Lietuvos matematikos rinkinys, Vol. 60, 2019
Proc. of the Lithuanian Mathematical Society, Ser. A
https://doi.org/10.15388/LMR.A.2019.14948

ISSN 0132-2818
eISSN 2335-898X
pages 7–14

Functional approach to analysis of daily tax revenues*

Jovita Gudan, Alfredas Račkauskas†

Institute of Applied Mathematics, Vilnius University
Naugarduko st. 24, LT-03225 Vilnius, Lithuania
E-mail: jovita.gudan@mif.vu.lt, alfredas.rackauskas@mif.vu.lt

Abstract. We present a functional data analysis approach to modeling and analyzing daily tax revenues. The main features of daily tax revenue we need to extract are some patterns within calendar months which can be used for prediction. As standard seasonal time series techniques cannot be used due to varying number of banking days per calendar month and presence of seasonality between and within months we interpret monthly tax revenues as curves obtained from daily data. Standard smoothing techniques and registration taking into account time variability are used for data preparation.

Keywords: functional data analysis, data smoothing, registration, prediction.

1 Introduction

The State Tax Inspectorate under the Ministry of Finance of the Republic of Lithuania (hereinafter referred to as STI) makes forecasts of daily tax revenues for State budget according to historical data trends, which are based on expert experience. It is a difficult task to predict and assess daily changes in revenue collection, which depend not only on the tax calendar, but also on tax payer behavior, especially in cases where the payment date of the obligation coincides with the weekend or holiday.

The main goal of this research is to test some statistical methodologies that could improve tax revenue forecasts.

As far as we know only few papers are devoted to time-series models of daily tax revenues. Koopman and Ooms [1] provides a detailed discussion of this problem and suggests to use a two-way mapping to transform irregular data to regular ones. They use a state space model to forecast daily tax revenues. The later articles by Koopman and Ooms [2, 3, 4] are the improved versions of the analysis of daily tax revenues.

In this paper we illustrate daily time series features using a series for Lithuanian aggregate tax revenues using functional data analysis tools. Functional data are often characterized by both shape and phase variability. Tax revenue is a typical

* The research supported by the Research Council of Lithuania, grant No. S-MIP-17-76.
† The authors would like to thank the State Tax Inspectorate for their assistance with the collection of tax revenue data and particularly Vytenis Zaskevičius for numerous discussions on this research.

example where these two sources of variation are clearly identified and interpreted. An overall pattern is observed that tax revenue accelerates around two fixed days. In this setting, phase variability is identified as variation in the calendar timing. Explicit consideration of phase variability is necessary in order to obtain consistent estimation of typical tax revenues patterns. This paper will focus on fitting a structural model and using it to forecast a tax revenues monthly patterns. By structural model we mean a model estimated from registered data. The main issue is to find appropriate wrapping functions. After accommodating prediction techniques to tax revenue data the most accurate predictions are considered to be derived from functional principal component regression and exponential smoothing.

In Section 2 we present preliminary analysis of daily series for Lithuanian tax revenues including smoothing and registration of data. In Section 3 we discuss some prediction tools.

2 Preliminary analysis

We use a daily series for Lithuanian tax revenues, i.e. taxes, fees and other payments paid by the tax payers that are paid to STI's budget revenue collection accounts. For the analysis data are taken from the period January 2011 to February 2019. The data have the form:

$$y_{k,j}, \quad j = 1, \ldots, N_k, \ k = 1, \ldots, n,$$

where k corresponds to regular time (months in our case) and index $j = 1, \ldots, N_k$ corresponds to a time grid within period k.

