VILNIAUS UNIVERISTETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

Magistro darbas

Mokesčių surinkimo analizė naudojant funkcinius duomenis

Analysis of a Tax Collection Using Functional Data

Jovita Gudan

VILNIUS 2017

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

EKONOMETRINĖS ANALIZĖS KATEDRA

Darbo vadovas ____

Darbo recenzentas ____

Darbas apgintas____

Darbas įvertintas ____

Registravimo Nr. ____

Gavimo data ____

Contents

Mokesčių surinkimo analizė naudojant funkcinius duomenis

Santrauka

Šio tyrimo tikslas yra ištirti mokesčių surinkimo duomenis naudojant funkcinę duomenų analizę. Šiame darbe nagrinėjami du praktiniai taikymai su mėnesiniais Lietuvos savivaldybių biudžeto pajamų ir dieninių mokestinių įplaukų duomenimis. Pirmosios dalies tikslas yra išanalizuoti prognozių elgesį naudojant iš viršaus į apačią, iš apačios į viršų ir pristatyto iš vidurio į viršų metodus, siekiant nustatyti, kokiomis sąlygomis ir kuris iš šių metodų yra pranašesnis už kitą mažesnės prognozavimo paklaidos atžvilgiu. Tuo tarpu antrosios dalies tikslas yra išanalizuoti korektiškumą, naudojant funkcinį duomenų modelį mokestinių įplaukų duomenims. Remiantis rezultatais, iš viršaus į apačią metodas rodos yra pranašesnis mėnesinių savivaldybių biudžeto pajamų prognozių atžvilgiu ir funkcinis tiesinis modelis yra tinkama priemonė modeliuoti mokestinių įplaukų duomenis.

Raktiniai žodžiai: funkcinė duomenų analizė, duomenų glodinimas, funkcinė pagrindinių komponenčių analizė, funkciniai tiesiniai modeliai

Analysis of a Tax Collection Using Functional Data

Abstract

The purpose of the study is to investigate tax collection data using functional data analysis (FDA) technique. This work consists of two applications with monthly municipal budget revenue of Lithuania and daily tax receipt data. The objective of the first application is to analyze the behavior of various forecasts under the Top-Down, Bottom-Up and introduced Middle-Up approaches in order to identify under which conditions one approach would be preferred instead of the other in terms of lower forecasting errors. Herewith the objective of the second application is to analyze the correctness of applying functional regression model for daily tax receipt data. The results show that Top-Down approach seems to be superior for monthly municipal budget revenue forecasts and functional linear model is appropriate tool for modeling tax receipt data.

Key words: functional data analysis, data smoothing, functional principal component analysis, functional linear models

Introduction

The objective of the Law on the Budget Structure (see [15]) is to ensure efficient use of monetary resources in the process of formation and implementation of the budget with a view to attaining the long-term, overall economic and social welfare for citizens of the Republic of Lithuania, sustainable long-term economic growth, employment without posing threats to the stability of prices. Lithuania has separate budgets for the state (central government), the municipalities (local government) and the social insurance funds. The state budget, as approved by parliament, covers the revenue and expenditures of the government ministries and other budgetary institutions, including state transfers to the municipalities. The municipal budgets are approved by the municipal councils, and cover municipal revenues and funds transferred from the state for delegated functions.

In this work, the effect of tax collection will be investigated by applying functional data analysis (FDA) methods. While classical statistics deals with the analysis of random scalars, vectors, and matrices, functional data analysis refers to the statistical analysis of random functions. Key aspects of FDA include the choice of smoothing technique, data reduction, adjustment for clustering, functional linear modeling and forecasting methods. A monograph on the functional data analysis by Ramsay and Silverman (2005) summarizes the typical models considered in the FDA and most of the popular FDA techniques.

The analysis of tax collection consists of two parts: applications of municipal monthly budget revenue and daily tax receipts data. On a daily basis only total gross tax receipts are available. Since the amount of budget revenue in each month and the amount of tax receipts in each day is a count that can only take non-negative values let's assume that the counts are assumed to arise from a nonhomogenous Poisson process. The key feature of the model is to use regression splines to model the distribution of the amount of budget revenues and tax receipts over time.

The objective of the first application is to analyze the behavior of various forecasts under the Top-Down, Bottom-Up and introduced Middle-Up approaches in order to identify under which conditions one approach would be preferred instead of the other in terms of lower forecasting errors. Herewith the objective of the second application is to analyze the correctness of applying functional regression model for daily tax receipt data.

The rest of the work is organized as follows. In Chapter 1, the nonhomogeneous Poisson process is introduced. In Chapter 2, the statistical methods are discussed, including data smoothing, functional ANOVA test, distance based clustering, functional principle component analysis and functional linear models. In Chapter 3, the variations of municipal budget revenue and tax receipt data are explored using functional methods described in Chapter 2. Conclusions are given in Chapter 4.

1 The Nonhomogeneous Poisson Process

The Poisson process is a stochastic process that counts the number of randomly occurring events over time. A nonhomogeneous Poisson process (NHPP) is a modification of a Poisson process which allows the intensity of the process to be time dependent. Let $N[t, t+h)$ be the number of events occurring in the time interval $[t, t+h)$, where $t \geqslant 0$ and $h \geqslant 0$ and let $N(t) = N[0, t)]$. Then a process $\{N(t), t \geqslant 0\}$ is said to be a nonhomogeneous Poisson process with intensity $\lambda(t)$ if

1) $N(0) = 0$

2) $\{N(t), t \geqslant 0\}$ has independent increments

3) $P\{N(t, t+h) = 0\} = 1 - \lambda(t)h + o(h)$

4) $P\{N(t, t+h) = 1\} = \lambda(t)h + o(h)$

where $\lambda(t) \geqslant 0$ is called the intensity function of the process and $o(h)$ denotes a remainder quantity $g(h)$ which approaches zero faster than h, i.e. $\lim\limits_{0} \frac{g(h)}{h} = 0$. The intensity function $\lambda(t)$ is continuous and

$$\Lambda(t) = \int\limits_0^t \lambda(u)du \tag{1}$$

is called the cumulative intensity function. Conditions 3) and 4) imply that $P\{N[t, t+h) \geqslant 2\} = o(h)$ and that the occurrence of events prior to time t does not affect those in $[t, t+h)$. This gives rise to the property that counts in non-overlapping intervals are independent of one another or more succinctly that the process is memoryless. It can also be shown (Ross, 1996) that 1)-4) imply that $N[t, t+h)$ follows a Poisson distribution

$$P\{N[t, t+h) = n\} = \frac{\Big(\int_t^{t+h} \lambda(u)du\Big)^n exp\Big\{-\int_t^{t+h} \lambda(u)du\Big\}}{n!}, n = 0, 1, 2, \ldots \tag{2}$$

Thus $N(t)$ is distributed as a Poisson random variable with mean $\Lambda(t)$.For this reason the cumulative intensity $\Lambda(t)$ is often referred to as the cumulative mean function.

2 Functional Data Analysis

In this chapter we aim to give a concise conceptual overview of the basic steps involved in functional data analysis (FDA). The emphasis will be on the relevant techniques and methods that are applied throughout this work.

In conventional data analysis the data consist of a set of measurements or observations. In the functional data analysis context, observed data are regarded as depicting an underlying function at various locations; hence each curve is treated as a single functional entity. The continuum of a function is often time, but can be any continuous domain.

The assumption in FDA is that the underlying process generating the data is smooth, which means that a function possesses a certain number of derivatives. Although the data are still observed at discrete time points and subject to measurement error, i.e. noise. The underlying process may typically be measured on as few as 20 or up to tens of thousands of discrete points on the continuum. Additionally the process may also be measured repeatedly, either multiple samples of a single process (within subjects), or samples from the process measured in multiple subjects (across subjects). In many data sets a given observation is dependent on adjacent observations, i.e. correlated. This situation violates the independence assumption in traditional multivariate analysis. In FDA we do not assume that adjacent observations are independent.

2.1 Data Smoothing: From Discrete Points to Smoothed Curves

The first step in FDA is to smooth discretely observed data points to obtain a functional datum or objects. Let t be a one-dimensional argument sometimes referred as time. Functions of t are observed over a discrete grid $\{t_1, \ldots, t_J\} \in \mathcal{T}$ at sampling values t_j, which may or may not be equally spaced. Generally, the observed data are filled with observational errors (or noise) that are superimposed on the underlying signal. In the real world, a typical scenario involves N processes beign observed at the same time. Let $\boldsymbol{y}$ be a vector of N functional data $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N]^T$, where each functional data are written as follows

$$y_{ij} = X_i(t_{ij}) + \epsilon_{ij} \tag{3}$$

where y_{ij} is a noisy observation of the stochastic process $X_i(t_{ij})$ and ϵ_{ij} is a random error with zero mean and variance function σ^2 associated with the i^{th} functional datum. Mathematically, $X_i(t)$ is the conditional expectation of y_{ij}, given $t_{ij} = t$. That is,

$$X_i(t) = \mathbb{E}(y_{ij}|t_{ij} = t), i = 1, 2, \ldots, N, j = 1, 2, \ldots, J_i.$$

There are a number of existing smoothing techniques that can be used to smooth the regression function $X_i(t)$ in (3). Different smoothing techniques have different strengths in one aspect or another. For example, smoothing splines may be good for handling sparse data, while local polynomial smoothers may be computationally advantageous for handling dense designs.

Often basis expansions are used and smoothness is imposed by either restricting the basis or by explicitly specifying a roughness penalty. Fourier bases, polynomial spline bases and B-spline bases are popular. Alternatively, free-knot splines and wavelets provide data-adaptive basis systems. Wavelets are especially useful for data with sharp peaks. Splines are well suited to cases where derivatives of functions are required.

2.1.1 Representing functions by basis functions

The sample curves are assumed to be observations of a stochastic process $X_i = \{X_i(t) : t \in \mathcal{T}\}$ whose sample functions belong to the Hilbert space $\mathcal{L}_2$ of square integrable functions with the inner product $\langle X_1, X_2 \rangle = \int_{\mathcal{T}} X_1(t)X_2(t)dt, \forall X_1, X_2 \in \mathcal{L}_2(\mathcal{T})$. In order to create a functional datum, a basis needs to be specified. The chosen basis is a linear combination of functions defining the functional object. A functional observation x_i is defined as follows

$$X_i(t) = \sum_{k=1}^{K} c_{ik}\phi_{ik}(t) = \sum_{k=1}^{K} \boldsymbol{c}_i\boldsymbol{\phi}(t), \forall t \in \mathcal{T}, i = 1, 2, \ldots, N \tag{4}$$

where $\boldsymbol{\phi}(t) = [\phi_1(t), \phi_2(t), \ldots, \phi_K(t)]^T$ and $\boldsymbol{c}_i = [c_{i1}, c_{i2}, \ldots, c_{iK}]^T$.

Equation (4) can be written in matrix notation as

$$X_i(\boldsymbol{t}_i) = \boldsymbol{\Phi}(\boldsymbol{t}_i)\boldsymbol{c}_i, \forall t \in \mathcal{T} \tag{5}$$

where $\boldsymbol{\Phi}(\boldsymbol{t}_i) = \begin{bmatrix} \phi_1(t_{i1}) & \ldots & \phi_K(t_{i1}) \\ \vdots & \ddots & \vdots \\ \phi_1(t_{iJ_i}) & \ldots & \phi_K(t_{iJ_i}) \end{bmatrix}$ is a $J \times K$ matrix of basis functions evaluated at each time point t_{ij} and $J = \max_{i=1,\ldots,N}(J_i)$.

Basis functions expansion represent the potentially infinite-dimensional universe of functions within the finite-dimensional framework of vectors like $\boldsymbol{c}$. A great deal depends on how the vector of basis functions $\phi(t)$ is chosen.

The Fourier Basis System

The most appropriate basis for periodic functions defined on an interval $\mathcal{T}$ is the Fourier Basis where the ϕ_k's take the following form

$$\phi_0(t) = \frac{1}{\sqrt{|\mathcal{T}|}}, \phi_{2r-1}(t) = \frac{\sin(rwt)}{\sqrt{|\mathcal{T}|/2}}, \phi_{2r}(t) = \frac{\cos(rwt)}{\sqrt{|\mathcal{T}|/2}} \tag{6}$$

for $r = 1, \ldots, \frac{K-1}{2}$, where K is the number of basis functions; notice that the K must be an odd number to compute Fourier Basis. The frequency w determines the period and the length of the interval $|\mathcal{T}| = 2\pi/w$. The function vector $\phi(t)$ has the form $\phi(t) = [\phi_0(t), \phi_1(t), \phi_2(t), \ldots, \phi_{2r}(t)]^T$ evaluated at discrete time points $t_j, j = 1, \ldots, J_i$.

If the values of t_j are equally spaced on τ and the period is equal to the length of interval τ, then the basis is orthogonal in the sense that the cross product matrix $\mathbf{\Phi}'\mathbf{\Phi}$ is diagonal, and can be made equal to the identity by dividing the basis functions by suitable constants, $\sqrt{n}$ for j $= 0$ and $\sqrt{2/n}$ for all other j.

The Fast Fourier transform (FFT) makes it possible to find all the coefficients extremely efficiently when n is a power of 2 and the arguments are equally spaced, and in this case we can find both the coefficients c_k and all n smooth values at $X(t_j)$ in $O(\log n)$ operations.

Derivative estimation in a Fourier basis is simple since

$$D \sin rwt = rw \cos rwt$$

$$D \cos rwt = -rw \sin rwt$$

This implies that the Fourier expansion of DX has coefficients

$$(0, c_1, -wc_2, 2wc_3, -2wc_4, \ldots) \tag{7}$$

and of D^2X has coefficients

$$(0, -w^2c_1, -w^2c_2, -4w^2c_3, -4w^2c_4, \ldots) \tag{8}$$

Similarly, we can find the Fourier expansions of higher derivatives by multiplying individual coefficients by suitable powers of rw, with sign changes and interchange of sine and cosine coefficients as appropriate.

The B-Spline Basis System

It is well-known that polynomials are not flexible in their ability to model data over a large range of the design time points. However, this is not the case when the range is small enough or when the big range is divided into some small subintervals or local neighborhoods. In regression spline smoothing, however, the local neighborhoods are specified by a group of locations, say,

$$\tau_0, \tau_1, \ldots, \tau_L, \tau_{L+1}, \tag{9}$$

in the range of interest, say, an interval $[a, b]$ where $a = \tau_0 < \tau_1 < \ldots < \tau_L < \tau_{L+1} = b$.

These locations are known as knots, and $\tau_l, l = 1, 2, \ldots, L$ are called interior knots or simply knots. These knots divide the interval of interest, $[a, b]$, into L subintervals: $[\tau_l, \tau_{l+1}), l = 0, 1, \ldots, L$, so that within any two neighboring knots, a Taylor expansion up to some degree is valid. Mathematically, a regression spline is defined as a piecewise polynomial that is a polynomial of some degree within any two neighboring knots τ_l and τ_{l+1} for $l = 0, 1, \ldots, L$ and is joined together at knots properly but allows discontinuous derivatives at the knots.

Let $B_{k,m}(t)$ denote the k^{th} B-Splines Basis function of order m defined for any value of t, for the non-decreasing sequence of knots $\{\tau_l\}_{l=0}^{L}$.

In this case, $\phi_k(t)$ is defined as follows:

$$\phi_k(t) = B_{k,m}(t), \forall t \in \mathcal{T}, k = 1, \ldots, m + L - 2 \tag{10}$$

Let $\xi_0 < \xi_1$ and $\xi_K < \xi_{K+1}$ be two boundary knots defining the domain over which the spline is evaluated. The augmented knot sequence τ is defined as

- $\tau_1 \leqslant \tau_1 \leqslant \ldots \leqslant \tau_M \leqslant \xi_0$;
- $\tau_{j+M} = \xi_j, j = 1, 2, \ldots, K$;
- $\xi_{K+1} \leqslant \tau_{K+M+1} \leqslant \tau_{K+M+2} \leqslant \ldots \leqslant \tau_{K+2M}$.

Any additional knots beyond the boundary are abitrary, and the usual scenario is to make them all the same an equal to ξ_0 and ξ_{K+1}. The set of basis functions $B_{k,m}(t)$ of order m for the knot-sequence τ (where m < M) is derived using a recursion formula as follows

$$B_{1,k}(t) = \begin{cases} 1, & t \in [\tau_l, \tau_{l+1}] \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

for $k = 1, \ldots, K + 2M - 1$. These functions are called Haar basis functions.

$$B_{k,m}(t) = \frac{t - \tau_l}{\tau_{k+m-1}} B_{k,m-1}(t) + \frac{\tau_{k+m} - t}{\tau_{k+m} + \tau_{k+1}} B_{k+1,m-1}(t), \forall t \in \mathcal{T}, m \geqslant 2 \tag{12}$$

for $k = 1, \ldots, K + 2M - m$. In this case, the function vector $\phi(t)$ defined in equation (10) has $K + 2M - m$ basis functions evaluated at discrete time points t_i, where $i = 1, \ldots, n$. In other words, the number of basis functions is defined by its order and its number of knots.

2.1.2 Smoothing functional data by least squares

This section describes an approach for model estimation when using basis functions, namely the Least Squares method (with and without penalty).

The regression can be re-expressed as $X_i(t) = \boldsymbol{\phi}(t)\boldsymbol{c}_i$, so that the model (3) can be approximately expressed as

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{c} + \boldsymbol{\epsilon} \tag{13}$$

where $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N)^T$, $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1(t), \boldsymbol{\phi}_2(t), \ldots, \boldsymbol{\phi}_K(t))^T$, and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \ldots, \boldsymbol{\epsilon}_N)^T$. As $\boldsymbol{\phi}(t)$ is a basis vector, $\boldsymbol{\Phi}$ is of full rank, and hence $\boldsymbol{\Phi}^T\boldsymbol{\Phi}$ is invertible whenever N > K. A natural estimator of $\boldsymbol{c}_i$, which solves the approximation linear model (13) by the ordinary least squares method, is

$$\hat{\boldsymbol{c}}_i = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{y}_i. \tag{14}$$

It follows that the regression fit of the function $x(t)$ in (3) is

$$\hat{X}_i(t) = \boldsymbol{\phi}(t)^T(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{y}, \tag{15}$$

which is often called a regression spline smoother of $x(t)$ and obviously $a(t) = \boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\phi}(t)$. In particular, the values of $\hat{X}_i(t)$ evaluated at the design time points t_{ij}, $i = 1, 2, \ldots, N, j = 1, 2, \ldots, J_i$ are collected in the following fitted response vector

$$\hat{\boldsymbol{y}}_i = \boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{y}_i = \boldsymbol{A}\boldsymbol{y}_i, \tag{16}$$

where $\hat{\boldsymbol{y}}_i = (\hat{y}_{i1}, \ldots, \hat{y}_{iJ_i})^T$ with $\hat{y}_{ij} = \hat{x}(t_{ij})$, $i = 1, 2, \ldots, N, j = 1, 2, \ldots, J_i$, and $\boldsymbol{A} = \boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T$ is called the regression smoother matrix.

Remark: The regression spline smoother matrix $\boldsymbol{A}$ is an idempotent matrix, satisfying $\boldsymbol{A}^T = \boldsymbol{A}$,$\boldsymbol{A}^2 = \boldsymbol{A}$ and $tr(\boldsymbol{A}) = K$. The trace of the smoother matrix $\boldsymbol{A}$ is often called the degrees of freedom of the regression smoother. It measures the complexity of the fitted regression model.

For the regression smoother, when the knot locating method is specified, the remaining task is to choose the number of knots, K. In general, K is smaller than the number of the measurements. As K must be an integer, choices for K are limited. Alternatively, we can use all the distinct design time points as knots. This may result in undersmoothing when there are too many distinct design time points. The resulting fit is usually quite rough. To overcome this problem, we then introduce a penalty to control the roughness of the fitted curve.

To be specific, without loss of generality, let us again assume that the range of interest of $x(t)$ in (3) is a finite interval, say, $[a, b]$ for some finite numbers a and b. The roughness of $x(t)$ is usually defined as the integral of its squared d-times derivative

$$\int_{\mathcal{T}} \left\{X_i^{(d)}(u)\right\}^2 du \tag{17}$$

for some $d \geqslant 1$. This quantity is large when the function $X_i(\cdot)$ is rough. The smoothing spline smoother of the function $X_i(t)$ in (3) is defined as the minimizer $\hat{X}_{i\lambda}(t)$ of the following penalized least squares (PLS) criterion:

$$\sum_{i=1}^{N}\sum_{j=1}^{J_i} \Big[y_{ij} - X_i(t_{ij})\Big]^2 + \lambda \int_{\mathcal{T}} \left\{X_i^{(d)}(t)\right\}^2 dt \tag{18}$$

over the following d^{th} order Sobolev space $W_2^k[\mathcal{T}]$:

$$\left\{X : X(s), \ldots, X^{(d)}(s) \text{ abs. continu.}, \int_{\mathcal{T}} \left\{X^{(d)}(t)\right\}^2 dt < \infty\right\} \tag{19}$$

where $\lambda > 0$ is a smoothing parameter controlling the size of the roughness penalty, and it is usually used to trade off the goodness of fit, represented by the first term in (18), and the roughness of the resulting curve. The $\hat{X}_{i\lambda}(t)$ is known as a natural smoothing of degree $(2d - 1)$.

The roughness (17) of $X_i(t)$ can be expressed in matrix term as

$$\int_{\mathcal{T}} \left\{X_i^{(d)}(u)\right\}^2 du = \int_{\mathcal{T}} \left\{\boldsymbol{c}_i^T \boldsymbol{\phi}^d(u)(\boldsymbol{\phi}^d(u))^T \boldsymbol{c}_i\right\} du = \boldsymbol{c}_i^T \int_{\mathcal{T}} \left\{\boldsymbol{\phi}^d(u)(\boldsymbol{\phi}^d(u))^T du\right\} \boldsymbol{c}_i = \boldsymbol{c}_i^T \boldsymbol{R} \boldsymbol{c}_i, \quad (20)$$

where the matrix $\boldsymbol{R} : K \times K$ is known as a roughness matrix. It follows that the PLS criterion (18) can be written as

$$||\boldsymbol{y}_i - \boldsymbol{\Phi}\boldsymbol{c}_i||^2 + \lambda \boldsymbol{c}_i^T \boldsymbol{R} \boldsymbol{c}_i, \quad (21)$$

where $||\cdot||$ denotes the usual $\mathcal{L}_2$-norm. Therefore, the estimate $\hat{\boldsymbol{c}}_{i\lambda}$ that minimizes (21) is

$$\hat{\boldsymbol{c}}_{i\lambda} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda \boldsymbol{R})^{-1}\boldsymbol{\Phi}^T\boldsymbol{y}_i \quad (22)$$

The estimated smooth function is then

$$\hat{X}_i(t) = \boldsymbol{\phi}(t)^T \hat{\boldsymbol{c}}_i. \quad (23)$$

2.2 Descriptive analysis of Functional Data

One of the most important parts in data analysis is the exploratory part: estimating means and standard deviations. Because the functional nature of the data, the associated descriptive statistics are therefore functional.

2.2.1 Mean and variance functions

The mean function is defined as $\mu(t) = \mathbb{E}(X(t)), \forall t \in \mathcal{T}$. And the sample functional mean is

$$\bar{X}(t) = N^{-1}\sum_{i=1}^{N} X_i(t), \forall t \in \mathcal{T} \quad (24)$$

where N is the number of curves or replications and $X_i(t)$ is the i^{th} function evaluated at time t.

Likewise, the estimation of the functional variance is defined as $\sigma^2 = \mathbb{E}[X(t) - \mu(t)]^2, \forall t \in \mathcal{T}$. And the sample functional variance is

$$Var(X(t)) = (N-1)^{-1}\sum_{i=1}^{N}(X_i(t) - \bar{X}(t))^2 \quad (25)$$

and the standard deviation function is the square root of the variance function.

2.2.2 Covariance and Correlation functions

The covariance function summarizes the dependence of records across different argument values. We define v to be the covariance function

$$v(t_1, t_2) = \mathbb{E}[(X(t_1) - \mu(t_1))(X(t_2) - \mu(t_2))], \forall t_1, t_2 \in \mathcal{T}, \tag{26}$$

and $\hat{v}$ to be the sample covariance function

$$\hat{v}(t_1, t_2) = \frac{1}{N}\sum_{i=1}^{N}[(X_i(t_1) - \bar{X}(t_1))(X_i(t_2) - \bar{X}(t_2))], \forall t_1, t_2 \in \mathcal{T}. \tag{27}$$

The associated correlation function is

$$corr(t_1, t_2) = \frac{\hat{v}(t_1, t_2)}{\sqrt{Var(t_1)Var(t_2)}} \tag{28}$$

2.2.3 Functional Depth

Although, most often are other measures used to summarize a set of functional data such as depth measures. The depth is a concept emerged in the literature of robustness which measures how deep (or central) is a datum respect to a population. In univariate data, the median would be the deepest point of clouds of points.

- Fraiman-Muniz Depth (Fraiman and Muniz (2001)): The depth measure is based on the median. For every $t \in [0, 1]$, let $F_{n,y}$ be the empirical distribution of the sample $X_1(t), ..., X_N(t)$ and let $Z_i(t)$ denote the univariate depth of the data $X_i(t)$ in this sample, given by $D_i(t) = 1 - |1/2 - F_{n,t}(X_i(t))|$. Then, define for $i = 1, \ldots, n$,

$$I_i = \int_0^1 D_i(t)dt \tag{29}$$

  and rank the observations $X_i(t)$ according to the values of I_i.

- Modal Depth (Cuevas et al. (2007)): The depth measure is based on how surrounded the curves are respect to a metric or a semi-metric distance, selecting the trajectory most densely surrounded by other trajectories of the process. The population h-depth of a datum z is given by the function

$$f_h(z) = \mathbb{E}(K_h(||z - X||)) \tag{30}$$

where X is the random element describing the population, $||\cdot||$ is a suitable norm and $K_h(t)$ is a re-scaled kernel and tuning parameter h. An given a random sample $X_1, ..., X_N$ of X, the empirical h-depth is defined as

$$\hat{f}_h(z) = N^{-1}\sum_{i=1}^{N}(K_h(||z - X||)) \tag{31}$$

- Random Projection Depth (Cuevas et al. (2007)): The depth measure is calculated through random projections (RP) based on the Tukey depth. Given a sample $X_1, ..., X_N$ let us take a random direction a (independent from the X_i) and project the data along this direction. Then, the sample depth of a datum X_i is defined as the univariate depth of the corresponding one-dimensional projection (expressed in terms of order statistics so that the median is the deepest point). When the sample is made of functional data, we will assume throughout that the X_i belong to the Hilbert space $\mathcal{L}_2[0,1]$ so that the projection of a datum X is given by the standard inner product $\langle a, X\rangle = \int_a^b a(t)X(t)dt$. In the finite-dimensional case the projection of $X = (\xi_1, \ldots, \xi_d)$ along the direction a is evaluated through the usual Euclidean inner product $a_1\xi_1 + \ldots + a_d\xi_d$, denoted also by $\langle a, X\rangle$.

- Random Projection Depth with derivatives (Cuevas et al. (2007)): The depth measure is calculated through random projections of the curves and theirs derivatives.

  Let $X_1, ..., X_N$ be a sample of differentiable functions defined on $[0,1]$. The basic idea is to use the method of random projections simultaneously (for the functions and their derivatives) thus incorporating the information on the function smoothness provided. The sample of functions is reduced to a sample in $\mathbb{R}^2$ defined by $\langle a, X_1\rangle, \langle a, X_1'\rangle, \ldots, \langle a, X_N\rangle, \langle a, X_N'\rangle$, where a is a randomly chosen direction.Then the depth of the bidimensional sample data is evaluate in a second step using the depth function defined above: Fraiman-Muniz Depth, Modal Depth or Random Projection Depth.

2.3 Functional ANOVA test

Let $X_{i1}(t), X_{i2}(y), \ldots, X_{in_i}(t), i = 1, \ldots, k$, denote k groups of random functions defined over a given finite interval $T = [a, b]$. Let $SP(\mu, \gamma)$ denote a stochastic process with mean function $m(t), t \in T$ and covariance function $\gamma(s,t), s, t \in T$. Assuming that $X_{i1}(t), X_{i2}(y), \ldots, X_{in_i}(t)$

are i.i.d. $SP(\mu_i, \gamma), i = 1, \ldots, k$ it is often interesting to test the equality of the k mean functions

$$H_0 : m_1(t) = \ldots = m_k(t). \tag{32}$$

against the alternative that its negation holds. This problem is known as the k-sample testing problem or the one-way ANOVA problem for functional data.

Cuevas et al. (2004) proposed to use the following test statistic for testing (32)

$$V_n = \sum_{1 \leqslant i \leqslant j \leqslant k} n_i \int_T (\bar{X}_i(t) - \bar{X}_j(t))^2 dt \tag{33}$$

Under the null hypothesis (32) and the assumptions that $n_i, n \to \infty$ in such a way that $n_i/n \to p_i > 0$ for $i = 1, \ldots, k$, they proved that the approximate distribution of V_n is that of the statistic

$$V = \sum_{1 \leqslant i \leqslant j \leqslant k} n_i \int_T (Z_i(t) - \sqrt{p_i/p_j} Z_j(t))^2 dt \tag{34}$$

where $Z_1(t), \ldots, Z_k(t)$) are independent Gaussian processes with mean 0 and covariance function $\gamma(s,t)$. . Cuevas et al. (2004) computed the p-value of V_n, or its empirical critical value, by resampling $Z_i(t), i = 1, \ldots, k,$, from Gaussian processes $GP(0, \hat{\gamma})$, where $\hat{\gamma}(s,t) = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij}(s) - \bar{X}_i(s))(X_{ij}(t) - \bar{X}_i(t))$, a large number of times.

