TI 2001-032/4
Tinbergen Institute Discussion Paper

Time Series Modelling of Daily Tax Revenues

Siem Jan Koopman
Marius Ooms

Department of Econometrics and Operations Research, Faculty of Economics and Business Administration, Vrije Universiteit Amsterdam

Tinbergen Institute
The Tinbergen Institute is the institute for economic research of the
Erasmus Universiteit Rotterdam, Universiteit van Amsterdam and
Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Keizersgracht 482
1017 EG Amsterdam
The Netherlands
Tel.: +31.(0)20.5513500
Fax: +31.(0)20.5513555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31.(0)10.4088900
Fax: +31.(0)10.4089031

Most TI discussion papers can be downloaded at
http://www.tinbergen.nl

Time Series Modelling of Daily Tax Revenues

Siem Jan Koopman and Marius Ooms

2 February 2001

vrije Universiteit *amsterdam*, The Netherlands

Abstract

We provide a detailed discussion of the time series modelling of daily tax revenues. The main feature of daily tax revenue series is the pattern within calendar months. Standard seasonal time series techniques cannot be used since the number of banking days per calendar month varies and because there are two levels of seasonality: between months and within months.

We start the analysis with a periodic regression model with time varying parameters. This model is then extended with a component for intra-month seasonality, which is specified as a stochastic cubic spline. State space techniques are used for recursive estimation and evaluation as they allow for irregular spacing of the time series.

The model is recently made operational and used for daily forecasting at the Dutch Ministry of Finance. For this purpose a front-end for model configuration and data input is implemented with Visual C++, while statistical tools and graphical diagnostics are built around Ox and `SsfPack`. We present the current model and forecasting results up to December 1999. The model and its forecasts are evaluated.

Address for correspondence: Marius Ooms, Department of Econometrics and Operations Research, Free University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands, E-mail: mooms@econ.vu.nl

1 Introduction

The production of daily forecasts of tax revenues is an important task of day-to-day cash management at the Treasury in the Netherlands. The main purpose of a statistical daily time series model is to process information of revenues of previous days systematically and efficiently. Dutch central government outlets are usually known at least one day ahead. Therefore, several days ahead forecasts of revenues can also be used to monitor the targets for the budget.

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 problem of irregular spacing can be mitigated by an auxiliary time transformation. We transform the data to regularly spaced data with the inclusion of missing values, such that standard Kalman Filter techniques can be applied. The features of a periodically varying variance of seasonal and irregular components are incorporated using a time-varying Kalman Filter in a way similar to Burridge and Wallis (1990).

Other problems for daily economic time series are, surprising as it may seem at first sight, small sample problems. Daily patterns show usually more structural breaks, due to institutional changes in the financial and tax system than monthly or yearly data. These breaks are often so large, that it does not make sense to combine pre-break and post-break data for the daily model. Since there are usually not many years of homogeneous daily data available, we cannot estimate long-term trends and monthly patterns very precisely. This means that a model for daily data is not well suited for long-term forecasting.

We illustrate daily time series features using a series for Dutch aggregate tax revenues. Frequent changes in the tax collection system occurred up to 1993. Therefore, our series starts in March 1993. It contains a (negative) component of tax restitutions up to 1997 which means that values close to zero and even negative values can occur. Tax revenues are only received on bank days: Mondays to Fridays.

Dutch total national daily tax revenues consist of several major components like income tax, social security premiums, corporate tax, value-added tax and a number of smaller categories, like special duties on gas and alcohol. All of these revenues are compiled per category on a yearly basis, while many revenue categories are compiled on a monthly basis. However, these figures are not immediately available after the turn of the period and they are mostly compiled on a net basis, i.e. revenues minus restitutions. On a daily basis only total gross tax revenues are available. Yesterday's figures can be used to forecast today's revenues. Restitution payments are currently exactly known a few days in advance. We lack relevant information from tax assessments on taxes that are due on a daily basis. However, monthly data on expected tax revenues are available and can be used to evaluate the (forecasts of) net monthly sums of revenues and restitutions.

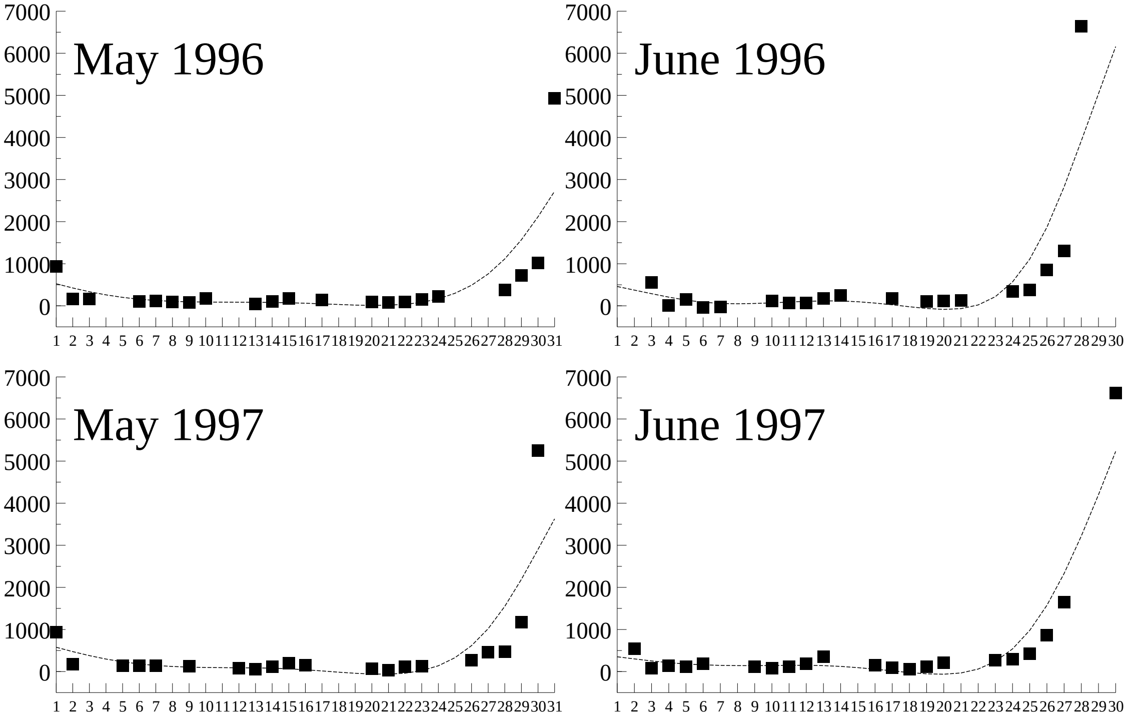

Figure 1: Daily Dutch national tax Revenues in millions of euro

Figure 1 for the daily Dutch central tax revenues in May and June of 1996 and 1997 illustrates the main features. The aim is to model the conditional mean and variance of this series for short-term forecasting. Many taxes are due on the last bank day of the month. The majority of revenues is collected on the last bank day, but the revenues on the four days leading up to this day are also substantial. The revenues on the last bank days vary clearly from May to June. Tax income on the first day of the month is also important, but here the seasonal effect is less pronounced, as we shall see below. The intra-monthly income pattern on the remaining days is not nearly as variable.

The mean of income clearly depends on the number of bank days that remain until the turn of the month and on the number of bank days after the turn of the month. The basic intramonthly pattern in the middle of each month is similar across months. This pattern does not seem to be affected by the number of bank holidays. The data for May illustrate this for the bank holidays on Ascension day (Thursday) and Whit Monday.

The original data, indexed by the calendar-day of the month, as in Figure 1 are irregularly spaced. Straightforward application of splines (depending on the calendar-day-of-the month) to fit the intramonthly pattern is not a good idea. This is illustrated by the estimated natural cubic splines (with 5 knot parameters) in Figure 1, see Doornik and Hendry

(1999) for computational details.

The splines describe most of the data well and they pick up a local maximum in revenues around the middle of the month. The pitfalls of this approach become apparent towards the ends of the months. The fitted income patterns for May and June vary across 1996 and 1997, whereas the observed pattern on the last bank days of the month is much more similar: the irregular spacing leads to an exaggerated time-variation across years. The estimated splines, which minimise the sum of squared deviations across all observations subject to a smooth penalty, also show that we want to vary the weights of the observations and the smoothness within the month. Smoothness can be imposed in the first half of the month, but the income pattern around the turn of the month is not smooth. In practice this means that we set up a prespecified "mesh" for a cubic spline function with some points in the first half of the month but more points around the turn of the month, cf. Harvey, Koopman, and Riani (1997).

We like to set up a model for regularly spaced observations that share the basic pattern within the month, so that the time distance between two turns of the month becomes constant. This hopefully enables us to model the data for months with varying numbers and spacing of bank days in a relatively parsimonious way.

Therefore we need a two-way mapping between our irregularly spaced observations in calendar time, y_τ, and approximately regularly spaced observations, y_t, for our model. These regularly spaced observations will be modelled in a discrete time linear state space model. We index these "model observations" by $t = 1, \ldots, n$. The mapping $t(\tau)$ defines the model index as a function of calendar time $\tau = 1, \ldots, T$. In our case we use the following functions of calendar time: Y_τ is Calendar Year; $1993, \ldots, 1999$, d_τ is Day of the Month, $1, \ldots, 31$; m_τ is Month of the Year, $1, \ldots, 12$; w_τ is Day of the Week, $1, \ldots, 7$; h_τ is (non)bank holiday, $0, 1$. The function h_τ can vary over time and has to be known in advance for forecasting. The other functions of τ are deterministic. In our case Saturdays and Sundays are bank holidays: $h_\tau = 0$ if $w_\tau = 1$ or $w_\tau = 7$.

Figure 2 presents the time transformation of the data of Figure 1 where we have chosen a constant underlying grid of 23 points each month. The pattern is now much more regular, both across years and across months. We have created other missing observations, but this should not present theoretical problems for our estimation procedure, see §3. A finer grid, leading to more missing observations, can be considered, at the cost of an increase in computation.

Figure 3 presents a complete picture of the revenues on the 1461 bank days used to specify the model of this paper. The period covers 2132 calendar days in 70 months, March 1993-December 1998.