2.1 Smoothing

We interpret the date as observations of random curves:

$$y_k(s), \quad s \in [0, 1], \ k = 1, \ldots, n.$$

Moreover we assume that the sample curves are observed at discrete instants of time with some noise, so that

$$y_{k,j} = y_k(j/N_k) + \varepsilon_{k,j}, \quad j = 1, \ldots, N_k.$$

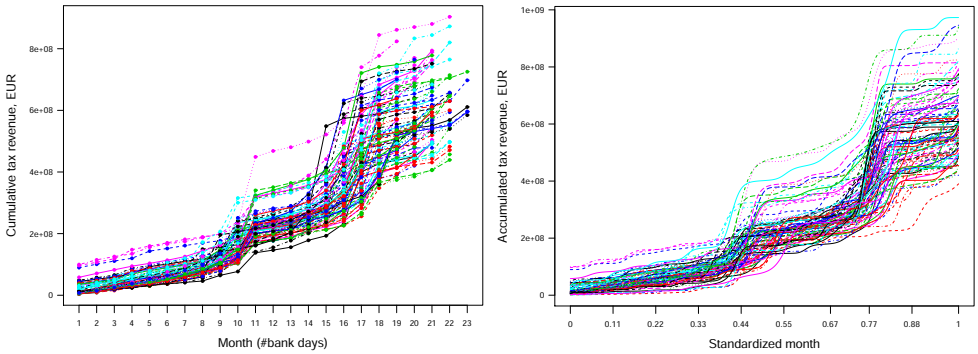
Figure 1(a) represents monthly patterns of accumulated tax revenue data, which are observed at a discrete time (see dots in Fig. 1(a) and for the visualization purposes the interpolated lines helps to recognize the trends within one month). Clearly, the number of bank days in a month is specific for each month as well as the calendar day at which the discrete data point is observed.

We reconstruct each function $(y_k(s), s \in [0, 1])$ by smoothing techniques thus obtaining functional sample

$$\hat{y}_k(s), \quad s \in [0, 1]; \ k = 1, \ldots, n$$

which we interpret as observations of random functions

$$Y_1 = \big(Y_1(s), s \in [0, 1]\big), \ldots, Y_n = \big(Y_n(s), s \in [0, 1]\big),$$



(a) Discrete observations (b) Smoothed functions

Fig. 1. Accumulated tax revenue data. Source: created by the authors.

with values in the classical Hilbert space $L_2(0,1)$. In order to have increasing and differentiable functions we need to choose appropriate smoothing technique.

Since the process behind the raw accumulated tax revenue data is always increasing, monotone transformation will be used to smooth the data. Suppose $W(t)$ is a conventional functional data object, and that is unconstrained in any way except for $W(t_0) = 0$ where t_0 is the lower boundary over which we are smoothing.

Given that there are two clearly visible peaks in the middle and end of the month (see Fig. 1(a)), by examining various smoothing techniques it was found that splines can track such features with satisfactory accuracy.

In order to define monotone smoothing, a B-spline basis will be used, and the constraint $W(t_0) = 0$ can easily be achieved by fixing the first coefficient to be zero. Then each smoothed function takes the following form:

$$\hat{y}(t) = \beta_0 + \beta_1 \int_{t_0}^{t} e^{W(u)} du + \epsilon(t) = \beta_0 + \beta_1 \int_{t_0}^{t} e^{\phi(u)' \mathbf{c}} du + \epsilon(t),$$

where ϕ denotes a vector containing the B-spline basis functions and the parameter $\boldsymbol{c}$ is a vector containing the coefficients of the B-spline expansion. Coefficients $\boldsymbol{c}$ are estimated by minimizing the sum of squared errors. Figure 1(b) shows in this way smoothed accumulated tax revenue data.

