Time-Series Modeling of Daily Tax Revenues

Björn de Groot*, Siem Jan Koopman** and Marius Ooms*

8 July 1999

Econometric Institute
Erasmus University Rotterdam
P.O. Box 1738
3000 DR Rotterdam, The Netherlands
email: ooms@few.eur.nl
URL: `http://www.eur.nl/few/ei/links/ooms`

Abstract

This paper discusses a time-series model for daily tax revenues. The main feature of tax revenue series is the pattern within calendar months. Standard seasonal time series techniques need to be modified, because the number of trading days per calendar month varies from month to month and from year to year. The model is an unobserved components model, with a trend and seasonal components that vary over time. The seasonality for inter-month and intra-month movements is modelled using stochastic cubic splines. The model is made operational and used to produce daily forecasts at the Dutch Ministry of Finance. A front-end for model configuration and data input is implemented with Visual C++, while the econometrics and graphical diagnostics are built around Ox, and `SsfPack`, which implements general procedures for the Kalman filter and state space models.

Acknowledgements We thank Robert Rooderkerk and especially 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 organization of the project.

* Econometric Institute, Erasmus University Rotterdam, The Netherlands.

** Free University Amsterdam, The Netherlands.

1 Introduction

Daily forecasts for tax revenues play an important role in day-to-day cash management at the Treasury. Money markets should not be disturbed by surprise shocks due to last-minute lending or borrowing amounts of the central government. The size of these shocks can be diminished using tax revenue forecasts that are as realistic as possible. The main purpose of the statistical daily time series model is to process information of revenues on previous days systematically, and as efficiently as possible. Central government outlets are usually known at least one day ahead. So, forecasts several steps ahead can also be used to monitor the monthly targets for the budget.

Forecasting daily time series is a difficult task that has to performed on an on-line basis. State space modeling with its recursive estimation techniques and corresponding statistical methodology is an attractive method. The dominating feature of many daily time series is a seasonal pattern that changes persistently over time. Many interesting seasonal models, including simple regression models and Holt-Winters-type smoothers, are suited for analysis in state space form. This allows for statistical testing of different models against each other. Proper use and monitoring of a sophisticated statistical model is not trivial. The model should be run online without the regular assistance of a recent Ph.D. in statistics. Fortunately, modern computer technology makes this possible. The basic procedures for estimation, filtering and smoothing algorithms are available in a well documented package, `Ssfpack`, Koopman, Shephard, and Doornik (1999) for the object-oriented matrix programming language Ox, Doornik (1998). Procedures to put well known structural and ARMA models into state space are also supplied in the package. Day-to-day forecasting only involves updating the database and pushing a button. The menus can be made in a Visual programming language, like Visual Basic or Visual C++.

Setting up the basic model and the diagnostics is the subject of this paper. Once the model is up and running one can modify and completely reestimate it in an interactive (point-and-click) environment. Here graphical output is of primary importance. For this purpose we use GiveWin 1.2, see Doornik and Hendry (1996). In order to avoid unwanted trial-and-error modelling we set up the model in a so-called structural form, following the approach Harvey, Koopman, and Riani (1997) introduced for weekly data. The basic idea is simply a periodic regression model with persistent time varying parameters. The basic assumptions of linearity and Gaussianity are implied. Non-Gaussian and non-linear extensions are not yet practically feasible, so outliers and regime shifts must be accounted for by the on-line modeller, e.g. by downweighting observations, or by adding dummy variables.

Basic econometric knowledge should suffice to keep the model in shape.

The modeling of daily time series involves several issues which are not well covered in the literature. We address these issues in §2. We illustrate them using our series of Tax Revenues, for which we also set up notation. We discuss the main sample characteristics that are needed for the identification of a basic daily time series model. The remainder of this paper is structured as follows. Section 3 addresses the specifics of structural time series modeling, the basic idea, the notation corresponding to `Ssfpack`, 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 and forecasts for time aggregates deserve special attention. Section 4 presents our model for daily tax revenues. In §5 we apply our model, estimating it for data upto 1997 and producing on-line forecasts for 1998. Section 6 suggests some extensions to our approach, including forecast testing by the comparison with other models, benchmarking forecasts for time aggregates with information from other models and concludes.

2 Daily time series features

Daily economic time series often have properties that makes them harder to model and 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: First, irregular spacing: the number of observations varies per month and per year and second, a variance that depends on the day-of-the month. Many aggregate economic transactions have a pattern with a clear peak once a month, e.g. salary payments, money circulation, and tax revenues. It is often not easy to stabilize the variance by taking logs: the (persistently changing) seasonal pattern is not simply multiplicative and the irregular 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 more regularly spaced data, such that standard Kalman Filter techniques kan be applied. The problem of a periodically varying variance of the seasonal pattern and the irregular is solved by applying a time varying Kalman Filter.

Other features of economic time series present more problems for data at a daily frequency, in particular, strange as it may seem at first sight, small sample problems. Daily

patterns show frequent structural breaks, due to important institutional changes in the financial and in the tax system, much more clearly than monthly or yearly data. These breaks are often so important that it does not seem to make sense to combine pre-break and post-break data for the daily model. Note that the structural breaks in the daily pattern do have an important effect on monthly data, when they occur around the turn of the month! Since there is usually not too many years of comparable daily data available, we cannot estimate long term trends and the monthly pattern in a very flexible way. 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. Upto 1997 this series also contains a (negative) component of tax restitutions 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, many revenue categories are compiled on a monthly basis. However, these figures are not immediately available after the turn of the period. 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. Relevant tax assessments that are due are not available on a daily basis. Monthly data on (expected) tax revenues per category can be used to evaluate the net monthly sums of revenues and restitutions. These monthly data can be be expected to play a more important role for the daily forecasts towards the end of the month. Given realizations upto the last few days, forecasts for the remaining days of the month will automatically imply a forecast for the monthly total.

Figure 1 for the daily Dutch central tax revenues in May and June of 1996 and 1997 illustrates the main features. We like 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 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 month to month. 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.