We plot daily revenues against the year to indicate the presence of trends. The evidence for trends is not significant at first sight. We plot daily revenues against month-of-the year to show the month-of-the year effect. The variance does not seem to depend on the year or on the month of the year. The variance does depend on the day of the month. The figures for the last day of each month, which are seen in the upper half of the plots, are clearly more volatile than the other days with lower revenues, as shown in the lower half of the

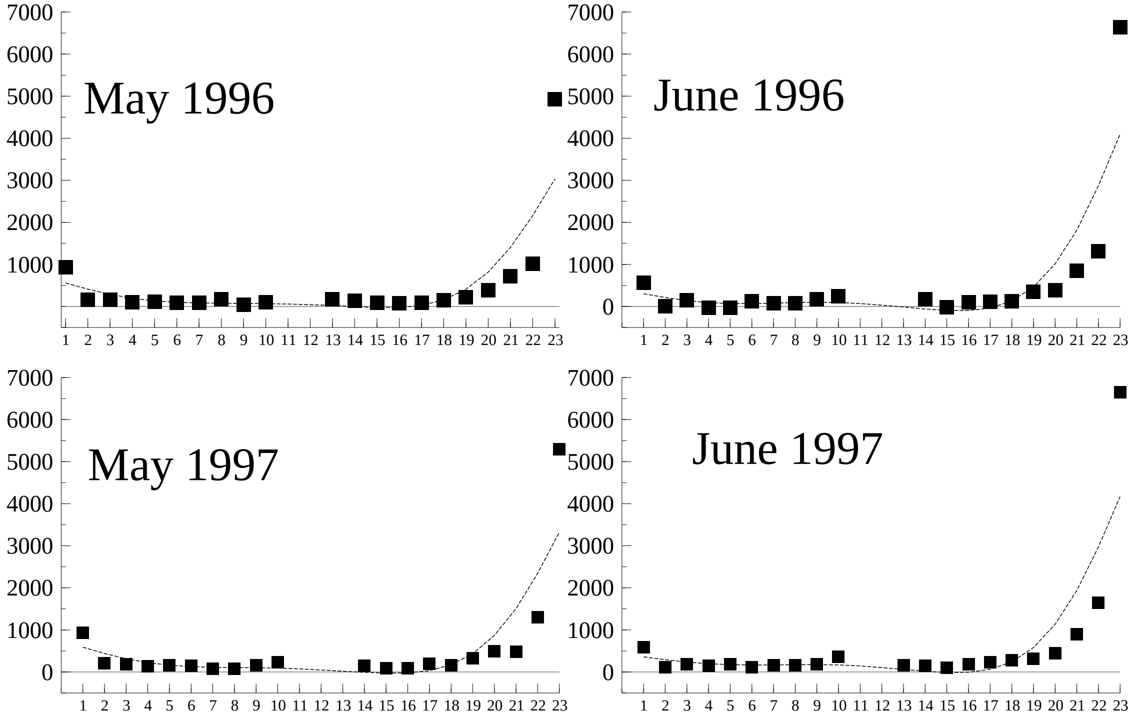

Figure 2: Daily Dutch national tax Revenues in millions of Euro after time transformation

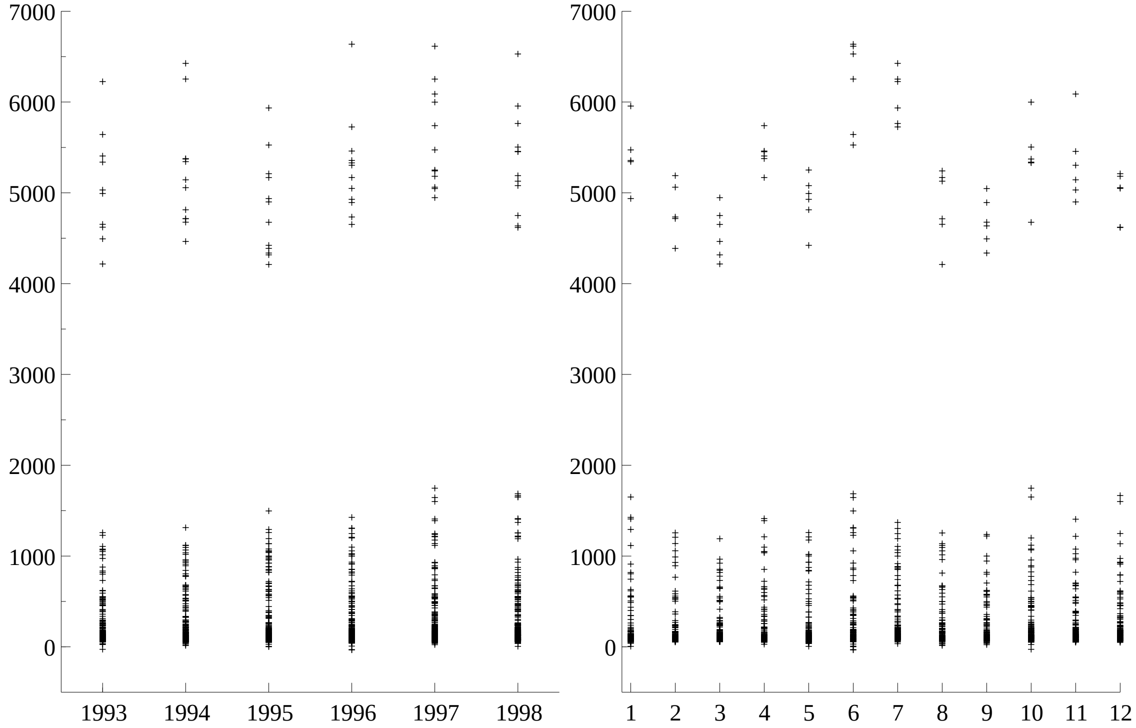

Figure 3: Daily Dutch national tax revenues in millions of Euro against year and against month-of-year.

plots. In the sequel of this paper we use “seasonal” to describe the month-of-the-year effect. A seasonal difference means the difference with the corresponding value one year before. We use “periodic” to describe the day-of-the-month effects in the mean, the variance and the autocovariances. Periodicity refers to the pattern that occurs once a month.

The upper graph of figure 4 illustrates the seasonal and the periodic movements of the tax revenues in one graph. It also shows the one-step-ahead out-of-sample forecasting performance of our model over 1998. The basic patterns are well matched and the coverage probability of the forecast intervals seems satisfactory. The lower graph shows the mean of the stochastic intramonthly spline function of our model as computed at the end of 1997, based on parameter estimates presented in Table 4 below.

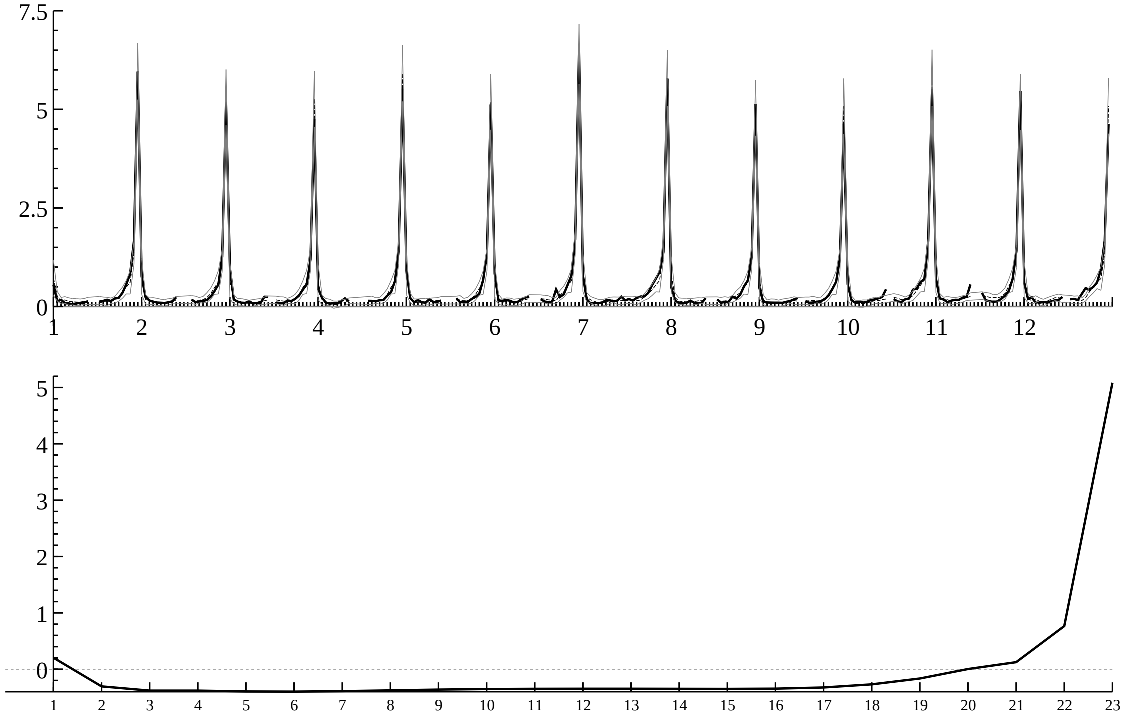

Figure 4: Top graph: 1998: One step ahead forecasts (dashed), 95% confidence intervals (dotted) and realisations (solid), Bottom graph: mean of intramonthly spline as computed at the end of 1997.

The remainder of this paper is structured as follows. In the next section we specify a simple periodic regression model to capture seasonal and periodic properties, before going into the details of the time transformation for Figure 2. Section 3 addresses the basic ideas of structural time series modelling, with the notation corresponding to the `SsfPack` documentation of Koopman, Shephard, and Doornik (1999), the state space formulation for trend and seasonal components, filtering, estimation and forecasting. In our case the treatment of time varying cubic splines, the occurrence of (artificial) missing observations, and the generation of several-steps-ahead forecasts deserve special attention. In section 4 we discuss the modelling strategy and its implementation in modern software. Section 5 presents the empirical results. We describe our model and present estimates for data up to 1997 which we used to produce on-line forecasts for 1998 of Figure 4 We adjust the model using these results, estimate it up to 1998, and produce online forecasts for 1999. We evaluate our model forecasts against naive predictions. Section 6 suggests some extensions to our approach and concludes.

2 Initial regression analysis

So far, we have mainly looked at the unconditional mean of the series as a periodic function of calendar time. In this subsection we use flexible regression models to summarise the main properties of this unconditional mean function. We use the residuals to estimate the periodic variances we would like to exploit in our statistical forecasting model.

The initial analysis shows a clear periodic variation in the mean of the series. The dominating effects are due to the month of the year and the bank day of the month. It is possible there is a nonstationary trend component. The variation from month to month is partly caused by a quarterly effect from corporate tax revenues, that one could label month of the quarter effect. This leads to a higher average for January, April, July and October, see Figure 3. In addition there is a yearly effect due to extra salary payments prior to the summer holidays. This additional month of the year effect is most clearly seen for June.

As discussed above and shown in Figure 2 there is a clear banking day of the month effect which displays clear similarities across months. The mean of the series is mainly determined by the number of days before the turn of the month.

We suggest a simple regression procedure to identify the main periodicities in the mean of the series. For the purpose of this preliminary analysis we introduce the following notation, based on calendar time τ and the position of bank days within a each month.

The bank-day index $b_\tau = -15, -14, \ldots, 14, 15$ equals the number of bank days from the beginning of the sample, $r_\tau = 1, \ldots, R$, minus the number of bank days up to the nearest turn of the month $l_\tau = 1, \ldots, L$, with $l_\tau = 0$ for $\tau < 15$. Therefore, bank days leading up to the turn of the month have a negative b_τ, bank days after the turn of the month have a positive b_τ and the last bank day of each month has index $b_\tau = 0$.

