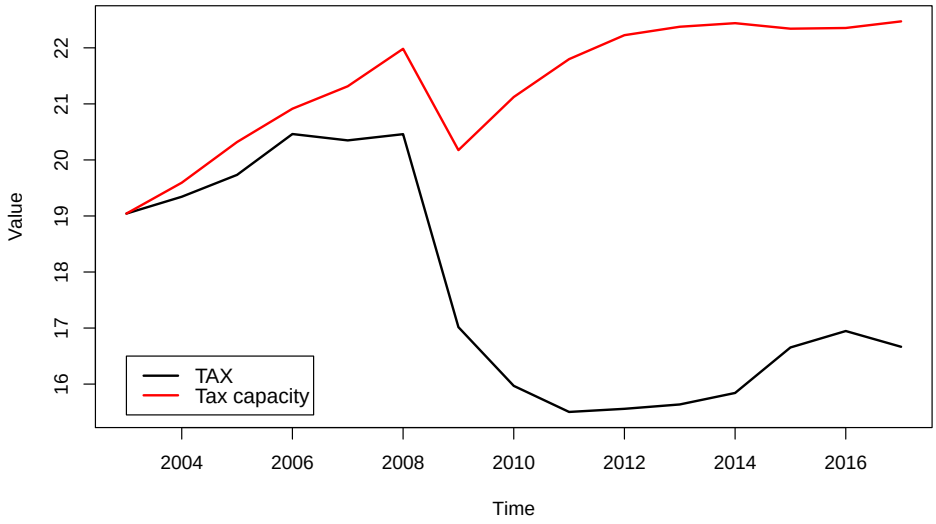

Figure 14: Lithuania's TAX fact and potential from the system GMM approach

analysis. Comparing these models is not trivial as both models are estimated very differently. Furthermore, any metrics, like MAPE, MSE or R^2 can not be used because a better model in this approach would be the one that fits the TAX data better, which is not the aim in these models. It is noteworthy to point, that the Sargan test for the system GMM model showed a p value of 0.051, which indicates that the instruments might not be fully valid, indicating that the MLE approach is better.

Analyzing the autocorrelation function estimates, 4 countries in the system GMM model and 6 countries in the MLE model appear to have significant lags. Since these numbers are low (99 countries in total), it can be concluded that the residuals of both models are not suffering from autocorrelation. Appendix F provides autocorrelation functions plots of those countries, who had significant lags.

In order to choose the correct model, 3 simulations are performed, which are done using either the coefficients and residual standard deviation from the MLE model, the system GMM or the mean from both. The following steps are carried out:

1. Take one of the 3 coefficient vectors and residual standard deviations;
2. Generate a new TAX for each country by using the coefficients obtained in step 1. The error term is generated using the normal distribution with 0 mean and the standard deviation of the residuals of either model (or the mean standard deviation from both);
3. Repeat the process in step 2 1000 times, thus obtaining 1000 TAX simulations for every country;
4. Fit both system GMM and MLE on the simulated data. This gives 1000 coefficient

vectors for each model;

5. Calculate the mean of all the coefficients from both models and then divide the results by the true coefficients with which the data is generated. This shows how much each model "captures" the true coefficient values;
6. Calculate the standard deviation of the coefficients for both models;
7. Choose the model who's deviations from the "true" coefficients are smaller and have lesser standard deviation.

In essence, this means that 3 scenarios are generated: one where the "true" process is from the MLE model, the other is from the system GMM and the last is the mean taken from them both (the coefficients and the standard deviations of residuals). The results from this simulations are provided below:

	Mean of both models		System GMM		MLE	
	System GMM	MLE	System GMM	MLE	System GMM	MLE
TAX lag	1.096	1.015	1.013	1.005	1.019	1.013
CONS	0.912	0.995	0.938	0.949	1.078	1.071
AGR	0.985	0.912	0.897	0.942	1.109	1.143
TRADE	0.913	0.978	1.065	1.074	0.871	0.888
GDPPC	0.889	0.956	0.854	0.873	1.091	1.072
CORRU	0.880	0.812	0.998	0.974	1.023	1.007
POP	1.178	1.059	1.060	1.082	0.983	0.909
CRISIS	1.021	1.071	0.989	0.994	1.040	1.023

Table 12: Simulation results: deviation from true parameters

As seen from table 12 both models are able to estimate the coefficients that are near their true value. When the mean of both models is taken, the MLE outperforms system GMM, as 5 out of 8 coefficients are closer to their "true" values. A similar situation is seen when the data is simulated using system GMM coefficients. Lastly, when the data is simulated using MLE coefficients, the latter once again shows better results, because 6 out of 8 coefficient values are closer to 1. This would indicate that MLE outperforms system GMM in these simulations.

Next, the standard deviations of the coefficients for each model in table 13 is shown. Both models perform similarly because when simulation is done with the mean of both coefficients. When simulating with system GMM, the latter performs better, with the same situation seen when simulating with MLE, as then MLE outperforms system GMM.

In accordance with tables 12, 13 and the fact that the Sargan's test showed that the

	Mean of both models		System GMM		MLE	
	System GMM	MLE	System GMM	MLE	System GMM	MLE
TAX lag	0.035	0.029	0.032	0.031	0.033	0.028
CONS	0.062	0.054	0.058	0.061	0.046	0.041
AGR	0.023	0.024	0.028	0.031	0.023	0.030
TRADE	0.034	0.035	0.041	0.039	0.035	0.037
GDPPC	0.116	0.073	0.115	0.120	0.095	0.101
CORRU	0.039	0.041	0.048	0.053	0.043	0.035
POP	0.272	0.185	0.254	0.287	0.197	0.248
CRISIS	0.011	0.014	0.014	0.012	0.015	0.010

Table 13: Simulation results: standard deviation of each coefficient

instruments might be weak for system GMM estimation, the MLE model is chosen for further analysis. It should be noted that in essence, these models are similar and using one or the other does not give very different results in further analysis.

Having chosen the model, the tax effort results are calculated for every country and presented below in a histogram. Each value is the mean of the tax effort time series.

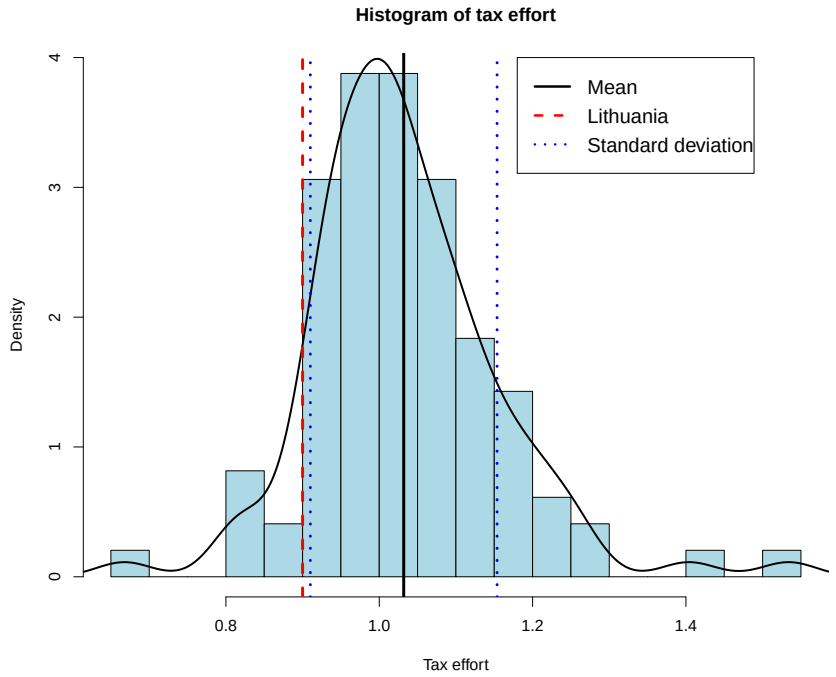

Figure 15: Histogram of tax effort

In figure 15, the red dotted line represents Lithuania's position (0.9) and the black line is the mean (1.05). Georgia was excluded from when making this plot as its efficiency greatly exceeded the rest of the countries and distorted the plot. It appears that Lithuania is more than one standard deviation away from the mean, making it a country which has a low tax

effort level. This puts the country in the seventh place from the bottom of all countries. Since the mean is 1.05, this means that countries are fulfilling their potential. Estonia's, Poland's and Latvia's efficiency is 1, 0.96 and 1.02 respectively, which means that Lithuania performs the worst among its neighbors. Other country tax effort is provided in Appendix C.