Provided that the n_i are large enough, hypothesis H_0 is rejected, at a level α, whenever $V_n > V_\alpha$ where $P_{H_0}\{V > V_\alpha\} = \alpha$.

2.4 Distance based clustering

The goal of cluster analysis is to group a collection of subjects into clusters, such that those falling into the same cluster are more similar to each other than those in different clusters. Therefore, a measure of similarity or dissimilarity between subjects is a necessary ingredient for clustering. A metric defined on the subject space is one way to obtain dissimilarities, simply using the distance between two subjects as a measure of dissimilarity.

This section collects several metric and semi-metric functions which allow to extract as much information possible from the functional variable.

The most simple spaces for functional data are the complete metric spaces where only the notion of distance between elements of the space is given. If the metric can be expressed as $d(X(t), Y(t)) = ||X(t) - Y(t)||$ with a norm $|| \cdot ||$ verifying the triangle inequality, we have a normed space (or a Banach space). In these spaces, there is also the notion of size of the elements of the space. If the norm verifies the paralelogram law ($||x+y||^2 + ||x-y||^2 = 2(||x||^2 + ||y||^2)$), an inner product can be defined in the space in the following way: $\langle x, y\rangle = \frac{1}{4}(||x+y||^2 + ||x-y||^2)$. A complete space with an inner product is called a Hilbert space which is a special kind of Banach spaces where $||X(t)|| = \sqrt{\langle X(t), Y(t)\rangle}$.

A collection of semi-metrics is described below:

- If we focused on $\mathcal{L}_p$ spaces (the set of functions whose absolute value raised to the p-th power has finite integral), Simpson's rule is used to compute distances between elements. Let $f(t) = X_1(t) - X_2(t)$.

  $$||f||^p = \Big(\frac{1}{\int_{\mathcal{T}} w(t)dt} \int_{\mathcal{T}} |f(t)|^p w(t)dt\Big)^{\frac{1}{p}} \tag{35}$$

  where w are the weight. The observed points on each curve t are equally spaced or not.

- Computes semi-metric distances of functional data based on Ferraty and Vieu (2006). The semi-metric given two curves X_1 and X_2 is:

  $$d(X_1, X_2) = \sqrt{\int_{\mathcal{T}} (X_1^{(q)}(t) - X_2^{(q)}(t))^2 dt} \tag{36}$$

  where $X_1^{(q)}(t)$ denote the q-th derivative of X.

  Using the equation (4) the derivatives of the approximated curves by B-spline are $\hat{X}_i^{(q)} = \sum_{k=1}^{K} \hat{c}_{ik} B_{m,k}^{(q)}(t)$ and the semi-metric can written now as

  $$d(X_1, X_2) = \sqrt{\int_{\mathcal{T}} (\hat{X}_1^{(q)}(t) - \hat{X}_2^{(q)}(t))^2 dt} \tag{37}$$

  where the integral is computing by "Gauss method" following next approximation

  $$\int_a^b f(t)dt \approx \frac{b-a}{2} \sum_{k=1}^{K} w_k f(\frac{b-a}{2} + \frac{b-a}{2}\delta_k) \tag{38}$$

  where w_k are the weights and δ_k are the gauss points (see Ferraty and Vieu (2006), pages 32-33).

- This metric computes proximities between curves based on the functional principal components analysis (FPCA) method. The FPCA reduces the functional data in a reduced dimensional space (q functional principal components).

$$d_q^{FPCA} = \sqrt{\sum_{k=1}^{q} \Big(\int [X_i(t) - f_i(t)]\xi_k(t)dt \Big)^2} \tag{39}$$

  where f_i is the score of the principal component $\int X_i \xi$.

  The FPCA semi-metric for two curves X_1 and X_2 is calculated as

$$d_q^{FPCA}(X_1, X_2) = \sqrt{\sum_{k=1}^{q} \Big(\sum_{j=1}^{J} w_j(f_1(t_j) - f_2(t_j))[\xi_k]_j \Big)^2} \tag{40}$$

  where ξ_i is the i^{th} orthonormal eigenvector of the covariance matrix $W = diag(w_1, \ldots, w_J)$ with quadrature weights, in this case we use $w_j = t_j - t_{j-1}$.

- This metric computes proximities between curves a PLS semi-metric based on the functional partial least-squared (FPLS) method. The FPLS uses a scalar response observed additionally to reduce the functional data in a q functional PLS components.

  We consider n scalar responses, the PLS semi-metric with q factors is defined as

$$d_q^{PLS}(X_1, X_2) = \sqrt{\sum_{k=1}^{p} \Big(\sum_{j=2}^{J} w_j(f_1(t_j) - f_2(t_j))[\xi_k^q]_j \Big)^2} \tag{41}$$

  with quadrature weights $w_j = t_j - t_{j-1}$.

  PCA and PLS semi-metrics can be used only if the curves are observed at the same discretized points and in a grid sufficiently fine.

2.5 Registration

The functional data often comes with lateral displacements in curves. There are two sources of variability present in smoothed curves that form the functional data. Amplitude variation is displayed in the different size of features between curves. Phase variation can be seen in the difference in the timing, or location on the continuum, of specific features between curves. Phase variation is often referred to as misalignment of curves. The aim of registration in functional data is to separate amplitude and phase variation by aligning curves. Functional registration is also called warping, time warping or alignment. The most well known registration methods in FDA are landmark registration and continuous registration. Landmark registration removes phase variation by monotonically transforming the domain for each curve so that points specifying the locations of shape features are aligned across curves. Continuous registration uses a measure of closeness to quantify the similarity between curves. The method aligns curves by warping their time (or horizontal axis) parameters by selecting the optimal warping function from a class of warping functions in order to maximise the similarity between curves. Note that the functional registration is always performed on curves and not on data points. It is essential to perform registration of the smooth functions before further analyses, since misalignment can have a serious effect on results. For example, inflates data variance, blurs underlying data structures, and distorts principal components.

The landmark registration will be used in this thesis and can be expressed formally as follows. The data that we consider are a sample of N smooth random functions $X_1, \ldots, X_N$ defined over a closed interval on the real line. Time warping or curve registration aims at eliminating the phase variation in a functional sample. It achieves this goal by applying transformations, the warping functions h_i, to the function arguments. The deformation functions $h_i(x), i = 1, \ldots, m$, called warping functions, must check the following properties:

- Initial conditions: $h_i(0) = 0, h_i(X) = X$.
- Landmarks alignment: $h_i(t_{0,j}) = t_{i,j}$.
- Strict monotonicity: $h_i(t_1) > h_i(t_2)$ for $t_1 > t_2$. That is, the function h_i is invertible so that for the same event the time points on two different timescales correspond to each other uniquely.

The registration problem is defined as the search for a set of smooth strictly monotonic functions h_i, called its warping function, such that the functions of the form

$$y(t) = X_i\{h_i(t)\} + \epsilon_i(t) \tag{42}$$

or

$$y = X_i \circ h_i + \epsilon_i, \tag{43}$$

where ϵ is small relative to X_i and roughly centred on 0. If, alternatively, the template y is defined by discrete values $y_j, j = 1, \ldots, n$, then the model becomes

$$y_j = X_i\{h_i(t_j)\} + \epsilon_{ij}. \tag{44}$$

The method of landmarks alignment which consists to determine, for each curve, a deformation function so that specific points called landmarks of the registered curves are aligned. Specific points defined as landmarks are generally the positions of maxima, minima, inflection points, or zero crossings. Then, the landmarks registration of m signals $X_1, \ldots, X_m$ defined on the same interval $[0, T]$ can be divided into the five following steps:

1. Definition of characteristic points to be used as landmarks (e.g., minimum, maximum, zero crossing, etc.).

2. Extraction of landmarks $t_{i,1}, \ldots, t_{i,K}$ from an observed sequence of each signal $X_i, i = 1, \ldots, m$. Note that since observed signals are noisy, the landmarks $t_{i,1}, \ldots, t_{i,K}$ are usually extracted from a estimator $\hat{X}_i$ of the signal X_i.

3. Identify landmarks reference $t_{0,1}, ..., t_{0,K}$, i.e. the points at which the curves must match.

4. Determine deformation functions $h_1, \ldots, h_m$ so that corresponding landmarks are matched, i.e. for all $i = 1, \ldots, m$, $h_i(t_{0,j}) = t_{i,j}, j = 1, \ldots, K$.

5. Deformation of the signals using transformations obtained in the previous step. The registered functions $\hat{X}_i(t) = X_i[h_i^{-1}(t)], i = 1, \ldots, m$, are the aligned at each points $t_{0,1}, \ldots, t_{0,K}$.

Comparing before and after registration

Kneip and Ramsay (2008) developed a useful way of quantifying the amount of these two types of variation by comparing results for a sample of N functional observations before and after registration. The notation X_i stands for the unregistered version of the i^{th} observation, y_i for its registered counterpart and hi for associated warping function. The sample means of the unregistered and registered samples are $\bar{x}$ and $\bar{y}$, respectively.

The total mean square error is defined as

$$MSE_{total} = N^{-1} \sum_{N}^{i=1} \int [x_i(t) - \bar{x}(t)]^2. \tag{45}$$

We define the constant C_R as

$$C_R = 1 + \frac{N^{-1} \sum_i^N \int [Dh_i(t) - N^{-1} \sum_i^N Dh_i(t)][y_i^2(t) - N^{-1} \sum_i^N y_i^2(t)]dt}{N^{-1} \sum_i^N \int y_i^2(t)dt} \tag{46}$$

The structure of C_R indicates that $C_R - 1$ is related to the covariation between the deformation functions Dh_i and the squared registered functions y_i^2. When these two sets of functions are independent, the number of the ratio is 0 and $C_R = 1$: The measures of amplitude and phase mean square error are, respectively,

$$MSE_{amp} = C_R N^{-1} \sum_{i}^{N} \int [y_i(t) - \bar{y}(t)]^2 dt$$
$$MSE_{phase} = C_R \int \bar{y}^2(t)dt - \int \bar{x}^2(t)dt \tag{47}$$

It can be shown that, defined in this way, $MSE_{total} = MSE_{amp} + MSE_{phase}$.

If we have registered our functions well, then the registered functions y_i will have higher and sharper peaks and valleys, since the main effect of mixing phase variation with amplitude variation is to smear variation over a wider range of t values. Consequently, the first term in MSE_{phase} will exceed the second and is a measure of how much phase variation has been removed from the y_i's by registration. On the other hand, MSE_{amp} is now a measure of pure amplitude variation to the extent that the registration has been successful. The decomposition does depend on the success of the registration step, however, since it is possible in principle for MSE_{phase} to be negative. From this decomposition we can get a useful squared multiple correlation index of the proportion of the total variation due to phase $R^2 = \frac{MSE_{amp}}{MSE_{phase}}$.

2.6 Functional Principal Component Analysis

Functional principal component analysis (FPCA) is being used to study the variations in the data. Before doing FPCA, the mean curve is usually subtracted. Let $\bar{y}(t) = \frac{1}{n}\sum_{i=1}^{N} y_i(t)$ and $r_i(t) = y_i(t) - \bar{y}(t)$. FPCA will be conduced on $r_i(t)$.

Let $\xi_1(t)$ be the first functional principal component (FPC). It is estimated by maximizing

$$\sum_{i=1}^{N} f_{i1}^2, \tag{48}$$

subject to $||\xi_1||^2 = \int \xi_1^2 ds = 1$, where $f_{i1} = \int \xi_1(s) r_i(s) ds$ is the first PC score of the i^{th} curve $r_i(t)$. Similarly, the second FPC $\xi_2(t)$ is estimated by maximizing $\sum_{i=1}^{N} f_{i2}^2$, subject to $||\xi_2||^2 = \int \xi_2^2 ds = 1$ and $\int \xi_1(s)\xi_2(s) = 0$, where $f_{i2} = \int \xi_2(s) r_i(s) ds$ is the second FPC score of the i^{th} curve $r_i(t)$. The subsequent FPCs, $\xi_3(t), \ldots, \xi_M(t)$, can be estimated similarly with additional constraints $\int \xi_u(s)\xi_v(s) = 0$ for all u, v where $1 \leqslant u \leqslant v \leqslant M$. Let

$$v(s,t) = \sum_{i=1}^{N} r_i(s) r_i(t) \tag{49}$$

be the variance-covariance function for $r_i(t)$. All FPCs can be calculated as the eigenfunctions of the following functional eigenequations

$$\int v(s,t)\xi_m(s)ds = \rho_m \xi_m(t), \tag{50}$$

where ρ_m is the corresponding eigenvalue, $q = 1, \dots, M$ and $\rho_1 \geqslant \dots \geqslant \rho_M$. Each eigenfunction $xi_m(t)$ takes account $\frac{\rho_m}{\sum\limits_{m=1}^{M} \rho_m}$ of the total variations among N curves.

2.7 Functional regression models

The aim of this section is to introduce functional models with a scalar or functional response from one or more functional covariates. Since functional linear regression modeling has its roots from multivariate multiple regression modelling, the final result of all derivations have the form

$$\boldsymbol{y} = \boldsymbol{Z\beta} + \boldsymbol{\epsilon}. \tag{51}$$

This chapter will review some key concepts related to the functional linear regression model. Like in multivariate analysis, functional linear regression model has appeared to be extremely useful in a broad range of applications including Time Series. A typical functional linear regression model intends to explore the variability of a scalar continuous (functional) response while considering how much of its variation is explainable by other variables. Linear regression models can be functional in one or both of two ways:

- The dependent or response variable is functional;
- One or more of the independent variables or covariates are functional.

Clearly, the functional-response case is an extension of the multivariate-response case with vectors converted into functions. The main change is that the regression coefficients now become regression functions with values $\beta_j(t)$ or $\beta_j(t,s)$ depending on the nature of the problem. Although the main focus of this chapter is on scalar response predicted by one or more functional covariates. It should be noted that all inferential tools for functional linear regression models have been developed under the assumption that the covariate/response pairs are independent.

2.7.1 Scalar response and functional independent variables

Let $\{y_i, i = 1, \ldots, N\}$ be an N-vector of scalar responses and $X_{im}(t), m = 1, \ldots, M$ are M functional predictors. Using the definitions from Chapter 2, the functions $X_{im}(t)$ can be obtained using the smoothing techniques. The regression model that evaluates the relationship between the vector of scalar responses and the functional covariates is given by

$$y_i = \beta_0 + \sum_{m=1}^{M} \int_{\mathcal{T}_m} X_{im}(t)\beta_m(t)dt + \epsilon_i, \forall i, m \tag{52}$$

where β_0 is the usual intercept term that adjusts for the origin, $\beta_m(t)$ are the coefficient functions and ϵ_i are the error terms which are independently and normally distributed with mean 0 and variance σ^2. Using the expansion in (3) to reduce the degrees of freedom in the model further using basis functions, the functional predictors $X_{im}(t)$ are expressed as

$$X_{im}(t) = \sum_{k=1}^{K_m^x} c_{imk}\phi_{mk}(t) = \boldsymbol{c}_{im}^T \boldsymbol{\phi}_m(t), \forall t \in \mathcal{T}_m. \tag{53}$$

In certain cases, $\phi_m(t)$ may differ depending on how different the functional predictors are among $m = 1, \ldots, M$. Furthermore, the coefficient functions are represented by linear combinations of K_m^β basis functions $\left\{\psi_{m1}(t), \ldots, \psi_{mK_m^\beta}(t)\right\}$, with the following form

$$\beta_m(t) = \sum_{k=1}^{K_m^\beta} \boldsymbol{b}_{ml}\boldsymbol{\psi}_{ml}(t), \forall t \in \mathcal{T}_m \tag{54}$$

Replacing equations (53) and (54) in equation (52) yields

$$y_i = \beta_0 + \sum_{m=1}^{M} \int_{\mathcal{T}_m} \boldsymbol{c}_{im}^T \boldsymbol{\phi}_m(t)\boldsymbol{\psi}_m^T(t)\boldsymbol{b}_m dt + \epsilon_i \tag{55}$$

$$= \beta_0 + \sum_{m=1}^{M} \int_{\mathcal{T}_m} \boldsymbol{c}_{im}^T \boldsymbol{J}_{\phi\psi}^{(m)} \boldsymbol{b}_m + \epsilon_i \tag{56}$$

wehre $\boldsymbol{J}_{\phi\psi}^{(m)} = \int_{\mathcal{T}_m} \boldsymbol{\phi}_m(t)\boldsymbol{\psi}_m^T(t)dt$ is the $K_m^x \times K_m^\beta$ cross-product matrix. Taking equation (56) one step further, it can be rewritten as

$$\boldsymbol{y} = \boldsymbol{Z\beta} + \boldsymbol{\epsilon} \tag{57}$$

where

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{z}_1^T \\ \vdots \\ \boldsymbol{z}_N^T \end{bmatrix} = \begin{bmatrix} 1 & \boldsymbol{c}_{11}^T \boldsymbol{J}_{\phi\psi}^{(1)} & \ldots & \boldsymbol{c}_{1M}^T \boldsymbol{J}_{\phi\psi}^{(M)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \boldsymbol{c}_{N1}^T \boldsymbol{J}_{\phi\psi}^{(1)} & \ldots & \boldsymbol{c}_{NM}^T \boldsymbol{J}_{\phi\psi}^{(M)} \end{bmatrix}, \tag{58}$$

$$\boldsymbol{B} = \begin{bmatrix} \beta_0 \\ \boldsymbol{b}_1 \\ \vdots \\ \boldsymbol{b}_M \end{bmatrix}, \tag{59}$$

$\boldsymbol{y}$ is the N-vector of scalar responses, $\boldsymbol{Z}$ is the $N \times (\sum\limits_{m=1}^{M} K_m^x + 1)$ matrix of functional covariatees, $\boldsymbol{B}$ is the $(\sum\limits_{m=1}^{M} K_m^\beta + 1) \times 1$ vector of functional coefficients, and $\boldsymbol{\epsilon}$ is the N-vector error terms.

Functional linear model with basis representation

This section assumes that the relationship between the scalar response Y and the functional covariate $X(t)$ has a linear structure. Thus, the functional linear model under the parametric approach is given by the expression:

$$y_i = \langle X, \beta \rangle + \epsilon_i = \int_{\mathcal{T}} X_i(t)\beta(t)dt + \epsilon_i \tag{60}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product on $\mathcal{L}_2$ and ϵ_i are random errors with mean zero and finite variance σ^2.

Ramsay and Silverman (2005) model the relationship between the scalar response and the functional covariate by basis representation of the observed functional data $X(t)$ and the unknown functional parameter $\beta(t)$. The functional linear model in Equation (60) is estimated by the expression

$$\hat{y}_i = \int_{\mathcal{T}} X_i(t)\beta(t)dt \approx \boldsymbol{C}_i^T \psi(\boldsymbol{t})\phi^T(\boldsymbol{t})\hat{\boldsymbol{b}} = \tilde{\boldsymbol{X}}\hat{\boldsymbol{b}} \tag{61}$$

where $\tilde{\boldsymbol{X}}_i(\boldsymbol{t})$ is the scores such $\tilde{\boldsymbol{X}}_i(\boldsymbol{t}) = \boldsymbol{C}_i^T \psi(\boldsymbol{t})\phi^T(\boldsymbol{t})$, and $\hat{\boldsymbol{b}} = (\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}y$ and so, $\hat{y} = \tilde{\boldsymbol{X}}\hat{\boldsymbol{b}} = \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}y = \boldsymbol{H}y$ where $\boldsymbol{H}$ is the hat matrix with degrees of freedom $df = trace(\boldsymbol{H})$.

Functional linear model with functional PCA basis

Similarly, Cardotet al.(1999) used a basis of functional principal components to represent the functional data $X(t)$ and the functional parameter $\beta(t)$ in the so-called functional principal components regression (FPCR).

Now, the estimation of β can be made by a few principal components (PC) of the functional data and the integral can be approximated by:

$$\hat{y}_i = \int_{\mathcal{T}} X_i(t)\beta(t)dt \approx \sum_{k=1}^{k_n} \gamma_{ik_n}\hat{\beta}_{k_n} \tag{62}$$

where $\hat{\beta}_{(1:k_n)} = \left(\frac{\gamma_{.1}^T y}{n\lambda_1}, \ldots, \frac{\gamma_{.k_n}^T y}{n\lambda_{k_n}}\right)$ and $\gamma_{(1:k_n)}$ is the $n \times k_n$ matrix with k_n principal components estimation of β scores and λ_i the eigenvalues of the PC.

The model of Equation (62) is expressed as $\hat{y} = \boldsymbol{H}y$ where $\boldsymbol{H} = \left(\frac{\gamma_{\cdot 1}\gamma_{.1}^T y}{n\lambda_1}, \ldots, \frac{\gamma_{\cdot k_n}\gamma_{.k_n}^T y}{n\lambda_{k_n}}\right)$ with degrees of freedom $df = trace(\boldsymbol{H}) = k_n$.

FLR with functional and non functional covariate

This section is presented as an extension of the previous linear regression models. Now, the scalar response Y is estimated by more than one functional covariate $X^j(t)$ and also more than one nonfunctional covariate Z^j. The regression model is given by

$$y_i = \alpha + \beta_1 Z_i^1 + \ldots + \beta_p Z_i^p + \int_{\mathcal{T}} X_i^1(t)\beta_1(t)dt + \ldots + \int_{\mathcal{T}} X_i^q(t)\beta_q(t)dt + \epsilon_i \tag{63}$$

where $Z = [Z^1, \ldots, Z^p]$ are the non functional covariates and $X(t) = [X^1(t_1), \ldots, X^q(t_q)]$ are the functional covariates.

The functional linear model (63) is estimated by the expression

$$\hat{y} = \tilde{\boldsymbol{X}}\hat{\boldsymbol{b}} = \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}y = \boldsymbol{H}y \tag{64}$$

where the first columns of $\tilde{\boldsymbol{X}}$ are the p non-functional covariates Z and the following columns are the q scores. This scores can be done by

1. basis expansion analogous to equation (61)

$$\tilde{X} = [Z^1, \ldots, Z^p, (\boldsymbol{C}^1)^T\psi(\boldsymbol{t}_1)\phi^T(\boldsymbol{t}_1), \ldots, (\boldsymbol{C}^q)^T\psi(\boldsymbol{t}_q)\phi^T(\boldsymbol{t}_q)] \tag{65}$$

2. functional principal components basis as in equation (62)

$$\tilde{X} = \left[Z^1, \ldots, Z^p, \{f_{i1}^1, \ldots, f_{ik_1^1}\}, \ldots, \{f_{i1}^q, \ldots, f_{ik_1^q}\}\right] \tag{66}$$

2.7.2 Functional response and functional independent variables

In the previous section, the scenario involved scalar responses and functional covariates. In this section, the linear model is a fully functional linear regression model in which both the response and covariates are functions. This is given below

$$y_i(t) = \beta_0(t) + \int\limits_{\mathcal{T}_m} X_{im}(s)\beta_m(s,t)ds + \epsilon_i(t), \forall s \in \mathcal{T}_m, \forall t \in \mathcal{T}. \tag{67}$$

The function $\beta_0(t)$ is a parameter function acting as the constant term in the standard regression setup, and allows for different functional origins for the functional response. The function $\beta_m(s,t)$ are bivariate coefficient functions which impose varying weights on $X_{im}(s)$ at arbitrary time $t \in \mathcal{T}_m$ and $\epsilon_i(t)$ are the error functions. Using the expansion in (3), the functional predictors $X_{im}(t)$ are expressed as

$$X_{im}(s) = \sum_{j=1}^{K_m^x} \tilde{c}_{imj}\phi_{mj}(s) = \tilde{\boldsymbol{c}}_{im}^T \boldsymbol{\phi}_m(s), \forall s \in \mathcal{T}_m, \tag{68}$$

the functional responses $y_i(t)$ are given by

$$y_i(t) = \sum_{k=1}^{K_m^y} \tilde{d}_{ik}\psi_k(t) = \tilde{\boldsymbol{d}}_i^T \boldsymbol{\psi}(t), \forall t \in \mathcal{T}_m. \tag{69}$$

The expression of β as a double expansion seems to be appropriate due to its double effect on both the predictors and response variables. The coefficient functions $\beta_m(s,t)$ are expressed as follows

$$\beta_m(s,t) = \sum_{j,k} b_{mjk}\phi_{mj}(s)\psi_k(t) = \boldsymbol{\phi}_m^T(s)\boldsymbol{B}_m\boldsymbol{\psi}(t), \tag{70}$$

where $\boldsymbol{B}_m$ is a $K_m^x \times K_y$ coefficient matrices. By centering the functional linear regression model (67) in the following way

$$\begin{aligned} X_{im}^*(s) &= X_{im}(s) - \bar{X}_{im}(s) \\ &= \tilde{\boldsymbol{c}}_{im}^T\boldsymbol{\phi}(s) - \bar{\boldsymbol{c}}_{im}^T\boldsymbol{\phi}(s) \\ &= \boldsymbol{c}_{im}^T\boldsymbol{\phi}(s), \end{aligned} \tag{71}$$

$$\begin{aligned} y_i^*(t) &= y_i(t) - \bar{y}_i(t) \\ &= \tilde{\boldsymbol{d}}_i^T\boldsymbol{\psi}(t) - \bar{\boldsymbol{d}}_i^T\boldsymbol{\psi}(t) \\ &= \boldsymbol{d}_i^T\boldsymbol{\psi}(t), \end{aligned} \tag{72}$$

equation (67) now become

$$y_i^*(t) = \sum_{m=1}^{K} \int_{\mathcal{T}_m} X_{im}^*(s)\beta_m(s,t)ds + \epsilon_i^*. \tag{73}$$

From equations (70), (71) and (72), equation (73) have the following form

$$\begin{aligned}\boldsymbol{d}_i^T\boldsymbol{\psi}(t) &= \sum_{m=1}^{K}\int\limits_{\mathcal{T}_m} \boldsymbol{c}_{im}^T\boldsymbol{\phi}(s)\boldsymbol{\phi}_m^T(s)\boldsymbol{B}_m\boldsymbol{\psi}(t)ds + \epsilon^*(t)\\ &= \sum_{m=1}^{K} \boldsymbol{c}_{im}^T\boldsymbol{J}_{\phi m}\boldsymbol{B}_m\boldsymbol{\psi}(t) + \epsilon^*(t)\\ &= \boldsymbol{z}_i^T\mathcal{B}\boldsymbol{\psi}(t) + \epsilon^*(t)\end{aligned}\tag{74}$$

where $\boldsymbol{z}_i = (\boldsymbol{c}_{i1}\boldsymbol{J}_{\phi 1}, \dots, \boldsymbol{c}_{iM}\boldsymbol{J}_{\phi M})^T$ is a vector of length $[\sum\limits_{m=1}^{M} K_m^x]$, $\boldsymbol{J}_{\phi m} = \int\limits_{\mathcal{T}_m} \boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^t ds$ which is $K_m^x \times K_m^x$ matrix, and $\mathcal{B} = (\boldsymbol{B}_1, \dots, \boldsymbol{B}_M)^T$ is a $[\sum\limits_{m=1}^{M} K_m^x \times K_y]$ matrix. Combining all the information above, the Functional Linear Regression model for all the observations is

$$\boldsymbol{D}\boldsymbol{\psi}(t) = \boldsymbol{Z}\mathcal{B}\boldsymbol{\psi}(t) + \boldsymbol{\varepsilon}(t)\tag{75}$$

where $\boldsymbol{D}$ is a $N \times K_y$ matrix and $\boldsymbol{Z}$ is a matrix with dimensions $N \times (\sum\limits_{m=1}^{M} K_m^x)$.

Model estimation by Least Squares

Ramsay and Silverman (2005) estimated $\mathcal{B}$ in the model (75) by minimizing the integrated residual sum of squares, the result is now

$$\begin{aligned}&\sum_{i=1}^{N}\int_{\mathcal{T}}\Big[Y_t^*(t) - \sum_{m=1}^{M}\int_{\mathcal{T}_m} X_{im}^*(s)\beta_m(s,t)ds\Big]^2 dt\\ &= \int_{\mathcal{T}} tr\Big\{(\boldsymbol{D}\psi(t) - \boldsymbol{Z}\mathcal{B}\psi(t))(\boldsymbol{D}\psi(t) - \boldsymbol{Z}\mathcal{B}\psi(t))^T\Big\}dt\\ &= \int_{\mathcal{T}} tr\Big\{(\boldsymbol{D} - \boldsymbol{Z}\mathcal{B})\psi(t)\psi^T(t)(\boldsymbol{D} - \boldsymbol{Z}\mathcal{B})^T\Big\}dt\\ &= tr\Big\{(\boldsymbol{D} - \boldsymbol{Z}\mathcal{B})\boldsymbol{J}_\psi(\boldsymbol{D} - \boldsymbol{Z}\mathcal{B})^T\Big\}dt\\ &= tr\Big\{\boldsymbol{D}\boldsymbol{J}_\psi\boldsymbol{D}^T - \boldsymbol{D}\boldsymbol{J}_\psi\mathcal{B}^T\boldsymbol{Z}^T - \boldsymbol{Z}\mathcal{B}\boldsymbol{J}_\psi\boldsymbol{D}^T + \boldsymbol{Z}\mathcal{B}\boldsymbol{J}_\psi\mathcal{B}^T\boldsymbol{Z}^T\Big\}\\ &= tr(\boldsymbol{D}\boldsymbol{J}_\psi\boldsymbol{D}^T) - 2tr(\boldsymbol{D}\boldsymbol{J}_\psi\mathcal{B}^T\boldsymbol{Z}^T) + tr(\boldsymbol{Z}\mathcal{B}\boldsymbol{J}_\psi\mathcal{B}^T\boldsymbol{Z}^T)\end{aligned}\tag{76}$$

where $\boldsymbol{J}_\psi = \int_{\mathcal{T}} \psi(t)\psi^T(t)dt$ is a $K_y \times K_y$ y matrix of basis functions. Computing the derivative of (76) with respect to $\mathcal{B}$ and set the result to zero gives

$$\begin{aligned}
&-2(\boldsymbol{Z}\mathcal{B}\boldsymbol{J}_{\psi}) + 2(Z\mathcal{B}\boldsymbol{J}_{\psi}\mathcal{B}^{T}) = 0\\
&\Rightarrow \boldsymbol{Z}\mathcal{B}\boldsymbol{J}_{\psi} = \boldsymbol{Z}\mathcal{B}\boldsymbol{J}_{\psi}\mathcal{B}^{T}\\
&\Rightarrow vec(\boldsymbol{Z}\mathcal{B}\boldsymbol{J}_{\psi}\mathcal{B}^{T}) = vec(\boldsymbol{Z}\mathcal{B}\boldsymbol{J}_{\psi})\\
&\Rightarrow (\boldsymbol{J}_{\psi} \otimes \boldsymbol{Z}^{T}\boldsymbol{Z})vec(\mathcal{B}) = vec(\boldsymbol{Z}\mathcal{B}\boldsymbol{J}_{\psi})\\
&\Rightarrow vec(\mathcal{B}) = (\boldsymbol{J}_{\psi} \otimes \boldsymbol{Z}^{T}\boldsymbol{Z})^{-1}vec(\boldsymbol{Z}\mathcal{B}\boldsymbol{J}_{\psi})
\end{aligned} \tag{77}$$

where $vec(\mathcal{B})$ is a column vector of length $(\sum_{m=1}^{K} K_m^x) \times K_y$ of $\mathcal{B}$.
Note: when the basis functions are orthonormal (e.g. Fourier, B-Splines), $\boldsymbol{J}_{\psi} = \mathbb{I}_{K_m^x}$.