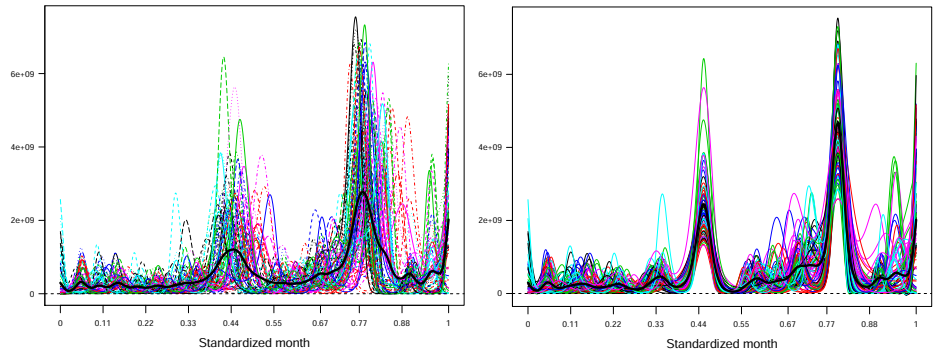
For $k = 1, \ldots, n$, let x_k be the derivative of the function $\widehat{y}_k$. We interpret the sample

$$x_1 = \big(x_1(s), s \in [0,1]\big), \ldots, x_n = \big(x_n(s), s \in [0,1]\big)$$

as observations of random curves

$$X_1 = \big(X_1(s), s \in [0,1]\big), \ldots, X_n = \big(X_n(s), s \in [0,1]\big)$$

again as random elements in the sample space $L_2(0,1)$. Figure 2(a) shows the derivative of smoothed accumulated tax revenue data. Clear peaks are visible around the 15th and 25th days of the month since the most important tax sources has to be paid on these days or the next business day if a due date falls on a Saturday, Sunday or legal holiday. The peak at the end of the month as well as the beginning of the month is due to the smoothing algorithm, since the assumptions that are necessary to calculate smoothing parameters are insufficient.



(a) Original curves (b) Registrated curves

Fig. 2. Derivative of accumulated tax revenue data and the mean functions (black solid line). Source: created by the authors.

2.2 Registration

We assume that for each $k = 1, \ldots, n$, there exists (probably a random) time transformation $v_k : [0,1] \to [0,1]$, such that

$$X_k(s) = \mu(v_k(s)) + \eta_k(v_k(s)) + \varepsilon_k(v_k(s)), \quad s \in [0,1], \tag{1}$$

where the non-random function $\mu(s)$, $s \in [0,1]$ can be interpreted as a structural mean, $\eta_k(s)$, $s \in [0,1]$ accounts a structural individual variation from μ, ε_k is an error process. We assume that η_k and ε_k are independent. Moreover, we assume that (ε_k) is a strong white noise.

The function $w_k = v_k^{-1}$ is called warping function and the random curve

$$X_k^*(s) = X_k\big(w_k(s)\big), \quad s \in [0,1]$$

is a registered or aligned version of X_k. The random sample $X_1^*, \ldots, X_n^*$ is a structural sample corresponding to $X_1, \ldots, X_n$ and is the main object of analysis within this paper. Clearly we have under the model (1)

$$X_k^*(s) = \mu(s) + \eta_k(s) + \varepsilon_k(s), \quad s \in [0,1]. \tag{2}$$

Hence, techniques developed so far in functional data analysis, can be applied for the statistical analysis of the structural sample.

There exists several constructions of warping functions proposed in the literature (see, e.g., Ramsay et al. [5, 6] and references therein). The first one we apply is the so-called landmark method which seems to be easiest. The method involves identifying the timings of specific features of the curves (deadlines for payments as defined by law in the tax revenue example), and then aligning the curves so that all these events occur at the same time for each curve.

Consider the situation with two landmarks per curve (that corresponds to the first and second obligations in the tax revenue calendar of i's month), say, τ_{i1} and τ_{i2}, on the interval $[0,1]$. Let

$$\tau_{01} = \frac{1}{n}\sum_{i=1}^{n} \tau_{i1}, \qquad \tau_{02} = \frac{1}{n}\sum_{i=1}^{n} \tau_{i2}$$

be the average landmarks.