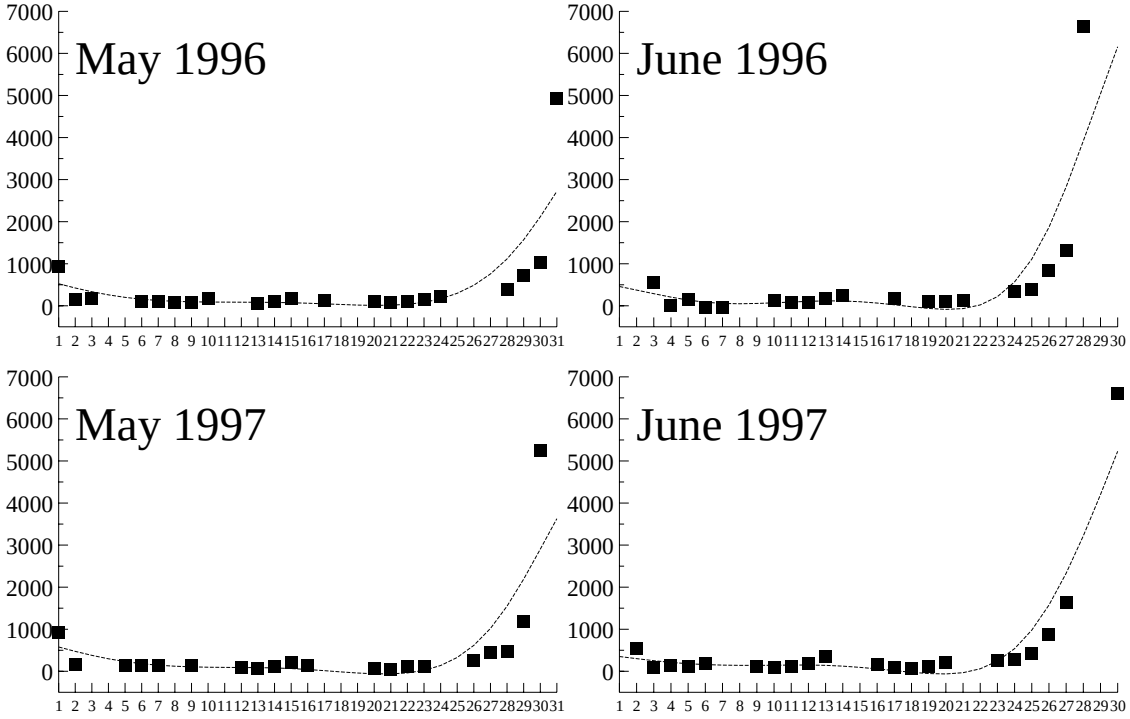

Figure 1: Daily Dutch national tax Revenues in billions of euro

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, see the natural cubic spline fits (with 5 equivalent parameters in Figure 1, c.f. Doornik and Hendry (1996)).

The splines seem to describe most of the data well and seem to pick up a local maximum in income around the middle of the month. The problems with this approach show most clearly towards the end of June 1996 and June 1997. The fitted income patterns vary across 1996 and 1997, whereas the observed income pattern at the last bank-days-of-the-month is very similar: the irregular spacing leads to an exaggerated time-variation across years. The spline estimates (which minimize the sum of squared deviations across all observations) also show that we might want to vary the weight of the observations and smoothness within the month. Smoothness can certainly be imposed in the first half of the month. 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 few points in the first half of the month and more points around the turn of the month, c.f. 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 modeled in a discrete time linear state space model. We index these "state space observations" by $t = 1, \ldots, T$. The mapping $t(\tau)$ defines the state space index as a function of calendar time $\tau = 1, \ldots, n$. In our case we use the following functions of calendar time: Y_τ: Calendar Year, $1993, \ldots, 1999$, d_τ: Day of the Month, $1, \ldots, 31$, m_τ: Month of the Year, $1, \ldots, 12$, w_τ: Day of the Week, $1, \ldots, 7$, h_τ: 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 gives the time transformation of the data of Figure 1 where we have chosen a constant underlying grid of 100 points each month. The pattern is now much more regular, both across years and across months. We have created more missing observations, but this does not present major problems for our estimation procedure, see §3.

Figure 3 presents a more complete picture of the revenues on the 1461 bank days used in 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. 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 which show in the lower half of the plots. In the sequel of this paper we use the adjective seasonal to describe the month-of-the-year effect. A seasonal difference means the difference with the corresponding value one year before. We use the adjective periodic to describe the day-of-the-month effects in the mean, the variance and the autocovariance. Periodicity refers to the pattern that occurs once a month. In the next subsection we specificy a simple periodic