3 Application

The data are going to be analyzed include two parts: monthly amount of Municipal budget revenue of Lithuania and daily amount of Lithuanian tax receipts. These data are the property of State Tax Inspectorate (STI), which are not publicly available. In this thesis, the aim is to investigate the effect of tax collection by applying functional data analysis (FDA) methods since FDA has never been used to investigate the tax collection (at least there is no publicly available information).

Since the amount of budget revenue in each month or tax receipts in each day is a count and let's assume that can only take non-negative integer values, another possible choice to smooth the data is using Poisson process model. Assume the amount of budget revenue that occur in year i follows an inhomogeneous Poisson process $N_i(t)$, which denotes the total amount of budget revenue occurring during time t in year i. A Poisson process is a stochastic process that counts the number of events (the amount of budget revenue or tax receipts) and the time points at which these events occur in a given time interval. The time to which the next event occurs is independent of the other events and the numbers of events occurred in disjoint intervals are independent of each other. The Poisson process is inhomogeneous in the sense that events occur at a variable rate as time t varies. Let the rate parameter $\lambda(t)$ be a function of t. The key feature of the model is the use of regression splines to model the distribution of the amount of budget revenues and tax receipts over time. The model assumes that the counts for each subject are generated by nonhomogeneous Poisson processes with smooth intensity functions modeled with penalized splines.

In the next subsections the applications of discussed data are presented.

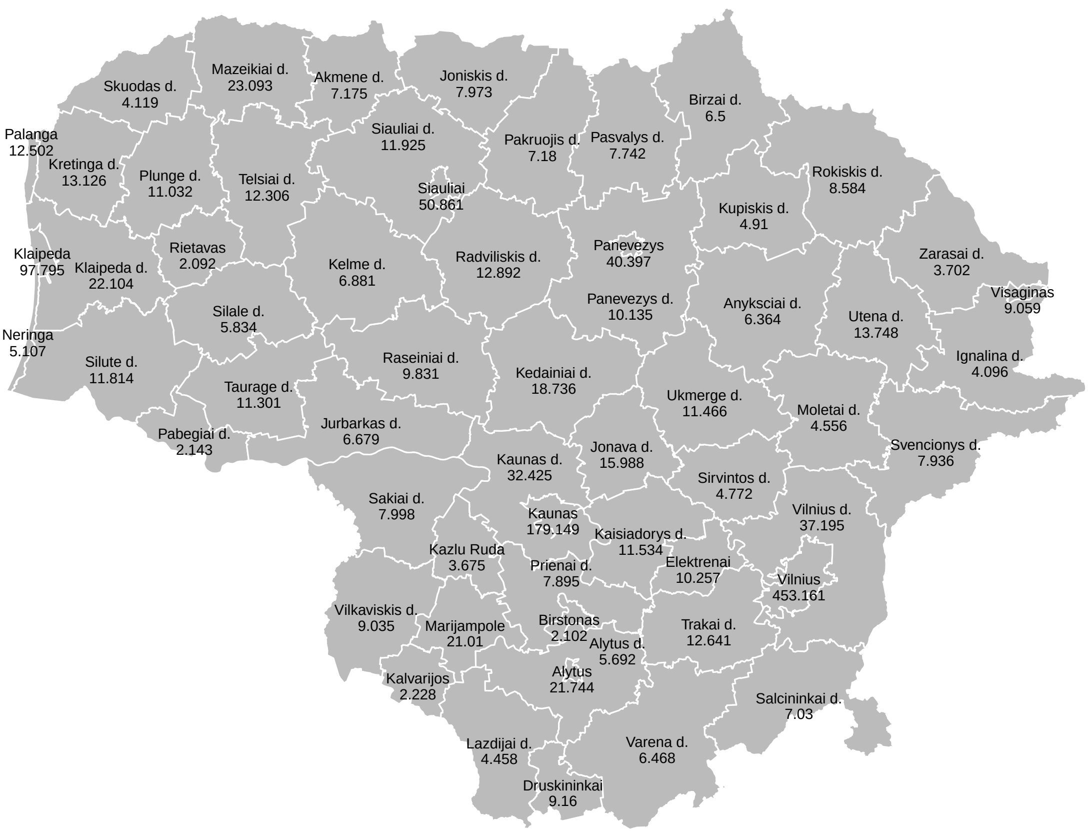
3.1 Application of budget revenue data

Before proceeding on application of functional data analysis, a short introduction of the structure of the municipal budget will be presented. There are various sources of income that can be used by municipalities to finance their expenditure. Common sources of municipal revenue include taxes – personal income tax, property taxes (land tax, inheritance tax), real estate taxes (natural personal tax, legal entities tax) and fees (member fees, local charges). Furthermore there are several other revenue sources including property income (interest on loans, dividends, member corporation tax, tax on state natural resources, hydrocarbon resources tax, rental fee for state land and inland waters fund), revenue for goods and services (income from rental, revenue for occasional service), income from fines and confiscations (revenue from fines for administrative violations, income from penalties for late payments, other revenue from fines and confiscations) and transactions for tangible and intangible assets and the assumption of financial liabilities

A local revenue structure is influenced by a municipality's size, geography, land use and coverage of government services. Other local determinants include numerous legal, political and economic influences, including historical precedent, national economic trends, state laws, intergovernmental relations, regional precedent, citizens' preferences and the city administration's preferences.

60 municipalities within Lithuania have records of the amount of revenue collected during 2001 January to 2016 November, but the observations up to 2016 are kept for analysis. The data for the year 2016 will be used for out-of-sample forecasting.

Figure 1 shows the location of municipalities where Municipal budget revenue data of 2015 are collected on the map of Lithuania. The biggest amount of taxes is collected by Vilnius municipality with the amount of 453 million euros. Notwithstanding in Kaunas municipality taxes are collected as well intensively with the amount of 179 million euros and with Klaipėda, Šiauliai and Panevėžys all these five municipalities collect 60% of total Municipal budget over the year 2015.



Source: created by the author

Figure 1: Municipal Budget of 2015 in Lithuania (million €)

3.1.1 Data smoothing

Let $y_i(t_{ij})$ represents the amount of budget revenue in a month, where $i = 1, 2, ..., 60$ corresponds to municipality and $j = 2001, \ldots, 2015$. The 10800 (60 municipalities and 15 years) discrete data points are smoothed by removing measurement errors and represent them as a continuous function of time t. y_{ij} can be modeled as (3), where $X_i(t_{ij})$ is the function and ε_i is the independent and identically distributed (i.i.d.) random error $(0, \sigma^2)$.

$X_i(t)$ is then approximated as a (4) linear combination of K basis functions, where $\boldsymbol{\phi}(t) = [\phi_1(t), \phi_2(t), \ldots, \phi_K(t)]^T$ is a vector containing K basis functions and and $\boldsymbol{c}_i = [c_{i1}, c_{i2}, \ldots, c_{iK}]^T$ contains corresponding coefficients of basis functions. The choice of the parameter number of basis K has no universal rule that would enable an optimal choice. Generally speaking, the more basis functions is being chosen, the closer the fitted curve will be compared to the discrete data points. However, if too many basis functions are chosen, the fitted curves may be too rough, thus the data will be overfitted. Overfitting makes it difficult to interpret results derived from rough curves. In addition, since a lot of random errors are included in the curves, the results become questionable. The objective of the smoothing is to catch the trend of data without overfitting it. In order to achieve that, instead of controlling K, a roughness penalty term is added.

Among different selection criteria to select the parameter $\nu = (K, \lambda)$, the following two is selected: Cross Validation (CV) and Generalized Cross Validation (GCV). Both criteria are defined as follows

$$CV(\nu) = \frac{1}{n}\sum_{i=1}^{N}\frac{(y_i - \hat{r}^{\nu}(X_i))^2}{1 - S_{ii}}w(X_i), \tag{78}$$

where $\hat{r}^{\nu}(X_i)$ is the prediction at point ti obtained by omitting the i pair (X_i, y_i), S_{ii} is the i diagonal element of the smoothing matrix S (with $\nu = trace(S)$) and $w(X_i)$ is the weight of data X at point t_i, and

$$GCV(\nu) = \frac{1}{n}\sum_{i=1}^{N}\frac{(y_i - \hat{r}^{\nu}(X_i))^2}{1 - S_{ii}}w(X_i)(1 - \frac{1}{n}trace(S))^{-2}. \tag{79}$$

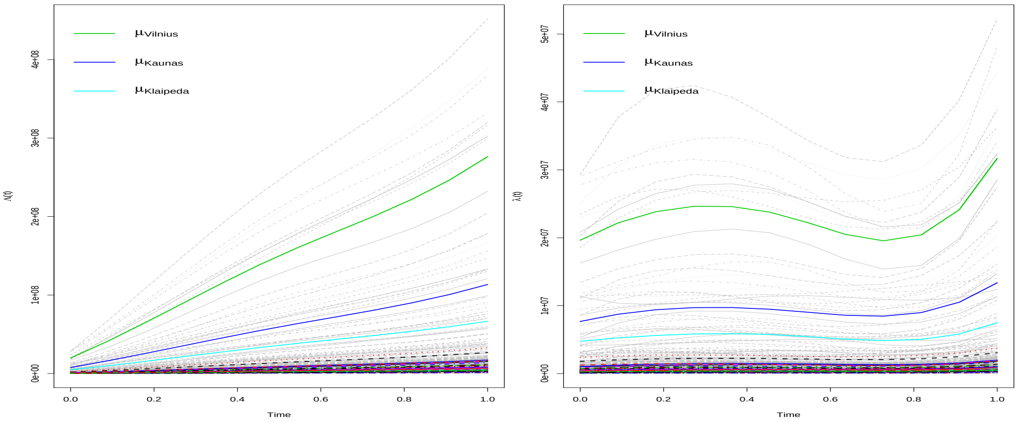
Table 1 shows the optimal values of $\hat{K}$ and $\hat{\lambda}$ that were found by computing each model criterion on the all 60 municipalities.

Type of Functional Basis	Model Criteria	$\hat{K}$	$\hat{\lambda}$
Fourier	GCV	5	0.001953125
	CV	5	0.001953125
B-Spline	GCV	5	0.0001220703
	CV	10	0.001953125

Source: created by the author

Table 1: Optimal choice of functional basis and penalty by cross-validation

B-spline basis system is chosen for the budget revenue data. The discrete budget revenue data are smoothed using 5 B-spline basis functions. The smoothing parameter is chosen to be $\lambda = 0.000122070$. Figure 2 shows the pooled results of smoothed budget revenue data for 60 municipalities over 15 years. Each curve corresponds to municipality $i, i = 1, \ldots, 60$ and year $j, j = 2001, \ldots, 2015$, and the month variable was normalized to interval [0,1], where 0 corresponds to January and 1 to December. Additionally, the mean curve for every municipality is added.



Source: created by the author

Figure 2: Accumulated budget revenue (left) and intensities of budget revenue (right) with mean curves for every municipality

Appendix 1 shows the results of smoothed budget revenue data for each municipality over 15 years. Additionally, the depth measures described in section 2.2.3 and confidence intervals using 2σ rule[1] were added.

3.1.2 Data clustering

According to the Figure 2, it would be reasonable to test whether the differences among municipality means are statistically significant. One-way ANOVA test for functional data (FANOVA) will be used to test the null hypothesis (32) at a significance level $\alpha = 0.05$, where k=60.

F Statsitic	p-value
2835016520	0

Table 2: Functional ANOVA test for Municipal budget revenue data

Table 2 represents the results of the one-way ANOVA test for functional data. According to the results, the p-values of the FANOVA test are less than the significance level 0.05, hence it can be concluded that location has an effect on the mean budget revenue curves of municipalities.

In order to get a more precise view of tax collection in municipalities, pairwise functional ANOVA tests have been performed. The results are shown in Appendix 2, where a gray area

[1] Approximately 95% of the data falls within two standard deviations of the mean ($\mu \pm 2\sigma$)

conform a strong evidence (p-value is less than 0.05) in favor of the hypothesis that both curves are actually different.

Now since functional ARIMA test have detected differences in tax collection among municipalities, hierarchical cluster analysis will be performed in order to group similar municipalities. Before accomplishing cluster analysis, the observations are centered, that is $y_i^*(t) = y_i(t) - \mu_i$, $\forall i$, $i = 1, \dots, 60$.

Hierarchical clustering algorithms are either top-down or bottom-up. Bottom-up algorithms treat each document as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all objects. Bottom-up hierarchical clustering is therefore called hierarchical agglomerative clustering or HAC. Top-down clustering requires a method for splitting a cluster. It proceeds by splitting clusters recursively until individual objects are reached. The methods differ in respect to how they define proximity between any two clusters at every step. Complete linkage will be used for further analysis, but there are several alternatives to complete linkage as a clustering criterion, such as single linkage, average linkage and others.

In the complete linkage, also called farthest neighbor, distance between groups is defined as the distance between the most distant pair of objects, one from each group. In the complete linkage method, D(r,s) is computed as

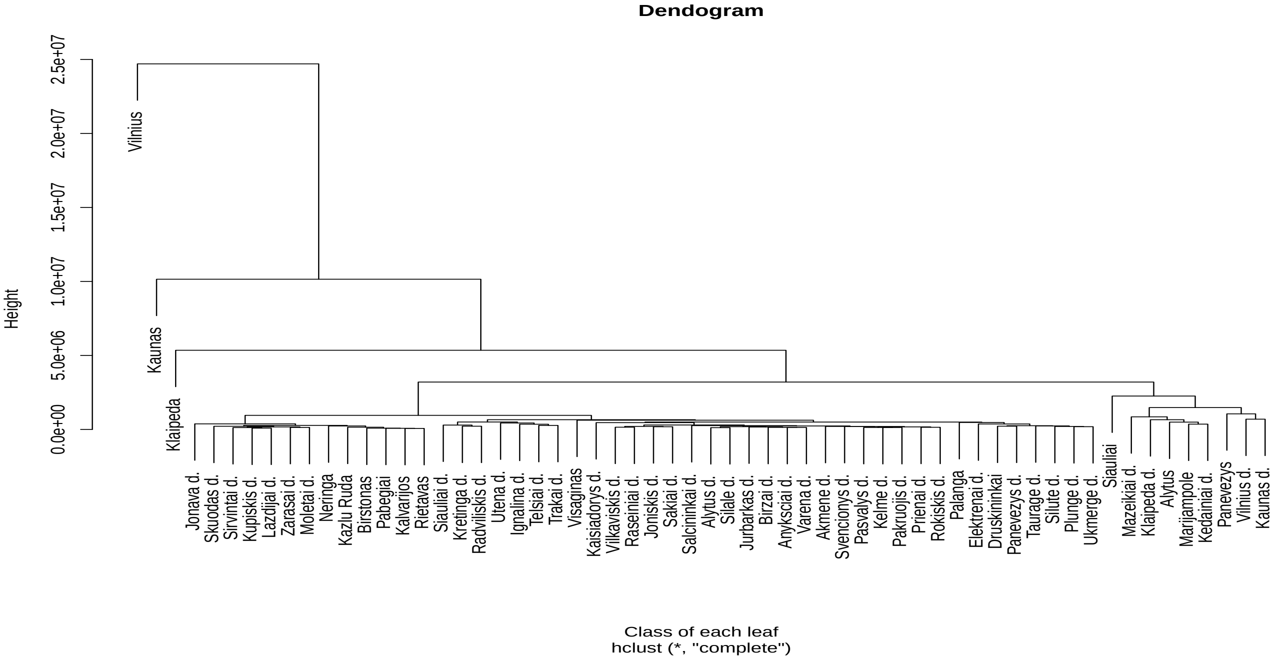
$$D(r, s) = \max\{d(i, j)|i \in r, j \in s\} \tag{80}$$

The distance between every possible object pair (i, j) is computed, where object i is in cluster r and object j is in cluster s and the maximum value of these distances is said to be the distance between clusters r and s. The distance between two clusters is given by the value of the longest link between the clusters. At each stage of hierarchical clustering, the clusters r and s, for which $D(r, s)$ is minimum, are merged.

The Simpson's rule is used to compute distances between elements on $\mathcal{L}_2$ space, where the form is

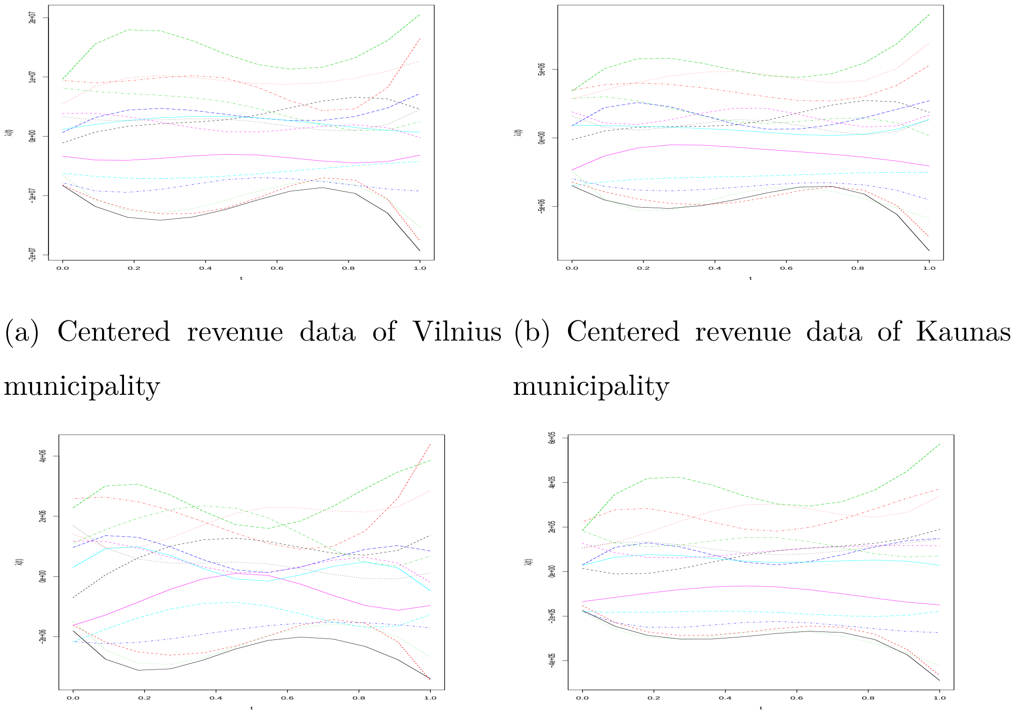
$$d(y_i^*, y_j^*) = \mathbb{E}\Big(\int\limits_0^1 (y_i^*(t) - y_j^*(t))^2 dt\Big)^{\frac{1}{2}} = \frac{1}{15}\sum_{k=1}^{15}\Big(\int\limits_0^1 (y_{ik}^*(t) - y_{jk}^*(t))^2 dt\Big)^{\frac{1}{2}}, \tag{81}$$

for all i, $i = 1, \dots, 60$ and j, j, $p = 1, \dots, 60$, and k, $k = 2001, \dots, 2015$. Figure 3 represents the result of hierarchical cluster analysis, where four clusters appear. More precisely, Vilnius, Kaunas, Klaipėda and, let's define, the rest clusters. Figure 4 shows centered smoothed tax receipt curves for every cluster.



Source: created by the author

Figure 3: Results of hierarchical clustering



(a) Centered revenue data of Vilnius municipality

(b) Centered revenue data of Kaunas municipality

(c) Centered revenue data of Klaipėda municipality

(d) Centered mean revenue data of rest municipalities

Source: created by the author

Figure 4: Centered revenue data of four clusters

3.1.3 Functional modeling and forecasting methods

The rest of the first part will be focused on the contrast between Top-Down model with a Bottom-Up macroeconomic model. In order to understand the nature of different macroeconomic models it is useful to make a distinction between Top-Down and Bottom-Up systems. A Top-Down approach may project a trend in a large aggregate such as GDP, then use historic relationships to derive the components of that total, such as personal consumption expenditures. A Bottom-Up method would go in the opposite direction, working from finer levels of detail (such as trends in population growth and business inventories) to generate a GDP projection.

In this context the finest level is assumed to be municipalities and a large aggregate is the total amount of Municipal budget revenue of Lithuania. In previous analysis municipalities were clustered and dimension of analyzed models decreased from sixty to four. Consequently clustered municipalities are not in the finest level and let's denote this modeling approach as Middle-Up. Accordingly the analysis is supplemented by this method. And now some pertinent issues are raised. More precisely, which of approaches, Top-Down or Bottom-Up, is superior? Or they are supplements? Or even middle-up approach is more preeminent than bottom-up, that is the sum of forecasts of clustered municipalities approximates total revenue forecast no worse than sum of distinct forecasts of municipalities?

In order to answer these questions, consider four functional regression models:

{1} **Functional linear model with scalar response and basis representation**

In this case the response variable is scalar and there is one functional covariate. More precisely,

$$y_i^* = \int\limits_0^1 y_{i-1}^*(t)\beta_1(t)dt + \epsilon_i, \tag{82}$$

where y_i^* are centered annual amount of budget revenue, $y_{i-1}^*(t)$ are functional covariate of centered budget revenue in previous period and ϵ_i are random errors with mean zero and finite variance σ^2.

{2} **Functional linear model with scalar response and functional PCA basis**

The estimation of β can be made by a few principal components of the funtional data and the integral can be approximated by

$$\hat{y}_i^* = \int\limits_0^1 y_{i-1}^*(t)\beta_1(t)dt \approx \sum_{k=1}^{k_n} \gamma_{ik_n}\hat{\beta}_{k_n}, \tag{83}$$

where $\hat{\beta}_{(1:k_n)} = \left(\frac{\gamma_{\cdot 1}^T y}{n\lambda_1}, \ldots, \frac{\gamma_{\cdot k_n}^T y}{n\lambda_{k_n}}\right)$ and $\gamma_{(1:k_n)}$ is the $n \times k_n$ matrix with k_n principal components estimation of β scores and λ_i the eigenvalues of the PC. In this case three principal components will be used.

{3} **Functional linear model with scalar response and functional and non functional covariate**

Now, the scalar response y_i^* is estimated by functional covariate $y_{i-1}^*(t)$ and more than one non functional covariate Z^j. More precisely,

$$y_i^* = \beta_1 Z_i^1 + \beta_2 Z_i^2 + \int_0^1 y_{i-1}^*(t)\beta_{11}(t)dt + \epsilon_i \tag{84}$$

where Z_i^1 are the non functional covariate of annual GDP in counties, Z_i^2 are the non functional covariate of annual number of registered unemployed in Sodra in municipalities and $y_{i-1}^*(t)$ are the functional covariates of budget revenue of previous period as before.

{4} **Functional linear model with functional response and functional independent variables**

$$y_i(t) = \int \beta_1(s,t)y_{i-1}(s)dt + \epsilon_i(t) \tag{85}$$

where $\beta_1(s,t)$ is non-random coefficient functions, the functional slopes. Model (85) is generally referred to as functional linear model (FLM), at any given time s, the value of $y(s)$ depends on the entire trajectory of X. It is a direct extension of traditional linear models with multivariate response and vector covariates by changing the inner product from the Euclidean vector space to $\mathcal{L}_2$ and the coefficient function varies with s, leading to a bivariate coefficient surface.

Now the results of three different approaches will be presented.

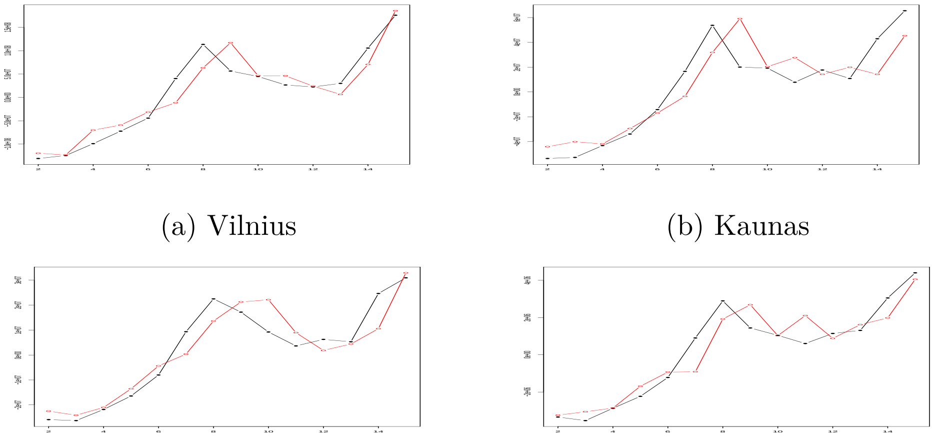
3.1.4 Middle-Up results

In this approach clusters obtained by distance based clustering (see Figure 4) are being modeled. First, functional linear model with scalar response and basis representation is estimated, where β is represented by 5 b-spline basis functions and the amount of annual budget revenue is selected as the scalar response. Table 15 in Appendix 3 shows the evaluated b-spline coefficients and standard error in brackets. According to the table, 89%, 87%, 88% and 88% of variation is explained by a functional linear model with basis representation for Vilnius, Kaunas, Klaipėda and rest budget revenue, respectively. Figure 15 in Appendix 3 represents the diagnostics for estimated model. Given a sample set, one can compute the standardized residuals and compare these to the expected frequency: points that fall more than 3 standard deviations from the norm are likely outliers and if there are many points more than 3 standard deviations from the norm, one likely has reason to question the assumed normality of the distribution. In this case, the square root standardized residuals were estimated and all standardized residuals are less than 1.5. Furthermore, Ljung-Box test is conducted to test whether there is autocorrelation in residuals. Since there is only 14 modeled observations, the first 13 autocorrelations were tested. And another test is performed - an univariate test for white noise under general weak dependent assumptions, which have been proposed by Lobato and Velasco (2004). A von Mises-type statistic is computed against a $\mathcal{N}(0, 4)$ distribution. Figure 3 represents test statistics and p-values of these two tests. Results of the test shows that there is no serial autocorrelation in the modeled residuals and all models are white noise with at least 95% confidence level.

	Ljung-Box Statistic	Ljung-Box p-value	von Mises Statistic	von Mises p-value
Vilnius	0.001	0.969	0.36	0.397
Kaunas	0.003	0.956	0.726	0.717
Klaipėda	0.609	0.435	0.596	0.593
Rest	0.007	0.932	0.735	0.725

Table 3: Tests for analysis of residuals of {2} model

Figure 5 represents fitted values (red) and true values (black) of the {1} model. According to the graphs, model doesn't take into account the period of global financial crisis, where sudden increase in GDP in 2007 was accompanied by higher amount of revenue collected was not perceived by a model and in all cases the annual amount of budget revenue is reduced. And by contrary, in the period from 2009 to 2013 the amount of budget revenue is modeled higher than actually collected. From the government point of view, such forecast would worsen even more the circumstances of the process of fund allocation as government should have borrowed funds thereby deepening deficit of Lithuania.