Appendix D provides tax capacity and TAX dynamics for Latvia, Estonia and Poland. It can be seen that Poland in general is under performing while Estonia and Latvia have spiked in later years.

Next, the results can be plotted on a two dimensional area where the x axis is the effort and the y axis is the TAX fact.

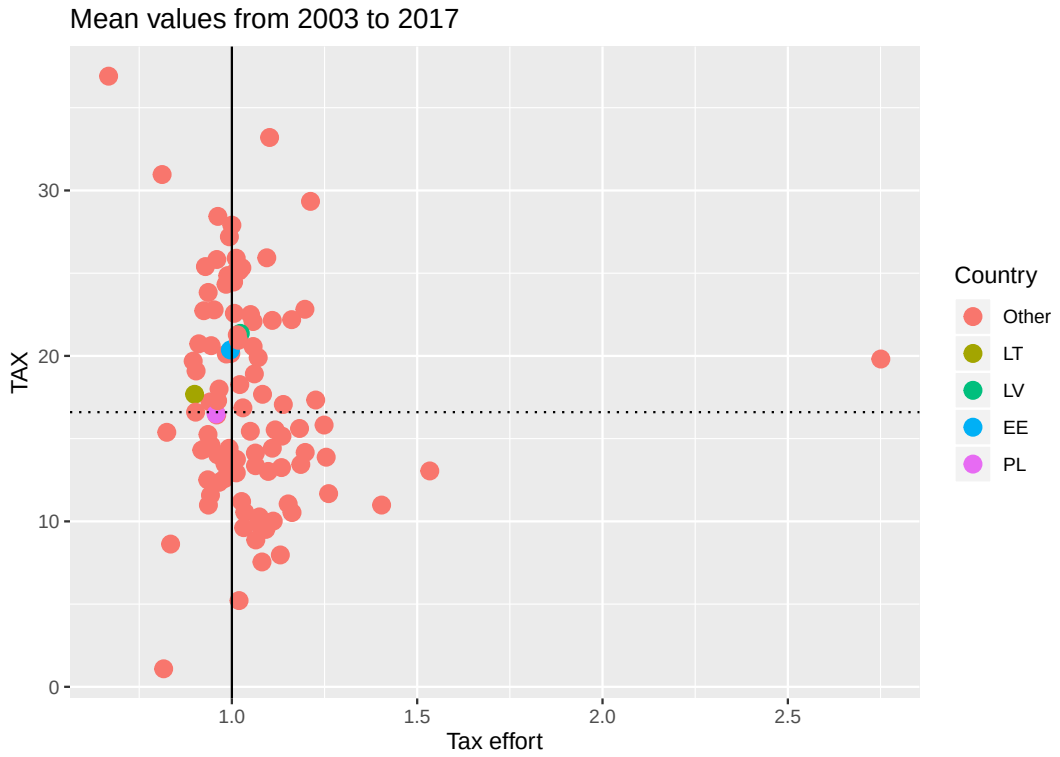

Figure 16: Results of TAX and tax effort

The vertical line represents tax effort when it is 1, and the horizontal dotted line is the mean of TAX. These two lines divide the plot into 4 groups: the one with low TAX and low effort (lower left), low TAX, high effort (lower right), high TAX, low effort (upper left) and high TAX, high effort (upper left). Lithuania belongs to the high TAX and low effort group. This means that in order for Lithuania to improve its tax effort, it should try to impose reforms in improving its tax collection rather than creating new taxes or widening its tax base. Perhaps the agriculture and foreign trade sectors could be a good starting point as the clustering analysis showed that it is possible that these sectors are inefficiently taxed.

The mean values (from 2003 to 2017) of these 4 groups in figure 16 by each variable is shown below in table 14. It appears that higher TAX is associated with larger GDPPC, TRADE, less corruption and smaller AGR. Also, figure 16 and table 14 shows that there appears to be no correlation between tax effort and TAX, meaning that a high taxation level in a country does not lead to its potential being reached.

	TAX	AGR	GDPPC	TRADE	CORRU	POP	CONS
Low TAX, low effort	13.51	6.54	13643.22	84.73	1.90	52.36	73.67
Low TAX, high effort	12.34	13.19	10704.37	85.85	1.80	62.02	81.51
High TAX, low effort	23.56	3.33	27252.31	112.42	3.12	52.13	78.07
High TAX, high effort	23.02	4.87	24443.03	100.86	2.76	52.32	79.07

Table 14: Mean values of each variable in every group

In order to have a better understanding, on how Lithuania's position has changed over time, the model can be estimated on different time periods. The period length, or window, was chosen to be 9 years long as this length provided stable models. The results are shown in figure 17 and the model coefficient estimates are provided in appendix G.

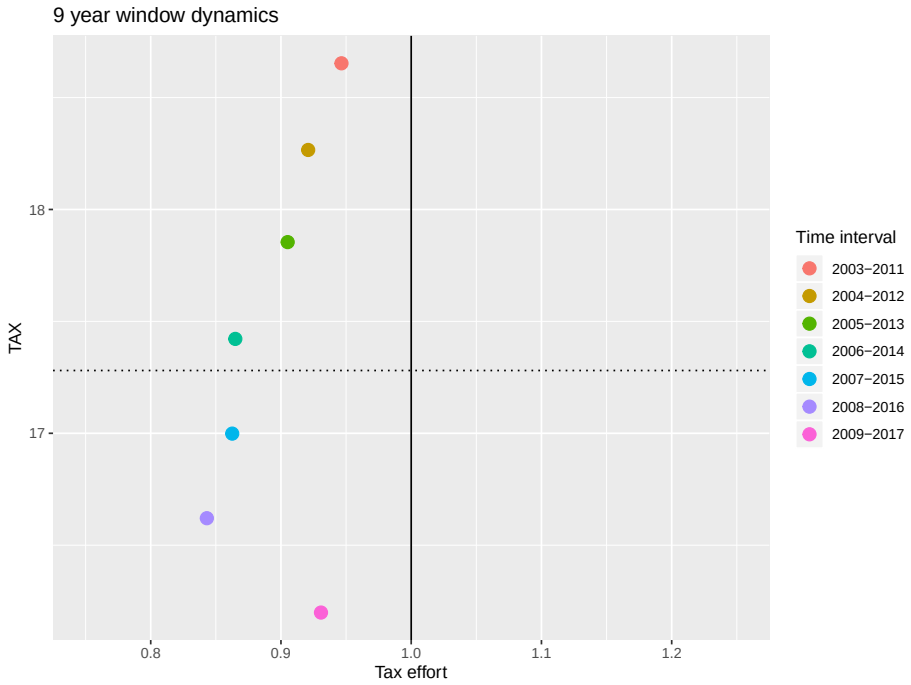

Figure 17: Lithuania's position through time

It seems that through the course of time, Lithuania has driven away from the mean TAX

level and widened the gap between its tax capacity and actual TAX. In the last period, however, there can be seen a rise in the level of tax effort. This is probably due to the rise in taxation in the last years, as seen in figure 13.

Next, the results can be plotted on a map, which gives a better understanding about the geographical distribution of tax effort. 28 European countries are shown with their color representing tax effort in figure 18 .

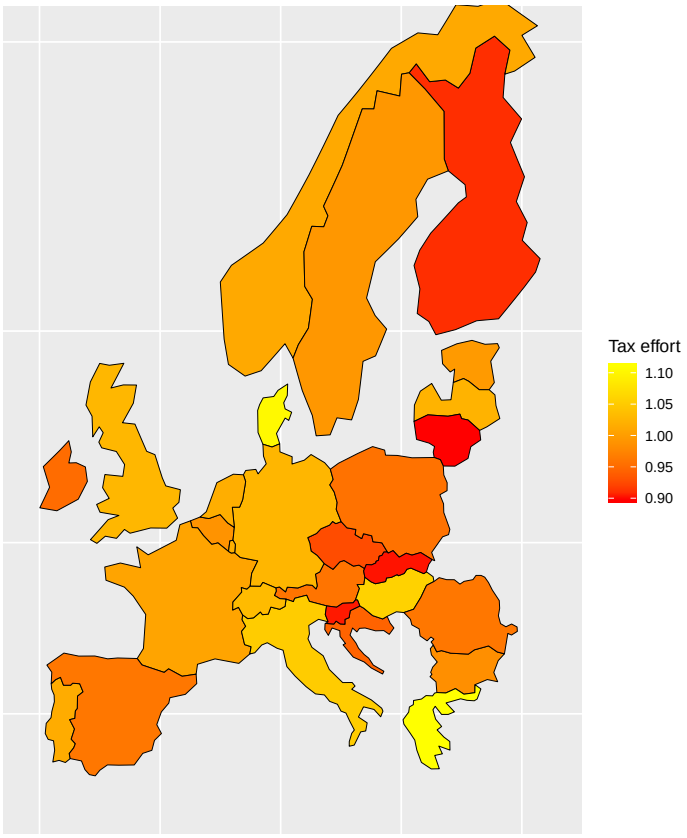

Figure 18: Map of tax effort

It appears that countries, located to the east of Europe seem to have a lower value of tax effort. Western countries like Germany, France, Italy and others characterize with higher tax effort. This might suggest the fact that there exists a "spillover" of tax efficiency among countries. Geographically related regions might implement similar fiscal policies which lead to akin tax efforts.