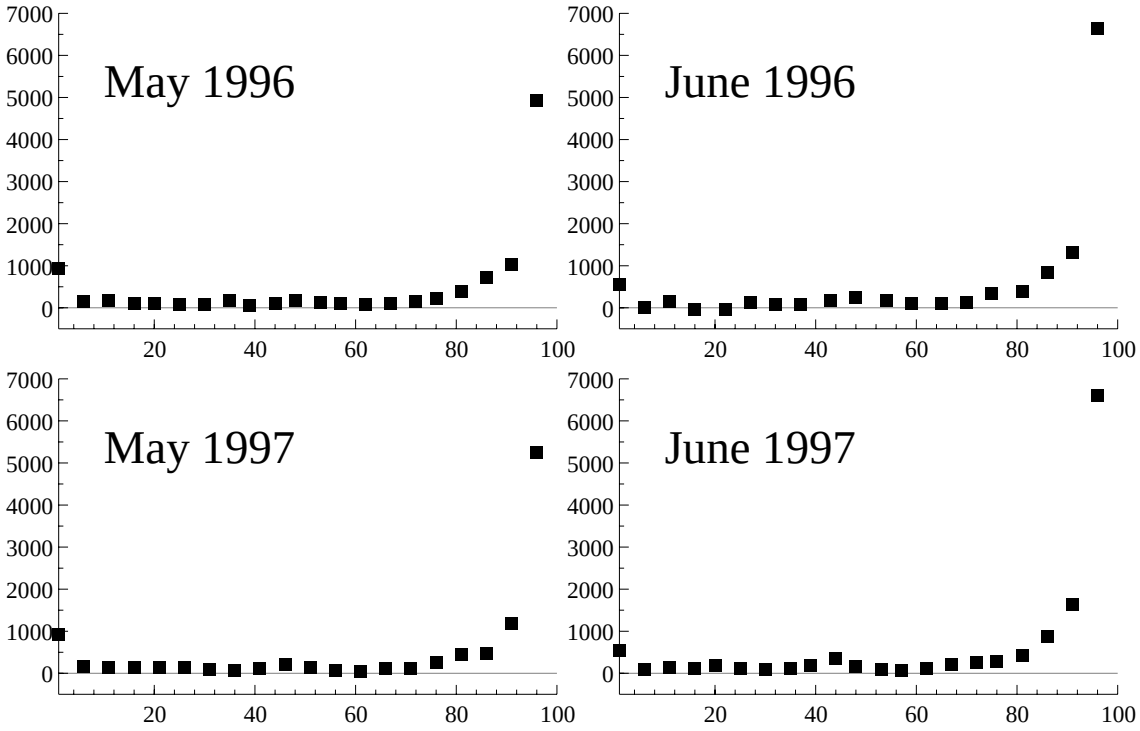

Figure 2: Daily Dutch national tax Revenues in billions of euro after time transformation

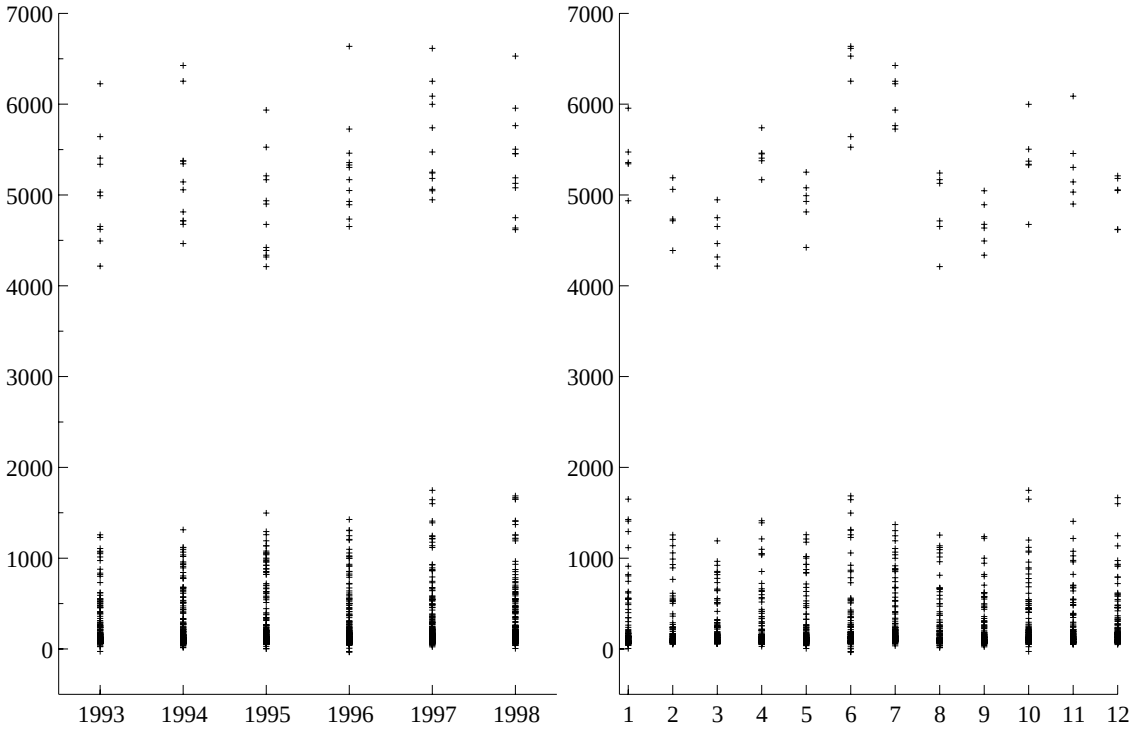

Figure 3: Daily Dutch national tax revenues in billions of euro against year and against month-of-year

regression model to capture seasonal and periodic properties, before going into the details of the time transformation for Figure 2.

2.1 Initial regression model

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 a flexible regression model to summarize the main properties of this unconditional mean-function. We use the residuals to estimate the periodic variances and covariances we would like to exploit in our statistical forecasting model.

The initial analysis showed 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 regression analysis we introduce the bank-day index $b_\tau = -15, -14, \ldots, 14, 15$, which equals the number of bank days since the beginning of the sample,$r_\tau = 1, \ldots, R$ minus the number of bank days until the nearest turn of the month $l_\tau = 1, \ldots, L$, where $l_\tau = 0$ for $\tau < 15$. Therefore, the last bank day of the month has $b_\tau = 0$.

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 analyze the variance and covariance function of these $70 \times 18 = 1260$ observations we basically regress them on 12×18 dummy variables, each dummy measuring the mean of y_t for a particular combination of b_τ and 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 centered seasonal dummies. In this way we allow automatically for periodic heteroskedasticity depending on b_τ. We present results in Table 1.

The first column summarizes the periodic mean function across all months. It reproduces the pattern seen in Figure 2 above. The function is smooth except at the exact turn of the month. The second 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 third column shows a simple estimate of (deterministic) seasonality for each bank day. Under a white noise assumption for the residuals a 5% critical value for $\widehat{R}$ of 0.5 could be used 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 estimate the serial correlation at daily intervals. Most large correlations are seen for the days at the end of each month. This could indicate the systematic presence of local trends. The only consistent series of negative (but small) correlations is seen for the revenues of the first day of each month. These revenues show a negative correlation with all 8 previous banking days. These periodic covariances determine also the periodic variance of the partial sum process towards the end of the month $\mathsf{Var} \sum_{b=-8}^{i} y_b$, see Table 2. The variance of the revenues increases the more days are aggregated, but the variance of the partial sum decreases when the revenues of the first day of the following month included in the sum. Forecasting the time aggregate including the first day of the month is easier than excluding the first day. In other words too low or too high aggregate revenues on the last days of the month are to some extent compensated by high or low revenues on the first day of the following month.