In our sample each month has at least 18 bank days. So each month has observations with index $b_\tau = 1, 2, \ldots, 9$ and $b_\tau = -8, -7, \ldots, -1, 0$. In order to analyse the variance and covariance function of these $70 \times 18 = 1260$ observations we regress them on 12×18 dummy variables, each dummy measuring the mean of y_t for a particular combination of b_τ and month of the year index m_τ. However, we do not pool all observations. We construct 18 monthly subseries for each bank-day index and regress each series on a constant and 11 centred seasonal dummies. In this way we allow automatically for periodic heteroskedasticity depending on b_τ. We present results in Table 1. In order to facilitate comparison with later results we present the regression in the order of our model-day-of-the-month index $p(t)$, where $p(t)$ is related to b_τ as: $p(t) = b_\tau + I_{[-P/2,0]}(b_\tau) \cdot P$, with $I_{[]}()$ an indicator function that equals 1 for negative b_τ. See also equation (1) below. The first 9 bank days of each month correspond to $p(t) = 1, \ldots, 9$, whereas $p(t) = 15, \ldots, 23$ for the last 9 bank days.

Table 1: *Descriptive statistics tax income by bank day of the month*

$p(t)$	b_τ	mean	$\hat{\sigma}_b$	$\hat{R}_b$	$c(1,b)$	$c(2,b)$	$c(3,b)$	$c(4,b)$	$c(5,b)$	$c(6,b)$	$c(7,b)$	$c(8,b)$	$c(9,b)$
1	1	685	208	0.52	-0.14	-0.09	-0.02	-0.11	-0.25	-0.06	-0.16	-0.17	-0.02
2	2	175	52	0.58	0.05	0.10	-0.05	0.02	-0.17	-0.24	-0.12	-0.03	0.07
3	3	115	36	0.44	0.14	-0.01	0.27	0.26	0.14	0.02	0.09	0.06	0.00
4	4	99	35	0.52	0.18	0.39	0.20	0.11	0.25	0.04	0.03	-0.06	0.02
5	5	88	36	0.39	0.49	0.32	0.18	0.18	0.32	0.25	-0.15	0.13	0.15
6	6	99	32	0.51	0.22	-0.08	0.04	-0.27	0.12	0.34	0.29	-0.03	0.32
7	7	91	40	0.47	-0.07	0.08	0.11	0.08	-0.05	-0.09	0.22	0.34	-0.10
8	8	101	41	0.42	0.25	0.01	0.25	0.27	0.30	-0.06	-0.04	0.20	0.59
9	9	102	43	0.36	0.45	0.43	0.34	0.36	0.18	0.36	-0.21	0.10	0.36
15	-8	111	44	0.49	0.09	0.22	0.05	0.11	0.22	0.20	0.01	0.20	-0.08
16	-7	127	45	0.76	0.30	0.38	0.25	0.19	0.17	0.40	0.32	0.40	0.32
17	-6	152	72	0.63	0.56	0.14	0.44	0.43	0.12	0.15	0.33	0.40	0.46
18	-5	193	75	0.69	0.56	0.60	0.14	0.42	0.39	0.31	0.02	0.23	0.48
19	-4	280	95	0.74	0.56	0.51	0.63	0.27	0.40	0.43	-0.03	0.19	0.23
20	-3	406	125	0.72	0.60	0.61	0.55	0.50	-0.02	0.52	0.39	0.21	0.25
21	-2	672	168	0.72	0.28	0.33	0.28	0.14	0.37	-0.14	0.12	0.00	0.06
22	-1	1167	248	0.65	0.28	0.68	0.66	0.70	0.60	0.66	0.25	0.59	0.47
23	0	5221	592	0.85	0.43	-0.19	0.29	0.24	0.09	0.20	0.14	0.19	0.34

$p(t)$ Model-day-of-the-month index, see equation (1) below
b_τ indexes position with respect to last bank day of the month.
mean: Estimate of constant in regression model per bank day with centred seasonal dummies for daily tax revenues with index b_τ.
$\hat{\sigma}_b$: regression standard error. Measurements in 10^6 Euro. Sample 1993.3-1998.12: 70 observations.
$\hat{R}_b$: correlation of fitted and dependent variable, 5% critical value: 0.53
$c(j,b)$ is a so-called periodic correlation at daily lags, cf. McLeod (1994): corr(ε_t,ε_{t-j},b_τ), where we assume there are 18 bank days in each month, so that modulo 18 arithmetic applies. The correlation depends only on the distance between the observations in bank days and on the index b_τ of the leading observation.

The third column of Table 1 summarizes the periodic mean function across all months. It reproduces the pattern seen in Figure 2 above. The function seems smooth except at the exact turn of the month. The fourth column averages the residual periodic variance function across all months. This function is also rather smooth. The periodic standard deviation is clearly not proportional to the periodic mean. For $b = 1$ and $b = -2$ one observes similar means, but very different variances. For $b = 1$ and $b = -1$ we observe similar variances but very different means. The fifth column shows the multiple correlation coefficient of each regression, which provides an estimate of (deterministic) seasonality for each bank day. Since there are 12 regressors and 70 observations, one could use a 5% critical value for $\hat{R}$ of 0.53, to test the null hypothesis of no seasonality. It is clear that the process generating the revenues is more seasonal towards the end of the month.

The last columns of Table 1 contain the estimated serial correlation coefficients at daily intervals. These are large for the days at the end of each month. This could indicate the systematic presence of local trends within the months. A consistent series of negative (but small) correlations corresponds to revenues of the first day of each month. These revenues show a negative correlation with all 8 previous banking days. The results of Table 1 motivate a periodic analysis.

Figures 1 and 2 show that time transformation may simplify the statistical model for our data, in the sense that we are better able to exploit the intermonthly similarity of the intramonthly pattern. The timing intervals for the statistical (state space) model will differ from the timing interval of the observations, not only when the distance between observations is measured in calendar days, but also when these are measured in bank days.

Let y_t denote the observations for the model. Since we have daily data and both seasonal and intramonthly effects, each observations has a three-way index: $j(t)$ is the year, $s(t)$ is the month of the year, and $p(t)$ is the day of the month. In our case, $j(t) = J_1, \dots, J_n$, $s(t) = 1, \dots, S$, $p(t) = 1, \dots, P$, where $j(t)$ and $s(t)$ are functions of the calendar indexes: $j(t) = Y_\tau$ and $s(t) = m_\tau$. The choice of the function $p(t)$ varies from application to application. In general we do *not* have $p(t) = d_\tau$. The index $p(t)$ serves as the explanatory variable of the periodic spline function. In general the series y_t will have more missing observations than the series y_τ. Table 2 summarises the information on all the indices and calendar variables for our sample. Next we discuss our time transformation choice.

Table 2: Indices and their Ranges, in Model Time and Calendar Time

Index	Name	Range	Sample 1993.3.1-1998.12.23
t	model time	$1, \dots, n$	$n = 1610$
$j(t)$	year	$J_1, \dots, J_n$	$J_1 = 1993,\ J_n = 1998$
$s(t)$	month of the year	$1, \dots, S$	$S = 12$
$p(t)$	day of the month	$1, \dots, P$	$P = 23$
$M(t)$	number of working days in a month	$18, \dots, 23$	
τ	Calendar time		
Y_τ	year	$1993, \dots, 1998$	
m_τ	month of the year	$1, \dots, 12$	
d_τ	calendar day of the month	$1, \dots, 31$	
b_τ	bank day of the month	$-13, \dots, 9$	
w_τ	day of the week	$1, \dots, 7$	
h_τ	working day	$0, 1$	

The time transformation leads to different timing intervals for the model and the observations. The timing interval for the model can be shorter than the observation interval. Harvey (1989) discusses statistical solutions to the problems of estimation and prediction for components of a linear dynamic model in the context of mixed timing intervals.

We first delete the weekends and (bank) holidays (h_τ=0) from our sample. Next we choose the number of days per month in model time $P = 23$: the maximum number of bank days in any month in our remaining sample. This introduces a minimal number missing values for the model data around the middle of the month. The timing interval for the observations and for the model is then still equal to one bank day, except for the observations in the middle of each month, where the timing interval for the model may vary from 1 to 6:

$$p(t) = b_\tau + I_{(-\infty,0)}(b_\tau) \cdot P, \tag{1}$$

with $I_{()}()$ an indicator function that equals 1 for negative b_τ and is zero elsewhere. The transformation is determined by the end conditions $p = 1$ if $b_\tau = 1$ and $p = P$ if $b_\tau = 0$ and the break near middle of the month where b_τ turns negative. For some months we have missing values for $p = 10, \ldots, 14$.

For different kinds of data sets a different time transformation function $p(t)$ may apply in connection with the observed intra-monthly pattern and its changes from month to month and from year to year. Note that we use $p(t)$ as the basis for our intramonthly spline function. This spline function is the basis for interpolation of the artificial missing values and for the forecasting of future values. We treat our data as a stock variable: the spline estimates the value of our variable at $p(t)$. If $y_{j(t),s(t),p(t)}$ corresponds to an observation, the spline will estimate y_{Y_τ,m_τ,d_τ}.

The simple time transformation with $P = 23$ and an equal timing interval for model and observations for the majority of the data does not pose major technical problems. Given the state space form of the dynamic regression model one can start forecasting from each non-missing observation, both one-step-ahead and multi-step-ahead, both for single days and for time aggregates. The next section presents more details. In the most simple case where we have white noise homoskedastic errors, and where we treat all regression coefficients as fixed this boils down to the application of recursive regression, where both one-step and multi-step forecast intervals take into account the parameter uncertainty due to estimation of the regression coefficients.

In the final stages of our analysis we want to translate non-missing model data back to observations in calendar time. We first translate $p(t)$ back to b_τ. Given the bank day number of the month, b_τ, and the calendar variables, w_τ and h_τ indicating the position of holidays, it is then straightforward to compute the calendar day of the month d_τ.

3 State space time modelling of time series

The purpose is to build a model for short-run-forecasting. The main problem is to estimate the recurring but persistently changing pattern within the months, averaging across months and across years in an efficient way for forecasting. Structural time series models provide a convenient statistical framework to solve this problem. Further, for our purposes it suits two aims: firstly, it decomposes the observed series into unobserved stochastic processes which provide (after estimation) a better understanding of the dynamic characteristics of the series; secondly, it generates optimal forecasts straightforwardly using the Kalman filter. The estimation of components and the forecasting of the series require first the estimation of parameters associated with unobserved components such as trend, seasonal and irregular. For this analysis we will use the *SsfPack* library of Koopman, Shephard, and Doornik (1999) which provides all Kalman filter related algorithms and is implemented for the object-oriented matrix language *Ox* of Doornik (1998). The basic aspects of structural time series modelling and the corresponding notation are introduced below.

3.1 Structural time series models

A univariate structural time series model is suitable for many economic time series data sets and is given by

$$y_t = \mu_t + \gamma_t + \varepsilon_t, \qquad \varepsilon_t \sim NID(0, \sigma^2_\varepsilon), \qquad t = 1, \ldots, n, \tag{2}$$

where μ_t, γ_t and ε_t are trend, seasonal and irregular components, respectively. The trend and seasonal components are modelled by dynamic processes which depend on disturbances. These components are formulated in a flexible way and they are allowed to change over time rather than being deterministic. The various disturbances are independent of each other and of the irregular component, ε_t. The definitions of the components are given below, but a full explanation of the underlying rationale can be found in Harvey (1989, Chapter 2). The effectiveness of structural time series models compared to ARIMA type models, especially when messy features in time series are present, is shown in Harvey, Koopman and Penzer (1998).