The simplest landmark registration method defines the warping functions such that $w_i(0) = 0$, $w_i(\tau_{01}) = \tau_{i1}$, $w_i(\tau_{02}) = \tau_{i2}$, $w_i(1) = 1$, and linearly interpolates in between. Then, the warping functions and their inverse functions have explicit forms:

$$\begin{aligned} w_i(t) = t &+ (\tau_{i1} - \tau_{01})\frac{t}{\tau_{01}}\mathbf{1}_{[0,\tau_{01}]}(t) \\ &+ \left[(\tau_{i1} - \tau_{01})\frac{\tau_{02} - t}{\tau_{02} - \tau_{01}} + (\tau_{i2} - \tau_{02})\frac{t - \tau_{01}}{\tau_{02} - \tau_{01}}\right]\mathbf{1}_{[\tau_{01},\tau_{02}]}(t) \\ &+ (\tau_{i2} - \tau_{02})\frac{1 - t}{1 - \tau_{02}}\mathbf{1}_{[\tau_{02},1]}(t). \end{aligned}$$

and its inverse is

$$\begin{aligned} v_i(s) = s &+ (\tau_{01} - \tau_{i1})\frac{s}{\tau_{i1}}\mathbf{1}_{[0,\tau_{i1}]}(s) \\ &+ \left[(\tau_{01} - \tau_{i1})\frac{\tau_{i2} - s}{\tau_{i2} - \tau_{i1}} + (\tau_{02} - \tau_{i2})\frac{s - \tau_{i1}}{\tau_{i2} - \tau_{i1}}\right]\mathbf{1}_{[\tau_{i1},\tau_{i2}]}(s) \\ &+ (\tau_{02} - \tau_{i2})\frac{1 - s}{1 - \tau_{i2}}\mathbf{1}_{[\tau_{i2},1]}(s). \end{aligned} \tag{3}$$

If we assume that the horizontal variation of the sample curves occurs randomly (due to some unexpected events), then it is reasonable to apply another time transformations, probably data driven ones.

3 Prediction

In this section we consider prediction of the tax revenue curve $(X_{n+h}(s), s \in [0,1])$ by using aligned functions $(X_k^*(s), s \in [0,1])$, $k = 1, \ldots, n$. The relation between the two predictions is defined by

$$\widehat{X}_{n+h}(s) = \widehat{X^*}_{n+h}(v_{n+h}(s)), \quad s \in [0,1],$$

where the function v_i is defined by (3). Generally speaking, one distinguishes two classes of prediction methods: empirical and model based. Actually the distinction is imprecise, as empirical methods often contain an underlying model for which the predictions are optimal. One step out-of-sample forecast procedure will be used to compare the accuracy of predictions, where validation sample is from 2011 January to 2019 January and prediction horizon is $h = 1$.