The results of Table 1 clearly motivate a periodic analysis. The results do not give directions for the specification of a model for seasonality and long term trends in the data. We present time series plots of the residuals for each of the regressions to further investigate these points. See Figures 5 and 4. Note the different scales on the plots. The first day and the last days of the month are clearly the most important from a practical financial forecasting point-of-view.

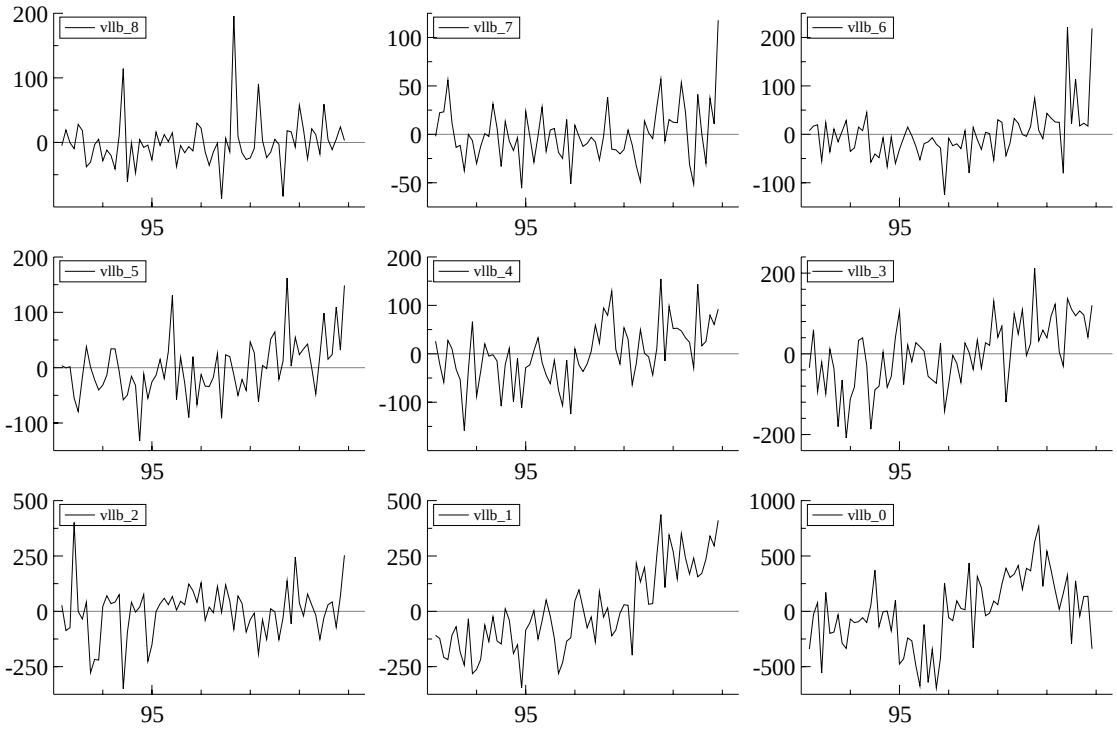

Figure 4: Residuals from initial regression model for bank days -8 to 0 of each month

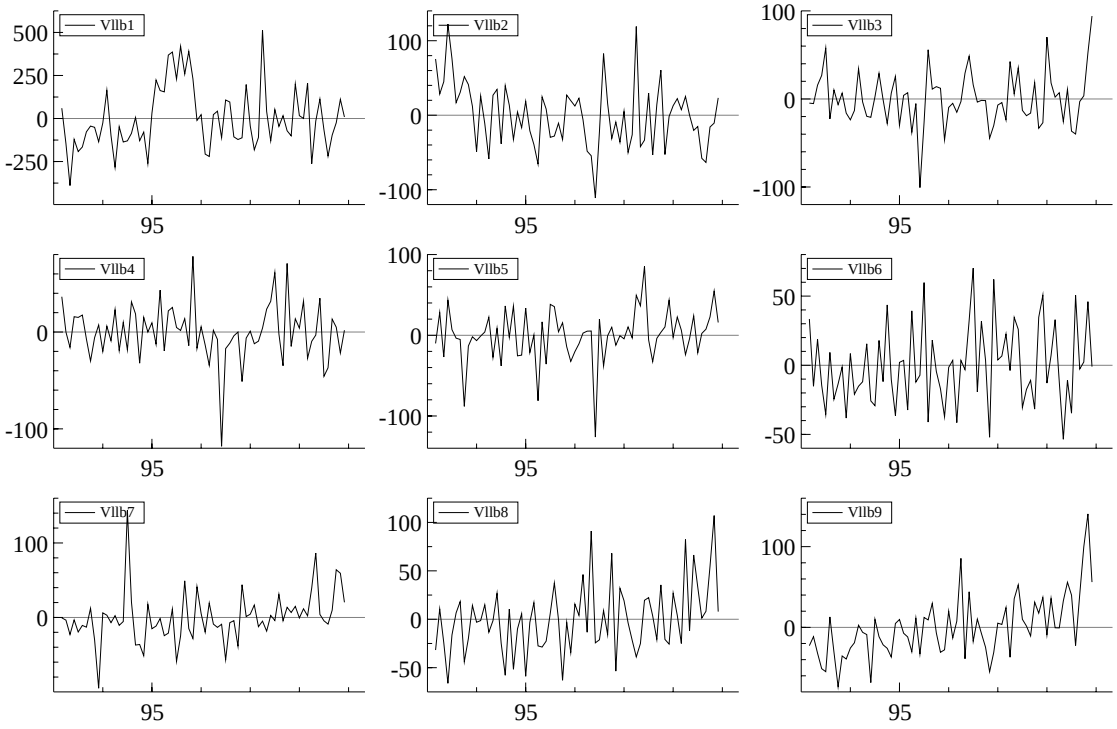

Figure 5: Residuals from initial regression model for bank days 1 to 9 of each month