Lastly, since every country now has its actual TAX and potential time series, FPCA can once again be applied in order to have a better understanding of these two variables. The same procedure is repeated for TAX capacity: every time series is smoothed with tha same base and parameters. Then, FPCA is performed. The first component, which explains for 99.8% of all variation is plotted together with the same first component of TAX. The results

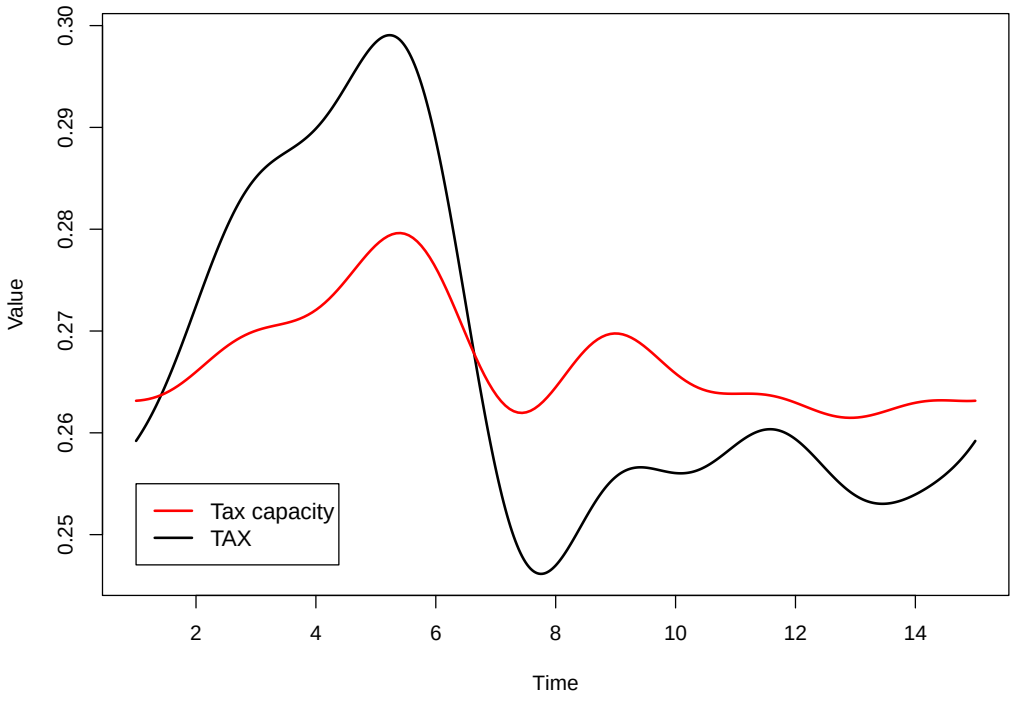

Figure 19: FPCA implemented on TAX and tax capacity

are shown below in figure 19.

It is visible that the general trend of tax capacity is considerably less variable than its fact. Furthermore, up to the financial crisis of 2008, countries tend to exceed their TAX potential and after the crisis they tend to under perform. In the last year however, a spike can be seen meaning that countries tend to catch up to their TAX potential. According to these findings, it is normal that Lithuania is not meeting its potential after the crisis. Probably the one of the reasons why Lithuania has a low tax effort overall is due to the fact that up to the crisis, the country did not exceed its tax capacity.

4 Conclusions

Several conclusions can be made from this thesis.

Clustering analysis showed that Lithuania's agricultural and trading sectors are respectively smaller and larger compared to other countries, that achieve a similar level of taxation as Lithuania. This can possibly mean that these sectors are inefficiently used taxwise. It is also possible, that the drop in tax to GDP ratio after the 2009 crisis is not natural in the sense of macroeconomics, meaning that fiscal policy imposers deviated from the taxation level which would be normal for Lithuania's macroeconomic climate.

Impulse response analysis showed that the factors governing tax revenue tend to act differently depending on the income level of the country. For example, positive shocks in the agricultural sector have a negative effect in countries with lower GDP per capita, but a positive effect in countries with high GDP per capita. This sort of "switching" can be for almost every variable that was included. Possibly, as a country progresses to a high income country, its institutional sectors also become more advanced, thus the standard effects of macroeconomic/institutional variables on taxation start to change.

Dynamic panel data modeling showed that from 2003 to 2017, Lithuania achieved 90% of its tax potential on average. Compared to other countries in terms of tax effort, this value puts Lithuania in the 7th place from the bottom out of 99 countries analyzed. The mean value of tax effort for the period of 2003 - 2017 is 105% meaning that on average, countries are fulfilling their tax potential. Despite this, Lithuania's tax to GDP ratio mean is higher then the mean of all countries, indicating that the level of taxation in the country is "good". This means that Lithuania should act through tax collection reforms in order to improve its overall state. Plotting the results on a map showed that there exists geographical groups with similar levels of tax effort.

Lastly, functional principal component analysis showed that prior to the 2009 crisis, countries were over performing and collecting more taxes than their potential. After the crisis, however, countries lowered their taxation levels and were collecting less then they could. This in turn gives insight that it is normal for Lithuania to not being able to achieve its taxation potential after the crisis.

Summing up, it can be concluded that Lithuania is not living up to its tax potential and that increasing tax revenue might not cause harm. These new increases can be done through new reforms oriented for a more effective tax collection. Perhaps the agriculture or foreign trade sectors would be a good starting point.

5 References

[1] Jeffrey M. Wooldridge (2010). *Econometric Analysis of Cross Section and Panel Data.* The MIT Press; second edition, Massachusetts.

[2] Tom S. Clark, Drew A. Linzer (2014). *Should I Use Fixed or Random Effects?*. Political Science Research and Methods, 3(2), 399-408. doi:10.1017/psrm.2014.32.

[3] Abhijit Sen Gupta (2007). *Determinants of Tax Revenue Efforts in Developing Countries.* IMF Working Papers, pp. 1-39.

[4] Gerardo Angeles Castro, Diana Berenice Ramírez Camarillo (2014). *Determinants of tax revenue in OECD countries over the period 2001-2011.* Contad. Adm [online]. 2014, vol.59, n.3, pp.35-59.

[5] Dioda Luca (2012). *Structural determinants of tax revenue in Latin America and the Caribbean, 1990-2009.* Sede Subregional de la CEPAL en México (Estudios e Investigaciones) 26103, Naciones Unidas Comisión Económica para América Latina y el Caribe (CEPAL).

[6] Muhammad Farhan Basheer, Aref Abdullah Ahmad, Saira Ghulam Hassan (2018). *Impact of economic and financial factors on tax revenue: Evidence from the Middle East countries.* Accounting. 53-60. 10.5267/j.ac.2018.8.001.

[7] Valeria Andreoni (2019). *Environmental taxes: Drivers behind the revenue collected.* Journal of Cleaner Production. 221. 10.1016/j.jclepro.2019.02.216.

[8] Mihn Le Tuan, Moreno-Dodson Blanca, Bayraktar Nihal. (2012). *Tax Capacity and Tax Effort: Extended Cross-Country Analysis from 1994 to 2009.* IMF working paper.

[9] M. Nagy Eltony (2002). *The Determinants of Tax Effort in Arab Countries.* Working Papers 0229, Economic Research Forum.

[10] Marcelo Piancastelli (2001). *Measuring the Tax Effort of Developed and Developing Countries: Cross Country Panel Data Analysis - 1985/95.* SSRN Electronic Journal. 10.2139/ssrn.283758.

[11] David A. Grigorian, Hamid Reza Davoodi (2007). *Tax Potential vs. Tax Effort: A Cross-Country Analysis of Armenia's Stubbornly Low Tax Collection.* IMF Working Papers. 07. 10.5089/9781451866704.001.