The trend component is defined here as

$$\begin{array}{rcll} \mu_t & = & \mu_{t-1} + \beta_{t-1} + \eta_t, & \eta_t \sim NID(0, \sigma^2_\eta), \\ \beta_t & = & \beta_{t-1} + \zeta_t, & \zeta_t \sim NID(0, \sigma^2_\zeta), \end{array} \tag{3}$$

where the level and slope disturbances, η_t and ζ_t are mutually uncorrelated. When σ^2_ζ is zero, we have a *random walk plus drift*, and when σ^2_η is zero as well, a deterministic linear trend is obtained. A relatively *smooth trend*, related to a cubic spline, results when a zero value of σ^2_η is coupled with a positive σ^2_ζ; Young (1984) calls this model an 'integrated random walk'.

3.2 State space analysis

The state space form provides a unified representation of a wide range of linear Gaussian time series models including the structural time series model; see, for example, Harvey (1989) and Kitagawa and Gersch (1996). The Gaussian state space form consists of a transition equation and a measurement equation; we formulate it, following De Jong (1991), as adopted inKoopman, Shephard, and Doornik (1999) as

$$\alpha_{t+1} = T_t\alpha_t + H_t\varepsilon_t, \qquad \alpha_1 \sim \mathrm{N}\left(\bar{a}, \overline{P}\right), \qquad t = 1, \ldots, n, \tag{4}$$

$$y_t = Z_t\alpha_t + G_t\varepsilon_t, \qquad \varepsilon_t \sim \mathrm{NID}\left(0, I\right), \tag{5}$$

where NID(μ, Ψ) indicates an independent sequence of normally distributed random vectors with mean μ and variance matrix Ψ, and, similarly, $\mathrm{N}(\cdot, \cdot)$ indicates a normally distributed variable. We treat the observed tax series y_t as a univariate time series for $t = 1, \ldots, n$. The $m \times 1$ state vector α_t contains unobserved stochastic processes and unknown fixed effects. The state equation (4) has a Markovian structure which is an effective way to describe the serial correlation structure of the time series. The initial state vector is assumed to be random with mean $\bar{a}$ and variance matrix $\overline{P}$ but some elements of the state can be diffuse which means that it has mean zero and variance κ where κ is large. The measurement equation (5) relates the observation y_t in terms of the state vector α_t through the signal $Z_t\alpha_t$ and the vector of disturbances ε_t. The deterministic matrices T_t, Z_t, H_t and G_t are referred to as system matrices and they usually are sparse selection matrices.

The Kalman filter is a recursive algorithm for the evaluation of moments of the normal distribution of state vector α_{t+1} conditional on the data set $Y_t = \{y_1, \ldots, y_t\}$, that is

$$a_{t+1} = \mathrm{E}\left(\alpha_{t+1}|Y_t\right), \qquad P_{t+1} = \mathrm{cov}\left(\alpha_{t+1}|Y_t\right),$$

for $t = 1, \ldots, n$; see Anderson and Moore (1979, page 36) and Harvey (1989, page 104). The Kalman filter is given by

$$\begin{array}{lcl} v_t & = & y_t - Z_t a_t \\ F_t & = & Z_t P_t Z_t' + G_t G_t' \\ K_t & = & \left(T_t P_t Z_t' + H_t G_t'\right) F_t^{-1} \\ a_{t+1} & = & T_t a_t + K_t v_t \\ P_{t+1} & = & T_t P_t T_t' + H_t H_t' - K_t F_t K_t' \end{array} \qquad (6)$$

for $t = 1, \ldots, n$, and with initialisations $a_1 = \bar{a}$, and $P_1 = \overline{P}$, and where v_t is the innovation and F_t is its variance. The derivative of the forecast function for the state with respect to the current innovation is the Kalman gain K_t. The initial state variance matrix $\overline{P}$ is given by

$$\overline{P} = P_* + \kappa P_\infty,$$

where κ is large; for example, $\kappa = 10^6$. The matrix P_* contains the variances and covariances between the stationary elements of the state vector (zeroes elsewhere) and P_∞ is a diagonal matrix with unity for nonstationary and deterministic elements of the state and zero elsewhere. The number of diffuse elements (that is the number of unity values in P_∞), is given by d.

The prediction error decomposition is the key result for computing the log-likelihood function for models in state space form, that is

$$\begin{array}{rcl} l & = & \log p\left(y_1, \ldots, y_n; \varphi\right) = \displaystyle\sum_{t=1}^{n} \log p\left(y_t|y_1, \ldots, y_{t-1}; \varphi\right) \\ & = & -\dfrac{n-d}{2}\log\left(2\pi\right) - \dfrac{1}{2}\displaystyle\sum_{t=d+1}^{n}\left(\log|F_t| + v_t' F_t^{-1} v_t\right) \end{array} \qquad (7)$$

where φ is the vector of parameters for a specific statistical model represented in state space form (6). The innovations v_t and its variances F_t are computed by the Kalman filter for a given vector φ. Note that the summation in (7) is from $d+1$ to n, since the first d summations will be approximately zero as F_t^{-1} will be very small for $t = 1, \ldots, d$.

Estimation of the unobserved components is usually referred to as signal extraction. The computation of $\hat{\alpha}_t = \mathrm{E}(\alpha_t|Y_n)$ and its variance matrix $V_t = \mathrm{var}(\alpha_t|Y_n)$ is referred to as moment state smoothing. The state smoothing algorithm we will employ is based on the work of De Jong (1988) and Kohn and Ansley (1989) and is given by

$$\hat{\alpha}_t = a_t + P_t r_{t-1}, \qquad V_t = P_t - P_t N_{t-1} P_t, \qquad t = n, \ldots, 1, \tag{8}$$

where r_{t-1} and N_{t-1} are evaluated by the backwards recursion

$$\begin{array}{lcl} e_t & = & F_t^{-1} v_t - K_t' r_t \\ D_t & = & F_t^{-1} + K_t' N_t K_t \\ r_{t-1} & = & Z_t' F_t^{-1} v_t + L_t' r_t \\ N_{t-1} & = & Z_t' F_t^{-1} Z_t + L_t' N_t L_t \end{array} \tag{9}$$

for $t = n, \ldots, 1$.

3.3 Diagnostic checking

The assumptions underlying a Gaussian model are that the disturbance vector ε_t is normally distributed and serially independent with unity variance matrix. On these assumptions the standardised one-step prediction errors

$$e_t = \frac{v_t}{\sqrt{F_t}}, \qquad t = d+1, \ldots, n, \tag{10}$$

are also normally distributed and serially independent with unit variance. To diagnose whether the normality assumption for e_t holds for a given model and a particular time series y_t, we usually compute

$$N = n\{\frac{S^2}{6} + \frac{(K-3)^2}{24}\},$$

where S is the sample skewness and K is the sample kurtosis. This test statistic has an asymptotic χ^2 distribution with 2 degrees of freedom. A simple test for heteroskedasticity is obtained by comparing the sum of squares of two exclusive subsets of the sample. For example, the statistic

$$H(h) = \frac{\sum_{n-h}^{n} e_t^2}{\sum_{t=1}^{h-1} e_t^2},$$

is $F_{h,h}$-distributed for some preset positive integer h, under the null hypothesis of homoskedasticity. Also, the standardised forecast errors should be serially uncorrelated. Therefore, the correlogram of the one-step prediction errors must not reveal serial correlation. A standard portmanteau test statistic for serial correlation is the Box-Ljung statistic and is given by

$$Q(p) = n(n+2)\sum_{j=1}^{p} \frac{c_j^2}{n-j},$$

for some preset positive integer p where c_j is the sample autocorrelation of lag j. This test is asymptotically χ^2 distributed with $p - q$ degrees of freedom, where q is the number of estimated parameters in ϕ.

Diagnostic tests can also be applied component by component using auxiliary residuals. Auxiliary residuals are constructed using the recursions in (9) and are defined by

$$\begin{array}{rclcrcl} G_t\hat{\varepsilon}_t & = & G_t e_t, & \quad & \mathrm{Var}(G_t\hat{\varepsilon}_t) & = & G_t D_t G_t', \\ H_t\hat{\varepsilon}_t & = & H_t r_t, & \quad & \mathrm{Var}(H_t\hat{\varepsilon}_t) & = & H_t N_t H_t', \end{array}$$

for $t = 1, \ldots, n$. The auxiliary residual $G_t\hat{\varepsilon}_t$ can be used to identify outliers in the time series. Large values in $G_t\hat{\varepsilon}_t$ indicate that the behaviour of the observed value cannot be appropriately represented by the model under consideration. The usefulness of $H_t\hat{\varepsilon}_t$ depends on the interpretation of the state elements in α_t. For the trend model (3) it is clear that η_t is the change of the trend for time t. It follows that structural breaks in the series y_t can be identified by detecting large values in the series for $\hat{\eta}_t$. Harvey and Koopman (1992) have formalised these ideas further for structural time series models and they constructed additional diagnostic tests for the auxiliary residuals.

3.4 Missing values and forecasting

When observations y_t for $t = t_0, \ldots, t_1 - 1$ are missing, the vector v_t and the matrix K_t of the Kalman filter are set to zero for these values, that is $v_t = 0$ and $K_t = 0$, and the Kalman updates become

$$a_{t+1} = T_t a_t, \qquad P_{t+1} = T_t P_t T_t' + H_t H_t', \qquad t = t_0, \ldots, t_1 - 1; \tag{11}$$

similarly, the backwards smoothing recursions become

$$r_{t-1} = T_t' r_t, \qquad N_{t-1} = T_t' N_t T_t, \qquad t = t_1 - 1, \ldots, t_0. \tag{12}$$

Other relevant equations for smoothing remain the same. This simple treatment of missing observations is one of the attractions of the state space methods for time series analysis.

Out-of-sample predictions, together with their mean square errors, can be generated by the Kalman filter by extending the data set $y_1, \ldots, y_n$ with a set of missing values. When y_{n+j} is missing, the Kalman filter step reduces to

$$a_{n+j+1} = T_{n+j} a_{n+j}, \qquad P_{n+j+1} = T_{n+j} P_{n+j} T_{n+j}' + H_{n+j} H_{n+j}',$$

which are the state space forecasting equations for $j = 1, \ldots, J$ where J is the forecast horizon; see also the treatment of missing observations in the previous section. The multi-step forecast of y_{n+j} is simply given by

$$\hat{y}_{n+j} = Z_{n+j} a_{n+j}, \qquad \mathrm{Var}(\hat{y}_{n+j}) = Z_{n+j} P_{n+j} Z_{n+j}', \qquad j = 1, \ldots, J.$$

A sequence of missing values at the end of the sample will therefore produce a set of multi-step forecasts.