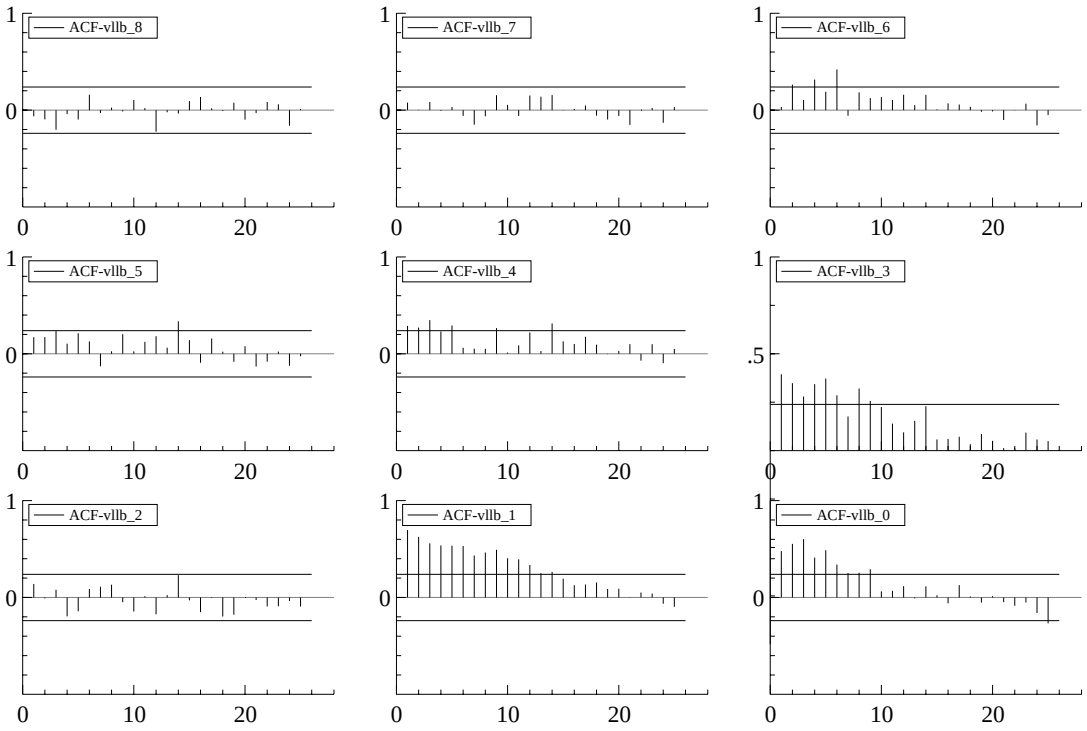

Figure 6: Residual correlation at monthly lags for bank days -8 to 0 of each month

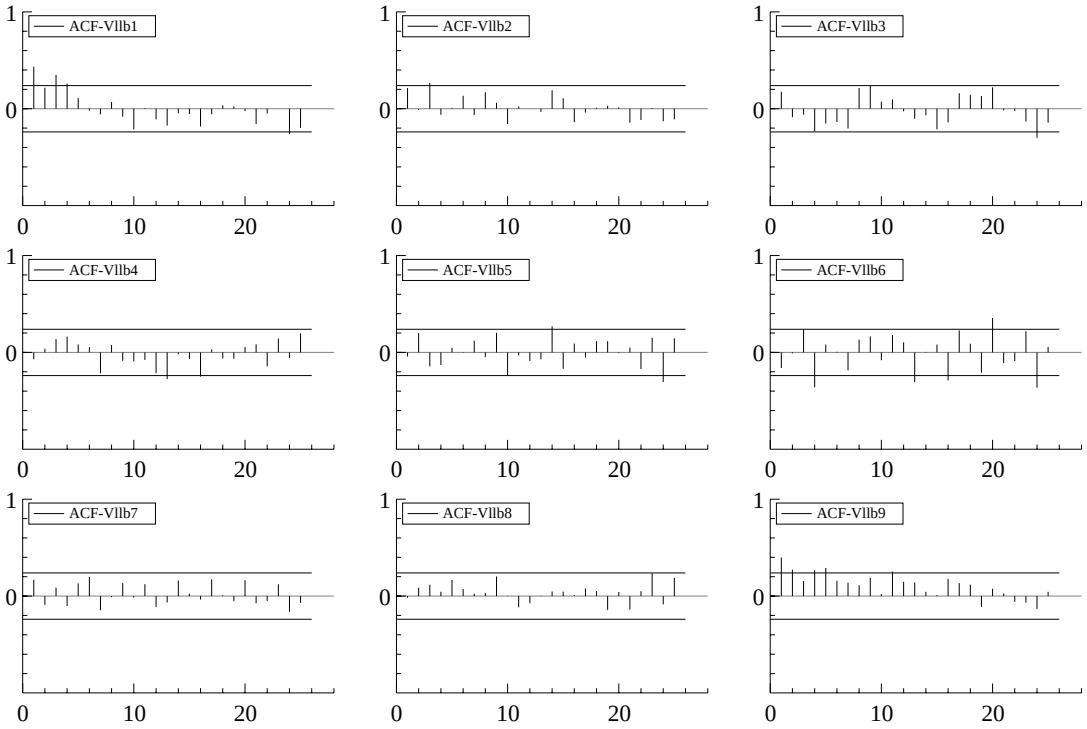

Figure 7: Residual correlation at monthly lags for bank days 1 to 9 of each month

These plots give a nearly complete picture of our sample. They represent 1260 out of 1461 available data points. The plots show local upward trends for a number of days-of-the-month. A number of outliers can be spotted as well, but these are not too severe. The residual correlations for the different bank days at monthly lags represented in Figures 7 and 6 enable us to investigate the presence of misspecification of the trend and the seasonal. Ooms and Franses (1998) used similar plots to discover periodically varying long memory persistence patterns. There is no clear indication of seasonal misspecification, there is no clear seasonal pattern in the residual correlation. The correlation function for $b = 0$, the most important series of all, increases up to a lag of 3 months before a comparatively large drop. This might indicate a "quarter effect", probably due to persistent changes in the relative weight of VAT (collected quarterly for many firms) in total tax revenues. The residuals indicate a clear misspecification of lower frequency components: A very slow decay in the autocorrelation function is seen for $b = 1, 9, -6, -5, -4, -3, -1, 0$.