[12] Carola Pessino, Ricardo Fenochietto (2010). *Determining Countries' Tax Effort.* Hacienda Pública Española. 195. 65-87.

[13] Sandhya Garg, Ashima Goyal, Rupayan Pal (2017). *Why Tax Effort Falls Short of Tax Capacity in Indian States: A Stochastic Frontier Approach.* Public Finance Review, Volume: 45 issue: 2, page(s): 232-259.

[14] Jaime Valles-Gimenez, Anabel Zarate-Marco (2017). *Tax Effort of Local Governments and its Determinants: The Spanish Case.* Annals of Economics and Finance, Society for AEF, vol. 18(2), pages 323-348.

[15] Basil Dalamagas, Panagiotis Palaios, Stefanos Tantos (2019). *A New Approach to Measuring Tax Effort.* Economies 2019, 7(3), 77.

[16] Saeid Mahdavi, Joakim Westerlund (2018). *Subnational government tax revenue capacity and effort convergence: New evidence from sequential unit root tests.* Economic Modelling, 2018, vol. 73, issue C, 174-183.

[17] Sergio Focardi, Frank Fabozzi (2004). *A methodology for index tracking based on time-series clustering.* Quantitative Finance, 2004, vol. 4, issue 4, 417-425.

[18] Iwo Augustynski, Pawel Laskos-Grabowski (2018). *Clustering Macroeconomic Time Series.* Econometrics, vol 22, Issue 2, p 74-88.

[19] Lin J, Li Y (2009). *Finding Structural Similarity in Time Series Data Using Bag-of-Patterns Representation.* Proceedings of the 21st International Conference on Scientific and Statistical Database Management, SSDBM 2009, pp. 461–477. Springer-Verlag, Berlin. ISBN 978-3-642-02278-4.

[20] Marcella Corduas (2010). *Mining Time Series Data: A Selective Survey.* Data Analysis and Classification, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 355–362. Springer-Verlag.

[21] Rakthanmanon T, Campana B, Mueen A, Batista G, Westover B, Zhu Q, Zakaria J, Keogh E (2012). *Searching and Mining Trillions of Time Series Subsequences under Dynamic Time Warping.* In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, pp. 262–270. ACM, New York.

[22] Pablo Montero, Jos'e A. Vilar (2014). *TSclust: An R Package for Time Series Clustering*. Journal of Statistical Software, vol 62, issue 1.

[23] Batista Geapa, Wang X, Keogh EJ (2011). *A Complexity-Invariant Distance Measure for Time Series*. In Proceedings of the Eleventh SIAM International Conference on Data Mining, SDM'11, pp. 699–710. SIAM, Mesa.

[24] Douzal C. A., Nagabhushan P. (2007). *Adaptive Dissimilarity Index for Measuring Time Series Proximity*. Advances in Data Analysis and Classification, 1(1), 5–21.

[25] James O. Ramsay, Bernard W. Silverman (2002). *Applied Functional Data Analysis: Methods and Case Studies*. Springer Series in Statistics.

[26] Han Lin Shang (2014). *A survey of functional principal component analysis*. AStA Advances in Statistical Analysis, 2014, vol. 98, issue 2, 121-142

[27] Janine B. Illian, James I. Prosser, Kate L. Baker, J. Ignacio Rangel-Castro (2009). *Functional principal component data analysis: A new method for analysing microbial community fingerprints*. Journal of Microbiological Methods 79 (2009) 89–95.

[28] Suyundykov, R., Puechmorel, S. and Ferre, L. (2010). *Multivariate functional data clusterization by PCA in Sobolev space using wavelets*. 42e'mes Journe'es de Statistique.

[29] Julien Jacques, Cristian Preda (2014). *Functional data clustering: a survey*. Advances in Data Analysis and Classification, Springer Verlag, 2014, 8 (3), pp.24.

[30] Badi H. Baltagi (2002). *Econometric Analysis of Panel Data*. Wiley; 4th edition.

[31] Nickell Stephen J. (1981). *Biases in Dynamic Models with Fixed Effects*. Econometrica, Econometric Society, vol. 49(6), pages 1417-1426.

[32] M. Nerlove (1967). *Experimental evidence on the estimation of dynamic economic relations from a time series of cross-sections*. Economic Studies Quarterly, 18:42–74.

[33] M. Nerlove (1971). *Further evidence on the estimation of dynamic economic relations from a time series of cross-sections*. Econometrica, 39:359–382.

[34] M. Arellano and S. Bond (1991). *Some tests of specification for panel data: Monte carlo evidence and an application to employment equations*. Review of Economic Studies, 58:277–297.

[35] R. Blundell and S. Bond (2000). *Initial conditions and moment restrictions in dynamic panel data models.* Journal of Econometrics, 87:115–143.

[36] R. Blundell and S. Bond (2000). *Initial conditions and moment restrictions in dynamic panel data models.* Journal of Econometrics, 87:115–143.

[37] C. Hsiao (2014). *Analysis of Panel Data.* Cambridge University Press, New York, NY, 3rd edition.

[38] C. Hsiao, M. H. Pesaran, and A. K. Tahmiscioglu (2002). *Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods.* Journal of Econometrics, 109:107–150.

[39] T. Lancaster (2002). *Orthogonal parameters and panel data.* Review of Economic Studies, 69:647–666.

[40] Mark Pickup, Paul Gustafson, Davor Cubranic, Geoffrey Evans (2017). *OrthoPanels: An R Package for Estimating a Dynamic Panel Model with Fixed Effects Using the Orthogonal Reparameterization Approach.* The R Journal Vol. 9/1.

[41] Holtz-Eakin Douglas, Newey Whitney, Rosen Harvey (1988). *Estimating Vector Autoregressions with Panel Data.* Econometrica, 1988, vol. 56, issue 6, 1371-95.

[42] Blundell, R., Bond, S. (1998). *Initial Conditions and Moment Restrictions in Dynamic Panel Data Models.* Journal of Econometrics, 87(1):115–143.

[43] Lutkepohl, H. (2007). *New Introduction to Multiple Time Series Analysis.* Springer-Verlag Berlin Heidelberg, 2 edition.

[44] Pesaran, H., Shin, Y (1998). *Generalized Impulse Response Analysis in Linear Multivariate Models.* Economics Letters, 58(1):17–29.

[45] Elena Ivona Dumitrescu, Christophe Hurlin (2012). *Testing for Granger non-causality in heterogeneous panels.* Economic Modelling, 2012, vol. 29, issue 4, 1450-1460.

[46] Kleiber C, Lupi C (2011). *punitroots: Tests for Unit Roots in Panels of (Economic) Time Series, With and Without Cross-sectional Dependence.* Online.

[47] Im KS, Pesaran MH, Shin Y (2003). *Testing for Unit Roots in Heterogeneous Panels.* Journal of Econometrics, 115(1), 53–74.

[48] Levin A, Lin CF, Chu CSJ (2002). *Unit Root Tests in Panel Data: Asymptotic and FiniteSample Properties.* Journal of Econometrics, 108(1), 1–24.

[49] Choi I (2001). *Unit Root Tests for Panel Data.* Journal of International Money and Finance, 20(2), 249–272.

[50] Maddala GS, Wu S (1999). *A Comparative Study of Unit Root Tests with Panel Data and a New Simple Test.* Oxford Bulletin of Economics and Statistics, 61(Supplement 1), 631–652.

[51] Rantou Kleio Elissavet (2017). *Missing Data in Time Series and Imputation Methods.* Online.

6 Appendices

6.1 Appendix A

The Minkowki distance is probably one of the most common metrics. It is also called an L_q distance and is calculated:

$$d(X_T, Y_T) = \left(\sum_{t=1}^{T} (X_t - Y_t)^q \right)^{1/q} \tag{1}$$

The Minkowski metric becomes the Euclidean distance, when $q = 2$ which is the one of the most common distances used. When $q = 1$, one deals with the Manhattan distance.

The Frechet distance is also quite common. Let M be the set of all possible sequences of m pairs preserving the observations order in the form

$$r = ((X_{a1}, Y_{b1}), ..., (X_{am}, Y_{bm})) \tag{2}$$

where $a_i, b_j \in \{1, ..., T\}$ such that $a_1 = b_1 = 1, a_m = b_m = T$, and $a_{i+1} = a_i$ or $a_i + 1$ and $b_{i+1} = b_i$ or $b_i + 1$,for $i \in \{1, ..., m - 1\}$. Then the Frechet distance is defined by:

$$d(X_T, Y_T) = \min_{r \in M} \left(\max_{i=1,,,m} |X_{ai} - Y_{bi}| \right) \tag{3}$$

Differing from the Minkowski distance, the Frechet distance requires to take into account the ordering of the observations. It also can be calculated on sequences with different sizes.