4 Methodology and implementation

Structural periodic models for daily data offer a wide range of possibilities for the online modeller and forecaster. Each specification produces different forecasts and forecast intervals. The identification of the model, i.e. the choice for the specification for a particular application, is done in several cycles. After a basic model is implemented and tested, the analysis of forecast errors and other diagnostics will lead to improvements. When sufficiently many new data points have been observed it is likely that the model has to be tuned again, either by reestimating it using a new sample, or by changing a number of its components. We present two stages of the model in this section.

In the implementation of a model for the Dutch ministry of Finance we found that a kind of integrated developer environment, IDE, for structural time series modelling is crucial if the model is to be used effectively on a day-to-day basis. The modelling and forecasting is performed simultaneously at different levels of sophistication using tools with different levels of user-friendliness. Today's software makes it possible to develop such an environment with a small number of people with a limited amount of programming time.

The next subsection recapitulates the tasks of the online modeller and forecaster of daily time series. Subsection 4.2 discusses the implementation of the developer environment for this task.

4.1 Methodology

Sections 2 and 3 described several aspects of structural time series modelling of daily time series. Here we recapitulate the modelling methodology. At each stage the forecaster has to choose from several options.

(a) *Time transformation*

It may be necessary to transform the timing interval of the observations from calendar time to a more "operational" model time. If there is a clear intramonthly pattern it is useful to work with three indices, where the last 2 indices are strictly periodic, indicating the model month and the model day respectively. This will often introduce artificial missing observations. Section 2 described some options.

(b) *Model for the mean of components*

One must specify models for the intra-monthly mean, i.e. a periodic component, a model for the intrayearly mean, i.e. a seasonal component, a model for the interyearly mean, i.e. a trend component, and finally a model for the irregular. The seasonal and periodic components can be modelled using deterministic or stochastic dummies, splines or trigonometric terms. The stochastic trends can have a fixed or a varying slope.

(c) *Knot positions of the spline*

Given the use of splines one must choose the number and positions of the knots. The choice depends on a priori ideas on local smoothness of the spline and on the familiar trade-off of bias and efficiency.

(d) *Variances and autocovariances of components*

A proper specification of the (time-varying variance) function for the innovations of the different components is needed to produce efficient estimators for the mean function and it is also required to provide realistic forecast error variances. In practice only a limited number of parameters modelling these variances can be estimated simultaneously. In the end it may be necessary to specify the "irregular" as a, possibly periodic, stationary ARMA-process to whiten the innovations of the measurement equation.

(e) *Additional regressors for different components*

Some variables may be available to explain changes in the different components. Effects for a single day of the week, or for single holidays may be captured in extra regressors. Innovation outliers can be modelled using single dummies. Dummies in the level or slope equation can model deterministic structural breaks.

(f) *Estimation and Diagnostics*

Estimation of the so-called hyperparameters, i.e. the free variances of the innovations of the different components, is performed by maximising the (prediction error decomposition) of the Gaussian likelihood. The current states of the conditional means and variances of the different components are available from the Kalman Filter output. The moments for previous time periods are estimated by smoothing. See §3 above. One can check for nonnormality, heteroskedasticity and serial correlation for the innovations, or for the auxiliary residuals of the different components, both intramonthly, intermonthly and interyearly. All other familiar and newly developed regression diagnostics can easily be programmed using a few lines of code.

Figure 5 presents two main in-sample diagnostics: autocorrelation function and density of one-step-ahead standardised forecast errors for the period 1993-1997.

4.2 Implementation

Many components of the menu of the previous section have been implemented in various well documented and tested software packages. The best known program is STAMP, see Koopman, Harvey, Doornik, and Shephard (2000), which is optimised for "standard" structural modelling of quarterly and monthly data. Although one can produce many useful results with STAMP, e.g. by specifying monthly models for the separate days of the week it is not really fit for day-to-day forecasting. First, it does not allow for data with 3 indices, therefore it does not allow for periodic models of the kind we are looking for. Second, it does not allow for the specification of time-varying splines. More importantly, it can only be used at one level of sophistication and user-friendliness.

Daily online forecasting requires programs at three levels of sophistication and user-friendliness. At the lowest level one needs a program to import and check new data and to

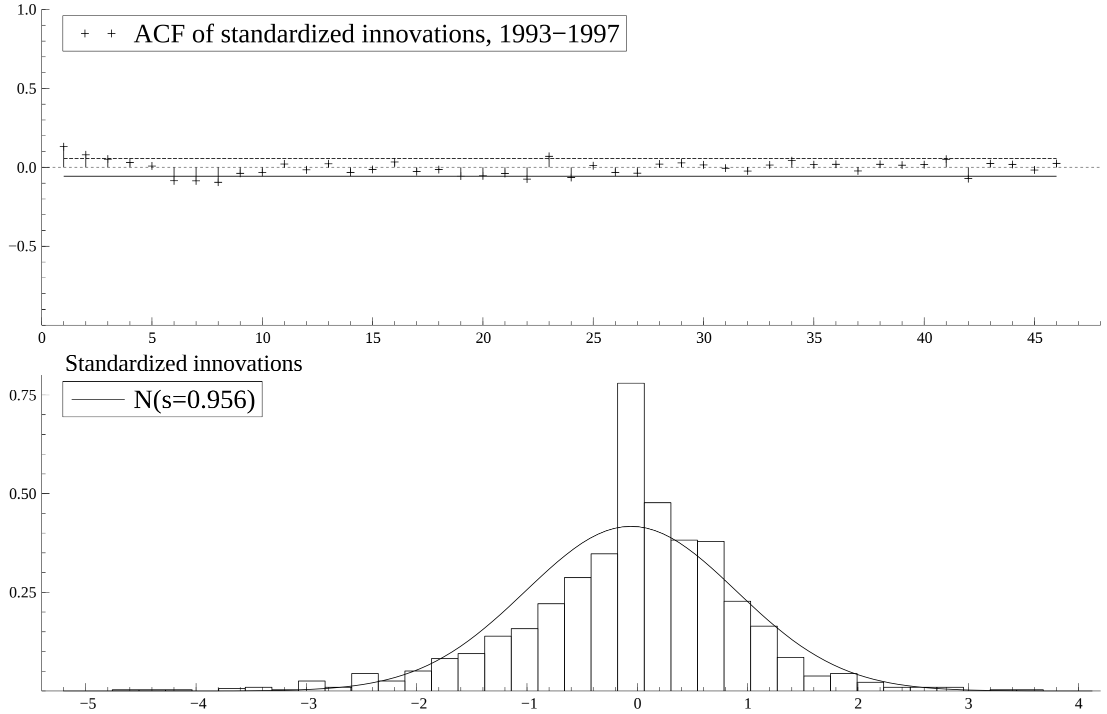

Figure 5: Diagnostics for standardised forecast errors in sample 1993-1997: $\hat{F}_t^{-1/2}\hat{v}_t$. Top panel: Autocorrelation up to lag 46. Bottom: histogram and normal density.

forecast using an existing model. The user records the observed revenue, puts it in an easily accessible database, which is specified in calendar time and updates the forecasts together with confidence intervals for the next few days. The forecasts are presented in calendar time and compared with the most relevant previous values (last month, or last year) and forecasts from other sources. Basic computer skills should suffice to operate at this level. We labelled this program `ETE`, Econometric Tax Estimator.

At the second level one may want to see more diagnostics, perform sensitivity analysis, and be able to fine-tune the model. This requires access to time series plots of components, standard errors, residuals, historical forecast records. The estimation sample and forecast sample for the states (in model time) can be changed. Standard components can be introduced or deleted and be made stochastic or deterministic. The number of knots and their positions can be changed. Individual observations may be downweighted or deleted. The hyperparameters can be reestimated occasionally. Regressors can be added. Basic computer skills and a practical knowledge of basic statistics and time series analysis should suffice to operate at this level. We labelled this program `STSM`, Structural Time Series Modeller.

At the highest level one may want to change the structure of the model, say a model with a strong intraweekly pattern, instead of a intramonthly pattern, introduce periodic seasonal heteroskedasticity, or a seasonal or periodic AR component, extend to forecasting for multivariate series, or introduce non-Gaussian errors. This level requires advanced practical and theoretical statistical knowledge and programming experience.

5 Empirical results

The project we describe in this paper first focussed on modelling and forecasting only the days around the turn of the month, since those are the day with the largest mean and variance and therefore the most relevant from a financial point of view. Even this first stage model performed at least as good as the existing method. That method was based on the distribution of the (remaining part) of a predicted value of the monthly total over the (remaining) days of the month, where monthly aggregate predictions were based on projections for the growth of the economy and (changes) in the different tax rates and collection policies. The parameters describing proportions were derived from a weighted average of the distribution measured for the same bank days, b_τ, in the same months, m_τ, in previous years, $Y_\tau - 1, Y_\tau - 2, Y_\tau - 3$.

Following the methodology described above, we have made the following modelling choices. First, for the time transformation from y_τ to $y_t = y_{j(t),s(t),p(t)}$, we picked $P = 23$ as described and motivated in §2. Second, we chose the following components. For the periodic intramonthly mean we chose a time-varying spline. For the periodic seasonal intrayearly movement we selected 3×2 deterministic dummy variables, 2 variables for each of the 3 days around the turn of the month, $p(t) = 22, 23, 1$: $b_\tau = -1, 0, 1$. An extra (long) spline function across the whole year, in our case depending on $P * (s(t) - 1) + p(t)$, see Harvey, Koopman, and Riani (1997), turned out to be insignificant. We also selected a nonperiodic stochastic trend, i.e. a trend that does not depend on $p(t)$, so that it can be taken to measure the overall level of tax revenues at a daily frequency. It was taken to have a fixed slope. Third, for the intramonthly spline we chose 10 knots at $p = (1, 2, 3, 5, 9, 15, 20, 21, 22, 23)$, thereby imposing smoothness only for the middle part of the month. Together with the 6 periodic seasonal dummies this makes a state vector of dimension 16 to describe the entire intrayearly pattern.

Fourth, we could identify four innovation variances, the so-called hyperparameters of the model, two for the intramonthly spline as discussed below, one for the level component and one for the irregular. The irregular itself has a periodic variance pattern as described below. This pattern was estimated using the residuals of the periodic regression model of Table 1, extended with a deterministic trend for each day of the month, for the sample 1993.3.1-1997.12.23. Fifth, we added three nonperiodic day-of-the week dummies $w_\tau = 3, 4, 5$, see Table 2 and a dummy measuring the length (in bank days) of the previous month, $M(t-P)$. The latter dummy could measure a trading day effect for VAT-revenues, which are collected after the month in which the value added is created. Sixth, we chose 1993.3.1 – 1997.12.23 as our estimation sample for the hyperparameters and 1997.1.1 – 1998.12.23 as a forecast period for one-step-ahead forecasts. Seventh, we estimated the model using maximum likelihood. The results are in the first row of Table 3.