2.2 Extensions of the regression model

The periodic regression model of Table 1 is of course overspecified. The periodic pattern of the mean and variance can be modeled using splines with a smaller number of equivalent parameters, c.f. §3 below. On the other hand the model is still too rigid. The model does not allow for long term changes in the mean that are clearly present in the data. We should allow for trends, especially for the days around the turn of the month. We will combine both ideas in a periodic structural time series model that we define below.

After modeling the bulk of the variation, we may be able to detect a day-of-the-week effect.

The practice of forecasting taxes often involves explanation by macroeconomic or institutional variables, which are usually only available at an aggregate monthly level. Although it possible to combine data with mixed observation timing intervals in a dynamic model, seeHarvey (1989, §6.3.7), this is beyond the scope of the analysis in this paper.

2.3 Procedures for time-transformation from observations to model

The graphs above showed that time transformation may simplify the statistical model for our data, in the sense that we are better able to exploit the intermontly similarity of the intramonthly pattern. The timing intervals for the statistical (state space) model will differ from the time 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, $j(t)$: year, $s(t)$: month of the year, and $p(t)$: day of the month. In our case, $j(t) = 1993, \ldots, J$, $j(t) = Y_\tau$, $s = 1, \ldots, S$, $s(t) = m_\tau$, $p(t) = 1, \ldots, P$. In general we do *not* have $p(t) = d_\tau$. $p(t)$ 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.

The time transformation leads to different timing intervals for the model and the observations. The timing interval for the model is shorter than the observation interval. Statistical solutions to the problems of estimation and prediction for time-invariant components of a linear dynamic model are discussed by Harvey (1989).

The most straightforward time-transformation was introduced earlier for the model of Table 1. There we skipped the (201) observations around the middle of the months with more than 18 working days: the timing interval for the model was effectively longer than the observation interval around the middle of the month, $P = 18, p(t) = b_\tau + I_{[-P/2,0]}(b_\tau) \cdot P$. Although this was not too big a problem in our application, this is clearly not a reasonable solution in more general cases.

As a first solution we extended the number of model days per month, P, from 18 to 23, the maximum number of bank days in any month in our sample. This introduces 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: $P = 23, p(t) = b_\tau + I_{[-P/2,0]}(b_\tau) \cdot P$, with $I_{[]}$ an indicator function that equals 1 for negative b_τ. Again, the transformation is determined by the end conditions $p = 1$ if $b_\tau = 1$ and $p = P$ if $b_\tau = 0$ and the break in the middle of the month where b_τ turns negative. For some months we have missing values for $p = 10, \ldots, 14$.

A general procedure for the construction of model days, would be to fix P to a large value first, say $P = 100$. Compute the number of bank days in each month, say, $M(Y_\tau, m_\tau)$. Then for the end of each observed bank day define $p(t) = [(b^*_\tau * P/M_\tau]$, where b^*_τ is the number of bank days since the turn of the month and [] denotes rounding to the nearest integer. The other model observations are treated as missing. Again we have the condition $p = P$ if $b^*_\tau = M_\tau$, $b_\tau = 0$. The number of missing model data points now varies from 82 to 77 per month. For Figure 2 we used $p(t) = P + 1 - [(M_\tau + 1 - b^*_\tau) \cdot /M_\tau]$, thereby imposing the restriction $p = 1$ if $b(\tau = 1)$. For different kinds of data sets a different function $p(t)$ may apply in connection with the observed intra-montly pattern and its changes from month to

month and from year to year. It is likely to be a good idea to have observations $p(t)$ for knot positions of the spline. The spline function will be the basis for interpolation of the missing model data and for forecasting of future values. Note that 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}. An in our case more natural, but technically still too demanding approach would be to treat our data as a flow variable, so that the model data time aggregate $\sum_i^{i+\delta} y_{j(t),s(t),p(i)}$ would correspond to an observation.

The simple time transformation with $P = 23$ and an equal time interval for model and observations for the majority of the data does not pose these technical problems. Given state space form of the dynamic regression model one can start forecasting from days with $p = 15, b_\tau = -8$ or later in the month, both one-step-ahead and multi-step-ahead, both for single days and for time aggregates right up to the next day in the following month with $p = 9, b_\tau = 9$. 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 multistep forecast intervals take into account the parameter uncertainty due to estimation.

In the end 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.

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

$b_r(\tau)$	mean	$\widehat{\sigma}_b$	$\widehat{R}_b$	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)
-8	105	43	.50	-.04	0.12	-.05	0.02	0.11	0.15	-.06	0.20	-.08
-7	118	31	.79	0.21	0.17	0.07	-.06	0.07	0.18	0.08	0.29	0.32
-6	146	58	.66	0.36	-.04	0.27	0.37	-.07	0.06	0.21	0.17	0.34
-5	188	58	.71	0.55	0.47	0.04	0.35	0.33	0.14	-.06	0.15	0.36
-4	273	67	.75	0.51	0.40	0.53	0.20	0.34	0.37	-.10	0.07	0.14
-3	401	80	.74	0.53	0.60	0.48	0.31	-.16	0.46	0.30	0.13	0.15
-2	666	127	.72	0.26	0.33	0.27	0.13	0.39	-.16	0.11	-.03	0.02
-1	1162	199	.67	0.25	0.66	0.60	0.64	0.52	0.41	0.08	0.51	0.33
0	5214	343	.85	0.40	-.24	0.25	0.18	0.03	0.13	-.10	0.05	0.25
1	679	196	.52	-.16	-.13	-.03	-.12	-.26	-.11	-.22	-.20	-.00
2	169	47	.59	0.05	-.00	-.20	0.04	-.26	-.29	-.20	-.13	0.10
3	110	33	.43	0.17	-.00	0.12	0.08	0.15	-.12	-.03	-.12	-.15
4	92	31	.56	0.09	0.33	0.21	-.00	0.02	0.05	-.17	-.22	-.17
5	85	35	.40	0.45	0.22	0.14	0.16	0.26	0.15	-.17	0.05	0.08
6	94	31	.51	0.18	-.19	-.05	-.33	0.12	0.24	0.12	-.06	0.25
7	87	36	.48	-.23	-.07	-.09	-.00	-.06	-.09	0.13	0.24	-.08
8	94	38	.46	0.08	-.10	0.11	0.13	0.22	-.16	-.09	0.04	0.48
9	98	41	.38	0.37	0.36	0.21	0.27	0.04	0.23	-.32	0.08	0.33