Next, the Dynamic time warping distance (DTW):

$$d(X_T, Y_T) = \min_{r \in M} \left(\sum_{i=1,...,m} |X_{ai} - Y_{bi}| \right) \tag{4}$$

Similarly to the Frechet distance, DTW is aimed to find a mapping between the series so that a specific distance measure between the coupled observations (X_{ai}, Y_{bi}) is minimized. Both Frechet and DTW distances allow to identify similar shapes, even when shifting or scaling is present in the time series. Compared to d_{Lp} distances, both d_F and d_{DTW} ignore the temporal structure of the values as the proximity is based on the differences $|X_{ai} - Y_{bi}|$ regardless of the behavior around these values.

Mowing forward, we have the complexity-invariant dissimilarity (CID) measure. Batista et al. (2011) argue that, under many dissimilarity measures, pairs of time series with high levels of complexity frequently tend to be further apart than pairs of simple series. This way, complex series are incorrectly assigned to classes with less complexity. In order to mitigate

this effect, the authors propose to use information about complexity difference between two series as a correction factor for existing dissimilarity measures.

$$d_{CID}(X_T, Y_T) = CF(X_T, Y_T) \cdot d(X_T, Y_T) \tag{5}$$

where $d(X_T, Y_T)$ denotes a convetional raw-data distance (e.g. Euclidean distance) and $CF(X_T, Y_T)$ is a complexity correction factor given by

$$CF(X_T, Y_T) = \frac{\max\{CE(X_T), CE(Y_T)\}}{\min\{CE(X_T), CE(Y_T)\}} \tag{6}$$

with $CE(X_T)$ a complexity estimator of X_T. If all series have the same complexity, then $d_{CID}(X_T, Y_T) = d(X_T, Y_T)$. Nevertheless, an important complexity difference between X_T and Y_T turns into an increase of the dissimilarity between them. The complexity estimator is very simple and cosists in computing:

$$CE(X_T) = \sqrt{\sum_{t=1}^{T-1}(X_t - X_{t+1})^2} \tag{7}$$

The CID method is intuitive, parameter-free, invariant to the complexity of time series, computationally efficient, and it has produced improvements in accuracy in several clustering experiments carried out.

The correlation dissimilarity measure uses the Pearson's correlation coefficient as a component:

$$COR(X_T) = \frac{\sum_{t=1}^{T}(X_t - \overline{X}_T)(Y_t - \overline{Y}_T)}{\sqrt{\sum_{t=1}^{T}(X_t - \overline{X}_T)^2}\sqrt{\sum_{t=1}^{T}(Y_t - \overline{Y}_T)^2}} \tag{8}$$

Golay, Kollias, Stoll, Meier, Valavanis, and Boesiger (2005) construct a fuzzy k-means algorithm using the following two cross-correlation-based distances:

$$d_{COR,1}(X_T, Y_T) = \sqrt{2(1 - COR(X_T, Y_T))} \tag{9}$$

and

$$d_{COR,2}(X_T, Y_T) = \sqrt{\left(\frac{1 - COR(X_T, Y_T)}{1 + COR(X_T, Y_T)}\right)^B}, with B \geq 0 \tag{10}$$

Note that $d_{COR,2}$ becomes infinite when $COR(X_T, Y_T) = 1$ and the parameter B allows regulation of the fast decreasing of the distance.

Lastly, we have a dissimilarity measure, that uses first order temporal correlation as a component:

$$CORT(X_T, Y_T) = \frac{\sum_{t=1}^{T-1}(X_{t+1} - X_t)(Y_{t+1} - Y_t)}{\sqrt{\sum_{t=1}^{T-1}(X_{t+1} - X_t)^2}\sqrt{\sum_{t=1}^{T-1}(Y_{t+1} - Y_t)^2}} \tag{11}$$

$CORT(X_T, Y_T)$ belongs to the interval [-1,1]. The value 1 would show that both series have a similar dynamic behavior through their growths. -1 would indicate a similar growth in the opposite direction. Lastly 0 implies that there is no similarity between X_T and Y_T. The dissimilarity index proposed by Douzal Chouakria and Nagabhushan (2007) is defined as follows:

$$d_{CORT}(X_T, Y_T) = \phi[CORT(X_T, Y_T)] \cdot d(X_T, Y_T) \tag{12}$$

Here, $\phi_k(\cdot)$ is an adaptive tuning function to automatically modulate a conventional raw-data distance $d(X_T, Y_T)$ according to temporal correlation. Instead of, for instance, a linear tuning function, Douzal Chouakria and Nagabhushan (2007) propose to use an exponential adaptive function given by

$$\phi_k(u) = \frac{2}{1 + exp(ku)}, k \geq 0 \tag{13}$$

6.2 Appendix B

		First group		Second group		Third group	
Variable	Information criteria	Value	Lag	Value	Lag	Value	Lag
AGR	AIC	-833.779	1	-833.422	1	-832.843	1
AGR	AIC	-821.841	2	-821.674	2	-820.922	2
AGR	BIC	-2566.58	1	-2566.22	1	-2565.64	1
AGR	BIC	-2499.22	2	-2499.06	2	-2498.3	2
AGR	HQIC	-1592.01	1	-1591.65	1	-1591.07	1
AGR	HQIC	-1558.67	2	-1558.51	2	-1557.75	2
CONS	AIC	-834.864	1	-835.304	1	-835.966	1
CONS	AIC	-822.935	2	-823.575	2	-824.067	2
CONS	BIC	-2567.67	1	-2568.11	1	-2568.77	1
CONS	BIC	-2500.32	2	-2500.96	2	-2501.45	2
CONS	HQIC	-1593.09	1	-1593.53	1	-1594.2	1
CONS	HQIC	-1559.77	2	-1560.41	2	-1560.9	2
CORRU	AIC	-834.637	1	-834.782	1	-836.259	1
CORRU	AIC	-822.701	2	-822.904	2	-824.297	2
CORRU	BIC	-2567.44	1	-2567.58	1	-2569.06	1
CORRU	BIC	-2500.08	2	-2500.29	2	-2501.68	2
CORRU	HQIC	-1592.87	1	-1593.01	1	-1594.49	1
CORRU	HQIC	-1559.53	2	-1559.74	2	-1561.13	2
GDPPC	AIC	-835.816	1	-835.517	1	-836.195	1
GDPPC	AIC	-823.89	2	-823.915	2	-824.255	2
GDPPC	BIC	-2568.62	1	-2568.32	1	-2569	1
GDPPC	BIC	-2501.27	2	-2501.3	2	-2501.64	2
GDPPC	HQIC	-1594.05	1	-1593.75	1	-1594.42	1
GDPPC	HQIC	-1560.72	2	-1560.75	2	-1561.09	2
POP	AIC	-836.38	1	-836.223	1	-836.4	1
POP	AIC	-824.457	2	-824.366	2	-824.433	2
POP	BIC	-2569.18	1	-2569.03	1	-2569.2	1
POP	BIC	-2501.84	2	-2501.75	2	-2501.82	2
POP	HQIC	-1594.61	1	-1594.45	1	-1594.63	1
POP	HQIC	-1561.29	2	-1561.2	2	-1561.27	2
TRADE	AIC	-834.024	1	-834.598	1	-835.503	1
TRADE	AIC	-822.14	2	-822.674	2	-823.666	2
TRADE	BIC	-2566.83	1	-2567.4	1	-2568.31	1
TRADE	BIC	-2499.52	2	-2500.06	2	-2501.05	2
TRADE	HQIC	-1592.25	1	-1592.83	1	-1593.73	1
TRADE	HQIC	-1558.97	2	-1559.51	2	-1560.5	2