Eighth, we present the following diagnostics: time series plots of in sample and out-of-estimation-sample one-step ahead forecasts errors, $\hat{v}_t$, and standardised forecast errors, $\hat{F}_t^{-1/2}\hat{v}_t$, both in the estimation sample and in the forecast sample, and the corresponding (nonperiodic) in-sample ACF, a normality test for the innovations and a CUSUM plot. The diagnostic graphs are presented in Figures 5 and 6.

Except for a single outlier in June 1998, our model fared very well up to the middle of October 1998, when an unexpected change in the pattern around the centres of the month appeared. This example illustrates that one should be able to make small but relevant changes to the model in a case like this, e.g. by changing the variances of the knots around the middle of the month. Figure 4 also illustrates the most important aspects of the component analysis: a plot of the intramonthly component at the end of 1997. This is the spline we wanted to identify when we started to look at time series plots like Figure 1. The state space analysis also allows us to estimate changes in the components over time.

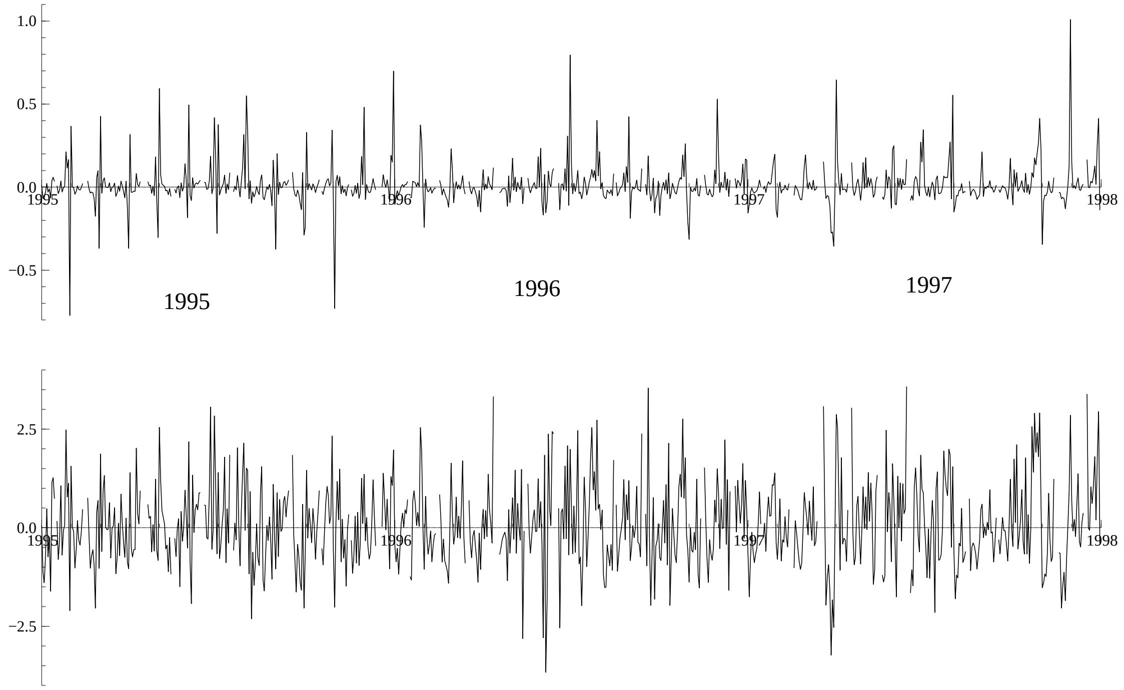

Figure 6: One-step (standardised) forecast errors in sample 1993.1.1-1997.12.23: $(\hat{F}_t^{-1/2})\hat{v}_t$. Top graph: nominal errors in 10^9 Euro. Bottom graph: standardised errors.

Periodic and seasonal heteroskedasticity tests and normality tests for auxiliary residuals of level and irregular, not reproduced here, indicate that this basic model does not fit all days and all months equally well. The variance seems to have increased during the sample period on some days of the month. Overall, the performance of mean and interval forecasts around the end of the month seems satisfactory. We can formulate the estimated model for y_t, $t = 1, \ldots, n$, up to 1998 as:

$$y_t = w_t' \lambda_t + \mu_t + x_t' \delta + G_t \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, \sigma_\varepsilon^2),$$

where the daily revenues y_t, are now measured in 10^9 Euro, where λ_t contains the 10 stochastic knots for intramonthly spline and where x_t contains 10 explanatory variables, 6 based on $s(t) = m_\tau$ for $p = 22, 23, 1$, 3 based on w_τ, and one based on $M(t - 23)$, see also Table 4 below. All regressors, except for the level, are demeaned such that they to have mean close to zero over the span of a year. The level component can therefore be interpreted as the current value of the mean across all bank days of the year.

The state space form of § 3 for knots of the spline is

$$\begin{aligned}&\lambda_{t+1} = \lambda_t + \nu_t, \qquad \nu_t \sim \mathrm{N}(0, \Sigma_\nu),\\ &\mathrm{diag}(\Sigma_\nu) = (\sigma_1^2, \sigma_1^2, \sigma_1^2, \sigma_2^2, \sigma_2^2, \sigma_1^2, \sigma_1^2, \sigma_1^2, \sigma_1^2, 0).\end{aligned}$$

The innovation variance for the last knot is also put to zero to avoid an identification problem for the level component. The level component is

$$\begin{aligned}\mu_t &= \mu_{t-1} + \beta_{t-1} + \eta_t, \qquad \eta_t \sim NID(0, \sigma_\eta^2),\\ \beta_t &= \beta_{t-1}.\end{aligned}$$

The periodic heteroskedasticity vector for the innovations with "basic" length $P = 23$ is estimated by periodic regression and normalised on the variance for $p(t) = 22$.

$$\begin{aligned}(G_1^2 \dots, G_{23}^2) = (&2.593, 0.153, 0.060, 0.072, 0.090, 0.059, 0.082, 0.068, 0.051, 0.1,\\ &0.1, 0.1, 0.1, 0.1, 0.121, 0.039, 0.079, 0.167, 0.250, 0.378, 1.123, 1, 5.766)\end{aligned}$$

The variances for $p(t) = 10, \dots, 14$ were interpolated from neighbouring values. We did not attempt full information maximum likelihood estimation of G_t. Sensitivity analysis did not show large effects of small changes in G_t on the other outcomes.

The variance estimates in Table 3 indicate a low variability of the spline near the middle of the month, a larger variability towards the end of the month, as expected from the results of Table 1. The moment estimates for the states of the different components at the end of 1997 are presented in Table 4. The estimate for the level of $.65 \cdot 10^9$ Euro per day is larger than the sample average, indicating an upward trend, which helped the Dutch government to reduce the budget deficit significantly. The recursive estimate of this trend (not reproduced here, but naturally available in the graphical output of the program) is relatively straight.

Table 3: *Estimated hyperparameters*

Sample	σ_1^2	σ_2^2	σ_η^2	σ_ε^2
1993.3.1 – 1997.12.23	2.82e-6	7.33e-7	3.69e-5	0.0179
1993.3.1 – 1998.12.23	4.62e-6	1.11e-6	7.20e-5	0.0186

Observations measured in 10^9 Euro. Variances normalised for the penultimate day of the month $p(t) = 22$, $b_\tau = -1$, Last row presents estimates for extended model

Table 4: *Estimated states at 1997.12.23*

State	mean	t-value
$\lambda(p=1)$	0.205	5.048
$\lambda(p=2)$	-0.305	-14.86
$\lambda(p=3)$	-0.380	-23.77
$\lambda(p=5)$	-0.395	-26.43
$\lambda(p=9)$	-0.361	-35.06
$\lambda(p=15)$	-0.350	-34.16
$\lambda(p=20)$	0.004	0.21
$\lambda(p=21)$	0.125	5.24
$\lambda(p=22)$	0.765	24.40
μ	0.652	15.35
β	1.02e-4	0.561
Tuesday	-0.019	-5.47
Wednesday	-0.015	-4.36
Thursday	-0.012	-3.39
$M(t-P)$	-0.004	-2.26
$p(t)=1, s(t) \mod 3 = 1$	0.098	1.59
$p(t)=22, s(t) \mod 3 = 1$	0.186	4.83
$p(t)=23, s(t) \mod 3 = 1$	0.708	7.63
$p(t)=1, s(t)=6$	0.166	1.61
$p(t)=22, s(t)=6$	0.366	5.67
$p(t)=23, s(t)=6$	1.284	8.19

Estimation sample 1993.3.1 – 1997.12.23
λ: spline (see also Figure 4), μ: level, β: slope

5.1 Model evaluation

Residual serial correlation in Figure 5 at lag 1 points to richer dynamics within the months and the residual serial correlation at lag 23 points to richer dynamics across months than allowed for by the model. From the analysis of §1 we know the extension of the model is likely to be of a periodic nature: the correlations are only important for some bank days of the month. Rather than switching to a complete periodic structure we included estimates of past forecast errors as extra periodic regressors in the model. These regressors are extended as time progresses, but past values are not revised. The parameter estimates of the coefficients of these regressors can be considered as first-step estimates of a periodic moving average structure in the error term, see Table 5. In this table $p(t) = 1, v(t-23)$ denotes a regressor consisting of one-step-ahead forecast errors, v_t, from the Kalman Filter, lagged one month, only applied to the first bank day of each month. The most important extra regressors are applied for the last bank day of each month, $p(t) = 23$. One regressor captures the correlation with cumulative forecast errors of the previous two days: $v(t-1) + v(t-2)$. The other captures the correlation with the the forecast error one month before, $v(t-23)$.

The second row of Table 3 presents the hyperparameters for the updated model, using data up to 1998.12.23. The variance ratios change considerably compared to the previous sample and corresponding model. The relative variance of the spline component and the level component increased.

5.2 Forecast results

The top graph of Figure 7 presents the forecasts for 1999 using the hyperparameter estimates up to 1998. A patch of huge outliers stands out around the end of April 1999. These outliers were caused by the introduction of the opening of the bank system and tax collection Queens day, the 30th of April, formerly a bank holiday, which was not anticipated by the taxpayers. The model produced forecast errors (realization - forecast) of 0.4, 1.0 and 3.3 billion Euro on the days leading up to the 30th and a “compensating” forecast error of -4.7 billion Euro on the 30th. The corresponding standardised forecast errors were about 4, 7, 20 and -13. Automatic indexing of these observations was clearly not appropriate, but simply ignoring April the 30th and treating it as missing would not have been adequate either. Two other very large standardised forecast errors are observed for the 16th of June and the 9th of July. These observations were also adjusted in order to reduce their influence on our subsequent forecast evaluation.

Figure 7 also presents the forecasts for 1999 using the outlier corrected data, which were obtained by subtracting the rounded forecast errors on the last four bank days of April. Similar corrections were made for June 16, 17 and 18 (-0.3 0.15, 0.15 billion) and for July 9 (-0.5 billion).