b_τ indexes position with respect to last bank day of the month. mean: Estimate of constant in regression model per bank day with centered seasonal dummies for daily tax revenues for b_τ. Measurements in 10^6 Euro. Sample 1993.3-1998.12. $\widehat{\sigma}_b$: regression standard error. $\widehat{R}_b$: correlation of fitted and dependent variable.
$c(l,b)$ is a so-called periodic correlation, cf. McLeod (1994): corr(ε_r,ε_{r-l},$b_r(\tau)$).
For simplicity we assume there are 18 bank days r in each month, so that modulo 18 arithmetic applies to b. The correlation depends only on the distance between the observations in bank days and on the index $b_{r(\tau)}$ of the leading observation.

Table 2: *Periodic variances of partial sums starting at* $b = -8$

i	-7	-6	-5	-4	-3	-2	-1	0	1
s.d.	52	56	93	138	201	256	397	568	558

s.d.: Standard deviation of partial sum of subsequent daily residuals of periodic regression model of Table 1 from $b = -8$ to $b = i$.
Sample 1993.4-1998.12. No degrees of freedom corrections applied.

3 Structural modeling: specification, estimation and forecasting

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 tool to solve this problem. For the problem at hand, the structural time series model 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 modeling and the corresponding notation are introduced in sections 3.1-...

3.1 Structural time series models

An univariate structural time series model which is particularly suitable for many economic data sets is given by

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

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 usually defined as

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

where the level and slope disturbances, η_t and ζ_t are mutually uncorrelated. When σ_ζ^2 is zero, we have a *random walk plus drift*, and when σ_η^2 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_η^2 is coupled with a positive σ_ζ^2; Young (1984) calls this model an 'integrated random walk'.

For the seasonal component we formulate a model which is based on a set of trigonometric terms which are made time-varying. This so-called trigonometric seasonal model for γ_t is given by

$$\gamma_t = \sum_{j=1}^{[s/2]} \gamma_{j,t}^+, \quad where \quad \begin{pmatrix} \gamma_{j,t+1}^+ \\ \gamma_{j,t+1}^* \end{pmatrix} = \begin{pmatrix} \cos \pi_j & \sin \pi_j \\ -\sin \pi_j & \cos \pi_j \end{pmatrix} \begin{pmatrix} \gamma_{j,t}^+ \\ \gamma_{j,t}^* \end{pmatrix} + \begin{pmatrix} \omega_{j,t}^+ \\ \omega_{j,t}^* \end{pmatrix}, \tag{3}$$

with $\pi_j = 2\pi j/s$ as the j-th seasonal frequency and

$$\begin{pmatrix} \omega_{j,t}^+ \\ \omega_{j,t}^* \end{pmatrix} \sim \text{NID} \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma_\omega^2 I_2 \right\}, \qquad j = 1, \ldots, [s/2].$$

Note that for s even $[s/2] = s/2$, while for s odd, $[s/2] = (s-1)/2$. For s even, the process $\gamma_{j,t}^*$, with $j = s/2$, can be dropped. The state space representation is straightforward and the initial conditions are $\gamma_{j,1}^+ \sim \text{N}(0, \kappa)$ and $\gamma_{j,1}^* \sim \text{N}(0, \kappa)$, for $j = 1, \ldots, [s/2]$. We have assumed that the variance σ_ω^2 is the same for all trigonometric terms. However, we can impose different variances for the terms associated with different frequencies; in the quarterly case we can estimate two different σ_ω^2 's rather than just one. We could also consider to drop a pair of sine-cosine terms at some frequency which appear not to be strongly present in the seasonal process. The trigonometric seasonal process evolves smoothly over time; it can be shown that the sum of the seasonals over the past 'year' follows an MA$(s-2)$ rather than white noise. More details on the trigonometric specification for the seasonal process can be found in Harvey (1989, page 56).

3.2 Statistical treatment

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 (1993, Chapter 4) and Kitagawa and Gersch (1996). The Gaussian state space form consists of

a transition equation and a measurement equation; we formulate it, following Koopman, Shephard, and Doornik (1999) as

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

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

where $\mathrm{NID}(\mu, \Psi)$ 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. The N observations at time t are placed in the vector y_t and the $N \times n$ data matrix is given by $(y_1, \ldots, y_n)$. We treat the tax series as coming from a univariate measurement equation: $N = 1$. 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 y_t. The initial state vector is assumed to be random with mean a and variance matrix 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 vector 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. When the system matrices are constant over time, we drop the time-indices to obtain the matrices T, Z, H and G. The resulting state space form is referred to as time-invariant. Note that the periodic regression model underlying Table 1 can also be written in state space form with $N = 18$, putting the regression coefficients in α_t and the corresponding time-varying regressors in Z_t. By allowing for heteroskedasticity through a time-varying G_t, one can also write the periodic regression model as a univariate model with $N = 1$. Koopman, Shephard, and Doornik (1999) give examples.

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

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

for $t = 1, \ldots, n$; see Anderson and Moore (1979, page 36) and Harvey (1989, page 104).

The Kalman filter is given by

$$\begin{aligned} v_t &= y_t - Z_t a_t \\ F_t &= Z_t P_t Z_t' + G_t G_t' \\ K_t &= (T_t P_t Z_t' + H_t G_t')\, F_t^{-1} \\ a_{t+1} &= T_t a_t + K_t v_t \\ P_{t+1} &= T_t P_t T_t' + H_t H_t' - K_t F_t K_t' \end{aligned} \tag{6}$$

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

$$P = P_* + \kappa P_\infty,$$

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.