Table 15: Information criterion of estimated PVAR models

6.3 Appendix C

TAX	TAX CAP	TAX EF	COUNTRY	TAX	TAX CAP	TAX EF	COUNTRY	TAX	TAX CAP	TAX EF	COUNTRY
7.55	6.98	1.08	AFG	20.36	20.49	1.00	EST	10.55	10.18	1.03	MDG
16.43	17.17	0.96	AGO	20.73	22.78	0.91	FIN	13.37	12.56	1.06	MLI
12.58	12.85	0.98	ARG	22.57	22.43	1.01	FRA	36.91	55.65	0.67	MLT
22.73	24.59	0.92	AUS	25.33	24.66	1.03	GBR	18.27	18.01	1.02	MNG
25.83	26.94	0.96	AUT	19.81	7.22	2.75	GEO	16.86	16.37	1.03	MUS
24.85	25.14	0.99	BEL	15.38	18.68	0.82	GHA	14.60	15.47	0.94	MYS
13.88	11.08	1.25	BFA	8.63	10.34	0.83	GNQ	29.34	24.19	1.21	NAM
7.97	7.05	1.13	BGD	22.15	20.05	1.11	GRC	14.17	11.83	1.20	NIC
20.12	20.42	0.98	BGR	10.98	11.72	0.94	GTM	20.94	20.58	1.02	NLD
11.68	9.27	1.26	BHS	15.53	13.91	1.12	HND	25.90	25.60	1.01	NOR
20.14	20.20	1.00	BIH	20.62	21.84	0.94	HRV	13.05	8.51	1.53	NPL
17.22	18.30	0.94	BLR	22.08	20.92	1.06	HUN	28.43	29.54	0.96	NZL
22.82	19.06	1.20	BLZ	11.58	12.28	0.94	IDN	15.15	13.35	1.14	PER
14.31	15.58	0.92	BRA	10.53	9.06	1.16	IND	13.02	11.86	1.10	PHL
25.11	24.84	1.01	BRB	22.78	23.91	0.95	IRL	16.49	17.22	0.96	POL
11.05	9.66	1.15	BTN	24.33	24.72	0.98	ISL	21.27	20.98	1.01	PRT
25.40	27.36	0.93	BWA	23.84	25.46	0.94	ISR	8.88	8.35	1.06	PRY
12.51	13.38	0.94	CAN	22.50	21.44	1.05	ITA	5.21	5.12	1.02	PSE
9.62	9.33	1.03	CHE	24.46	24.37	1.00	JAM	17.29	18.01	0.96	ROU
17.68	16.32	1.08	CHL	18.00	18.59	0.97	JOR	13.46	13.69	0.98	RUS
9.51	8.72	1.09	CHN	10.02	9.01	1.11	JPN	12.93	12.78	1.01	SGP
14.42	13.00	1.11	CIV	14.13	13.28	1.06	KAZ	15.82	12.67	1.25	SLV
9.87	9.16	1.07	COG	10.98	7.85	1.40	KHM	16.60	18.44	0.90	SVK
13.44	11.33	1.19	COL	14.42	14.53	0.99	KOR	19.09	21.13	0.90	SVN
19.68	21.98	0.90	CPV	1.09	1.33	0.82	KWT	27.20	27.38	0.99	SWE
13.76	13.89	0.99	CRI	15.25	16.31	0.94	LBN	27.90	27.97	1.00	SYC
30.96	38.05	0.81	CYP	19.90	18.59	1.07	LCA	15.61	13.20	1.18	TGO
14.40	15.56	0.93	CZE	12.36	12.80	0.97	LKA	15.44	14.71	1.05	THA
11.20	10.92	1.03	DEU	17.68	19.64	0.90	LTU	20.57	19.47	1.06	TUN
33.20	30.15	1.10	DNK	25.14	24.66	1.02	LUX	17.33	14.15	1.23	UKR
13.25	11.69	1.13	DOM	21.36	20.91	1.02	LVA	18.91	17.84	1.06	URY
13.75	13.56	1.01	EGY	22.19	19.10	1.16	MAR	10.27	9.56	1.07	USA
14.02	14.57	0.96	ESP	17.07	14.97	1.14	MDA	25.93	23.70	1.09	ZAF

Table 16: 2003-2017 period mean TAX, tax capacity and tax effort

6.4 Appendix D

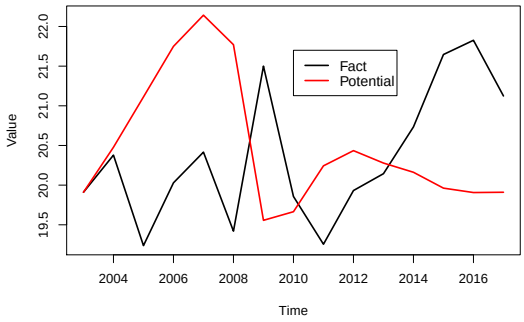

Figure 20: Estonia's tax effort and TAX

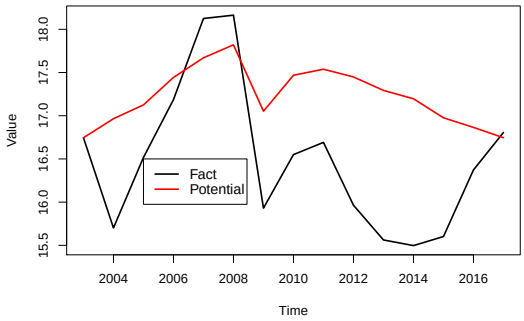

Figure 21: Polands tax effort and TAX

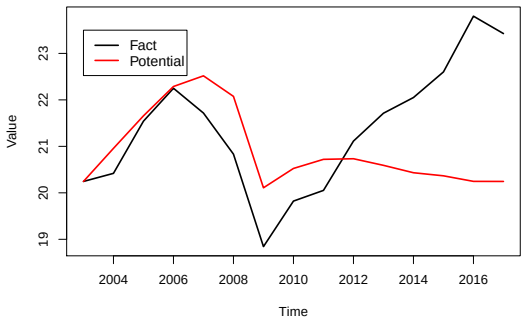

Figure 22: Latvia's tax effort and TAX

6.5 Appendix E

Le, Dodson and Bayraktar (2012) Developed and Developing Countries	Bahl (2003) OECD and Underdeveloped Countries	Alm, Martinez - Vazquez and Schneider (2004) Developed and Developing Countries	Teera (2002) Developed and Developing Countries	Piancastelli (2001) Developed and Developing Countries	Stotsky and Wolde Mariam (1997) Sub-Saharan African Countries	Tanzi (1992) Developing Countries	Leuhold (1991) African Countries	Bahl (1971) Developing Countries	Shin (1969) Developed and Developing Countries	Lotz and Morss (1967) Developed and Developing Countries
Ratio of tax revenue to GDP and that of fiscal revenue to GDP	Ratio of tax revenue to GDP	Ratio of tax revenue to GDP	Ratio of tax revenue to GDP	Total tax revenue/GDP	Ratio of tax revenue to GDP	Ratio of tax revenue to GDP	Ratio of tax revenue to GDP	Tax capacity	Tax burden	Ratio of tax revenue to GNP
Portion of agriculture in GDP (negative, statistically significant)	Non- agricultural portion of GDP (positive, statistically significant)	Agriculture/GNP (negative, statistically insignificant)	Portion of agriculture in GDP (negative or assumed to be positive, strongly negative effect in low-income countries.	Agriculture/GDP (negative and positive, negative and statistically significant in panel data analysis)	Portion of agriculture (negative, statistically significant)	Portion of agriculture in GDP (negative, statistically significant)	Portion of agriculture in GDP (negative, but not always statistically significant)	Portion of agriculture (negative, statistically significant)	GNP per capita (positive, statistically significant for samples and sub-samples of all high-income and low- income countries)	GNP per capita (positive, statistically significant for all the samples and the samples from low-income countries, but insignificant for high-income countries)
The ratio of the sum of exports and imports to the GDP (positive, statistically significant)	Openness ratio (ratio of the sum of imports and exports to GDP) (positive, statistically significant)	Mining/ GNP (positive, statistically significant)	Portion of manufacturing sector in GDP (negative, statistically insignificant)	Industry/ GDP (positive, statistically significant in time series analysis)	Portion of mining (negative, statistically significant)	Income per capita (positive, statistically significant, but statistically insignificant for some years)	Mining/ GDP (positive, negative, statistically insignificant)	Portion of mining (positive, statistically significant)	Foreign trade ratio (positive statistically insignificant)	Ratio of the sum of imports and exports to GNP (positive, statistically significant for all the samples and the samples from low-income countries, but insignificant for high-income countries)
Population growth rate (age sensitivity ratio) (negative, statistically insignificant in sub-periods)	GNP per capita (negative, statistically significant)	GDP per capita (negative and positive, not always statistically significant)	Portion of service sector in GDP (positive, not always statistically significant)	GNP per capita (positive/negative trend: positive, but not always significant)	Import/GDP (positive, statistically significant)	Ratio of foreign trade (ratio of the sum of imports and exports to GDP) (positive, statistically significant)	Income per capita (positive, but statistically insignificant)	Ratio of agricultural income (negative, statistically insignificant)	Inflation rate (positive, statistically significant for only the low- income countries)	
GNP per capita (positive, statistically insignificant with variable of institutional quality)	Population growth rate (positive, statistically significant)	Taxes on foreign trade/GNP (negative, statistically insignificant)	Ratio of the sum of imports and exports to GDP (negative and positive, statistically insignificant, high positive effect in low- and mid-level income countries)	Income per capita (positive, statistically significant)	Portion of manufacturing (positive, negative, statistically insignificant)	Ratio of foreign debts to GDP (Positive, but not always significant for all the assumptions)	The ratio of foreign aids and debts to income (positive, statistically significant)	Export rate (positive, but not always significant)		
Management Quality (Bureaucratic efficiency and Corruption Index) (negative, statistically significant for sub-periods)	Underground economy / GDP (negative, statistically significant)	Underground Economy (positive, but not always significant, negative and statistically significant only for OECD countries)	Trade/ GDP (positive, statistically significant)		Import/GDP (positive/negative, statistically insignificant)			Population growth rate (negative, statistically significant for all the samples and the samples from low-income countries)		
Underground Economy /GDP (negative, statistically insignificant when bureaucratic efficiency is added)	Simple correlation between tax effort and underground economy (negative, statistically insignificant)		Other determinants: Foreign aid rate (trend: Negative effect) Ratio of expenditures to GDP (trend: positive) Ratio of total expenditures (trend: Negative and positive)		Export/GDP (positive, statistically significant)					
Total consumption/GDP (positive, statistically significant)										