(a) Vilnius

(b) Kaunas

(c) Klaipėda

(d) Rest

Source: created by the author

Figure 5: Fitted values (red) of {1} model vs true smoothed values (black)

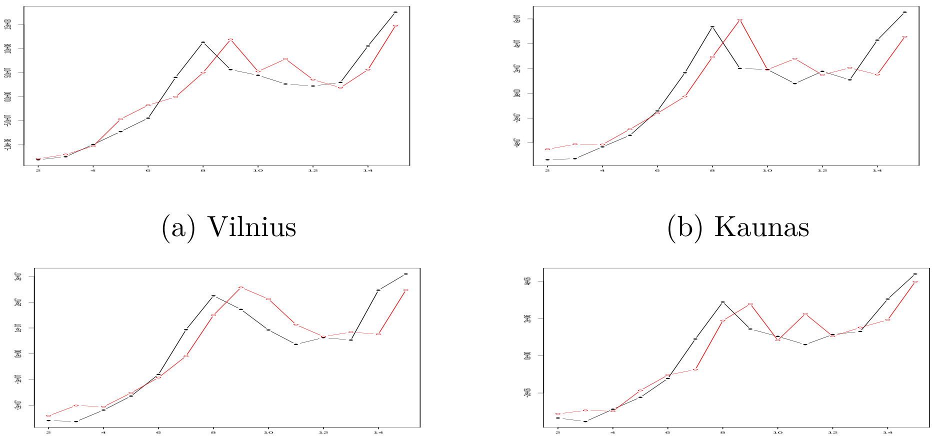
Second, the functional linear model with scalar response and PCA basis is evaluated. The amount of annual budget revenue is selected as the scalar response and the 1^{st}, 2^{nd} and 3^{rd} principal components were selected to estimate coefficients of β. Figure 31 in Appendix 3 shows the evaluated PCA coefficients and standard error in brackets. According to the table, 85%, 77%, 83% and 87% of variation is explained by a functional linear model with PCA representation for Vilnius, Kaunas, Klaipėda and rest budget revenue, respectively. Futhermore, the 1^{st} principal component accounts for 92.22%, 92.47%, 88.14% and 93% of the overall variability for Vilnius, Kaunas, Klaipėda and rest budget revenue, respectively. Other principal components explains a small part of the variation. Figure 31 in Appendix 3 shows the diagnostics for estimated model. The square root standardized residuals were estimated and all standardized residuals are less than 1.5. Figure 4 represents test statistics and p-values of Ljung-Box and white noise tests. Results of the test shows that there is no serial autocorrelation in the modeled residuals and all models are white noise with at least 95% confidence level.

	Ljung-Box Statistic	Ljung-Box p-value	von Mises Statistic	von Mises p-value
Vilnius	0.128	0.721	0.488	0.498
Kaunas	0.018	0.893	0.796	0.787
Klaipėda	2.242	0.134	1.203	0.788
Rest	0.036	0.850	0.831	0.823

Source: created by the author

Table 4: Tests for analysis of residuals of {2} model

Figure 6 represents fitted values (red) and true values (black) of the {2} model. According to the graphs, resembling estimates of the previous model is performed by {2} model and similar interpretation is applied and with this model, that enlarged forecast for the crisis period would made to accept the wrong political decisions.



(a) Vilnius

(b) Kaunas

(c) Klaipėda

(d) Rest

Source: created by the author

Figure 6: Fitted values (red) of {2} model vs true smoothed values (black)

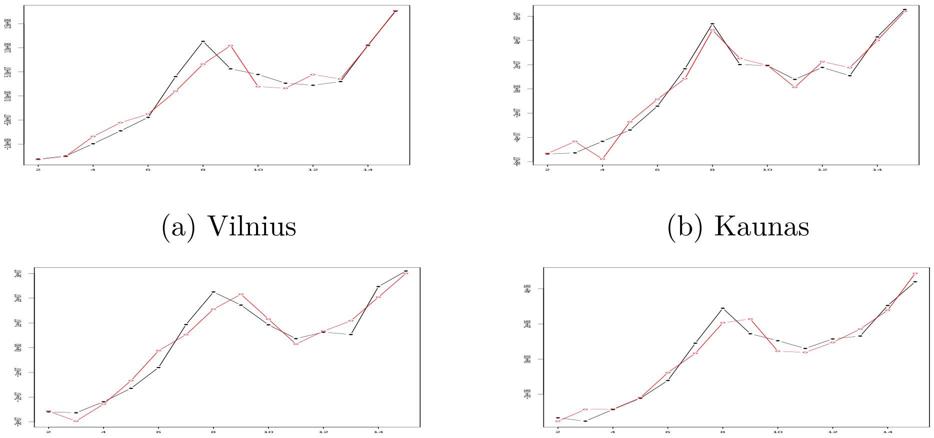
Third, functional linear model with scalar response and both functional and non functional covariates is estimated. GDP in county and the number of unemployed in municipality were selected as non functional covariates and the amount of previous period budget revenue was selected as functional covariate, which is represented with 5 b-spline basis functions. And the amount of annual budget revenue is selected as the scalar response. Figure 31 in Appendix 3 shows the evaluated b-spline and non functional coefficients and standard error in brackets. According to the table, 93%, 87%, 96% and 96% of variation is explained by a functional linear model with non functional covariates for Vilnius, Kaunas, Klaipėda and rest budget revenue, respectively. Non functional covariate *GDP* is statistically significant for Kaunas and Klaipėda budget revenue models and non functional covariate *Unemployed* is statistically significant for Kaunas and rest bidget revenue models. Figure 5 represents test statistics and p-values of Ljung-Box and white noise tests. Results of the test shows that there is no serial autocorrelation in the modeled residuals and all models are white noise with at least 95% confidence level.

	Ljung-Box Statistic	Ljung-Box p-value	von Mises Statistic	von Mises p-value
Vilnius	0.996	0.318	0.399	0.426
Kaunas	2.583	0.108	2.272	0.092
Klaipėda	0.193	0.660	1.635	0.401
Rest	2.310	0.129	1.57	0.451

Source: created by the author

Table 5: Tests for analysis of residuals of {3} model

Figure 7 represents fitted values (red) and true values (black) of the {3} model. According to the graphs, fitted values are far more closer to actual values. Non functional covariates of *GDP* and *Unemployed* help to explain annual budget revenue and to engross the crises effect.



(a) Vilnius

(b) Kaunas

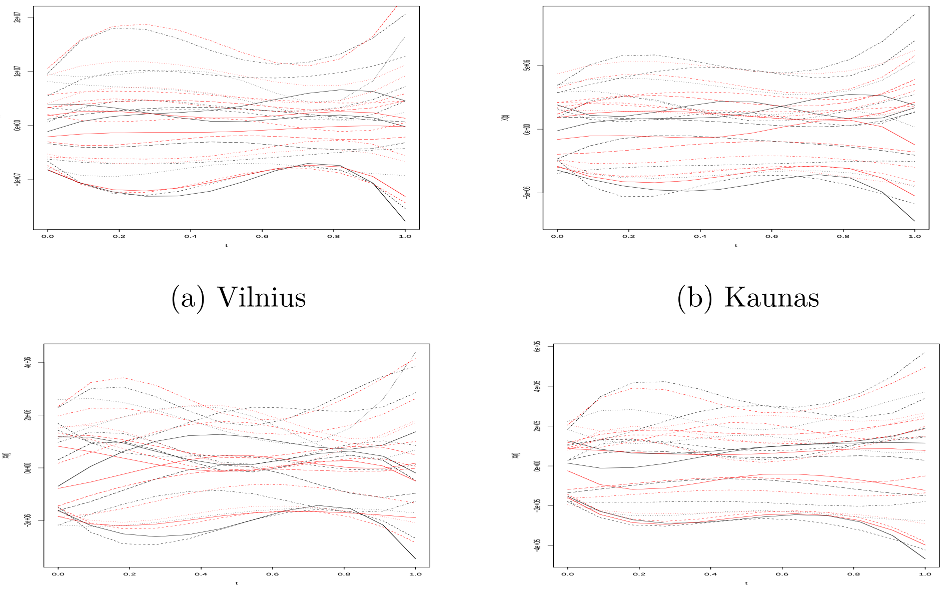
(c) Klaipėda

(d) Rest

Source: created by the author

Figure 7: Fitted values (red) of {3} model vs true smoothed values (black)

And fourth, functional linear model with functional response and functional independent variables is evaluated. The amount of previous period budget revenue was selected as functional covariate and the amount of current period budget revenue was selected as functional response. Figure 8 represents smoothed centered budget revenue data vs fitted values of {4} model. The difference in this respect is that response variable is a function which represents months in horizontal axis. Visually budget revenue data is estimated rather precisely, but to ascertain that model is adequate analysis of residuals is performed.



(a) Vilnius

(b) Kaunas

(c) Klaipėda

(d) Rest

Source: created by the author

Figure 8: Centered revenue data vs fitted values of {4} model

Table 6 represents test statistics and p-values of Ljung-Box and white noise tests for municipal budget revenue of Vilnius. Since residuals are functional, pointwise tests were performed. Results of the test shows that there is serial autocorrelation in 2004, 2006, 2007, 2008, 2009 and 2012 at least at 95% confidence level. It might be due to accession to the European Union and global financial crisis and probably some specific events occurred in Vilnius in 2012. According to the Table 20 in Appendix 4, proportion of the collected budget revenue in January, February and March of 2012 is abnormally high compared with other years. In those months 17.14%, 17.37% and 14.76% proportions consisted of annual budget revenue while 10.99%, 9.36% and 10.07% are the highest proportion collected in others years discarding 2009, respectively. Moreover, 0.91% is collected in December throughout the year 2012, while over other years in average 9.9% is collected in December discarding 2009. On contrary, white noise test accepts the null hypothesis for 2004, but reject for those mentioned years at least 95% confidence level.

	Ljung-Box Statistic	Ljung-Box p-value	von Mises Statistic	von Mises p-value
Year 2002	0.845	0.358	0.161	0.304
Year 2003	2.246	0.134	0.412	0.472
Year 2004	8.714	0.003	1.05	0.951
Year 2005	1.682	0.195	0.898	0.901
Year 2006	10.411	0.001	−0.848	0.024
Year 2007	8.093	0.004	−0.909	0.019
Year 2008	4.080	0.043	−0.957	0.017
Year 2009	4.056	0.044	−0.942	0.017
Year 2010	3.263	0.071	−0.785	0.029
Year 2011	2.904	0.088	2.149	0.159
Year 2012	5.953	0.015	−0.891	0.021
Year 2013	3.223	0.073	0.115	0.279
Year 2014	0.882	0.348	1.5	0.54
Year 2015	1.029	0.310	1.303	0.711

Source: created by the author

Table 6: Tests for analysis of residuals for Vilnius

Table 7 represents the same test statistics and p-values for municipal budget revenue of Kaunas. Results of the test shows that there is serial autocorrelation in 2004, 2006, 2007, 2012 and 2013 at least at 95% confidence level. It might be due to the same conditions as for Vilnius. Table 21 in Appendix 4 affirms aberrant behavior collected proportions of centered budget revenue in Kaunas, except for year 2004. Though white noise test accepts the null hypothesis for the year 2004 and rejects the null hypothesis for 2002, 2007, 2008, 2009, 2011 and 2015, but still the crisis effect is detected.

	Ljung-Box Statistic	Ljung-Box p-value	von Mises Statistic	von Mises p-value
Year 2002	2.825	0.093	−0.673	0.040
Year 2003	1.156	0.282	−0.025	0.210
Year 2004	5.694	0.017	0.196	0.325
Year 2005	3.385	0.066	−0.581	0.053
Year 2006	6.003	0.014	−0.179	0.149
Year 2007	4.686	0.030	−0.965	0.016
Year 2008	2.682	0.101	−0.857	0.023
Year 2009	2.362	0.124	−0.791	0.028
Year 2010	1.961	0.161	0.93	0.932
Year 2011	1.989	0.158	−0.743	0.033
Year 2012	4.582	0.032	1.03	0.971
Year 2013	3.998	0.046	1.755	0.355
Year 2014	1.972	0.160	0.58	0.607
Year 2015	4.542	0.033	−0.912	0.019

Source: created by the author

Table 7: Test for analysis of residuals for Kaunas

Table 8 represents test statistics and p-values of residual analysis for municipal budget revenue of Kaunas. Results of the test shows that there is serial autocorrelation in 2003, 2004, 2005, 2006, 2008, 2009, 2012, 2013 and 2015 at least at 95% confidence level. It might be due to the same conditions as for Vilnius and additionally some other effects. According to the Table 22 in Appendix 4, proportions of the centered budget revenue collected during the year in Klaipėda suspect aberrant behavior only for 2012 and 2013. Though white noise test rejects the null hypothesis only for 2004.

	Ljung-Box Statistic	Ljung-Box p-value	von Mises Statistic	von Mises p-value
Year 2002	1.038	0.308	0.094	0.267
Year 2003	9.481	0.002	−0.24	0.129
Year 2004	4.610	0.032	−0.921	0.019
Year 2005	7.398	0.007	−0.179	0.149
Year 2006	9.663	0.002	0.241	0.353
Year 2007	1.231	0.267	−0.252	0.125
Year 2008	6.757	0.009	−0.121	0.170
Year 2009	7.918	0.005	−0.42	0.082
Year 2010	0.872	0.350	0.897	0.899
Year 2011	0.730	0.393	1.652	0.424
Year 2012	7.683	0.006	0.029	0.234
Year 2013	5.476	0.019	1.029	0.972
Year 2014	0.886	0.346	0.348	0.424
Year 2015	7.235	0.007	−0.012	0.215

Source: created by the author

Table 8: Test for analysis of residuals for Klaipėda

Table 9 represents test statistics and p-values of residual analysis for municipal budget revenue of other municipalities in Lithuania. Results of the test shows that there is serial autocorrelation in 2003, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012 and 2014 at least at 95% confidence level. Consequently smaller municipalities have suffered from the crisis more severe than bigger municipalities, such as Vilnius, Kaunas and Klaipėda. According to the Table 23 in Appendix 4, proportions of the centered budget revenue collected during the year in the rest municipalities suspect aberrant behavior only for 2007, 2008, 2009, 2010 and 2011. Though white noise test rejects the null hypothesis for 2004, 2007, 2008, 2009, 2011 and 2014. Clearly the accession to the European Union and the global financial crisis had an effect on tax collection.

	Ljung-Box Statistic	Ljung-Box p-value	von Mises Statistic	von Mises p-value
Year 2002	2.712	0.100	−0.655	0.043
Year 2003	10.286	0.001	−0.523	0.062
Year 2004	0.302	0.582	2.911	0.019
Year 2005	6.160	0.013	1.562	0.491
Year 2006	4.761	0.029	−0.12	0.170
Year 2007	4.989	0.026	−0.96	0.016
Year 2008	6.264	0.012	−0.996	0.014
Year 2009	6.719	0.010	−0.998	0.014
Year 2010	4.511	0.034	0.088	0.264
Year 2011	3.956	0.047	−0.864	0.022
Year 2012	5.583	0.018	−0.112	0.173
Year 2013	3.602	0.058	0.806	0.812
Year 2014	7.356	0.007	4.064	0.0002
Year 2015	1.631	0.202	−0.242	0.128

Source: created by the author

Table 9: Tests for analysis of residuals for rest

Forecasting results of annual budget revenue data is represented in Table 10. Unfortunately at this time of the moment, the amount of actually collected revenues for 2016 December isn't available due to the deadline of submitting the thesis, which is earlier than the information is published. Thus actual amount of budget revenue is introduced only for January and November months of 2016. According to the results, the amount of budget revenue of Vilnius forecasted by {1} and {3} models is inaccurate, because the amount collected in 11 months is greater than models forecasted for the year 2016. The same effect is observed for Klaipėda. Other forecasts for Kaunas and rest municipalities appear to be tenable.

	True value 2016.01-2016.11	$\{1\}model$	$\{2\}model$	$\{3\}model$
Vilnius	$451, 430, 000$	$400, 569, 307$	$460, 298, 614$	$428, 569, 406$
Kaunas	$178, 444, 700$	$186, 838, 140$	$186, 575, 974$	$202, 496, 577$
Klaipėda	$94, 406, 600$	$94, 406, 600$	$97, 957, 956$	$93, 243, 436$
Rest	$10, 921, 933$	$12, 265, 094$	$12, 343, 845$	$12, 339, 128$

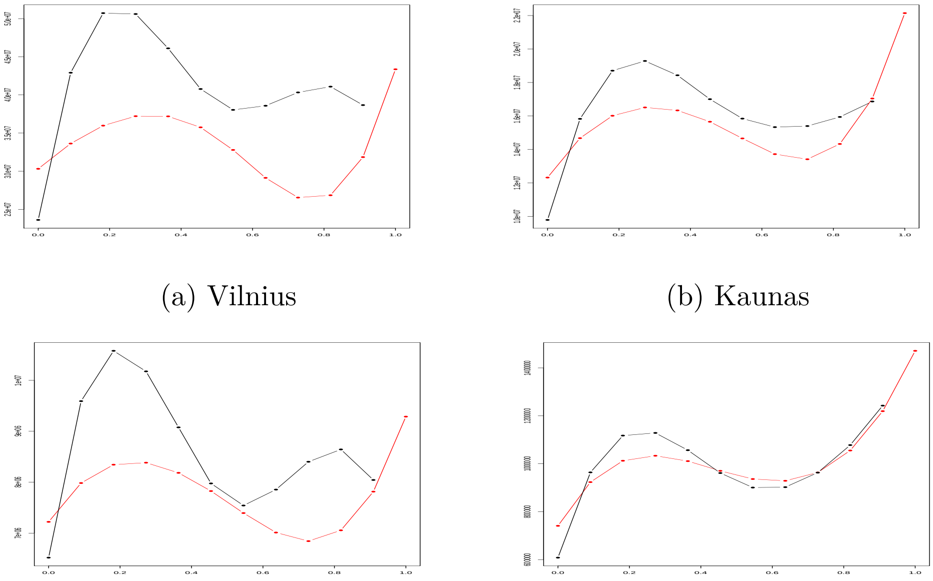
Source: created by the author

Table 10: Forecasting results of annual amount of budget revenue for 2016

In order to compare the performance of the models, the assumptions are being made about the proportions of the centered budget revenue collected in December in Vilnius, Kaunas , Klaipėda and rest municipalities. Table 25, Table 26, Table 27 and Table 28 in

Appendix 4 are being employed for calculating the proportions, respectively. The mean of the proportion collected in December over the fifteen years is calculated and added to the true value of revenue collected during the eleven months. In average 11.34%, 11.66%, 11.2% and 11.88% is collected in December during the year in Vilnius, Kaunas, Klaipėda and rest municipalities, respectively. Therefore, tentatively 502,605,233€; 199,256,741€; 104,984,746€ and 12,218,947€ should be collected in 2016 in Vilnius, Kaunas, Klaipėda and rest municipalities, respectively. For Vilnius and Klaipėda three forecasts of {1}, {2} and {3} models underestimates preliminary amount for 2016. For Kaunas {3} model appears to be the closest to the preliminary amount for 2016. And for the rest municipalities forecast of {1} model seems to be the closest to the preliminary amount for 2016. Consequently, models for Vilnius and Klaipėda should be improved in order to forecast more accurate amount of municipal budget revenue.

Figure 9 represents forecasts of the {4} model. The same implications are applied to this case that models, which is designed for Vilnius and Klaipėda should be refined.



(a) Vilnius

(b) Kaunas

(c) Klaipėda

(d) Rest

Source: created by the author

Figure 9: Monthly forecast (red) of {4} model vs true smoothed values (black) for 2016

However, the monthly information characterizing municipalities is not publicly available. The most reference is about some yearly performance of the municipalities, but still very few. Merely because of the advantages of the functional data analysis allowed to construct models, which could elucidate municipal budget revenue data.

3.1.5 Bottom-Up results

In order to not to stack with redundant information, only the forecasts of the {1}, {2} and {3} models will be presented. Table 18 in Appendix 3 shows the annual forecast for every municipality of 2016. Some forecasts of Birstonas, Vilnius district, Elektrėnai district, Jonava district, Kazlų Rūda, Prienai district, Rietavas, Telšiai district, Utena district, Šakiai district, Širvintai district municipalities underestimate actual 11 months amount of the budget revenue, therefore reasonable model should be selected. Though models for Vilnius, Klaipėda, Neringa, Palanga municipalities should be improved because all of three models underestimate the amount of budget revenue for 2016.

Preliminary results comparison will not be conducted.

3.1.6 Top-Down results

Analogous analysis is done for the total municipal budget revenue data as in section 3.1.4. Figure 11 represents Ljung-Box and white noise tests results. Figure 19 in Appendix 3 shows the evaluated coefficients of {1}, {2}, {3} models and standard error in brackets. According to the table, 86%, 85% and 93% of variation is explained by {1}, {2} and {3} models, respectively. Moreover, the 1^{st} principal component accounts for 92.77% of the overall variability of total budget revenue data of {2} model.

Table 11 represents the results of residual analysis. Both tests accepts the null hypothesis that there is no serial correlation in residuals and they follow a white noise with assumptions of general weak dependent.

	Ljung-Box Statistic	Ljung-Box p-value	von Mises Statistic	von Mises p-value
{1} model	0.006	0.937	0.626	0.621
{2} model	0.010	0.921	0.748	0.739
{3} model	3.225	0.073	1.759	0.315

Source: created by the author

Table 11: Tests for analysis of residuals for total municipal budget revenue

The results of residual analysis is presented in Table 12 for {4} model. Results of the Ljung-Box test shows that there is serial autocorrelation in 2006, 2007, 2009 and 2015 at least at 95% confidence level. Though the white noise test rejects the null hypothesis for 2006, 2007, 2008, 2009, that is for the pre-crisis and crisis period. This is confirmed by Table

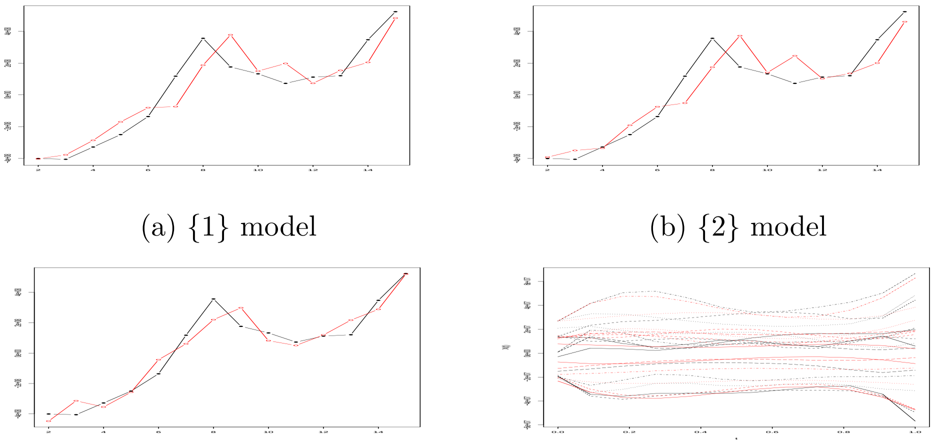
24 in Appendix 4, which shows aberrant behavior of centered budget revenue data for the years of 2006, 2007 and 2009.

	Ljung-Box Statistic	Ljung-Box p-value	von Mises Statistic	von Mises p-value
Year 2002	0.006	0.940	0.3	0.391
Year 2003	0.169	0.681	1.503	0.538
Year 2004	0.042	0.838	0.567	0.596
Year 2005	0.071	0.791	0.321	0.406
Year 2006	8.181	0.004	−0.653	0.043
Year 2007	3.846	0.050	−0.979	0.015
Year 2008	1.202	0.273	−0.886	0.021
Year 2009	5.993	0.014	−0.919	0.019
Year 2010	0.282	0.596	−0.064	0.193
Year 2011	1.751	0.186	0.0444	0.242
Year 2012	0.527	0.468	0.838	0.842
Year 2013	0.036	0.850	0.622	0.644
Year 2014	0.780	0.377	1.233	0.775
Year 2015	4.893	0.027	0.296	0.387

Source: created by the author

Table 12: Tests for analysis of residuals for total municipal budget revenue of {4} model

Figure 10 represents fitted values (red) and true values (black) of four models. According to the graphs, again fitted values are far more closer to actual values for the {3} model, which incorporates explanatory variables, where non functional covariates of *GDP* and *Unemployed* help to explain annual budget revenue and to engross the crises effect.



(a) {1} model

(b) {2} model

(c) {3} model

(d) {4} model

Source: created by the author

Figure 10: Fitted values (red) vs true smoothed values (black) of total municipal budget revenue models

3.1.7 A comparison of approaches

Now the comparison of bottom-up, middle-up and top-down approaches will be conducted. Forecasting results of total annual budget revenue data is represented in Table 13. Unfortunately, as was mentioned before the amount of actually collected revenues for 2016 December isn't available. Thus actual amount of total budget revenue is introduced only for January and November months of 2016. According to the results,

	True value 2016.01-2016.11	Bottom-up	Middle-up	Top-down
{1} model	1,346,831,000	1,380,925,807	1,380,924,405	1,407,915,195
{2} model	1,346,831,000	1,432,841,198	1,448,431,709	1,527,078,704
{3} model	1,346,831,000	1,417,102,432	1,427,612,715	1,643,354,521

Source: created by the author

Table 13: Comparison of approaches with forecast for 2016

In order to compare the performance of the models, the assumptions are being made about the proportions of the centered budget revenue collected in December in Lithuania. Table 29 in Appendix 4 is being employed for calculating the proportions. The mean of the proportion collected in December over the fifteen years is calculated and added to the true value of revenue collected during the eleven months. In average 11.59% is collected in December during the fifteen years in Lithuania. Therefore, tentatively 1,502,884,267€ should be collected in 2016 in Lithuania. According to he table, only Top-down approach with {2} and {3} models has forecasted the closest amount to the preliminary. So it can be stated that Top-down approach seems to be superior to Bottom-Up or Middle-Up approaches with functional linear models with PCA representation and with non functional covariates.

3.1.8 Further analysis

Despite this optimistic suggestion of the role of functional data analysis in application of municipal budget revenue data of Lithuania, further analysis is necessary to confirm its efficacy. The data set used in this study consisted of 2001-2015 years of municipal budget revenue data and 2016 was left for out-of-sample forecasting. Further research should analyze a wider range of explanatory variables, especially for models, which design Vilnius, Klaipėda, Neringa and Palanga municipalities. But inclusion of such variables is limited, because the characteristics of municipalities are not publicly available. Furthermore, other smoothing methods could be analyzed for transforming discrete data to smooth curves. For example, nonparametric kernel methods, such as Nadaraya-Watson method or local linear regression method, where after the selection of the type of kernels, bandwidth is being estimated as smoothing parameter. Also, an investigation could be pursued to find if functional principal components can differentiate among the months and municipalities of Lithuania. This type of analysis was not accomplished with the given data set and it is left for further research.

3.2 Application of tax receipt data

Daily economic time series often have properties that make them harder to model and to forecast than monthly or quarterly data for which numerous standard solutions exist. In addition to the well known features typical of monthly data - trend, season, trading day and calendar effects - there are two major problems with daily data. First, the number of observations varies per month and per year which leads to a time series with irregular spacing. Second, we need to take account of daily heteroskedasticity since the variance may depend on the day-of-the month. Many aggregate economic transactions have patterns with a clear peak once a month, e.g. salary payments, money circulation, and tax revenues. It is often not easy to stabilise the variance by taking logs: the (persistently changing) seasonal pattern is not simply multiplicative and the irregular component is not either. Moreover, very small (or even negative in cases of net series) values can be part of a daily time series.

The illustration of daily time series features is presented using a series for Lithuanian aggregate tax receipts from 2009-01-01 to 2016-10-31. Since the data are not publicly available, they are transformed. Lithuanian total national daily tax receipts consist of several major components like value-added tax, personal income tax, corporate income tax, excise and a number of smaller categories, like property tax or real estate tax[2]. Many taxes are due on the 15^{th} and 25^{th} workdays of the month. The majority of revenues is collected on the 15^{th} and 25^{th} workdays.

3.2.1 Data smoothing

Although one of the functional data analysis features allows evaluation of record at any time point (especially if observation times are not the same across records), practically if the number of observation points per curve and per variable vary, curves and variables must be smoothed individually. Table 14 represents the optimal values of $\hat{K}$ and $\hat{\lambda}$ that were found by computing each model criterion on the all months with different calendar days.