Table 5 shows the estimates of the state vector at the end of 1999, both with and without outlier correction. The effect of the outliers is largest for the state-variables that relate closely to the observations with $p(t) = 21, 22, 23$. There is a significant effect of the outliers on the coefficient of quarterly regressors for $p(t) = 22$ and $p(t) = 23$, positive and negative respectively. Note that the effective number of available observations for each of these regressors is only 27. There is a negligible effect of the outliers on $\hat{u}_t$. Comparing Table 4 with Table 5 one notes that the overall tax level, $\hat{u}_t$, increased considerably from 1998.12.23 to 1999.12.23

The bottom graph of Figure 7 shows that the outlier correction leads to good forecasts for $p(t) = 22$ and $p(t) = 23$ in July and October. Table 5 also shows that the estimates of the time-varying spline for the knots $\lambda(p = 21)$ and $\lambda(p = 22)$ are significantly affected by the outlier, even 8 months after they occurred.

Comparing the forecast errors of 1998 in Figure 4 and the forecast errors of 1999 in Figure 7, we see a reoccurrence of the positive deviations towards the middle of October and November. The pattern is less prominent in 1999, so without external information on the change in the tax collection procedures, such as a continued increase in the amount of tax revenues due around the middle of the month, it is not clear that this already merits a change in the model.

5.3 Comparison with naive forecasts

Table 6 compares our model forecasts over the years 1998 and 1999 with simple seasonal random walk forecasts. We use our model as specified and estimated up to 1997.12.23. We use the outlier corrected data.

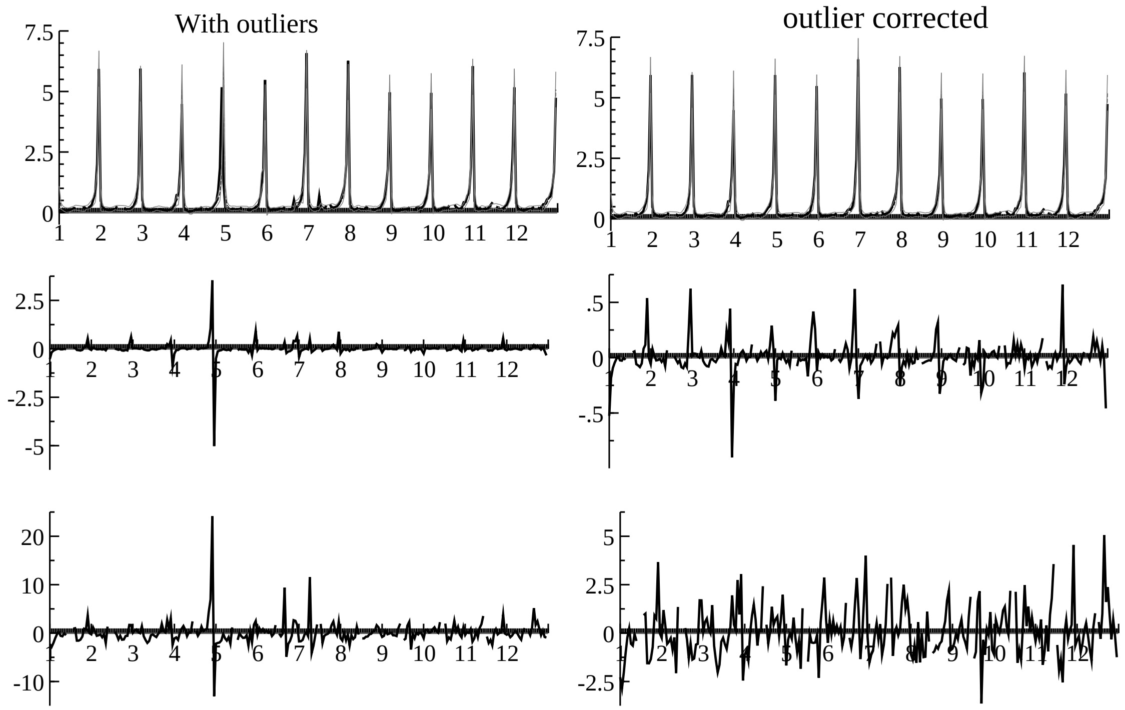

Figure 7: Forecasts for 1999 of Model with hyperparameters of second row of Table 3. Left hand side graphs: Top: One step ahead forecasts, 95% forecast intervals, realisations. Middle: nominal forecast errors. Bottom: standardised forecast errors. Right hand side graphs: as Left hand side, but with outliers in 1999 corrected

In the seasonal random walk (SRW) forecast we consider seasonality with period

$$S \cdot P = 12 \cdot 23 = 276, \quad \hat{y}_{(t+S \cdot P)} = y_t.$$

The estimated revenues equal those of the corresponding day in the previous year, provided the relevant observations are not missing. The underlying model consists of 276 independent random walk processes. The seasonal random walk forecasting procedure is robust to permanent changes in the tax level and its seasonal patterns. It provides competitive forecasts for most days of the year, although the effective forecast horizon is one year, instead of one day.

Comparing the root mean-squared-errors in Table 6 we see that the one-day-ahead forecast errors of the model outperform the one year ahead forecast errors of the seasonal random walk model by twelve percent.

Figure 8 presents the out-of-sample forecast errors by day of the month. Each graph shows the forecast errors for January 1998 to December 1999. It shows how the variance differs from day to day. These plots can be used to detect model misspecifications at monthly lags. They show that the model was slow to pick up a “local” trend in 1999 for days 21 and 22 of the month, ($p = 22$), which is reflected in the large forecast RMSE for these days, presented in Table 6. The seasonal random walk did not do better in this respect, since it adjusts its trend only with a lag of 12 months.

Table 5: *Estimated states at 1999.12.23*

	With outliers in 1999		Outlier corrected	
State	mean	t-value	mean	t-value
$\lambda(p=1)$	0.164	3.56	0.171	3.70
$\lambda(p=2)$	-0.344	-14.51	-0.344	-14.50
$\lambda(p=3)$	-0.418	-22.28	-0.412	-21.98
$\lambda(p=5)$	-0.428	-24.57	-0.428	-24.54
$\lambda(p=9)$	-0.348	-29.77	-0.344	-29.43
$\lambda(p=15)$	-0.354	-30.36	-0.356	-30.59
$\lambda(p=20)$	0.063	3.07	0.058	2.80
$\lambda(p=21)$	0.304	11.07	0.251	9.13
$\lambda(p=22)$	1.306	36.57	1.175	32.89
μ	0.759	14.47	0.767	14.64
β	1.09e-4	0.56	1.08e-4	0.55
Tuesday	-0.018	-5.94	-0.016	-5.10
Wednesday	-0.019	-6.37	-0.017	-5.71
Thursday	-0.015	-5.12	-0.014	-4.66
$M(t-P)$	-0.006	-3.05	-0.007	-3.24
$p(t)=1, s(t) \mod 3 = 1$	0.114	2.13	0.116	2.15
$p(t)=22, s(t) \mod 3 = 1$	0.306	9.25	0.183	5.53
$p(t)=23, s(t) \mod 3 = 1$	0.523	6.47	0.687	8.50
$p(t)=1, s(t)=6$	0.138	1.55	0.144	1.62
$p(t)=22, s(t)=6$	0.402	7.22	0.405	7.27
$p(t)=23, s(t)=6$	1.26	8.80	1.422	9.87
$p(t)=1, v(t-23)$	-0.174	-2.62	-0.177	-2.64
$p(t)=21, v(t-23)$	-0.365	-4.91	-0.523	-5.82
$p(t)=23, v(t-1)+v(t-2)$	0.052	2.28	-0.028	-1.17
$p(t)=23, v(t-23)$	0.003	3.10	0.002	2.36

Estimation sample states 1993.3.1–1999.12.23. See also Table 4
Estimation sample hyperparameters 1993.3.1–1998.12.23.

6 Summary and Conclusion

In this paper we have presented a modelling strategy for daily time series with a clear intramonthly pattern and a changing number of observations per month. We have applied the strategy and implemented an online one-step-ahead forecasting model for daily tax revenues.

The methodology consists of three stages. First, we need to choose an appropriate time indexing with three levels: year, season, and period within the season. Second, we require to specify a regression model with time-varying parameters that captures the year-to-year, seasonal and periodic movements. In our case the dominating periodic movements are modelled using time varying cubic splines and the seasonal movements are modelled by standard techniques. Third, recursive estimation and forecasting using state space techniques is carried out together with extensive graphic diagnostic analysis.

We have applied the technique to Dutch tax revenues and have produced out-of-sample forecasts for 1998 and 1999 with satisfactory results. However, the diagnostics indicate that our model can be extended. For example, modelling of cumulative sums of daily tax revenues can be considered. This extension is technically more demanding and requires a treatment based on flow variables. It allows external information on monthly totals to be

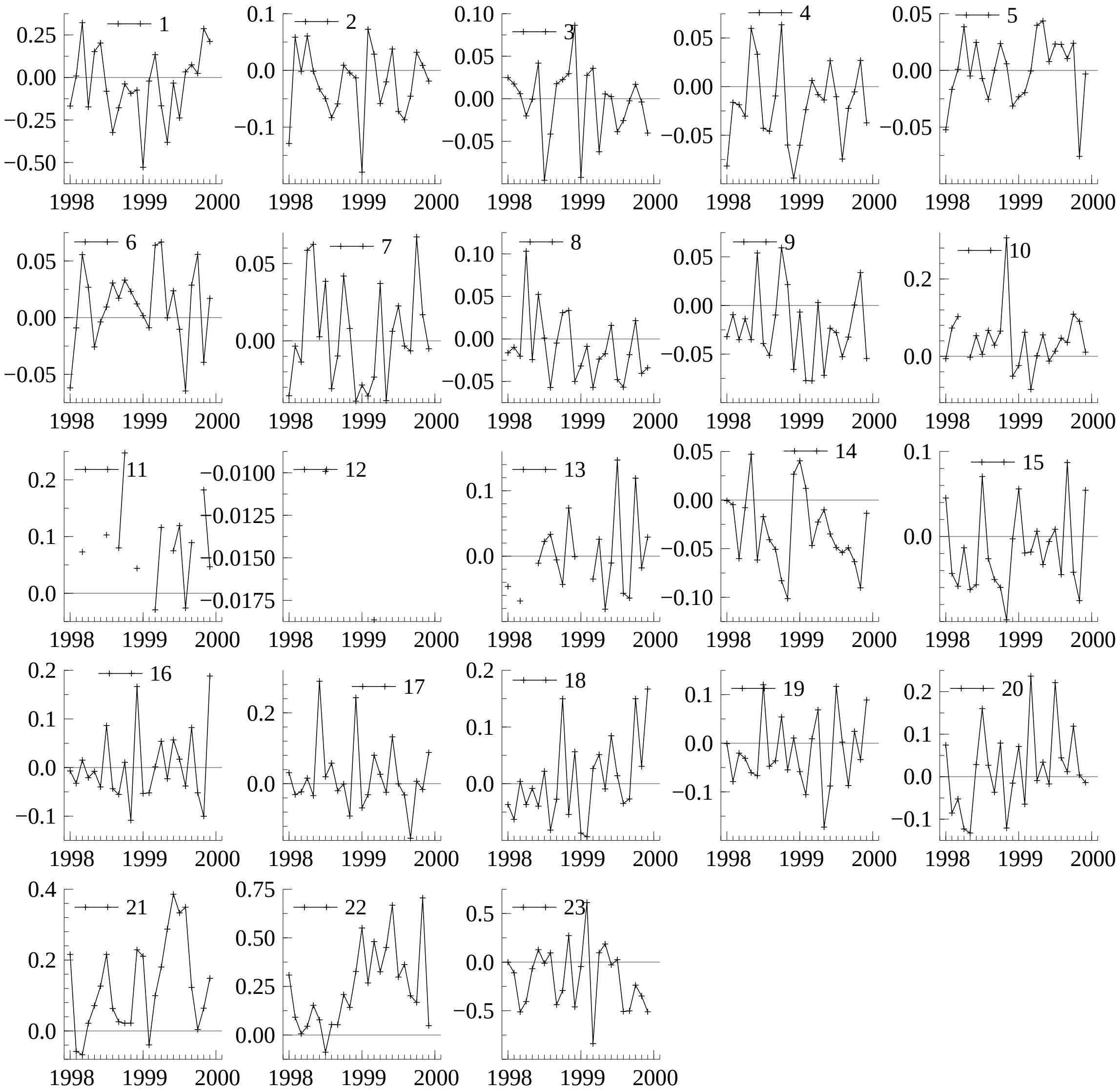

Figure 8: One-step-ahead recursive forecast errors for outlier corrected data 1998.1-1999.12, monthly intervals, by day of the month p

incorporated: first, by adjusting the forecasts over longer horizons; second by testing the viability of these external forecasts online; and third by examining to what extent daily modelling can improve forecasts for monthly totals and vice versa.