3.1 Empirical methods of prediction

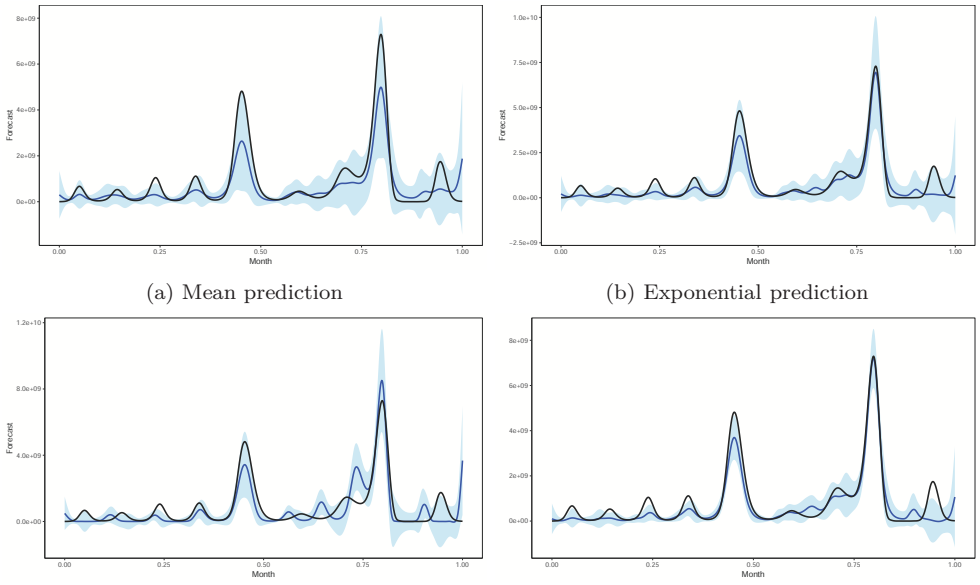
3.1.1 The empirical mean

Assuming that the time transformation v_{n+h} is know, one defines

$$\widehat{X^*}_{n+h}(s) = \widehat{\mu}(s), \quad s \in [0,1].$$

This predictor has good properties for a model of the form:

$$X_k^* = \mu + \varepsilon_k, \quad t \in N,$$



(a) Mean prediction

(b) Exponential prediction

(c) Naive prediction

(d) FPCR

Fig. 3. Forecasting results (blue line) and it's 95% prediction interval (blue ribbon) with true function (black line) for February 2019. Source: created by the authors.

where $\mu \in L_2(0,1)$ and (ε_t) is a (strong) white noise. In this case $E(X_{n+h}(s)|\mathcal{F}_n) = \mu(v_{n+h}(s))$. See Fig. 3(a) of mean prediction for the derivative of accumulated tax revenues. The 95% prediction interval is quite wide and the peaks of forecasts around the due dates are underestimated. On other days the tendency is grasped except in the end of the month, where true intensity of the accumulated tax revenue where observed before the last day, but the prediction shows that intensity is growing the few days before the end of the month and spikes at the last day.

3.1.2 Exponential smoothing

This method, which is widely used in practice, consists of assigning weights to the observations that tend to 0 at an exponential rate:

$$\widehat{X^*}_{n+h} = c\big(X_n^* + qX_{n-1}^* + \cdots + q^{n-1}X_1^*\big),$$

where $0 < q < 1$ and c is a normalization constant. Usually, we choose $c = 1 - q$ and $0.7 \leqslant q \leqslant 0.95$. After careful consideration $c = 0.1$ was chosen and Fig. 3(b) represents exponential smoothing prediction for the analyzed period. For the due date around 25th day the prediction is very accurate and for the second largest peak the prediction is underestimated, but is more accurate than by mean prediction. On other days the tendency is quite similar to the real curve except in the end of the month.

3.1.3 Naive predictors

These predictors are defined by:

$$\widehat{X^*}_{n+h} = X^*_n$$

and for initial time series gives

$$\widehat{X}_{n+h}(s) = X^*_n\big(v_{n+h}(s)\big) = X_n\big(w_n(v_{n+h}(s))\big), \quad s \in [0,1].$$

These are good predictors when the observed phenomenon varies a little, or rarely. In fact, X^*_n is the best predictor if and only if $(X^*_t, t \geq 1)$ is a martingale. Since of an analyzed period previous month is 2019 January thus naive prediction is illustrated in Fig. 3(c). The biggest peak is overestimated whereas the second largest peak is underestimated and the reduction of the intensity is predicted to be faster than it actually happened. Moreover, the tendency of intensity on other days is not captured precisely as on actual curve.

3.1.4 Prediction by scores

Consider

$$X^*_k(s) = \mu(s) + \sum_{j=1}^{J} \lambda^*_{kj}\psi^*_j(s) + \varepsilon_k(s),$$

where λ^*_{kj} are principal component scores and ψ^*_j are structural eigenfunctions. Univariate ARIMA models are fitted for $(\lambda^*_{kj},\ k = 1,\dots,n)$ time series for each $j = 1,\dots,J$. Then h-step-ahead forecasts are derived from

$$\hat{X}^*_{n+h}(s) = \mu(s) + \sum_{j=1}^{J} \hat{\lambda}^*_{n+h,j}\psi^*_j(s),$$

where $\hat{\lambda}^*_{n+h,j}$ denotes the h-step-ahead forecasts of $\lambda^*_{n+h,j}$ using a univariate time series. After performing functional principal component analysis, it has shown that to account for 87% of variation 8 harmonics are needed, so $J = 8$ was chosen and Fig. 3(d) represents $h = 1$ prediction using functional principal component regression (FPCR). The peak around 25th day is predicted very precisely and the peak around 15th day underestimated slightly. This model's prediction is very similar to exponential smoothing prediction, except for the tendency at the end of the month.