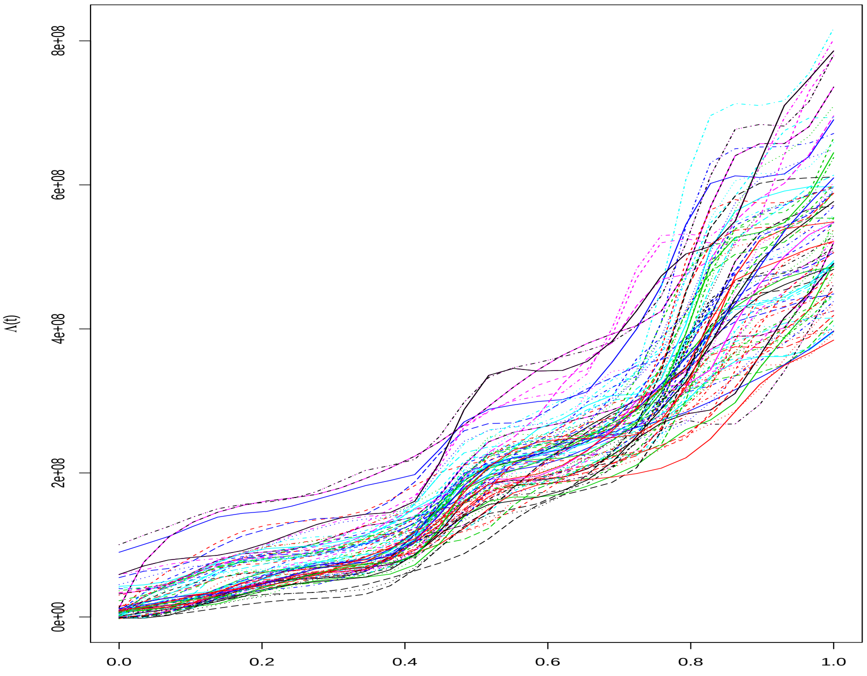
[2] Full tax contribution list can be found at https://www.vmi.lt/cms/imoku-kodai

Month	Type of Functional Basis	Model Criteria	$\hat{K}$	$\hat{\lambda}$
January	Fourier	GCV	8	$6.103516 \cdot 10^{-5}$
		CV	8	$6.103516 \cdot 10^{-5}$
January	B-Spline	GCV	19	$3.051758 \cdot 10^{-5}$
		CV	19	$6.10351 \cdot 10^{-5}$
February (non-leap years)	Fourier	GCV	10	$3.051758 \cdot 10^{-5}$
		CV	10	$3.051758 \cdot 10^{-5}$
February (non-leap years)	B-Spline	GCV	23	$9.536743 \cdot 10^{-5}$
		CV	14	$3.051758 \cdot 10^{-5}$
February (leap years)	Fourier	GCV	8	$3.051758 \cdot 10^{-5}$
		CV	8	$6.103516 \cdot 10^{-5}$
February (leap years)	B-Spline	GCV	19	$3.051758 \cdot 10^{-5}$
		CV	19	0.0001220703
March	Fourier	GCV	8	$6.103516 \cdot 10^{-5}$
		CV	8	$6.103516 \cdot 10^{-5}$
March	B-Spline	GCV	19	$3.051758 \cdot 10^{-5}$
		CV	19	$6.10351 \cdot 10^{-5}$
April	Fourier	GCV	9	$6.103516 \cdot 10^{-5}$
		CV	9	$6.103516 \cdot 10^{-5}$
April	B-Spline	GCV	17	$3.051758 \cdot 10^{-5}$
		CV	15	$3.051758 \cdot 10^{-5}$
May	Fourier	GCV	9	$6.103516 \cdot 10^{-5}$
		CV	9	$6.103516 \cdot 10^{-5}$
May	B-Spline	GCV	29	$7.629395 \cdot^{-6}$
		CV	29	$1.525879 \cdot 10^{-5}$
June	Fourier	GCV	10	$6.103516 \cdot 10^{-5}$
		CV	10	$6.103516 \cdot 10^{-5}$
June	B-Spline	GCV	17	$7.629395 \cdot 10^{-5}$
		CV	17	0.0001220703
July	Fourier	GCV	9	$6.103516 \cdot 10^{-5}$
		CV	9	$6.103516 \cdot 10^{-5}$
July	B-Spline	GCV	29	$7.629395 \cdot 10^{-6}$
		CV	15	$6.103516 \cdot 10^{-5}$
August	Fourier	GCV	9	$6.103516 \cdot 10^{-5}$
		CV	9	$6.103516 \cdot 10^{-5}$
August	B-Spline	GCV	23	$7.629395 \cdot 10^{-6}$
		CV	19	$3.051758 \cdot 10^{-5}$
September	Fourier	GCV	9	$6.103516 \cdot 10^{-5}$
		CV	9	$6.103516 \cdot 10^{-5}$
September	B-Spline	GCV	26	$7.629395 \cdot 10^{-6}$
		CV	15	$3.051758 \cdot 10^{-5}$
October	Fourier	GCV	9	$6.103516 \cdot 10^{-5}$
		CV	9	$6.103516 \cdot 10^{-5}$
October	B-Spline	GCV	29	$7.629395 \cdot 10^{-6}$
		CV	15	$6.103516 \cdot 10^{-5}$
November	Fourier	GCV	9	$6.103516 \cdot 10^{-5}$
		CV	9	$6.103516 \cdot 10^{-5}$
November	B-Spline	GCV	26	$7.629395 \cdot 10^{-6}$
		CV	27	$1.525879 \cdot 10^{-5}$
December	Fourier	GCV	10	$3.051758 \cdot 10^{-5}$
		CV	10	$3.051758 \cdot 10^{-5}$
December	B-Spline	GCV	26	0.0002441406
		CV	26	$7.629395 \cdot 10^{-6}$

Source: created by the author

Table 14: Optimal values of functional basis and penalty by cross-validation

The discrete tax receipt data are smoothed using B-spline basis with parameters displayed in Table 14 by GCV criteria. Since after smoothing the functional object is obtained, we can evaluate functional data object at specified argument values and join individual curves. 30 days were selected for estimation of functional data object. Figure 11 shows the pooled results of smoothed accumulated tax receipt data over 7 years. Each curve corresponds to a year $i, i = 2009, \ldots, 2016$ and a month $j, j = 1, \ldots, 12$, and the day variable was normalized to interval [0,1], where 0 corresponds to first day of the month and 1 to the last day of the month.



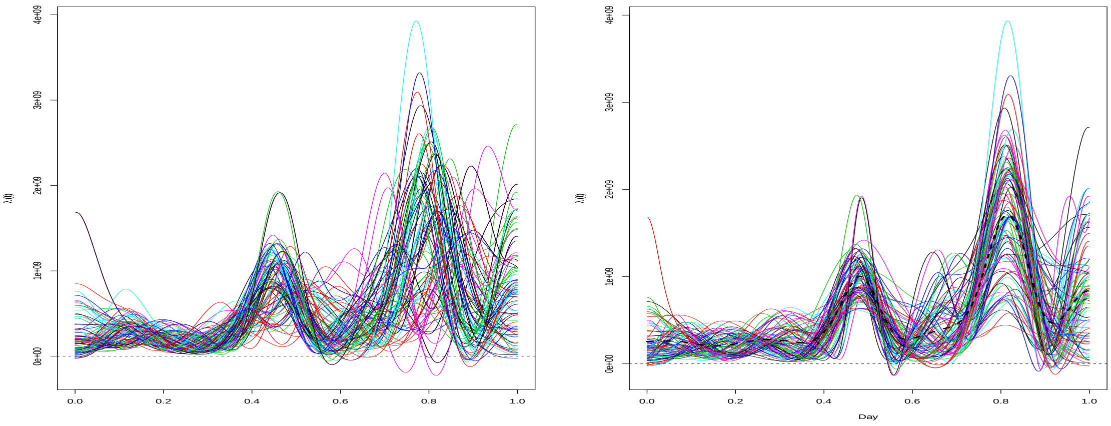
Source: created by the author

Figure 11: Smoothed accumulated tax receipts

Figure 12a represents the first derivative of accumulated tax receipt curves and it shows quite scattered curves around the day 25^{th}. But since month may have 28, 29, 30 or 31 day and in a smoothing process curves were transformed to have 30 model days, the curves are misaligned. In order to remove phase variation in the smoothed tax receipt data the registration process should be performed.

3.2.2 Registration

We have chosen to define landmarks as the positions of due dates of main taxes administrated by STI, namely due dates of submitting and paying value-added, personal income and excise taxes. Thus, the landmarks, corresponding to maximum at collecting the tax receipts, are extracted from the estimated profiles obtained with the smoothing procedure, and are matched with the reference landmarks defined by the average position of the increased intensity at this two days of the month. The results are shown in Figure 12b.



(a) Unregistrated curves

(b) Registrated curves

Source: created by the author

Figure 12: Tax receipt intensities

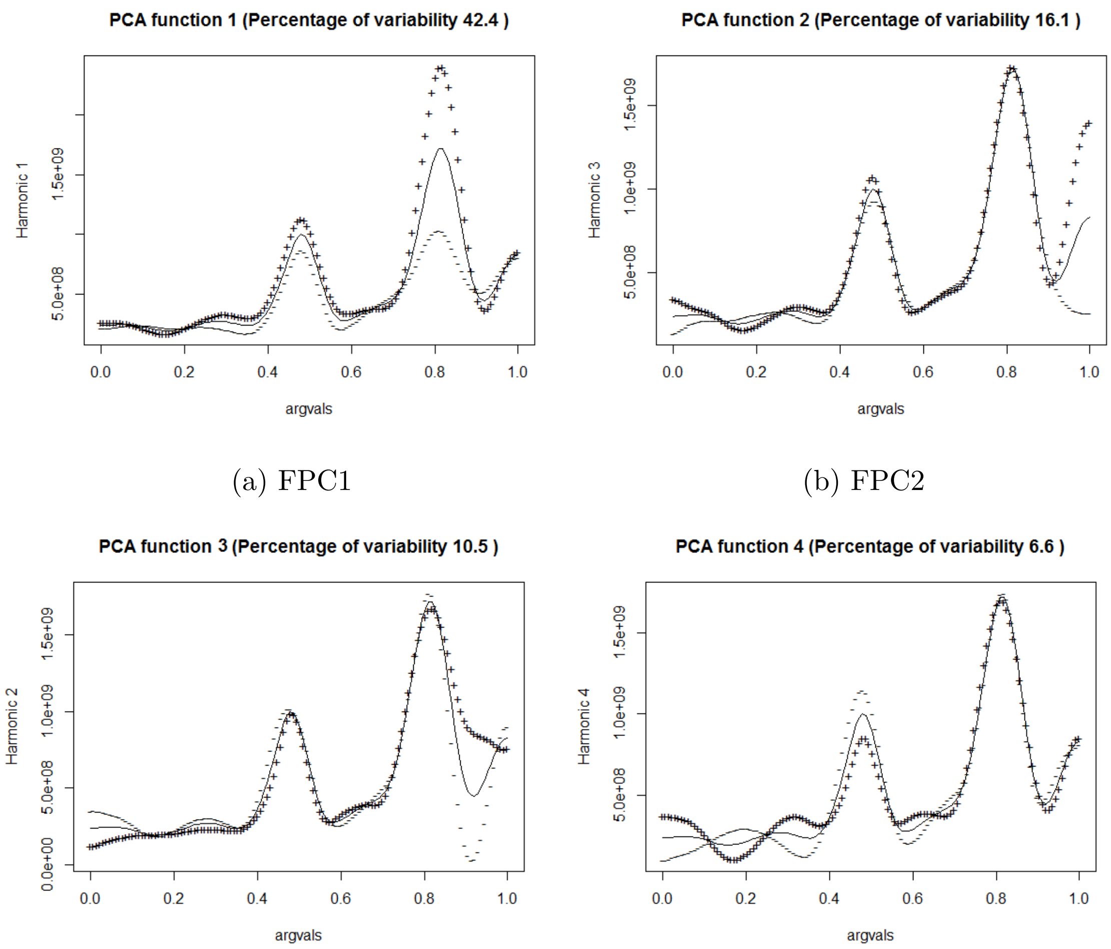
In order to quantify the amount of two types (amplitude and phase) of variation by comparing results for a sample of N functional observations before and after registration, MSE, which is determined in section 2.5, should be computed. After landmark registration of the tax receipts curves yields the value $R^2 = 0.357$. That is, nearly 36% of the variation in intensity over this period is due to phase. Other 64% of the variation is due to amplitude.

3.2.3 Functional principal component analysis

Functional principal component analysis (FPCA) is used to detect the variations in the amount of tax receipts recorded within State Tax Inspectorate among 7 years.

It is well known in classical multivariate analysis that an appropriate rotation of the principal components can, on occasion, give components of variability more informative than the original components themselves. A rotation method constructs new components based on the first k principal components, for some relatively small k. The idea is that k is chosen to include all the components that convey meaningful information, but not those that are just noise. In the present example, we concentrate on the first five components and set $k = 4$. The varimax method is often a useful approach. The method chooses components to maximize the variability of the squared principal component weights.

Figure 13 shows the four varimax-rotated FPCs by displaying the mean curve (solid line) along other two curves indicating the consequences of adding and subtracting a small amount of each principal component. Each of them takes account 42.4%, 16.1%, 10.5% and 6.6 % of the total variations, respectively. The first quantifies the variation corresponding to the period around the day 25^{th}, which is the due date of paying and submitting the value added tax return, and therefore captures the intensity of this event. The second and the third components measures the variability in the beginning and the end of the month, where several times a year is the deadline for paying and submitting the advance corporate income and corporate income tax returns. The fourth component indicates a mode of variability corresponding to the period around the day 15^{th}, which is the due date of paying and submitting the excise and personal income tax returns, and the period before this due date.



(a) FPC1

(b) FPC2

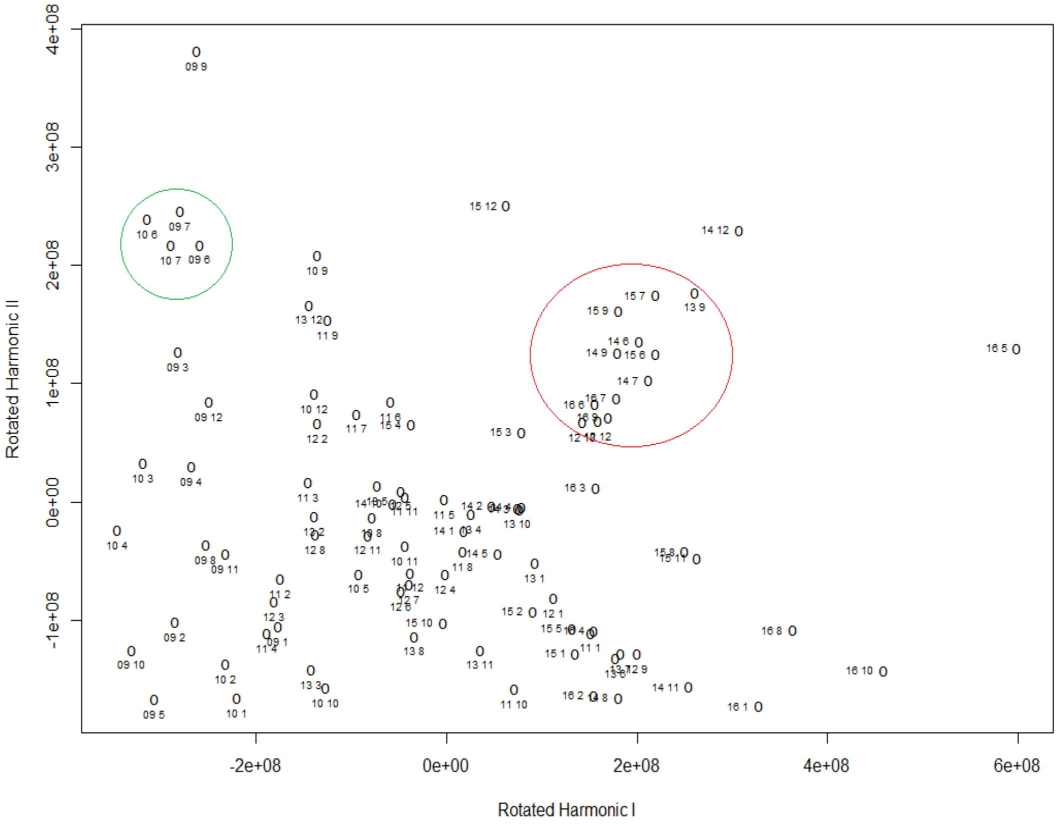
(c) FPC3

(d) FPC4

Source: created by the author

Figure 13: The rotated principal component functions

The score of each of the 94 months in the sample on these four principal components, by integrating the weight function against the functional datum in each case, which shows how curves cluster and otherwise distribute themselves within the K-dimensional subspace spanned by the eigenfunctions. This gives each month a score on each of the attributes. Figure 14 displays the scatterplot of PFC2 score versus PFC1 score for each month. First two numbers denote years and others correspond to months. For example, "15 2" labels 2015 February. 2009 September, 2012 December, 2014 December and 2016 May have extreme FPC1 or FPC2 scores, which means they have large difference among other months in sense of tax collection, and 2016 May has the largest FPC1 score, which indicates large amount of tax receipts in May. More precise, this increase is due to the increase of advance corporate income tax, which was collected 35.7 % more compared with the previous year. In addition, FPC scores moves obliquely from 2009 to 2016, which means that the amount of tax receipts increases over the years. Furthermore, there is an effect in June, July and September, which might be due to extra salary payments prior to the holidays and the number of vacations have increase over the 7 years.



Source: created by the author

Figure 14: The scores for the two rotated principal component functions

3.2.4 Functional linear models

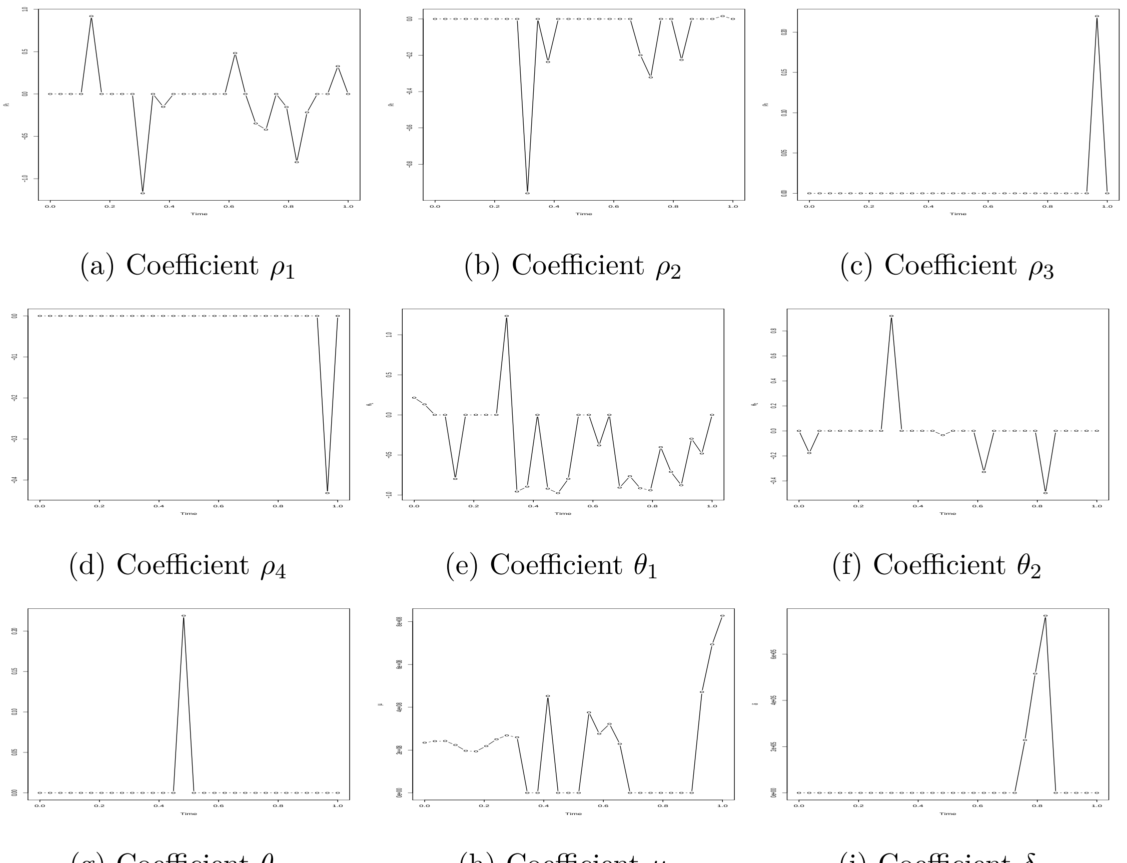
In this part of the section the main objective is to check whether coefficients from pointwise ARIMA(p,d,q) is constant. This will conclude if functional regression is appropriate type of method used to model tax receipts data. As was introduced earlier, curves were transformed to have 30 model days. Therefore thirty poitntwise ARIMA models will be conducted for each of 30 days and estimated coefficients will be presented to confirm or deny the raised hypothesis. The full model can be written as

$$y_t = c + \delta t + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \ldots + \rho_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t, \tag{86}$$

were p defines order of the autoregressive part, d - degree of first differencing involved, and q - order of the moving average part. In order to model observations the stationary tests were performed, the differences were taken if needed and the order of ARIMA(p,d,q) were selected using the Akaike information criterion (AIC), which is defined as

$$AIC = -2Log(\mathcal{L}) + 2(p + q + k + 1), \tag{87}$$

where $\mathcal{L}$ is the likelihood of the data, $k = 1$ if $c \neq 0$ and $k = 0$ if $c = 0$. In this case, it is assumed that sample follows a Gaussian distribution. Figure 15 and Appendix 5 represents estimated coefficients from pointwise ARIMA.



(a) Coefficient ρ_1

(b) Coefficient ρ_2

(c) Coefficient ρ_3

(d) Coefficient ρ_4

(e) Coefficient θ_1

(f) Coefficient θ_2

(g) Coefficient θ_3

(h) Coefficient μ

(i) Coefficient δ

Source: created by the author

Figure 15: Coefficient estimates of the pointwise ARIMA

Primarily, it can be assuredly assumed that the coefficients describing the model are not constant. In the beginning of the months, 13^{th}, 17^{th}, 18^{th}, 20^{th} and 30^{th} days tax receipts is explained by nothing else like the mean, which is different for every day.

There is an increasing tendency for the 23^{th}, 24^{th} and 25^{th} days, where the time coefficient is observed. These days correspond to due date of value-added tax return. It signifies the fact that the amount collected from VAT is increasing two days before due date and at the due day.

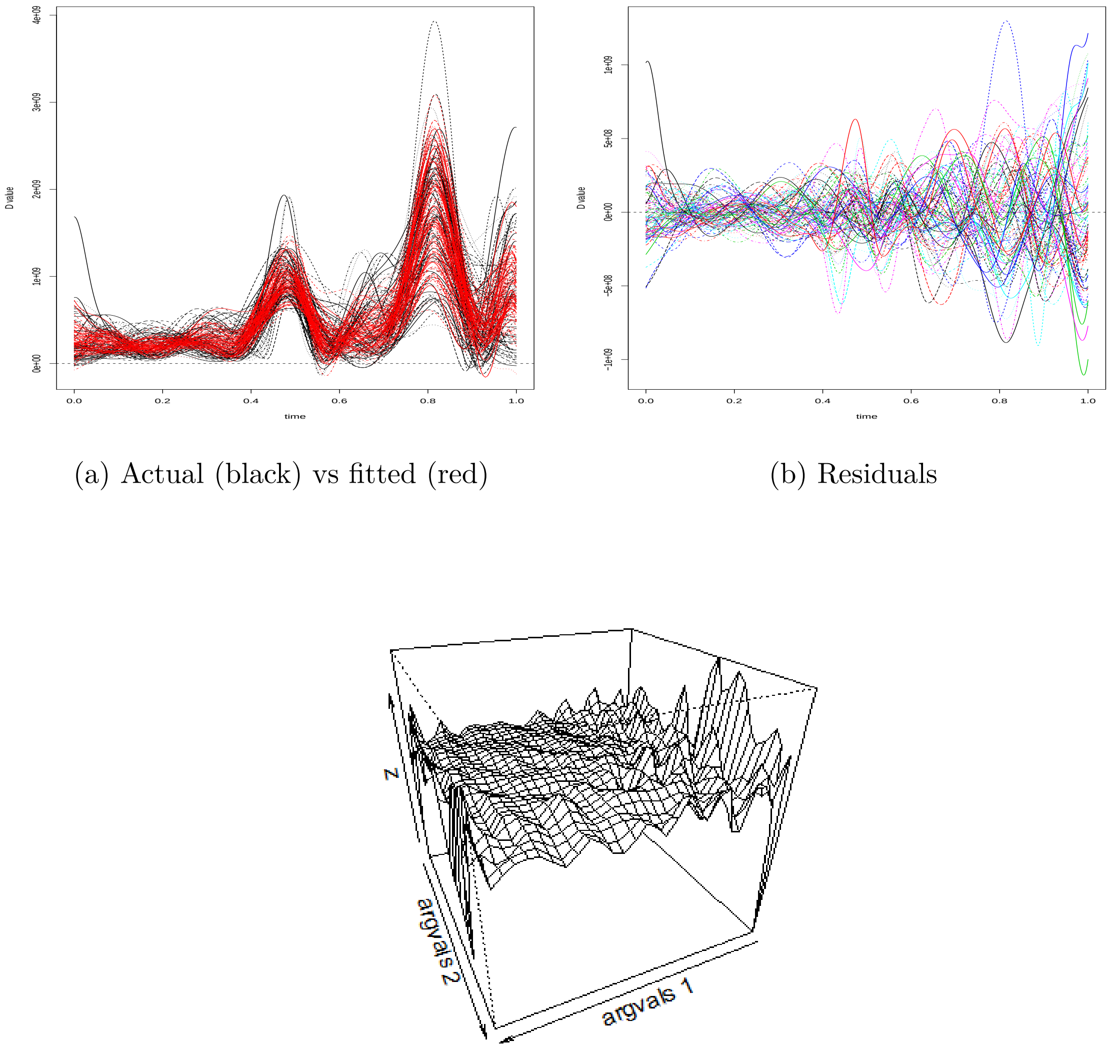
The amount of ax receipt collected in rest of the days are explained by integrated ARMA process, where order of the ARMA varies among days. Mostly the amount of tax receipts is explained by MA(1) or combination of autoregressive and moving average processes.

In order to ascertain the hypothesis consider the model

$$y_i(t) = \beta_0(t) + \int \beta_1(s,t) y_{i-1}(s) dt + \epsilon_i(t) \tag{88}$$

where $\beta_0(t)$ is non-random function that play the role of functional intercept, and $\beta_1(s,t)$ is non-random coefficient functions, the functional slopes.

Figure 31 represents the results of FAR(1) process. FAR(1) process modeled approximately 95% of the information, but more prominent curves have not been elucidated (see Figure 31). It is seen in the residuals of the FAR(1) process, where quite large degree of dispersion is observed in the beginning of the month, and increasing effect since the day 15. Consequently, model should be improved by adding more past values of the response variables or incorporate moving average process, or even include explanatory variables. But since the aim of this part is to ascertain that the coefficient of the FAR process isn't constant, so further analysis will not be conducted.



(a) Actual (black) vs fitted (red)

(b) Residuals

(c) Coefficient

Source: created by the author

Figure 16: FAR(1) model estimates

Accordingly Figure 31 represents the estimate of the first order autoregressive coefficient, which confirms the assumption that functional regression is adequate approach to model tax receipt data.

3.2.5 Further analysis

In the context of daily tax receipt data application, further analysis is also necessary to improve the efficiency. The data set used in this study consisted of 2009.01.01-2016.10.31 daily tax receipt data. Further research could analyze the effect of the other smoothing methods.

Moreover, the registration effect should be analyzed as only the landmark registration was performed. It might be that continuous registration would improve the results. As it was mentioned earlier, explanatory variables should be incorporated to the model and the order of ARMA process should be analyzed. That is

$$y_i(t) = \beta_0(t) + \gamma(s)\boldsymbol{Z}_i + \sum_{k=1}^{p} \int \beta_k(s,t) y_{i-k}(s) dt + \epsilon_i(t). \tag{89}$$

For example, one of the explaining variable could be the tax calendar. It should explained at least the some proportions of the peaks. Other variables depend on the structure of the nature as daily observations mostly probable only are available by Tax authority.

4 Conclusions

The common theme underlying the chapters of this thesis is the use of functional data analysis as a valuable tool for statistical modeling of tax collection data. While many theoretical economic models build upon smooth functions, tax collection data are usually observed discretely with some additional uninformative noise components. It is demonstrated that functional data analysis can bridge this gap in a most natural way. Consequently this thesis covers conceptual work, which shows the strength of functional data methods, when applied to economic contexts.

Regarding the conceptual part, municipal monthly budget revenue and daily tax receipt data are introduced. The tax collection data has been interpreted as noisy discretization points of smooth random functions. In contrast to classical time series models, this approach provides a much more convenient way for the development of statistical models, which are well-interpretable in the context of tax collection data. Yet separate objectives are promoted depending of the nature of the data.

The objective of the first application was to analyze the behavior of various forecasts under the Top-Down, Bottom-Up and introduced Middle-Up approaches in order to identify under which conditions one approach would be preferred instead of the other in terms of lower forecasting errors. However, at the moment of the analysis there were no available information of the amount of municipal budget revenue collected in December 2016. Thus preliminary analysis was performed. Practically, Top-Down approach appeared to be superior to Bottom-Up or Middle-Up approaches, since the latter had underestimated the preliminary amount for 2016 with assumption that 11.59% of the budget revenue will be collected in December 2016. Additionally, distance based clustering was performed in order to group municipalities for Middle-Up approach. Herewith the objective of the second application was to analyze the correctness of applying functional regression model for daily tax receipt data. After performing pointwise ARIMA, it was concluded that ARIMA coefficients are not constant. Therefore, functional data analysis is an appropriate tool for statistical modeling of tax receipt data. Additionally, in order to modify irregular time spacing registration has been performed. Moreover, in order to detect variations in the tax receipt data functional principal component analysis has been accomplished.

Bibliography

[1] Andrieu, C., Saint Pierre, G. and Bressaud, X. (2013) A functional analysis of speed profiles: smoothing using derivative information, curve registration, and functional boxplot. <`https://hal.archives-ouvertes.fr/hal-00915475/document`> (accessed: December, 31, 2016)

[2] Belitsera, E., Paulo Serrab, P., and Zantenc, H. (2013) Rate-optimal Bayesian intensity smoothing for inhomogeneous Poisson processes. *Journal of Statistical Planning and Inference*, 166: 24-35.

[3] Cardot, H., Ferraty, F., Sarda, P. (1999) Functional Linear Model. *Statistics & Probability Letters*, 45(1), 11–22.

[4] Cho, H., Goude, Y. ,Brossat, X. and Yao, Q. (2012) Modeling and Forecasting Daily Electricity Load Curves: A Hybrid Approach. *Journal of the American Statistical Association*, 108(501): 7-21.

[5] Cuevasa, A., Febrerob, M. and Fraimanc, R. (2004) An anova test for functional data. *Computational Statistics & Data Analysis*, 47: 111-122.

[6] Cuevas, A., Febrero-Bande, M., Fraiman R. (2007) Robust Estimation and Classification for Functional Data via Projection-Based Depth Notions. *Computational Statistics*, 22(3), 481–496.

[7] Febrero-Bande, M. and Fuente, M.O. (2012) Statistical Computing in Functional Data Analysis: The R Package fda.usc. *Journal of Statistical Software.* 51(4): 1-28.

[8] Ferraty, F., Vieu, P. (2006) *Nonparametric Functional Data Analysis.* Springer Series in Statistics. Springer-Velag, New York. Theory and practice.

[9] Fraiman, R., Muniz, G. (2001) Trimmed Means for Functional Data. *Test*, 10(2), 419–440.

[10] Górecki, T. and Smaga, L. (2015) A comparison of tests for the one-way ANOVA problem for functional data. *Computational Statistics*, 30(4): 987-1010.