Acknowledgements We thank Robert Rooderkerk and especially Björn de Groot and Dennis Fok for their significant contributions to the development of the computer programs. We thank Frans Hooijmans of the Dutch Ministry of Finance for providing us with the data, for comments and suggestions, and for the initiation and organisation of the project. We thank seminar participants at Nuffield College, Oxford, for constructive remarks.

Table 6: *Out-of-sample recursive forecasts RMSE*

p	STSM	n (STSM)	SRW	n (SRW)
1	209	24	275	24
2	64	24	46	24
3	42	24	54	24
4	44	24	34	24
5	28	24	38	24
6	36	24	34	24
7	33	24	41	24
8	39	24	46	24
9	44	24	44	24
10	84	23	116	21
11	112	13	95	7
12	15	2	5	1
13	60	19	78	15
14	49	24	54	24
15	50	24	47	24
16	72	24	51	24
17	96	24	83	24
18	72	24	65	24
19	73	24	71	24
20	98	24	120	24
21	180	24	181	24
22	321	24	382	24
23	360	24	384	24
overall	133	513	151	500

p_τ indexes model day of the month.
Measurements in 10^6 Euro.
Estimation sample: 1993.3-1997.12.23.
Forecast sample 1998.1.1-1999.12.23.
STSM: Model forecasts errors: see Tables 3 and 4
n: number of forecasts
SRW: Forecasts errors Seasonal Random Walk

Appendix

Time varying cubic splines

The regression spline function is defined as a smooth function through the data points y_t which are a response to the scalar series x_t, for which $x_t < x_{t+1}$ and $t = 1, \ldots, N$. In the daily tax model, x_t is the day of the month, $p(t)$, and N is chosen as P. Harvey, Koopman, and Riani (1997) used the calendar day of the year as explanatory variable. In our notation, see Table 2, this would correspond to setting $x(t) = (s(t) - 1) \cdot P + p(t)$ and selecting $N = S \cdot P$. The spline model is

$$y_t = \theta(x_t) + \varepsilon_t, \qquad \mathrm{E}\,(\varepsilon_t) = 0, \qquad \mathrm{Var}\,(\varepsilon_t) = \sigma^2,$$

where $\theta\,(\cdot)$ is a smooth function which is based on $k+1$ knot points $(x_0^\dagger, y_0^\dagger), \ldots, (x_k^\dagger, y_k^\dagger)$. The smoothness of $\theta\,(\cdot)$ is created by setting its second derivative with respect to x as a linear function of $k+1$ coefficients, that is

$$\theta_i''(x) = [(x_i^\dagger - x)/d_i]a_{i-1} + [(x - x_{i-1}^\dagger)/d_i]a_i$$

with $d_i = x_i^\dagger - x_{i-1}^\dagger$ and $\theta(x) = \theta_i(x)$ for $x_{i-1}^\dagger < x < x_i^\dagger$ and $i = 1, \ldots, k$. The $k+1$ coefficients a_i are assumed fixed and they can be identified by solving a linear set of equations.

These unknown coefficients are obtained as follows: (i) consider $\theta_i''(x)$ and use standard integration rules to obtain expressions for $\theta_i(x)$; (ii) enforce the spline function $\theta_i(x)$ at $x = x_i^\dagger$ to be equal to its known value of $y_i^\dagger$; (iii) restrict the first derivative to be continuous by enforcing $\theta_i'(x_i^\dagger) = \theta_{i+1}'(x_i^\dagger)$ for $i = 1, \ldots, k-1$. Step (ii) leads to a linear expression for $\theta_i(x)$ in terms of $y_i^\dagger$ and a_i, for $i = 0, \ldots, k$. Step (iii) leads to $k-1$ linear equations of $k+1$ coefficients $a_0, \ldots, a_k$ in terms of $y_0^\dagger, \ldots, y_k^\dagger$. The 'natural' restrictions $a_0 = a_k = 0$ allow solving this linear system with respect to the remaining coefficients a_i for $i = 1, \ldots, k-1$. The spline function can now be written in terms of $y_0^\dagger, \ldots, y_k^\dagger$ by

$$\theta(x_t) = \theta_i(x_t) = w_{0,t}y_0^\dagger + \ldots + w_{k,t}y_k^\dagger, \qquad x_{i-1}^\dagger < x_t < x_i^\dagger, \qquad , t = 1, \ldots, N,$$

where the weights $w_{0,t}, \ldots, w_{k,t}$ depend on the knot positions $x_0^\dagger, \ldots, x_k^\dagger$ and the value for (or implicitly the position of) x_t. For a given set of values $y_0^\dagger, \ldots, y_k^\dagger$, the spline function can be computed for any $x_0^\dagger < x < x_k^\dagger$. The regression spline also can be expressed as

$$\theta(x_t) = w_t' y^\dagger,$$

where $w_t = (w_{0,t}, \ldots, w_{k,t})'$ and $y^\dagger = (y_0^\dagger, \ldots, y_k^\dagger)'$.

In the case that $y_0^\dagger, \ldots, y_k^\dagger$ are not known, we can replace them by the coefficients $\lambda_0, \ldots, \lambda_k$ which may be estimated by least squares methods. For a given set of data points and a set of knot positions $x_0^\dagger, \ldots, x_k^\dagger$, the spline model can be expressed by the standard regression model

$$y_t = w_t' \lambda + \xi_t,$$

where the parameter vector $\lambda = (\lambda_0, \ldots, \lambda_k)'$ is estimated by $\left(\sum(w_t w_t')\right)^{-1} \sum w_t y_t$ and standard regression inference applies. Poirier (1976) gives more details.

The generalisation of time-varying regression splines within the state space framework is developed by Harvey and Koopman (1993). Time-varying splines are obtained by letting parameter vector λ evolve slowly over time, for example

$$\lambda_{t+1} = \lambda_t + \nu_t, \qquad \nu_t \sim \mathrm{N}(0, \Sigma_\nu),$$

where Σ_ν is a diagonal variance matrix.

The spline function can be used as a seasonal component within the structural time series model. A periodic variance function for the error term can be added. The summing-to-zero constraint, which eliminates the collinearity with the trend component, for a time-varying spline can also be implemented; details are given by Harvey and Koopman (1993). See Figure 4 for an example.

References

Anderson, B. D. O. and J. B. Moore (1979). *Optimal Filtering*. Englewood Cliffs: Prentice-Hall.

Burridge, P. and K. F. Wallis (1990). Seasonal adjustment and Kalman filtering - extension to periodic variances. *Journal of Forecasting 9*, 109–118.

De Jong, P. (1988). A cross validation filter for time series models. *Biometrika 75*, 594–600.

De Jong, P. (1991). The diffuse Kalman filter. *Annals of Statistics 19*, 1073–1083.

Doornik, J. A. (1998). *Object-oriented Matrix Programming using Ox*. London, U.K.: Timberlake Consultants Ltd.

Doornik, J. A. and D. F. Hendry (1999). *GiveWin: An Interface to Empirical Modelling* (2nd ed.). London, U.K.: Timberlake Consultants Ltd.

Harvey, A. and S. J. Koopman (1993). Forecasting hourly electricity demand using time-varying splines. *Journal of the American Statistical Association 88*, 1228–1237. nog te kopiëren.

Harvey, A., S. J. Koopman, and M. Riani (1997). The modeling and seasonal adjustment of weekly observations. *Journal of Business and Economic Statistics 15*, 354–368. [SE.PS. Festival effect increases over time. August and Easter/Spring/May holidays. One-step-ahead prediction errors in-sample (filtered estimates) mostly less than 0.5% Multistep predictions a year ahead also OK.

Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman Filter*. Cambridge, UK: Cambridge University Press. [4.[3.[2.UR.SE.CI.MS.

Harvey, A. C. and S. J. Koopman (1992). Diagnostic checking of unobserved components time series models. *Journal of Business and Economic Statistics 10*, 377–389. SA.SE.[ST. only applicab on innovations (white noise null applies). Comparison structural and canonical decomposition (a ala Pierce: maximize vairance irregular) residuals: simple ARIMA(0,1,1) Irr. same. Level:Can.res. is MA(2) of S.res.

Harvey, A. C., S. J. Koopman, and J. Penzer (1998). Messy time series: A unified approach. In T. B. Fomby and R. Carter Hill (Eds.), *Advances in Econometrics, Volume 13*, pp. 103–143. New York, NY, USA: JAI Press.

Kitagawa, G. and W. Gersch (1996). *Smoothness Priors Analysis of Time Series*. New York: Springer Verlag.

Kohn, R. and C. F. Ansley (1989). A fast algorithm for signal extraction, influence and cross-validation in state space models. *Biometrika 76*, 65–79. [OL.IA.STSP.

Koopman, S. J., A. C. Harvey, J. A. Doornik, and N. Shephard (2000). *STAMP, Structural Time Series Analyser, Modeller and Predictor*. London, U.K.: Timerlake Consultants Press.

Koopman, S. J., N. Shephard, and J. A. Doornik (1999). Statistical algorithms for models in state space using SsfPack 2.2. *The Econometrics Journal 2*, 107–160.

McLeod, A. I. (1994). Diagnostic checking of periodic autoregression models with application. *Journal of Time Series Analysis 15*, 221–233. PAR.SE.

Poirier, D. (1976). *The Econometrics of Structural Change: With Special Emphasis on Spline Functions*. Amsterdam: North-Holland. nog te kopiëren.

Young, P. C. (1984). *Ecursive Estimation and Time Series Analysis*. New York, NY, USA: Springer-Verlag.