Since the biggest attention is given to the due dates of paying taxes because around those days the largest amount of money is collected to STI revenue collection accounts, therefore it is very important to account for the considerable prediction error. The accurate prognosis is needed in order to forecast the flows of the state's cash resources that are needed not only to monitor the execution of the state budget task within a given month, but also to ensure the repayment for the taxpayers from the same revenue collection accounts. As the main focus is around the due dates that are 15th and 25th of the month, the predictions from exponential smoothing and functional principal component regression are the most accurate at these peaks.

4 Conclusions

In this paper we have investigated several modelling strategies for daily time series in the functional data analysis context, with the objective of short term forecasting. Since most of parametric models are not able to model irregularly spaced data or at least they need to be transformed, functional data analysis technique is introduced to overcome challenges linked to daily time series such as a changing number of observations per month or year. The empirical results were based on a series of Lithuanian daily tax revenues, which embodied three modelling stages. First, the choice of suitable basis function, which transforms data from discrete –to functional observations. Second, alignment of monthly curves since they might have the same shape, but individual curves have been deformed due to tax calendar. And lastly, comparison of several modelling techniques using one step out-of-sample forecast procedure. It is shown that all of strategies exponential smoothing and functional principal component regression are the most accurate at the peaks of 15th and 25th days of the month when taxes are most collected. Although predictions in other periods can be improved. This work was intended as an attempt to motivate public sector to improve daily tax revenue predictions with the tendencies using functional data analysis.

References

[1] S.J. Koopman and M. Ooms. *Time Series Modelling of Daily Tax Revenues*. Tinbergen Institute Discussion Papers 01-032/4, Tinbergen Institute, 2001.

[2] S.J.Koopman and M. Ooms. Time-series modeling of daily tax revenues. *Stat. Neerl.*, **57**(4): 439–469, 2003.

[3] S.J. Koopman and M. Ooms. *Forecasting Daily Time Series using Periodic Unobserved Components Time Series Models*. Tinbergen Institute Discussion Papers TI 2004-135/4, Tinbergen Institute, 2004.

[4] S.J. Koopman and M. Ooms. Forecasting daily time series using periodic unobserved components time series models. *Comput. Stat. Data Anal.*, **51**(2): 885–903, 2006.

[5] J. Ramsay, G. Hooker and S. Graves, *Functional Data Analysis with R and MATLAB*. Springer, London, 2009.

[6] J.O. Ramsay and X. Li. Curve registration. *J. R. Statist. Soc., Ser. B (Stat. Meth.)*, **60**:351–363, 1998.

REZIUMĖ

Dieninių mokestinių pajamų analizė funkciniu metodu

J. Gudan, A. Račkauskas

Šiame straipsnyje yra pateikiami funkcinių duomenų analizės metodai analizuojant ir modeliuojant dienines mokestines pajamas. Pagrindiniai bruožai, nusakantys dienines mokestines pajamas, yra kalendorinių mėnesių struktūros, kurių pagalba yra prognozuojami duomenys. Mėnesinės mokestinės pajamos yra interpretuojamos kaip funkcijos, kurios yra gautos iš dieninių duomenų, kadangi standartiniai sezoniniai laiko eilučių modeliai negali būti pritaikyti dėl skirtingų darbo dienų skaičiaus kalendoriniame mėnesyje ir dėl sezoniškumo tarp mėnesių ir mėnesio viduje. Duomenų paruošimui taikomi standartiniai glodinimo ir duomenų registracijos metodai.

Raktiniai žodžiai: funkcinė duomenų analizė, duomenų glodinimas, registracija, prognozavimas.