[11] Horváth, L., Kokoszka, P. (2012) *Inference for Functional Data with Applications.* Springer Series in Statistics. Springer-Verlag, New York.

[12] Kneip, A. and Ramsay, J.O. (2008) Combining registration and fitting for functional models. *Journal of the American Statistical Association* , 20, 1266–1305.

[13] Kosiorowski, D. (2014) Functional regression in short-term prediction of economic time series. *Statistics in Transition.* 15(4): 611-626.

[14] Kuhl, M., Damerdji, H. and Wilson, J. (1998) Least squares estimation of nonhomogeneous Poisson processes. In the Proceedings of the 1998 Winter Simulation Conference, Piscataway, New Jersey, 637-645.

[15] Lietuvos Respublikos biudžeto sandaros įstatymas. Valstybės žinios, 1990-08-31, No. 24-596.

[16] Lillestøl, J. and Ollmar, F. (2003) Functional data analysis: Introduction and applications to financial electricity contracts. <`https://brage.bibsys.no/xmlui/bitstream/handle/11250/163761/lillestol%20jostein%200603.pdf?sequence=1`> (accessed: December, 31, 2016)

[17] Lobato, I., and Velasco, C. (2004) A Simple and General Test for White Noise. *Econometric Society* , Latin-America Meetings, Paper No. 112.

[18] Pillow, J.W. (2009) Time-rescaling methods for the estimation and assessment of non-Poisson neural encoding models. *Advances in Neural Information Processing Systems* , 22: 1473-1481.

[19] Ramsay, J.O. and Li, X. (1998) Curve registration. *Journal of the Royal Statistical Society*, 60(2): 351-363.

[20] Ramsay, J.O., Silverman, B.W. (2005) *Applied Functional Data Analysis: Methods and case studies.* Springer Series in Statistics. Springer-Verlag, New York.

[21] Ramsay, J. O., Hooker, G., Graves, S. (2009) *Functional Data Analysis with R and MATLAB.* Springer Series in Statistics. Springer-Verlag, New York.

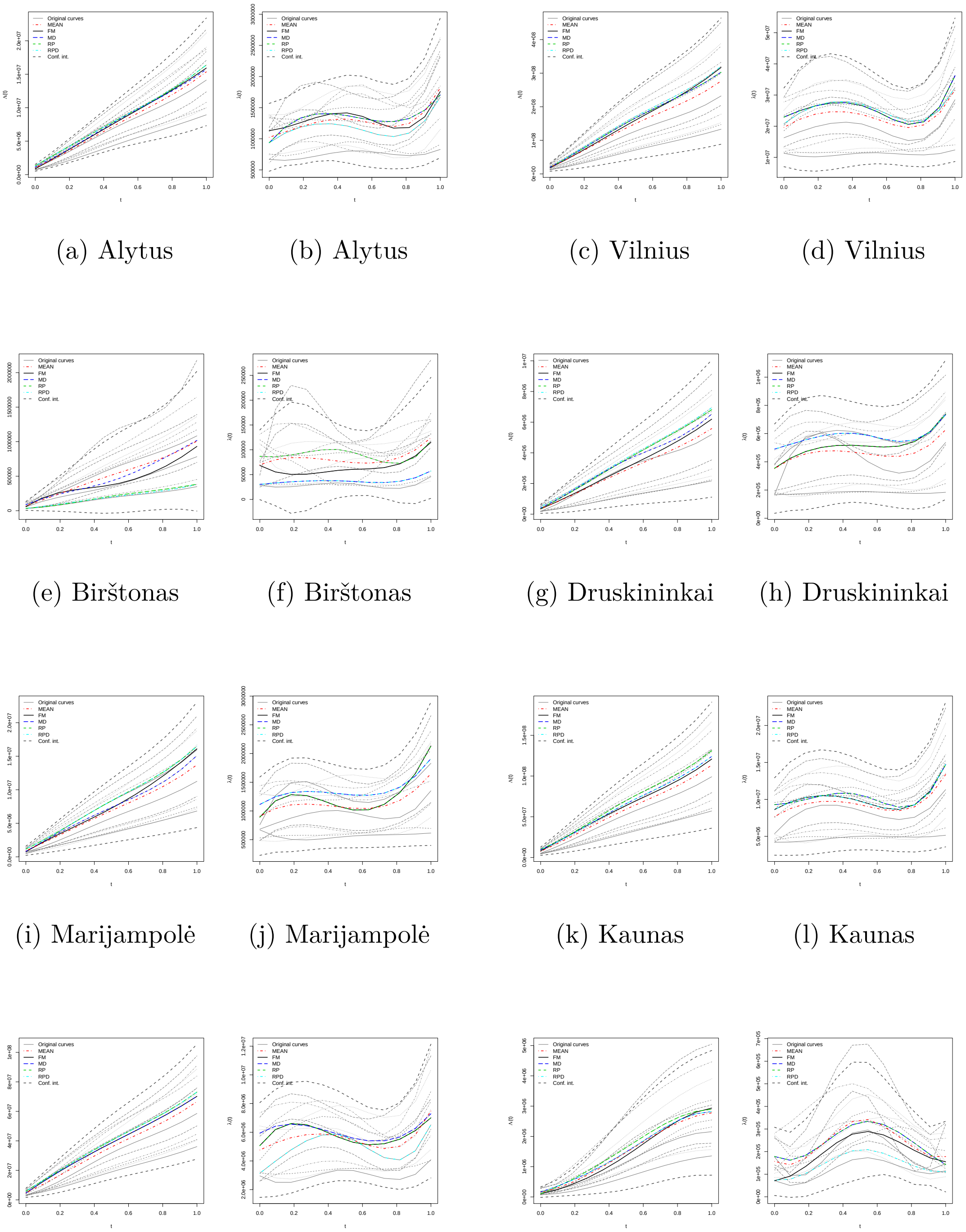
[22] Ramsay, J. O., Gribble, P. and Kurtek, S. (2014) Analysis of juggling data: Landmark and continuous registration of juggling trajectories. *Electronic Journal of Statistics*, 8: 1835–1841.

[23] Marron, J.S., Ramsay, J.O., Sangalli, L.M. and Srivastava, A. (2015) Functional Data Analysis of Amplitude and Phase Variation. *Statistical Science*, 30(4): 468–484.

[24] Vantini, S. (2012) On the definition of phase and amplitude variability in functional data analysis. *Test*, 21(4): 676–696.

[25] Zhang, Jin-Ting (2013) *Analysis of Variance for Functional Data.* Chapman & Hall Book. Taylor & Francis Group, Florida.

1 Appendix



(a) Alytus (b) Alytus (c) Vilnius (d) Vilnius

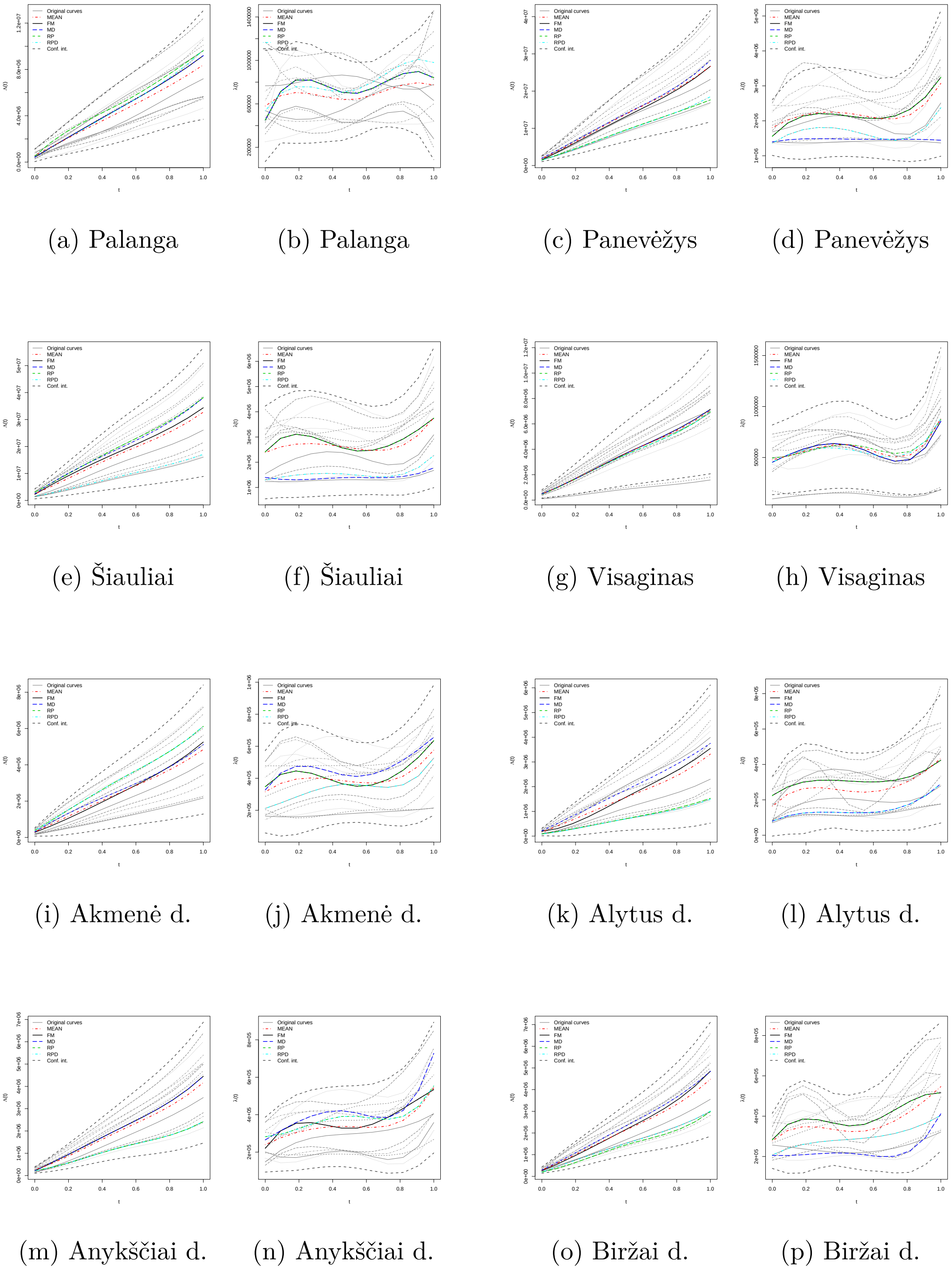
(e) Birštonas (f) Birštonas (g) Druskininkai (h) Druskininkai

(i) Marijampolė (j) Marijampolė (k) Kaunas (l) Kaunas

(m) Klaipėda (n) Klaipėda (o) Neringa (p) Neringa

Source: created by the author

Figure 17: The amount of budget revenue among municipalities and the intensities of revenue, and descriptive statistics



(a) Palanga (b) Palanga (c) Panevėžys (d) Panevėžys

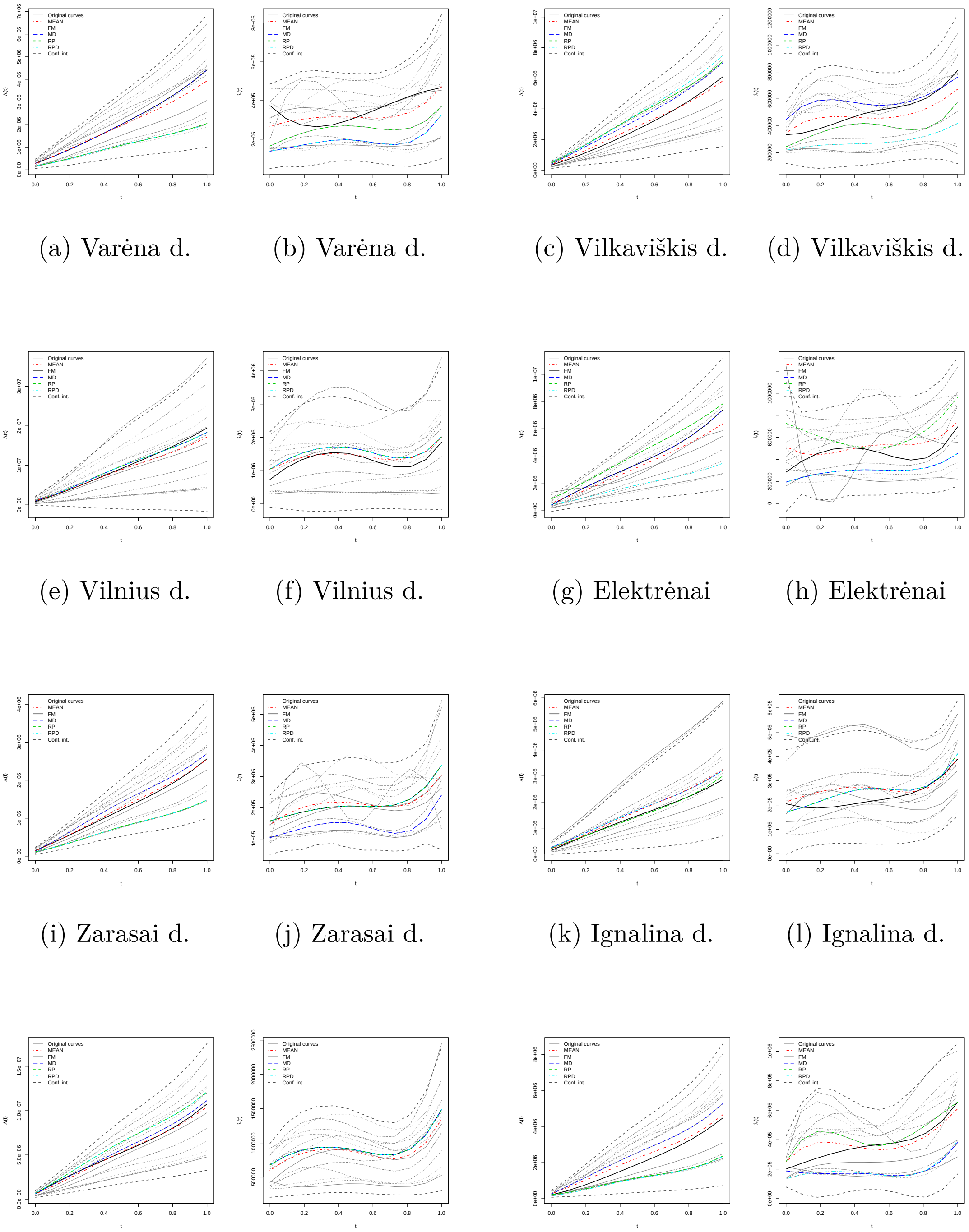
(e) Šiauliai (f) Šiauliai (g) Visaginas (h) Visaginas

(i) Akmenė d. (j) Akmenė d. (k) Alytus d. (l) Alytus d.

(m) Anykščiai d. (n) Anykščiai d. (o) Biržai d. (p) Biržai d.

Source: created by the author

Figure 18: The amount of budget revenue among municipalities and the intensities of revenue, and descriptive statistics



(a) Varėna d.

(b) Varėna d.

(c) Vilkaviškis d.

(d) Vilkaviškis d.

(e) Vilnius d.

(f) Vilnius d.

(g) Elektrėnai

(h) Elektrėnai

(i) Zarasai d.

(j) Zarasai d.

(k) Ignalina d.

(l) Ignalina d.

(m) Jonava d.

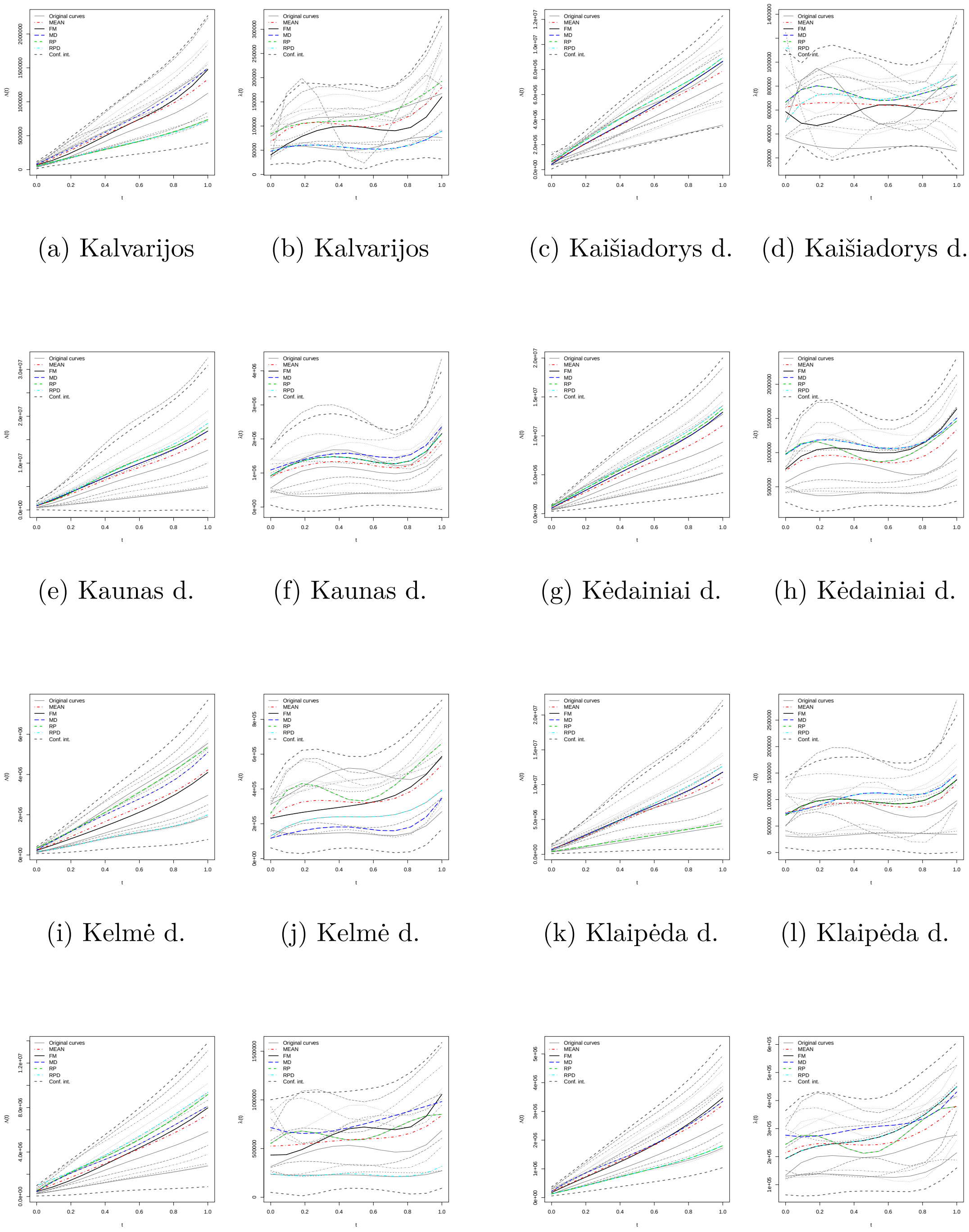
(n) Jonava d.

(o) Joniškis d.

(p) Joniškis d.

Source: created by the author

Figure 19: The amount of budget revenue among municipalities and the intensities of revenue, and descriptive statistics



(a) Kalvarijos (b) Kalvarijos (c) Kaišiadorys d. (d) Kaišiadorys d.

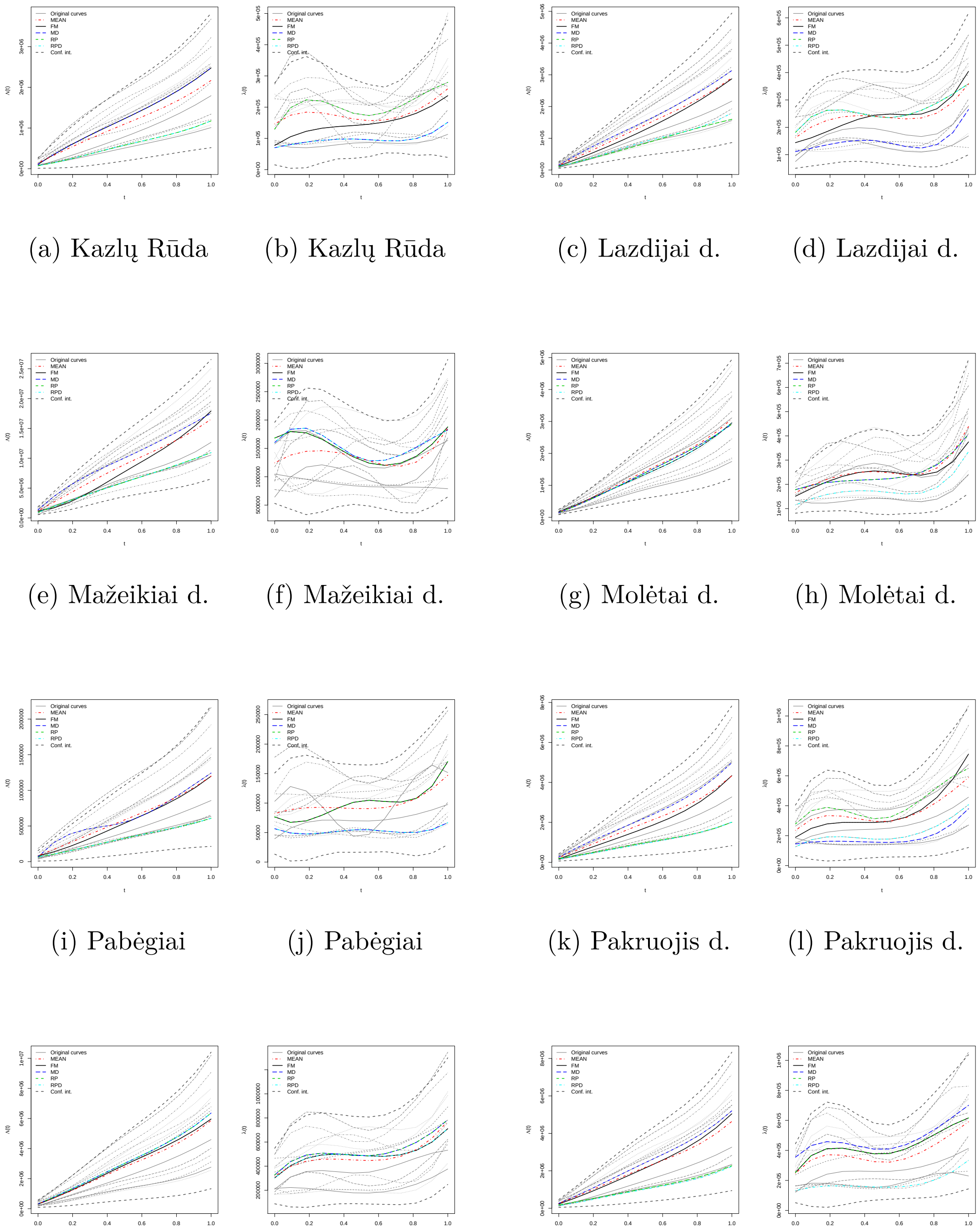
(e) Kaunas d. (f) Kaunas d. (g) Kėdainiai d. (h) Kėdainiai d.

(i) Kelmė d. (j) Kelmė d. (k) Klaipėda d. (l) Klaipėda d.

(m) Kretinga d. (n) Kretinga d. (o) Kupiškis d. (p) Kupiškis d.

Source: created by the author

Figure 20: The amount of budget revenue among municipalities and the intensities of revenue, and descriptive statistics



(a) Kazlų Rūda (b) Kazlų Rūda (c) Lazdijai d. (d) Lazdijai d.

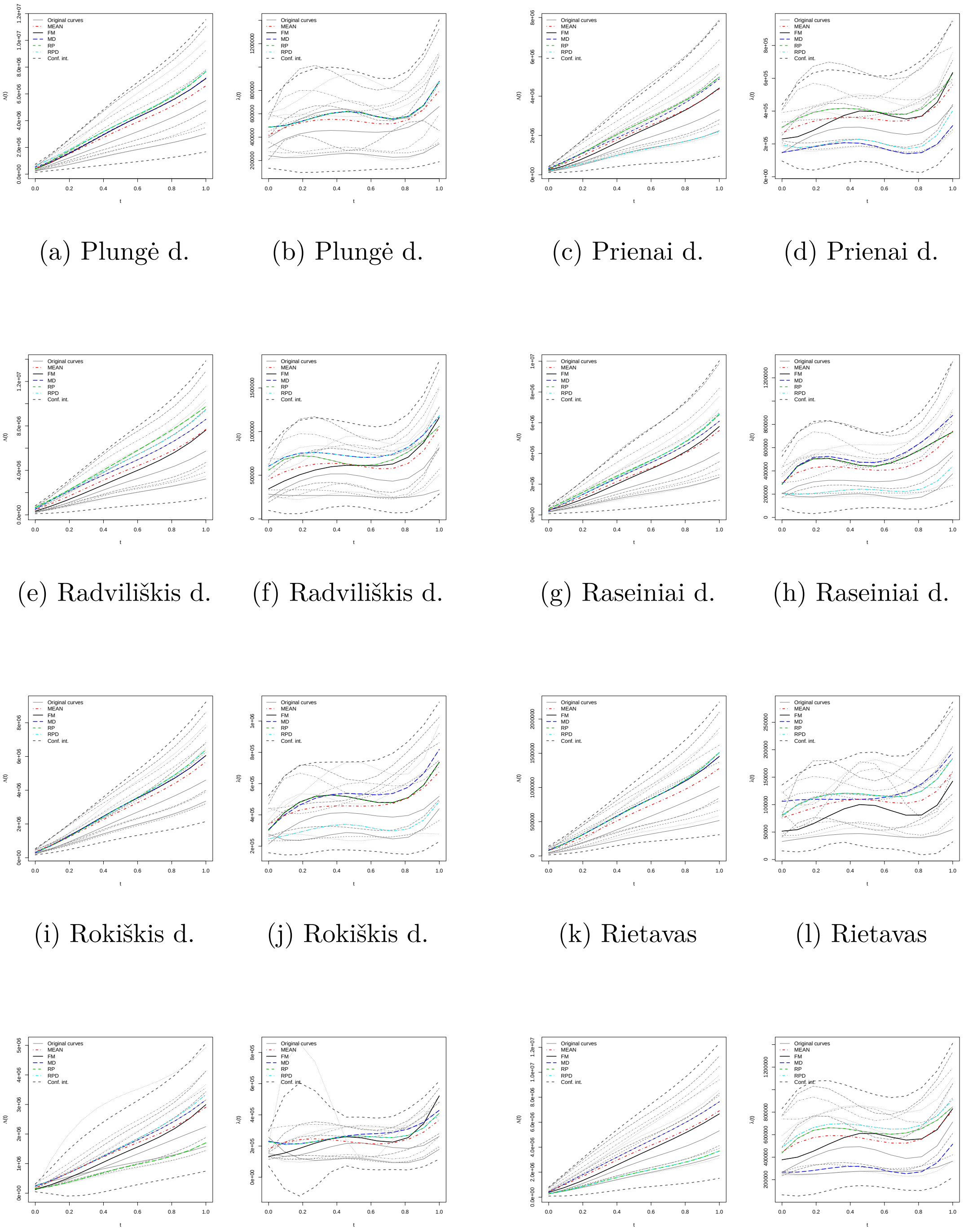
(e) Mažeikiai d. (f) Mažeikiai d. (g) Molėtai d. (h) Molėtai d.

(i) Pabėgiai (j) Pabėgiai (k) Pakruojis d. (l) Pakruojis d.

(m) Panevėžys d. (n) Panevėžys d. (o) Pasvalys d. (p) Pasvalys d.

Source: created by the author

Figure 21: The amount of budget revenue among municipalities and the intensities of revenue, and descriptive statistics



(a) Plungė d. (b) Plungė d. (c) Prienai d. (d) Prienai d.