6.6 Appendix F

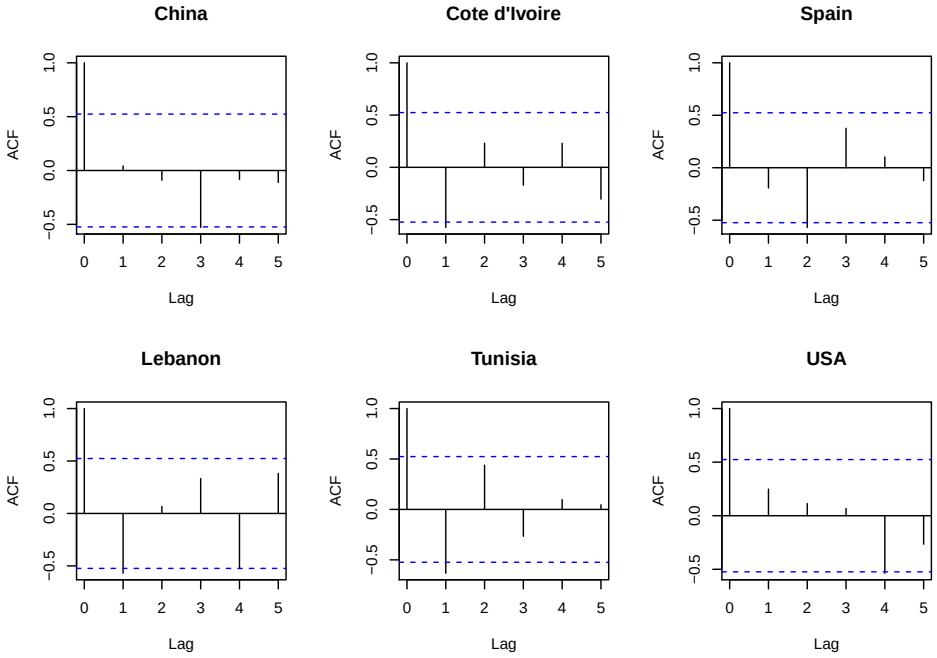

Figure 23: Residual autocorrelation function plots from MLE model

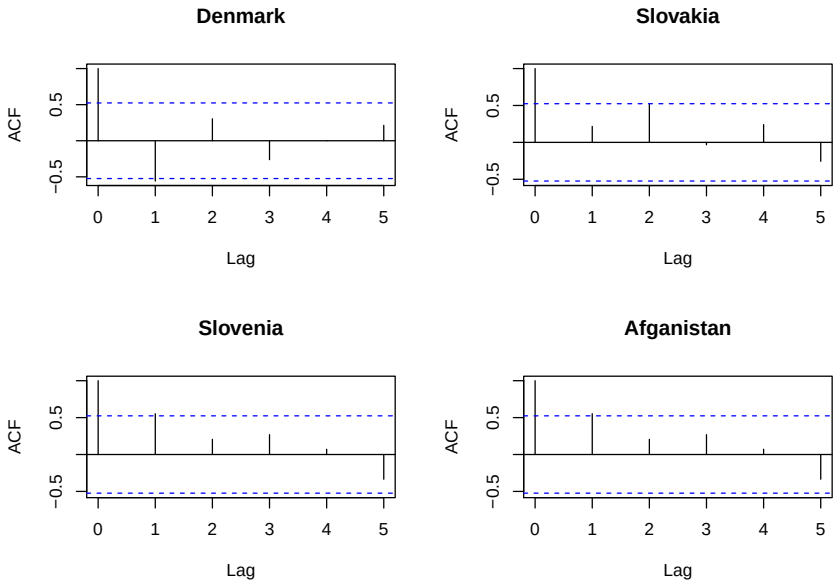

Figure 24: Residual autocorrelation function plots from system GMM model

6.7 Appendix G

	2003 - 2011	2004 - 2012	2005 - 2013	2006 - 2014	2007 - 2015	2008 - 2016	2009 - 2017
TAX lag	0.029	0.005	-0.024	-0.072*	-0.085*	-0.174*	-0.213*
AGR	0.156	0.019	0.021	0.008	0.33*	-0.001	0.002
GDPPC	0.526*	0.548*	0.543*	0.582*	0.67*	0.712*	0.198
TRADE	0.117*	0.12*	0.118*	0.085*	0.109*	0.14*	0.087*
POP	-0.767	-0.47	0.273	0.742*	0.897*	0.831*	0.315
CORRU	0.1*	0.054	0.048	0.028	0.76	0.078	-0.071
CONS	0.157*	0.16*	0.126*	0.207*	0.448*	0.508*	0.291*

Table 18: Model coefficient estimates and significance at $\alpha = 0.9$

Coefficients with the star sign mean that they are significant.

6.8 Appendix H

	Canberra distance						
	TAX	AGR	GDPPC	TRADE	CORRU	POP	CONS
clust1	12.227	9.398	13768.099	76.010	1.840	58.012	78.258
clust2	22.348	4.371	24943.065	117.927	2.787	51.073	76.127
clust3	23.562	4.225	25711.803	100.982	2.945	52.533	79.468
clust4	13.775	13.895	7622.650	104.987	1.851	60.274	79.645
	Euclidean distance						
	TAX	AGR	GDPPC	TRADE	CORRU	POP	CONS
clust1	9.275	10.510	16972.502	76.706	1.942	61.790	75.738
clust2	15.020	10.346	9011.120	90.676	1.869	56.346	80.171
clust3	26.820	2.555	37624.671	116.782	3.372	52.636	75.567
clust4	20.916	5.203	17781.835	96.498	2.550	51.357	81.178
	Manhattan distance						
	TAX	AGR	GDPPC	TRADE	CORRU	POP	CONS
clust1	9.337	11.463	16840.163	76.492	1.942	62.821	76.817
clust2	14.991	9.899	9073.205	90.776	1.869	55.863	79.664
clust3	26.820	2.555	37624.671	116.782	3.372	52.636	75.567
clust4	20.916	5.203	17781.835	96.498	2.550	51.357	81.178
	Maximum distance						
	TAX	AGR	GDPPC	TRADE	CORRU	POP	CONS
clust1	9.275	10.510	16972.502	76.706	1.942	61.790	75.738
clust2	14.926	10.407	8950.592	91.647	1.835	56.321	80.150
clust3	26.820	2.555	37624.671	116.782	3.372	52.636	75.567
clust4	20.803	5.343	17454.663	94.232	2.584	51.664	81.169
	Minkowski distance						
	TAX	AGR	GDPPC	TRADE	CORRU	POP	CONS
clust1	9.337	11.463	16840.163	76.492	1.942	62.821	76.817
clust2	14.991	9.899	9073.205	90.776	1.869	55.863	79.664
clust3	26.820	2.555	37624.671	116.782	3.372	52.636	75.567
clust4	20.916	5.203	17781.835	96.498	2.550	51.357	81.178