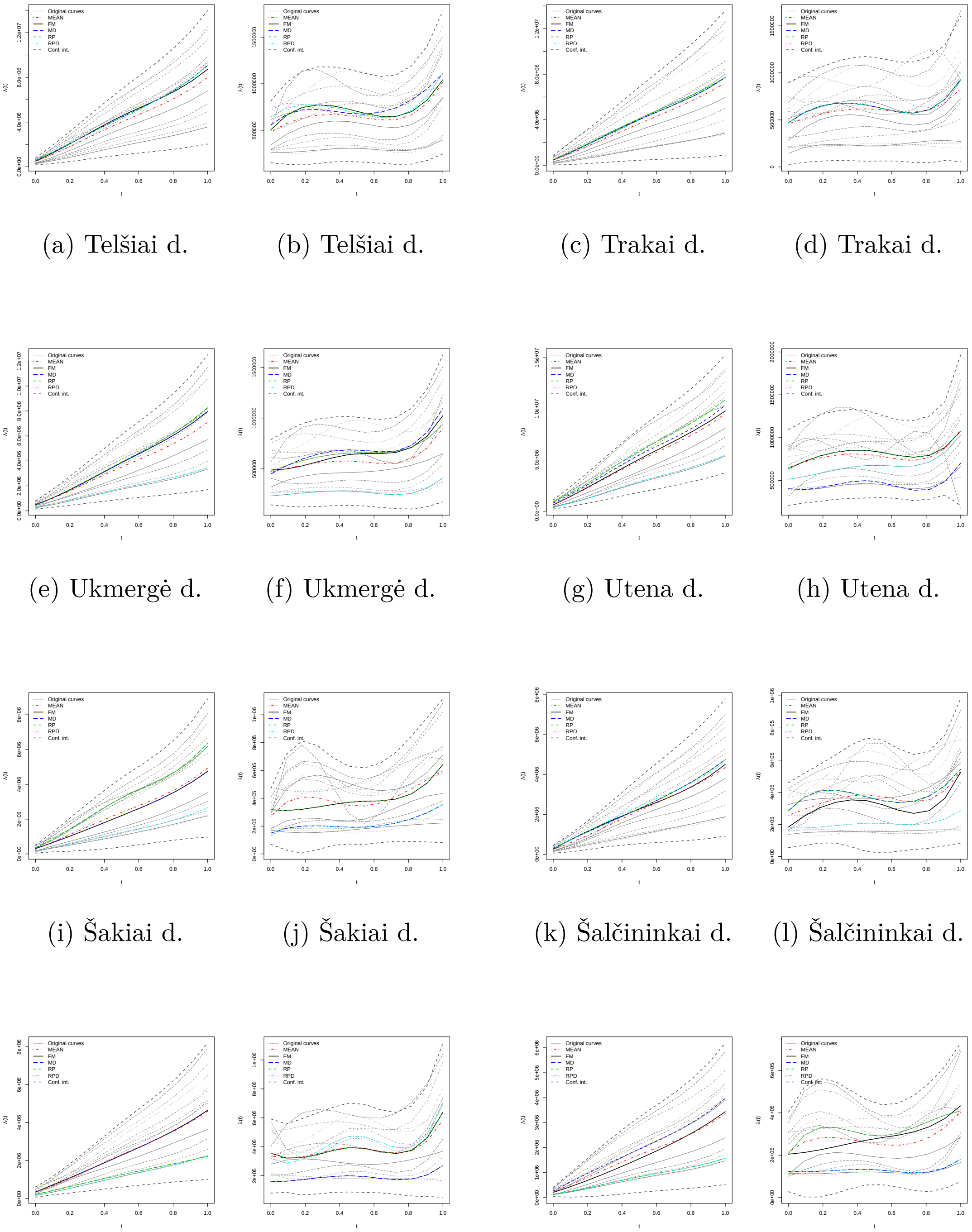
(e) Radviliškis d. (f) Radviliškis d. (g) Raseiniai d. (h) Raseiniai d.

(i) Rokiškis d. (j) Rokiškis d. (k) Rietavas (l) Rietavas

(m) Skuodas d. (n) Skuodas d. (o) Tauragė d. (p) Tauragė d.

Source: created by the author

Figure 22: The amount of budget revenue among municipalities and the intensities of revenue, and descriptive statistics



(a) Telšiai d. (b) Telšiai d. (c) Trakai d. (d) Trakai d.

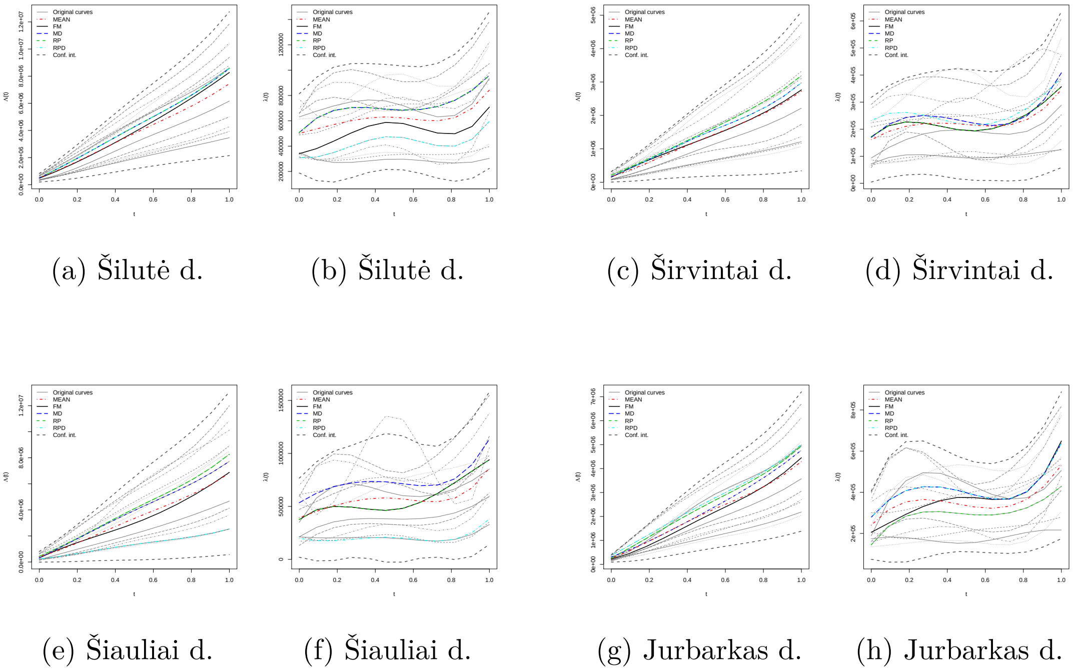
(e) Ukmergė d. (f) Ukmergė d. (g) Utena d. (h) Utena d.

(i) Šakiai d. (j) Šakiai d. (k) Šalčininkai d. (l) Šalčininkai d.

(m) Švenčionys d. (n) Švenčionys d. (o) Šilalė d. (p) Šilalė d.

Source: created by the author

Figure 23: The amount of budget revenue among municipalities and the intensities of revenue, and descriptive statistics



(a) Šilutė d.

(b) Šilutė d.

(c) Širvintai d.

(d) Širvintai d.

(e) Šiauliai d.

(f) Šiauliai d.

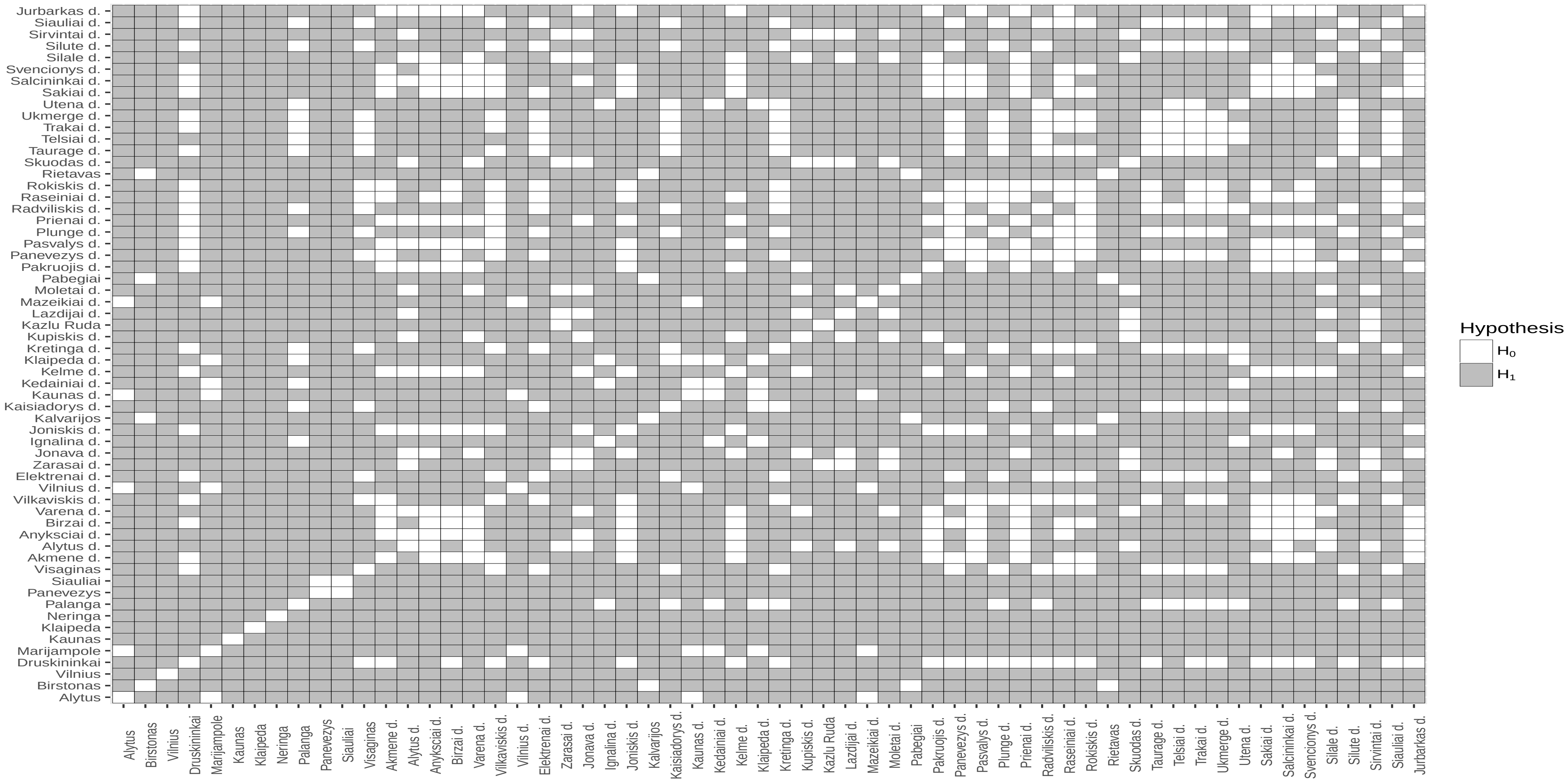
(g) Jurbarkas d.

(h) Jurbarkas d.

Source: created by the author

Figure 24: The amount of budget revenue among municipalities and the intensities of revenue, and descriptive statistics

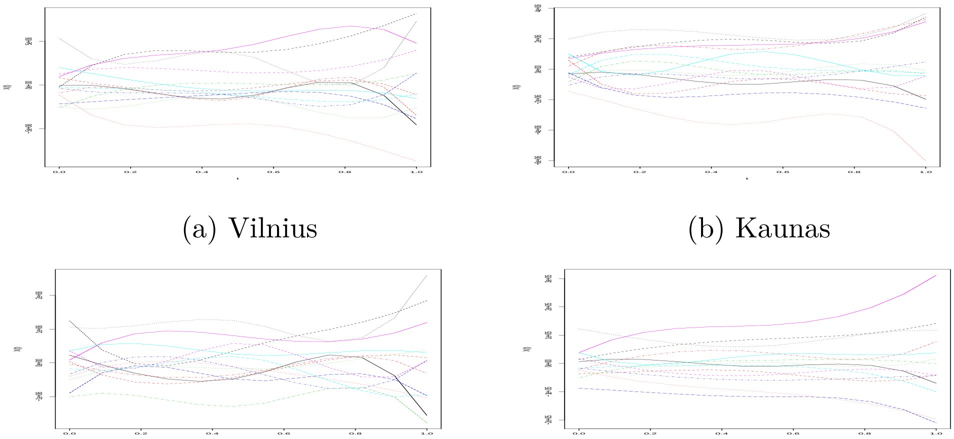
2 Appendix



Source: created by the author

Figure 25: Pairwise functional ANOVA tests

3 Appendix



(a) Vilnius

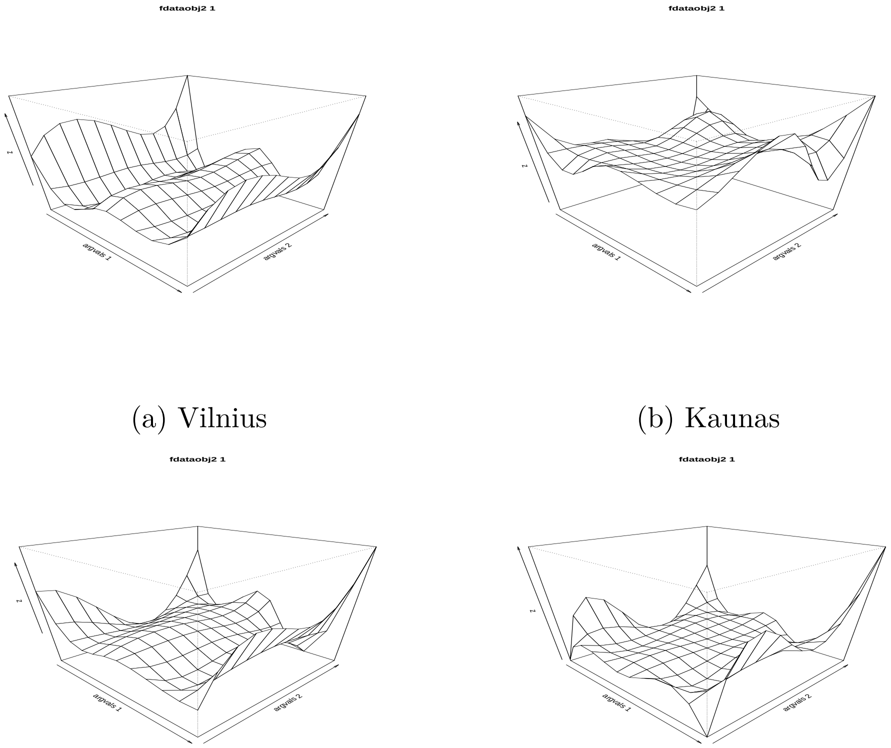
(b) Kaunas

(c) Klaipėda

(d) Rest

Source: created by the author

Figure 26: Residuals of {4} model



(a) Vilnius

(b) Kaunas

(c) Klaipėda

(d) Rest

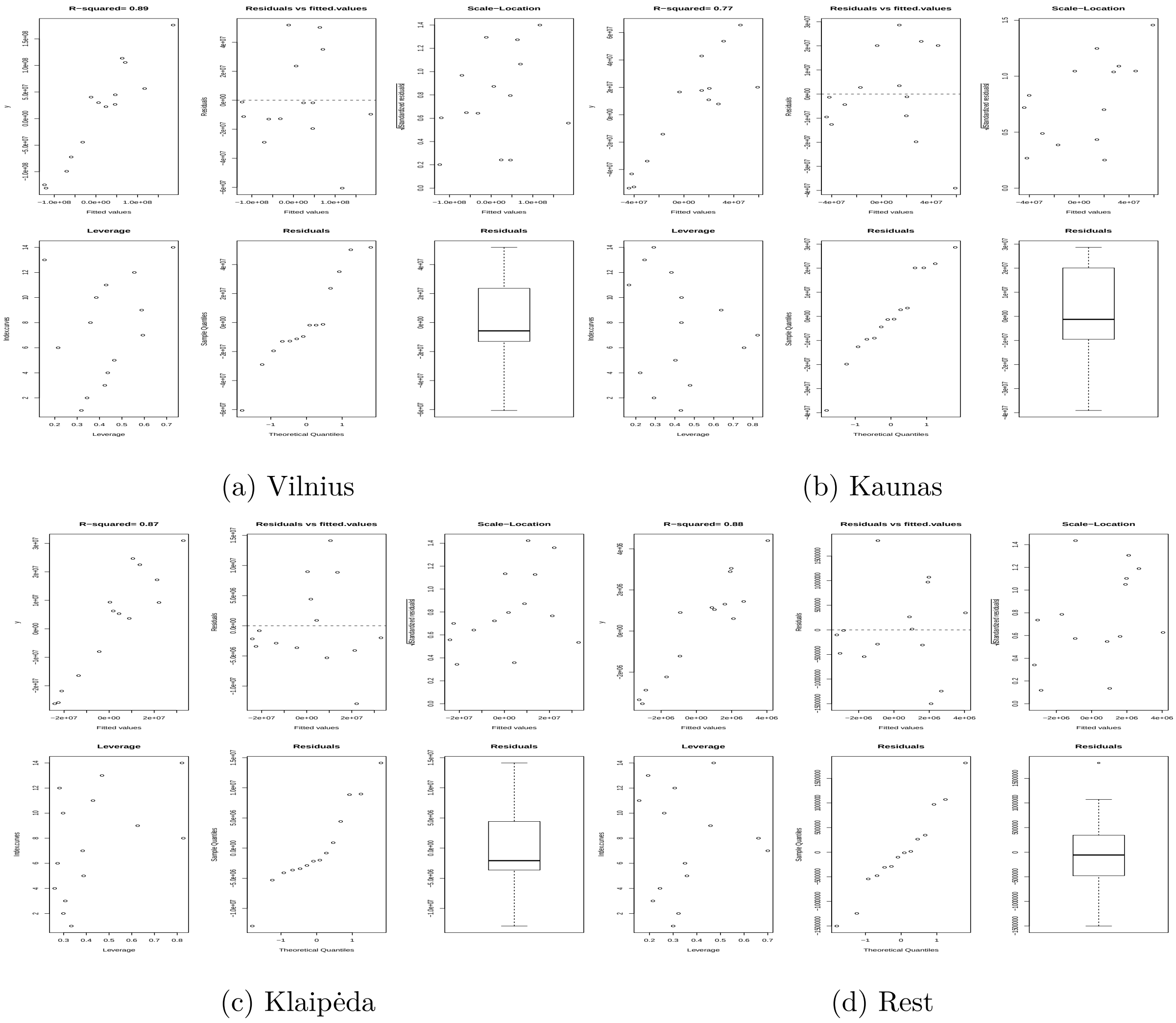
Source: created by the author

Figure 27: Coefficient estimates of {4} model

	Cluster:			
	Vilnius	Kaunas	Klaipėda	Rest
const	10321458.8	4044218.92	2183576.70	256378.23
	(1054222)	(626767)	(237816)	(29995)
ϕ_1	315.5	-31.00	49.11	54.12
	(196.0)	(185.42)	(93.33)	(203.46)
ϕ_2	-508.0	38.81	-70.84	-45.34
	(279.0)	(210.39)	(121.40)	(252.35)
ϕ_3	612.1^{*}	-33.56	181.76	70.17
	(314.3)	(239.38)	(158.31)	(355.17)
ϕ_4	-341.5^{*}	33.33	-163.60	-107.94
	(150.8)	(192.04)	(116.68)	(312.44)
ϕ_5	263.9^{**}	34.43	124.48	178.30
	(105.2)	(162.31)	(74.82)	(208.79)
R^2	0.89	0.87	0.88	0.88
Adjusted R^2	0.82	0.78	0.82	0.8

Note: *p<0.1; **p<0.05; ***p<0.01

Table 15: Estimates of {1} model



(a) Vilnius

(b) Kaunas

(c) Klaipėda

(d) Rest

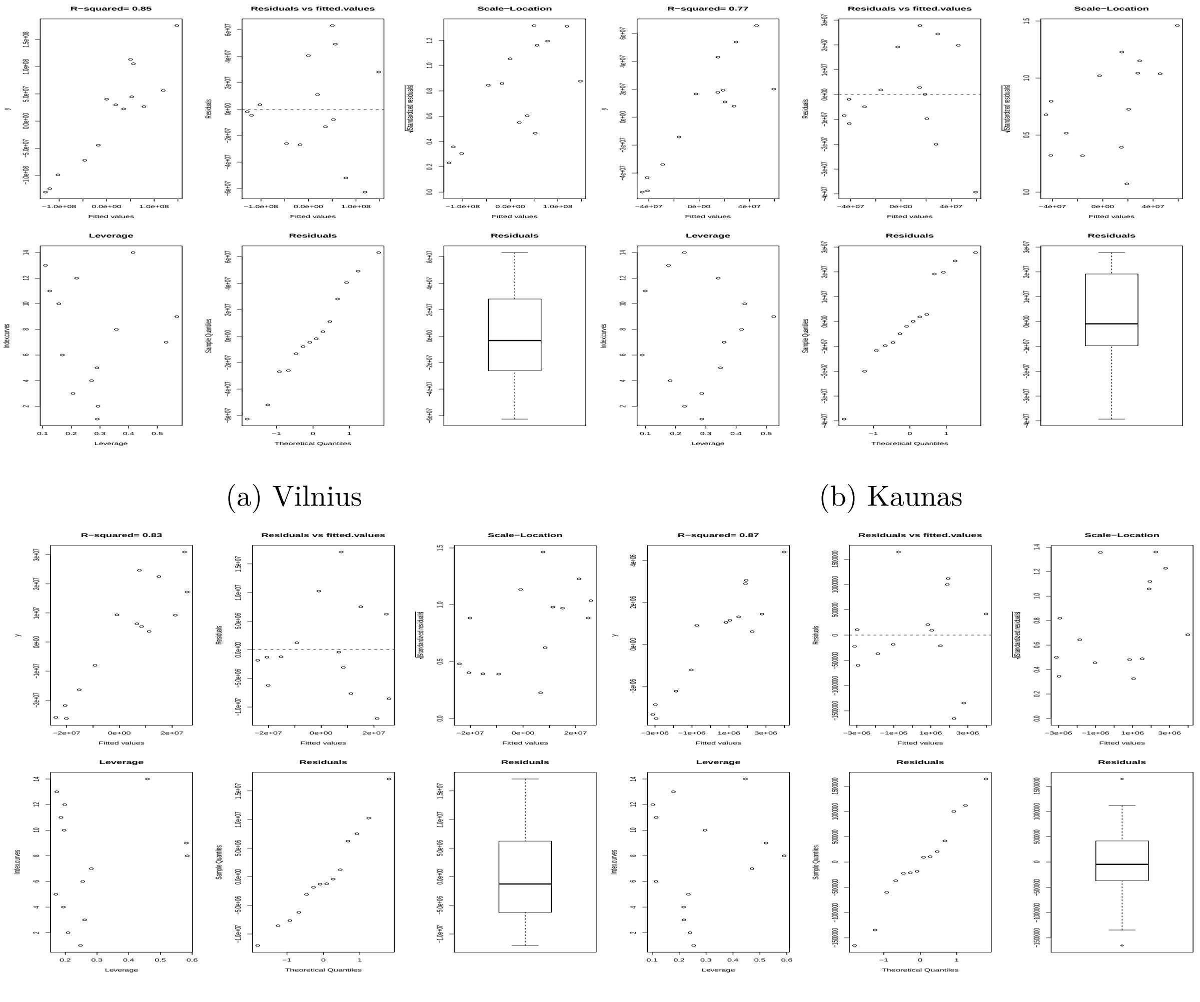
Source: created by the author

Figure 28: Adequacy results of {1} model

	Cluster:			
	Vilnius	Kaunas	Klaipėda	Rest
const	$1.032 \cdot 10^{7}$	$4.044 \cdot 10^{6}$	$2.184 \cdot 10^{6}$	$2.564 \cdot 10^{5}$
	$(1.111 \cdot 10^{5})$	$(5.615 \cdot 10^{4})$	$(2.426 \cdot 10^{4})$	$(2.718 \cdot 10^{3})$
PC1	17.06^{***}	11.96^{***}	11.36^{***}	14.82^{***}
	(3.53)	(2.181)	(2.392)	(2.257)
PC2	9.323	13.26	4.644	13.73
	(13.35)	(22.68)	(9.166)	(16.57)
PC3	32.52	−5.623	−1.421	25.64
	(19.72)	(22.78)	(11.70)	(15.44)
R^2	0.85	0.77	0.83	0.87
Adjusted R^2	0.8	0.7	0.77	0.84
Variability explained by (%):				
PC1	92.22	92.47	88.14	93.00
PC2	1.62	0.96	3.57	1.06
PC3	3.12	0.81	3.45	1.57

Note: *p<0.1; **p<0.05; ***p<0.01

Table 16: Estimates of {2} model



(a) Vilnius

(b) Kaunas

(c) Klaipėda

(d) Rest

Source: created by the author

Figure 29: Adequacy results of {2} model

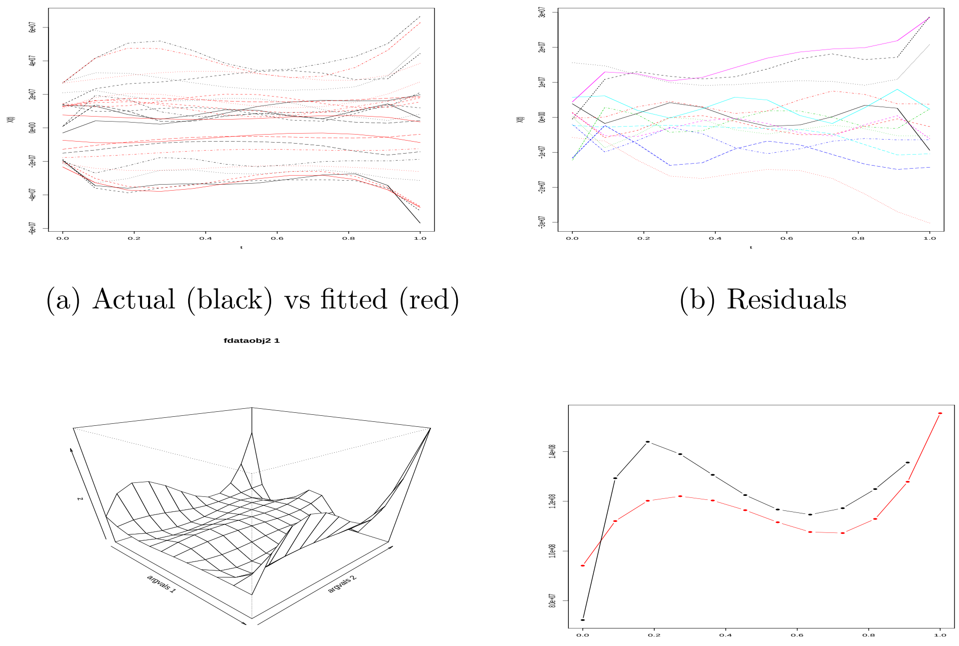
	Cluster:			
	Vilnius	Kaunas	Klaipėda	Rest
const	$-1.586 \cdot 10^{8}$	$-1.469 \cdot 10^{8***}$	$-4.859 \cdot 10^{7}$	$-1.610 \cdot 10^{6}$
	$(19.211 \cdot 10^{6})$	$(2.942 \cdot 10^{6})$	$(3.186 \cdot 10^{5})$	$(2.913 \cdot 10^{4})$
ϕ_1	122	-88.85	-5.38	21.07
	(199.8)	(97.83)	(68.67)	(135.1)
ϕ_2	-277.6	288.3^{**}	-75.13	175.6
	(284.4)	(105.1)	(98.69)	(188.1)
ϕ_3	393	-520.3^{***}	180	-348.6
	(312.2)	(135.4)	(145.2)	(283.9)
ϕ_4	-232	361.6^{**}	-158.8	241.2
	(150.8)	(105.8)	(104.9)	(241.3)
ϕ_5	145.1	-196.3^{*}	110.9	-110
	(113.3)	(82.72)	(71.87)	(164.6)
GDP	$1.883 \cdot 10^{-2}$	$3.542 \cdot 10^{-2***}$	$1.736 \cdot 10^{-2**}$	$1.538 \cdot 10^{-3}$
	$(9.350 \cdot 10^{-3})$	$(6.043 \cdot 10^{-3})$	$(7.955 \cdot 10^{-3})$	$(8.698 \cdot 10^{-4})$
Unemployed	$-8.031 \cdot 10^{2}$	$-1.758 \cdot 10^{3**}$	$-2.672 \cdot 10^{2}$	$-1.197 \cdot 10^{3**}$
	$(1.024 \cdot 10^{2})$	$(6.053 \cdot 10^{2})$	$(1.121 \cdot 10^{3})$	$(4.517 \cdot 10^{2})$
R^2	0.93	0.87	0.96	0.96
Adjusted R^2	0.88	0.78	0.9	0.91
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 17: Estimates of {3} model

	Municipality	True value of 2016.01-2016.11	Annual forecast with {3} model	Annual forecast with {1} model	Annual forecast with {2} model
1	Alytus	20, 704, 200	22, 441, 235	22, 313, 172	22, 185, 102
2	Birštonas	1, 966, 300	1, 488, 671	1, 510, 925	2, 223, 529
3	Vilnius	451, 430, 000	428, 569, 406	400, 569, 307	460, 298, 614
4	Druskininkai	9, 433, 700	10, 065, 782	9, 722, 715	10, 015, 202
5	Marijampolė	20, 423, 100	22, 098, 742	22, 214, 727	21, 775, 278
6	Kaunas	178, 444, 700	202, 496, 577	186, 838, 140	186, 575, 974
7	Klaipėda	94, 406, 600	93, 243, 436	93, 353, 504	97, 957, 956
8	Neringa	6, 102, 600	3, 099, 976	4, 042, 267	5, 402, 324
9	Palanga	13, 490, 700	12, 420, 424	12, 318, 353	12, 541, 600
10	Panevėžys	38, 484, 400	53, 742, 907	44, 649, 528	42, 621, 120
11	Šiauliai	47, 287, 000	56, 186, 029	51, 083, 770	50, 533, 358
12	Visaginas	8, 127, 500	9, 225, 458	9, 508, 311	8, 788, 972
13	Akmenė d.	6, 786, 500	7, 804, 330	6, 831, 299	7, 212, 885
14	Alytus d.	5, 611, 400	5, 667, 414	6, 123, 543	6, 294, 668
15	Anykščiai d.	6, 215, 500	6, 863, 732	7, 578, 451	6, 670, 273
16	Biržai d.	6, 367, 900	7, 157, 841	7, 225, 651	6, 709, 926
17	Varėna d.	6, 058, 100	6, 645, 173	6, 348, 743	6, 664, 077
18	Vilkaviškis d.	8, 616, 200	10, 087, 021	9, 878, 806	9, 397, 068
19	Vilnius d.	35, 739, 300	34, 831, 933	39, 201, 318	39, 771, 879
20	Elektrėnai d.	10, 452, 000	11, 751, 836	10, 335, 443	10, 289, 090
21	Zarasai d.	3, 335, 400	3, 696, 254	3, 966, 051	3, 994, 185
22	Jonava d.	3, 520, 700	3, 357, 182	3, 066, 119	3, 747, 148
23	Ignalina d.	14, 664, 300	18, 169, 015	18, 507, 828	18, 662, 625
24	Joniškis d.	7, 161, 500	8, 043, 814	9, 647, 684	8, 930, 689
25	Kalvarijos	2, 083, 600	2, 212, 257	2, 243, 456	2, 272, 537
26	Kaišiadorys d.	10, 010, 000	12, 273, 790	12, 683, 963	11, 301, 352
27	Kaunas d.	30, 056, 700	47, 662, 225	44, 983, 609	34, 457, 000
28	Kėdainiai d.	18, 382, 100	21, 079, 934.	20, 423, 539	20, 619, 839
29	Kelmė d.	6, 275, 900	7, 418, 097	7, 689, 263	7, 586, 186
30	Klaipėda d.	21, 062, 400	25, 685, 608	25, 082, 459	25, 139, 512
31	Kretinga d.	13, 099, 800	13, 829, 486	14, 381, 450	14, 592, 528
32	Kupiškis d.	4, 835, 700	5, 033, 779	4, 930, 235	5, 024, 901
33	Kazlų Rūda	3, 140, 700	3, 488, 856	3, 847, 098	4, 043, 392
34	Lazdijai d.	4, 296, 300	4, 647, 354	4, 547, 950	4, 535, 702
35	Mažeikiai d.	22, 399, 100	23, 064, 642	24, 148, 214	23, 810, 689
36	Molėtai d.	3, 963, 500	5, 373, 750	4, 858, 786	4, 932, 051
37	Pagėgiai	2, 073, 900	2, 304, 584	2, 213, 358	2, 404, 348
38	Pakruojis d.	6, 634, 000	7, 298, 121	7, 688, 46	8, 440, 691
39	Panevėžys d.	10, 591, 500	11, 034, 150	11, 036, 999	11, 378, 647
40	Pasvalys d.	7, 203, 100	7, 696, 780	8, 511, 055	8, 618, 833
41	Plungė d.	10, 408, 900	11, 368, 946	10, 981, 868	12, 020, 409
42	Prienai d.	7, 193, 000	8, 405, 719	6, 965, 407	8, 208, 784
43	Radviliškis d.	11, 686, 100	13, 779, 240	13, 350, 617	14, 388, 180
44	Raseiniai d.	8, 772, 300	11, 117, 228	11, 030, 690	11, 507, 606
45	Rokiškis d.	8, 412, 300	9, 680, 137	9, 749, 959	9, 530, 913
46	Rietavas	1, 918, 300	1, 263, 644	1, 312, 391	2, 296, 270
47	Skuodas d.	3, 862, 400	4, 668, 205	4, 370, 355	4, 418, 788
48	Tauragė d.	11, 329, 500	11, 564, 249	11, 658, 893	11, 468, 241
49	Telšiai d.	12, 012, 800	10, 294, 601	15, 093, 149	12, 159, 975
50	Trakai d.	12, 053, 400	15, 931, 972	14, 435, 940	14, 009, 548
51	Ukmergė d.	11, 298, 400	15, 664, 824	14, 097, 061	13, 230, 059
52	Utena d.	13, 472, 300	14, 790, 508	10, 578, 904	12, 823, 772
53	Šakiai d.	8, 583, 500	8, 490, 676	9, 055, 465	8, 238, 022
54	šalčininkai d.	7, 283, 400	8, 427, 532	7, 558, 583	7, 285, 795
55	Švenčionys d.	7, 495, 800	9, 330, 358	9, 124, 353	8, 792, 275
56	Šilalė d.	5, 796, 300	6, 174, 197	6, 431, 797	6, 143, 360
57	Šilutė d.	10, 992, 800	11, 756, 661	12, 612, 898	12, 279, 767
58	Širvintai d.	4, 965, 300	5, 014, 248	4, 810, 321	4, 642, 112
59	Šiauliai d.	11, 943, 300	13, 058, 447	13, 961, 727	13, 904, 830
60	Jurbarkas d.	6, 413, 500	7, 644, 455	7, 639, 882	7, 070, 743

Source: created by the author

Table 18: Forecasting results with Bottom-Up approach for 2016



(a) Actual (black) vs fitted (red)

(b) Residuals

(c) Estimate of unctional coefficient

(d) Forecast (red) with true smoothed values (black) for 2016

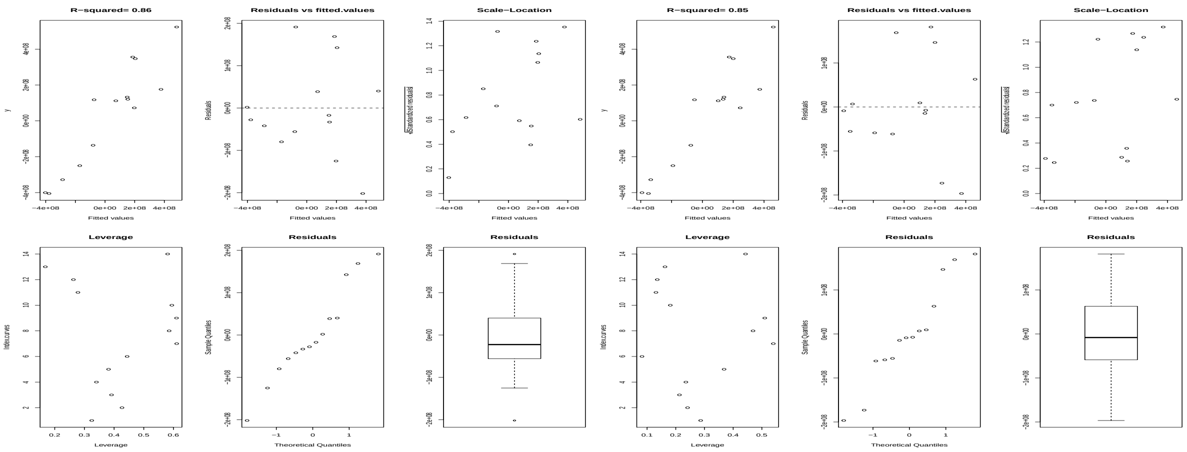
Source: created by the author

Figure 30: Results of {4} model for total centered Municipal budget revenue

	Model:		
	{1}	{2}	{3}
const	31214275.1	$3.121 \cdot 10^7$	$-2.805 \cdot 10^8$
	(3760576)	$(3.448 \cdot 10^5)$	$(3.781 \cdot 10^6)$
ϕ_1	137.6		−62.55
	(224.4)		(190.7)
ϕ_2	−196.3		361.6
	(320.7)		(360.1)
ϕ_3	285.8		−557.4
	(405.2)		(515.2)
ϕ_4	−252.3		313.5
	(283.7)		(345.2)
ϕ_5	252.3		−141
	(187.4)		(221.7)
PC1		17.01^{***}	
		(3.867)	
PC2		18.96	
		(18.44)	
PC3		−27.64	
		(21.27)	
GDP			$2.433 \cdot 10^{-2}$
			$(1.297 \cdot 10^{-2})$
Unemployed			$-1.855 \cdot 10^3$
			$(9.864 \cdot 10^2)$
R^2	0.86	0.85	0.93
Adjusted R^2	0.77	0.8	0.86
Variability explained by (%):			
PC1		92.77	
PC2		1.16	
PC3		2.40	

Note: *p<0.1; **p<0.05; ***p<0.01

Table 19: Estimates for total centered Municipal budget revenue models



(a) {1} model

(b) {2} model

Source: created by the author

Figure 31: Adequacy results for total centered Municipal budget revenue models

4 Appendix

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2001,13	5.750	8.186	9.462	9.800	9.425	8.560	7.433	6.409	5.986	6.668	8.959	13.363
2002,13	6.283	8.133	9.372	9.986	9.962	9.287	7.964	6.390	5.352	5.655	8.106	13.510
2003,13	5.295	8.500	10.068	10.373	9.785	8.679	7.427	6.407	6.002	6.594	8.567	12.303
2004,13	7.904	9.297	9.546	9.048	8.200	7.400	7.033	7.197	7.708	8.366	8.972	9.329
2005,13	8.593	9.364	9.766	9.852	9.679	9.300	8.770	8.147	7.489	6.854	6.300	5.886
2006,13	7.570	9.027	9.116	8.404	7.458	6.846	7.109	8.186	9.417	10.114	9.591	7.161
2007,13	2.749	1.752	4.153	5.264	5.892	6.848	8.898	11.890	14.748	16.355	15.595	11.353
2008,13	4.833	7.414	8.690	9.017	8.747	8.236	7.832	7.731	7.979	8.617	9.684	11.220
2009,13	14.389	13.350	12.740	12.246	11.553	10.348	8.339	5.746	3.300	1.754	1.861	4.375
2010,13	1.447	7.072	9.861	10.523	9.768	8.307	6.844	5.938	6.003	7.448	10.680	16.108
2011,13	4.380	7.530	9.988	11.691	12.575	12.576	11.638	9.918	7.778	5.594	3.740	2.590
2012,13	17.141	17.371	14.755	10.664	6.468	3.538	3.177	5.130	7.580	8.645	6.442	0.911
2013,13	10.987	9.170	8.854	9.312	9.814	9.632	8.083	5.551	3.484	3.376	6.722	15.017
2014,13	8.921	8.535	8.836	9.368	9.677	9.307	7.839	5.689	4.104	4.369	7.769	15.588
2015,13	5.488	8.843	10.196	10.112	9.157	7.895	6.882	6.433	6.623	7.517	9.179	11.673

Source: created by the author

Table 20: Proportion (%) of the centered budget revenue collected during the year in Vilnius

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2001,19	6.156	7.983	8.895	9.075	8.707	7.976	7.072	6.327	6.218	7.227	9.836	14.528
2002,19	5.993	7.312	8.287	8.884	9.068	8.806	8.076	7.160	6.643	7.122	9.194	13.458
2003,19	4.569	8.467	10.032	9.964	8.960	7.719	6.922	6.848	7.374	8.357	9.656	11.131
2004,19	6.868	8.147	8.779	8.917	8.716	8.329	7.908	7.604	7.560	7.919	8.826	10.425
2005,19	10.230	9.423	8.915	8.620	8.450	8.321	8.148	7.915	7.674	7.478	7.383	7.443
2006,19	16.379	9.195	5.171	3.566	3.642	4.660	5.901	7.079	8.345	9.870	11.822	14.371
2007,19	0.739	3.054	4.680	5.055	5.097	5.722	7.803	11.141	14.474	16.491	15.883	11.338
2008,19	5.335	6.496	7.554	8.404	8.938	9.050	8.646	7.946	7.480	7.793	9.428	12.932
2009,19	14.402	14.979	13.530	10.948	8.126	5.958	5.294	6.021	7.058	7.285	5.579	0.819
2010,19	4.643	11.543	13.432	11.936	8.680	5.290	3.348	3.414	5.028	7.687	10.884	14.116
2011,19	12.013	11.537	10.913	10.054	8.873	7.281	5.216	3.184	2.255	3.526	8.094	17.055
2012,19	10.740	6.147	5.587	7.478	10.239	12.290	12.117	9.730	6.671	4.543	4.953	9.505
2013,19	14.733	7.130	5.250	6.926	9.992	12.281	11.721	8.383	4.485	2.335	4.244	12.520
2014,19	8.107	9.060	9.350	9.132	8.561	7.793	6.986	6.377	6.288	7.039	8.954	12.354
2015,19	5.236	7.727	8.814	8.888	8.339	7.559	6.933	6.742	7.159	8.354	10.496	13.755

Source: created by the author

Table 21: Proportion (%) of the centered budget revenue collected during the year in Kaunas

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2001,21	5.892	8.974	10.168	10.019	9.070	7.865	6.938	6.568	6.783	7.598	9.030	11.094
2002,21	6.227	8.415	9.650	10.059	9.767	8.901	7.597	6.246	5.488	5.977	8.366	13.306
2003,21	5.616	9.275	10.930	11.081	10.225	8.862	7.485	6.452	5.985	6.303	7.623	10.162
2004,21	9.887	10.181	9.988	9.465	8.768	8.052	7.472	7.105	6.954	7.020	7.303	7.804
2005,21	13.201	10.628	8.352	6.562	5.448	5.197	5.982	7.544	9.198	10.241	9.969	7.678
2006,21	20.175	15.927	10.605	5.259	0.940	1.302	0.473	3.112	7.829	12.000	13.945	11.984
2007,21	7.424	0.614	6.655	10.779	13.067	13.601	12.485	10.343	8.321	7.587	9.311	14.660
2008,21	6.310	4.278	4.318	5.679	7.610	9.362	10.212	10.106	9.658	9.507	10.296	12.665
2009,21	6.472	9.043	11.285	12.914	13.646	13.198	11.311	8.317	5.137	2.718	2.008	3.953
2010,21	10.467	14.712	14.126	10.613	6.074	2.411	1.453	3.369	6.666	9.778	11.142	9.190
2011,21	8.415	25.440	27.065	19.114	7.412	2.217	4.191	1.497	9.274	13.329	7.847	12.985
2012,21	18.842	18.575	15.104	10.081	5.160	1.993	2.144	5.187	8.698	10.169	7.092	3.044
2013,21	31.641	17.907	10.869	8.434	8.509	9.002	7.893	4.871	1.329	1.266	1.445	2.258
2014,21	10.483	10.677	10.059	8.879	7.384	5.823	4.451	3.676	4.058	6.161	10.552	17.798
2015,21	7.374	9.711	9.897	8.734	7.026	5.574	5.155	5.945	7.518	9.424	11.212	12.430

Source: created by the author

Table 22: Proportion (%) of the centered budget revenue collected during the year in Klaipėda

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2001,0	4.850	6.833	7.968	8.439	8.432	8.132	7.729	7.461	7.622	8.506	10.407	13.621
2002,0	4.566	6.822	8.088	8.584	8.529	8.145	7.652	7.321	7.467	8.409	10.465	13.952
2003,0	5.073	7.355	8.387	8.541	8.193	7.715	7.473	7.640	8.197	9.117	10.373	11.936
2004,0	6.245	7.983	8.699	8.717	8.359	7.951	7.804	8.009	8.430	8.923	9.340	9.539
2005,0	8.042	8.193	8.174	8.076	7.991	8.010	8.219	8.574	8.906	9.037	8.790	7.989
2006,0	11.068	9.512	7.970	6.628	5.671	5.285	5.647	6.706	8.185	9.797	11.257	12.276
2007,0	1.541	1.186	0.907	1.384	4.690	8.016	10.404	11.730	12.703	14.070	16.579	20.976
2008,0	3.584	4.502	6.033	7.784	9.362	10.374	10.450	9.723	8.830	8.430	9.182	11.745
2009,0	12.694	8.909	7.749	8.273	9.538	10.603	10.559	9.275	7.390	5.583	4.528	4.901
2010,0	2.851	10.407	12.356	10.642	7.208	3.996	2.887	4.310	7.241	10.594	13.284	14.224
2011,0	4.737	10.639	12.598	11.901	9.838	7.694	6.714	7.092	7.976	8.466	7.666	4.678
2012,0	11.018	7.341	5.603	5.314	5.986	7.130	8.266	9.149	9.770	10.130	10.228	10.065
2013,0	8.208	10.053	10.173	9.190	7.723	6.391	5.797	6.112	7.073	8.403	9.823	11.054
2014,0	7.317	9.088	9.285	8.490	7.282	6.242	5.931	6.501	7.688	9.214	10.799	12.162
2015,0	4.184	7.900	9.510	9.614	8.814	7.710	6.891	6.682	7.141	8.314	10.249	12.992

Source: created by the author

Table 23: Proportion (%) of the centered budget revenue collected during the year in the rest municipalities

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2001	4.980	8.573	9.016	8.422	8.144	8.050	7.850	7.506	7.231	7.347	9.021	13.859
2002	4.851	8.594	8.991	8.413	8.323	8.416	8.207	7.621	7.004	6.830	8.565	14.185
2003	4.642	8.871	9.541	8.853	8.291	7.944	7.742	7.642	7.624	7.760	8.880	12.211
2004	6.377	9.995	9.218	7.719	7.655	8.240	8.389	7.938	7.639	8.164	9.116	9.551
2005	7.887	10.770	9.031	7.074	7.371	8.647	9.249	8.767	8.032	7.812	7.878	7.483
2006	11.052	9.674	7.976	6.623	5.988	5.870	6.018	6.524	7.826	10.159	11.886	10.405
2007	2.516	3.591	3.025	1.923	3.873	7.508	10.929	13.138	14.038	13.744	13.883	16.862
2008	3.936	6.579	7.411	7.707	8.263	8.937	9.481	9.651	9.202	8.108	8.248	12.477
2009	11.952	12.735	10.857	9.214	9.329	10.048	9.955	8.558	6.299	3.772	2.554	4.726
2010	0.751	14.644	12.937	7.309	5.140	5.458	6.376	6.969	7.280	7.568	9.609	15.958
2011	1.316	19.079	13.910	5.550	5.578	9.713	12.028	10.586	7.445	4.754	3.947	6.094
2012	12.050	11.644	7.210	4.672	6.709	9.696	9.478	6.325	4.934	9.082	12.854	5.346
2013	11.650	8.198	8.396	9.672	10.151	9.303	6.820	3.911	3.304	7.116	11.579	9.899
2014	7.893	9.470	9.426	8.646	7.785	7.054	6.609	6.492	6.632	7.080	8.960	13.951
2015	5.117	7.968	9.672	9.920	8.836	7.378	6.568	6.703	7.349	8.144	9.611	12.733

Source: created by the author

Table 24: Proportion (%) of the centered budget revenue collected during the year in the Lithuania

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2001,13	8.536	7.811	7.667	7.890	8.270	8.593	8.656	8.461	8.215	8.134	8.434	9.331
2002,13	7.800	7.894	7.915	7.903	7.898	7.940	8.067	8.287	8.574	8.902	9.245	9.577
2003,13	8.544	7.604	7.406	7.673	8.127	8.491	8.501	8.195	7.911	8.003	8.823	10.723
2004,13	6.628	7.289	8.077	8.796	9.246	9.229	8.567	7.500	6.689	6.816	8.561	12.603
2005,13	6.552	7.528	8.192	8.545	8.587	8.320	7.753	7.125	6.900	7.552	9.557	13.389
2006,13	6.991	7.812	8.504	8.978	9.141	8.905	8.192	7.240	6.603	6.847	8.539	12.247
2007,13	5.835	7.212	8.037	8.426	8.493	8.356	8.130	7.963	8.031	8.512	9.583	11.423
2008,13	6.430	7.834	8.628	8.924	8.836	8.477	7.964	7.490	7.322	7.732	8.991	11.371
2009,13	8.320	8.911	9.303	9.454	9.326	8.875	8.073	7.113	6.418	6.419	7.549	10.238
2010,13	6.300	7.877	8.777	9.113	8.997	8.538	7.855	7.189	6.907	7.381	8.982	12.083
2011,13	6.847	7.965	8.724	9.132	9.197	8.926	8.336	7.613	7.117	7.214	8.272	10.658
2012,13	7.836	8.706	9.062	9.019	8.692	8.199	7.657	7.222	7.093	7.465	8.539	10.511
2013,13	7.465	8.120	8.627	8.927	8.964	8.679	8.025	7.212	6.705	6.980	8.514	11.783
2014,13	7.591	8.152	8.667	9.019	9.094	8.777	7.969	6.922	6.239	6.542	8.449	12.579
2015,13	6.464	8.332	9.221	9.362	8.983	8.311	7.577	7.019	6.886	7.427	8.891	11.526

Table 25: Proportion (%) of the budget revenue collected during the year in the Vilnius

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2001,19	7.352	7.398	7.653	8.021	8.402	8.699	8.818	8.765	8.647	8.575	8.659	9.009
2002,19	7.438	8.027	8.258	8.244	8.096	7.923	7.835	7.900	8.147	8.602	9.291	10.240
2003,19	8.651	7.016	6.747	7.318	8.203	8.876	8.837	8.159	7.493	7.511	8.889	12.301
2004,19	6.688	7.408	7.960	8.318	8.455	8.346	7.973	7.518	7.362	7.894	9.502	12.576
2005,19	5.281	6.953	7.999	8.515	8.598	8.347	7.864	7.396	7.337	8.084	10.036	13.591
2006,19	5.381	7.474	8.715	9.258	9.256	8.865	8.241	7.618	7.308	7.622	8.877	11.385
2007,19	5.802	7.099	7.815	8.102	8.114	8.006	7.930	8.008	8.333	8.997	10.090	11.705
2008,19	6.300	7.306	8.042	8.500	8.677	8.567	8.172	7.678	7.451	7.868	9.304	12.135
2009,19	7.903	8.783	9.061	8.906	8.490	7.982	7.550	7.321	7.380	7.811	8.695	10.117
2010,19	6.451	8.247	9.018	9.036	8.572	7.898	7.283	6.953	7.089	7.872	9.482	12.099
2011,19	7.096	7.938	8.442	8.643	8.575	8.271	7.772	7.269	7.103	7.621	9.171	12.100
2012,19	7.293	7.482	7.910	8.402	8.781	8.872	8.510	7.845	7.334	7.451	8.667	11.454
2013,19	7.452	7.641	8.008	8.405	8.680	8.683	8.277	7.623	7.180	7.418	8.809	11.825
2014,19	7.126	8.065	8.567	8.706	8.556	8.190	7.685	7.229	7.122	7.667	9.165	11.921
2015,19	6.201	7.703	8.470	8.671	8.476	8.054	7.577	7.255	7.335	8.068	9.702	12.488

Table 26: Proportion (%) of the budget revenue collected during the year in Kaunas

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2001,21	8.163	6.932	6.919	7.629	8.570	9.247	9.190	8.467	7.684	7.471	8.457	11.273
2002,21	7.693	7.518	7.616	7.875	8.185	8.433	8.516	8.456	8.402	8.513	8.943	9.850
2003,21	8.104	6.951	6.765	7.190	7.872	8.454	8.598	8.343	8.108	8.325	9.431	11.859
2004,21	5.781	6.741	7.636	8.361	8.814	8.888	8.494	7.838	7.426	7.777	9.407	12.837
2005,21	5.141	6.965	8.422	9.427	9.892	9.730	8.870	7.616	6.643	6.644	8.311	12.337
2006,21	5.344	6.767	8.105	9.195	9.870	9.967	9.337	8.210	7.196	6.920	8.007	11.083
2007,21	5.341	6.977	8.191	8.974	9.322	9.226	8.690	7.935	7.400	7.536	8.792	11.616
2008,21	6.919	6.962	7.376	7.956	8.499	8.803	8.677	8.230	7.873	8.027	9.116	11.562
2009,21	6.991	8.107	8.995	9.580	9.791	9.553	8.805	7.745	6.835	6.544	7.345	9.709
2010,21	7.530	8.697	9.099	8.952	8.468	7.862	7.346	7.085	7.198	7.802	9.013	10.948
2011,21	7.191	8.774	9.370	9.259	8.727	8.055	7.521	7.283	7.375	7.829	8.674	9.941
2012,21	8.128	8.784	8.979	8.839	8.487	8.047	7.644	7.392	7.394	7.755	8.580	9.971
2013,21	8.929	8.605	8.587	8.701	8.777	8.643	8.140	7.398	6.834	6.882	7.971	10.533
2014,21	8.030	8.624	8.851	8.764	8.417	7.862	7.160	6.541	6.405	7.160	9.213	12.972
2015,21	7.203	8.450	8.878	8.726	8.237	7.652	7.208	7.075	7.350	8.129	9.508	11.583

Table 27: Proportion (%) of the budget revenue collected during the year in Klaipėda

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2001,0	8.153	7.939	7.918	8.011	8.138	8.222	8.187	8.077	8.051	8.273	8.909	10.122
2002,0	8.191	7.875	7.811	7.906	8.069	8.204	8.227	8.164	8.164	8.378	8.957	10.054
2003,0	7.873	7.388	7.491	7.913	8.387	8.647	8.442	7.884	7.456	7.651	8.968	11.901
2004,0	6.605	6.945	7.417	7.891	8.240	8.333	8.054	7.587	7.412	8.023	9.914	13.580
2005,0	5.713	6.986	7.836	8.303	8.429	8.254	7.826	7.379	7.326	8.091	10.095	13.762
2006,0	5.483	6.919	7.938	8.569	8.842	8.786	8.439	7.984	7.756	8.096	9.345	11.843
2007,0	5.895	6.393	6.932	7.448	7.877	8.158	8.232	8.215	8.388	9.042	10.467	12.955
2008,0	5.616	6.531	7.384	8.100	8.603	8.819	8.683	8.335	8.123	8.404	9.533	11.868
2009,0	7.523	7.634	7.910	8.238	8.502	8.591	8.397	8.019	7.755	7.913	8.798	10.720
2010,0	5.988	7.766	8.517	8.544	8.148	7.632	7.293	7.312	7.754	8.679	10.147	12.219
2011,0	6.320	7.632	8.314	8.523	8.412	8.137	7.853	7.707	7.842	8.399	9.518	11.342
2012,0	7.099	7.367	7.614	7.820	7.965	8.029	7.996	7.956	8.104	8.637	9.756	11.658
2013,0	6.733	7.794	8.295	8.382	8.200	7.894	7.612	7.501	7.711	8.394	9.701	11.782
2014,0	6.718	7.893	8.351	8.309	7.983	7.588	7.338	7.378	7.787	8.642	10.019	11.993
2015,0	5.579	7.576	8.549	8.765	8.492	7.996	7.542	7.344	7.564	8.362	9.899	12.333

Table 28: Proportion (%) of the budget revenue collected during the year in the rest municipalities

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2001	8.483	6.913	7.484	8.485	8.830	8.647	8.207	7.819	7.832	8.507	9.344	9.449
2002	8.316	7.027	7.628	8.488	8.618	8.278	7.865	7.693	7.983	8.868	9.719	9.516
2003	8.541	6.762	7.120	8.093	8.650	8.701	8.279	7.675	7.436	8.054	9.448	11.240
2004	6.893	6.372	7.674	8.909	8.993	8.406	7.799	7.486	7.454	7.760	9.219	13.034
2005	6.205	6.519	7.950	9.022	8.937	8.217	7.520	7.203	7.316	7.957	9.716	13.438
2006	5.871	7.398	8.321	8.799	8.951	8.810	8.405	7.874	7.468	7.490	8.667	11.945
2007	5.583	7.257	7.633	7.665	7.924	8.242	8.376	8.321	8.312	8.619	9.747	12.323
2008	5.885	7.413	8.015	8.234	8.416	8.517	8.453	8.244	8.017	7.972	8.906	11.928
2009	7.590	8.607	8.707	8.583	8.624	8.633	8.354	7.812	7.316	7.210	8.051	10.512
2010	5.898	8.683	8.892	8.300	8.032	7.956	7.804	7.567	7.492	7.868	9.237	12.271
2011	6.276	8.643	8.706	8.227	8.253	8.449	8.337	7.888	7.519	7.668	8.771	11.262
2012	7.314	8.209	8.144	8.016	8.275	8.499	8.193	7.505	7.226	8.050	9.604	10.965
2013	7.307	7.814	8.282	8.604	8.686	8.461	7.876	7.197	7.004	7.816	9.476	11.477
2014	7.041	8.253	8.600	8.509	8.280	7.971	7.617	7.323	7.268	7.674	9.116	12.349
2015	6.097	7.839	8.799	9.010	8.615	7.977	7.472	7.297	7.458	8.002	9.343	12.091

Table 29: Proportion (%) of the budget revenue collected during the year in Lithuania

5 Appendix

	Day:														
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)*	(12)*	(13)	(14)*	(15)*
μ	234602238	241635089	242393405	223725145	196657445	194010496	218603001	250383247	268420806	259781710			452777202		
δ															
ρ_1					0.9192					−1.1732		−0.1517			
ρ_2										−0.9601		−0.2369			
ρ_3															
ρ_4															
θ_1		0.1312			−0.7998					1.2350	−0.9573	−0.8964		−0.9215	−0.9766
θ_2		−0.1776								0.9193					−0.0339
θ_3															0.2189
AIC	3184.64	3061.08	2987.96	2943	2902.04	2917.75	2927.51	2884.78	2884.77	2953.7	2965.07	2947.49	2959.93	2996.57	2994.58
AICc	3184.92	3061.56	2988.1	2943.14	2902.52	2917.9	2927.51	2884.92	2884.91	2954.73	2965.21	2947.98	2960.07	2996.72	2995.06
BIC	3192.07	3070.99	2992.91	2947.96	2911.95	2922.71	2932.47	2889.74	2889.72	2968.56	2970	2957.36	2964.88	3001.5	3004.44

Note: *Integrated series

Source: created by the author

Table 30: Results of pointwise ARIMA

	Day:														
	(16)*	(17)	(18)	(19)	(20)	(21)*	(22)*	(23)*	(24)*	(25)*	(26)*	(27)*	(28)	(29)	(30)
μ		375936470	276110272	322154171	228798930								471340016	693781595	827845398
δ								228237.40	515664.60	766353.25					
ρ_1				0.4830		−0.3483	−0.4212	−0.1543	−0.8052	−0.2171				0.3281	
ρ_2						−0.1995	−0.3219			−0.2256				0.0154	
ρ_3														0.2198	
ρ_4														−0.4321	
θ_1	−0.7986			−0.3817		−0.9082	−0.7659	−0.9157	−0.9408	−0.4030	−0.7117	−0.8783	−0.2977	−0.4829	
θ_2				−0.3291						−0.4979					
θ_3															
AIC	2960.29	3060.5	3025.24	3002.52	3073.73	3065.63	3038.7	2997.43	3057.03	3105.03	3090.7	3090.12	3140.32	3143.74	3264.52
AICc	2960.43	3060.64	3025.38	3003.25	3073.87	3066.12	3039.18	2997.72	3057.51	3106.08	3090.99	3090.27	3140.61	3145.14	3264.66
BIC	2965.22	3065.45	3030.19	3014.9	3078.68	3075.5	3048.56	3004.83	3066.89	3119.82	3098.1	3095.06	3147.75	3161.08	3269.48

Note: *Integrated series

Source: created by the author

Table 31: Results of pointwise ARIMA