Table 19: Cluster centers from FPCA clustering

6.9 Appendix I

	CID distance						
	TAX	AGR	GDPPC	TRADE	CORRU	POP	CONS
clust1	12.024	10.969	12848.572	81.661	1.868	59.541	78.059
clust2	18.102	8.240	8952.786	104.449	2.089	53.062	80.265
clust3	23.821	3.270	30026.544	101.210	3.037	52.395	78.306
clust4	31.114	3.108	32138.266	158.337	3.323	46.764	78.381
	COR distance						
	TAX	AGR	GDPPC	TRADE	CORRU	POP	CONS
clust1	16.503	9.901	16254.363	93.001	2.192	58.965	80.248
clust2	19.219	5.781	21724.720	92.594	2.468	54.218	76.440
clust3	17.333	4.765	15283.436	103.942	2.035	52.184	70.955
clust4	17.839	6.431	20293.623	94.494	2.675	50.496	81.432
	CORT distance						
	TAX	AGR	GDPPC	TRADE	CORRU	POP	CONS
clust1	9.450	11.587	16153.696	75.211	1.910	62.236	76.440
clust2	15.062	9.886	9133.725	92.801	1.835	56.067	79.840
clust3	21.122	4.817	22225.066	94.222	2.784	51.604	79.810
clust4	27.663	2.532	35742.119	122.459	3.307	52.679	76.305
	DTWARP distance						
	TAX	AGR	GDPPC	TRADE	CORRU	POP	CONS
clust1	10.578	12.372	14141.310	70.332	1.899	63.259	79.198
clust2	15.503	8.644	9253.055	100.333	1.886	53.485	78.362
clust3	26.820	2.555	37624.671	116.782	3.372	52.636	75.567
clust4	20.916	5.203	17781.835	96.498	2.550	51.357	81.178
	Euclidean distance						
	TAX	AGR	GDPPC	TRADE	CORRU	POP	CONS
clust1	9.450	11.587	16153.696	75.211	1.910	62.236	76.440
clust2	15.175	9.813	9072.301	91.777	1.888	55.822	80.179
clust3	21.474	4.483	23558.570	96.283	2.773	51.650	79.170
clust4	27.663	2.532	35742.119	122.459	3.307	52.679	76.305
	Frechet distance						
	TAX	AGR	GDPPC	TRADE	CORRU	POP	CONS
clust1	9.209	10.018	17591.524	67.219	1.968	61.251	76.721
clust2	14.731	10.736	8869.145	94.040	1.825	56.772	79.741
clust3	23.280	3.858	28135.259	101.819	2.997	52.561	78.066
clust4	34.558	1.913	25752.083	194.503	2.951	44.745	78.840

Table 20: Cluster centers from discrete time series clustering

6.10 Appendix J

PVAR(1) model variables	First group	Second group	Third group
GDPPC and TAX	0.921	0.893	0.892
	0.857	0.446	0.727
CONS and TAX	0.880	0.832	0.783
	0.582	0.440	0.749
TRADE and TAX	0.841	0.759	0.797
	0.841	0.477	0.676
AGR and TAX	0.889	0.795	0.690
	0.784	0.455	0.690
POP and TAX	0.909	0.857	0.991
	0.898	0.479	0.753
CORRU and TAX	0.821	0.832	0.740
	0.730	0.476	0.740

Table 21: Eigenvalues of PVAR(1) models

Since no eigenvalue exceeds 1, all PVAR models are stable.

6.11 Appendix K

Model	DIC
$\Delta ln(TAX_{i,t}) = \gamma \Delta ln(TAX_{i,t-1}) + \beta_1 \Delta ln(AGR_{i,t})' + \beta_2 \Delta ln(TRADE_{i,t}) + \beta_3 \Delta ln(GDPPC_{i,t}) + \beta_4 \Delta ln(CONS_{i,t}) + \beta_5 \Delta ln(POP_{i,t}) + \beta_6 \Delta ln(CORRU_{i,t})' + \phi CRISIS + \alpha_i^* + \epsilon_{i,t}$	-16045.65
$\Delta ln(TAX_{i,t}) = \gamma \Delta ln(TAX_{i,t-1}) + \beta_1 \Delta ln(AGR_{i,t-1}) + \beta_2 \Delta ln(TRADE_{i,t}) + \beta_3 \Delta ln(GDPPC_{i,t}) + \beta_4 \Delta ln(CONS_{i,t}) + \beta_5 \Delta ln(POP_{i,t})' + \beta_6 \Delta ln(CORRU_{i,t})' + \phi CRISIS + \alpha_i^* + \epsilon_{i,t}$	-16031.61
$\Delta ln(TAX_{i,t}) = \gamma \Delta ln(TAX_{i,t-1}) + \beta_1 \Delta ln(AGR_{i,t})' + \beta_2 \Delta ln(TRADE_{i,t}) + \beta_3 \Delta ln(GDPPC_{i,t}) + \beta_4 \Delta ln(CONS_{i,t}) + \beta_5 \Delta ln(POP_{i,t})' + \beta_6 \Delta ln(CORRU_{i,t-1})' + \phi CRISIS + \alpha_i^* + \epsilon_{i,t}$	-16023.97
$\Delta ln(TAX_{i,t}) = \gamma \Delta ln(TAX_{i,t-1}) + \beta_1 \Delta ln(AGR_{i,t-1}) + \beta_2 \Delta ln(TRADE_{i,t}) + \beta_3 \Delta ln(GDPPC_{i,t}) + \beta_4 \Delta ln(CONS_{i,t}) + \beta_5 \Delta ln(POP_{i,t})' + \beta_6 \Delta ln(CORRU_{i,t-1})' + \phi CRISIS + \alpha_i^* + \epsilon_{i,t}$	-16029

Table 22: Estimated models and their DIC

Here, DIC is defined as $D(\theta) = -2log(p(Y|\theta)) + C$, with C being a constant that cancels out in all calculations that compare different models, and which therefore does not need to be known.

The $'$ near a variable indicates that it was insignificant at $\alpha = 0.95$.

6.12 Appendix L

Model	Sargan's test p value
$\Delta ln(TAX_{i,t}) = \gamma\Delta ln(TAX_{i,t-1}) + \beta_1\Delta ln(AGR_{i,t})' + \beta_2\Delta ln(TRADE_{i,t}) + \beta_3\Delta ln(GDPPC_{i,t}) + \beta_4\Delta ln(CONS_{i,t}) + \beta_5\Delta ln(POP_{i,t}) + \beta_6\Delta ln(CORRU_{i,t})' + \phi CRISIS + \alpha_i^* + \epsilon_{i,t}$	0.051
$\Delta ln(TAX_{i,t}) = \gamma\Delta ln(TAX_{i,t-1}) + \beta_1\Delta ln(AGR_{i,t-1})' + \beta_2\Delta ln(TRADE_{i,t}) + \beta_3\Delta ln(GDPPC_{i,t}) + \beta_4\Delta ln(CONS_{i,t}) + \beta_5\Delta ln(POP_{i,t}) + \beta_6\Delta ln(CORRU_{i,t})' + \phi CRISIS + \alpha_i^* + \epsilon_{i,t}$	0.067
$\Delta ln(TAX_{i,t}) = \gamma\Delta ln(TAX_{i,t-1}) + \beta_1\Delta ln(AGR_{i,t})' + \beta_2\Delta ln(TRADE_{i,t}) + \beta_3\Delta ln(GDPPC_{i,t}) + \beta_4\Delta ln(CONS_{i,t}) + \beta_5\Delta ln(POP_{i,t}) + \beta_6\Delta ln(CORRU_{i,t-1})' + \phi CRISIS + \alpha_i^* + \epsilon_{i,t}$	0.068
$\Delta ln(TAX_{i,t}) = \gamma\Delta ln(TAX_{i,t-1}) + \beta_1\Delta ln(AGR_{i,t-1})' + \beta_2\Delta ln(TRADE_{i,t}) + \beta_3\Delta ln(GDPPC_{i,t}) + \beta_4\Delta ln(CONS_{i,t}) + \beta_5\Delta ln(POP_{i,t}) + \beta_6\Delta ln(CORRU_{i,t-1})' + \phi CRISIS + \alpha_i^* + \epsilon_{i,t}$	0.091

Table 23: Estimated models and their Sargan's test p value

The $'$ near a variable indicates that it was insignificant at $\alpha = 0.95$.