It is well-known that the Kalman filter can compute of the Gaussian log-likelihood function via the prediction error decomposition for models in state space form; see Schweppe (1965), Jones (1980) and Harvey (1989, section 3.4). The log-likelihood function is given by

$$\begin{aligned} l &= \log p\left(y_1, \ldots, y_n; \varphi\right) = \sum_{t=1}^{n} \log p\left(y_t | y_1, \ldots, y_{t-1}; \varphi\right) \\ &= -\frac{nN}{2} \log\left(2\pi\right) - \frac{1}{2} \sum_{t=1}^{n} \left(\log |F_t| + v_t' F_t^{-1} v_t\right) \end{aligned} \tag{7}$$

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 1 to n, but usually the first d summations will be approximately zero as F_t^{-1} will be very small for $t = 1, \ldots, d$. For more details on diffuse initialisation; see Koopman (1997).

3.3 Signal extraction

Estimation of the unobserved components is usually referred to as signal extraction. The evaluation of $\hat{\alpha}_t = \mathrm{E}(\alpha_t | Y_n)$ and variance matrix $V_t = \mathrm{var}(\alpha_t | Y_n)$ is referred to as moment

state smoothing. The state smoothing algorithm we employ is based on 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

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

for $t = n, \ldots, 1$. When only the smoothed state $\hat{\alpha}_t$ is required, more efficient methods of calculation are available; see Koopman (1993).

3.4 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

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

are also normally distributed and serially independent with unit variance. We can check that these properties hold by means of the following diagnostic tests:

- Normality

  The first four moments of the standardised prediction errors are given by

  $$\begin{aligned} m_1 &= \frac{1}{n} \sum_{t=1}^{n} e_t, \\ m_q &= \frac{1}{n} \sum_{t=1}^{n} (e_t - m_1)^q, \qquad q = 2, 3, 4. \end{aligned}$$

  Skewness and kurtosis are denoted by S and K, respectively, and are defined as

  $$S = \frac{m_3}{\sqrt{m_2^3}}, \qquad K = \frac{m_4}{m_2^2},$$

  and it can be shown that when the model assumptions are valid they are asymptotically normally distributed as

  $$S \sim \mathrm{N}(0, \frac{6}{n}), \qquad K \sim \mathrm{N}(3, \frac{24}{n});$$

see Bowman and Shenton (1975). Standard statistical tests can be used to check whether the observed values of S and K are consistent with their asymptotic densities. They can also be combined as

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

which asymptotically has a χ^2 distribution with 2 degrees of freedom on the null hypothesis that the normality assumption is valid.

- Heteroskedasticity

  A simple test for heteroskedasticity is obtained by comparing the sum of squares of two exclusive subsets of the sample. For example, the statistic

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

  where e_t is defined in (10), is $F_{h,h}$-distributed for some preset positive integer h, under the null hypothesis of homoskedasticity.

- Serial correlation

  When the model holds, the standardised forecast errors are serially uncorrelated. Therefore, the correlogram of the one-step prediction errors should reveal no serial correlation. A standard portmanteau test statistic for serial correlation is based on the Box-Ljung statistic; see Ljung and Box (1978). This 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 j-th correlogram value

$$c_j = \frac{1}{nm_2}\sum_{t=j+1}^{n}(e_t - m_1)(e_{t-j} - m_1).$$

  This test is asymptotically χ^2 distributed with p degrees of freedom.

3.5 Missing values

When observations y_t for $t = \tau, \ldots, \tau^* - 1$ 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

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

similarly, the backwards smoothing recursions become

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

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.

3.6 Forecasting

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.

3.7 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 mainly the bank-day-of the month. Harvey, Koopman, and Riani (1997) used the calendar-day-of-the-year as x_t. 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 $\theta_i(x) = \theta(x)$ 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 regression spline equations are obtained as follows: (i) by considering $\theta_i''(x)$ and using standard integration rules, we get expressions for $\theta_i(x)$; (ii) we enforce the spline function $\theta_i(x)$ at $x = x_i^\dagger$ to be equal to the known value of $y_i^\dagger$; (iii) we 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 for the $k+1$ 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 fully expressed in terms of $y_0^\dagger, \ldots, y_k^\dagger$ by

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

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 can be expressed as

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

where $w_t = (w_{0,t}, \ldots, w_{k,t})\prime$ 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 can be estimated by least squares. 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 parameter vector $\lambda = (\lambda_0, \ldots, \lambda_k)'$ is estimated by least squares techniques. More details are given by Poirier (1973, 1976).

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. The summing-to-zero constraint, which avoids the colinearity with the trend component, for a time-varying spline can be implemented; the details are given by Harvey and Koopman (1993). (to be added more lately)

4 Model for daily Tax Revenues

4.1 The main model

The model for daily Tax Revenues y_t, $t = 1, \ldots, n$, is given by

$$y_t = w_t'\lambda_t + x_t'\delta + \xi_t, \qquad \xi_t \sim \mathrm{N}(0, \sigma_\xi^2), \tag{13}$$

where the time-varying spline function $w_t'\lambda_t$ takes account of trend and seasonal variations, the regression function $x_t'\delta$ allows for deterministic effects and ξ_t is the irregular. The spline function depends on a set of knot points which in our case should be placed within the interval of one month. For ease of exposition, we assume temporarily that one month consists of a fixed number of bank days, say 20. Further we recall that most of the tax revenue is received during the last three bank days of the month (day 18, 19 and 20) and to a lesser extent on the first bank day of the month (day 1). It seems therefore sensible to place knot points at these days and to place another knot at a day in the middle of the month, say at bank day 10. In this case we have a total five knots. The time-varying parameters $\lambda_{1,t}, \ldots, \lambda_{5,t}$ corresponding with the five knots are not known and since it is argued in section 2 that the effect within the month is periodic, we model each parameter by a basic structural time series model with components trend and seasonal, that is

$$\lambda_{i,t} = \mu_{i,t} + \gamma_{i,t}, \qquad i = 1, \ldots, 5, \tag{14}$$

where $\mu_{i,t}$ and $\gamma_{i,t}$ are the local linear trend and trigonometric seasonal associated with the ith knot. The subscript i is also attached to β_t, η_t, ζ_t and the various γ_t's and ω_t's.

The stochastic process for the trend is given by (2) with possibly $\beta_{i,t} = 0$ and different variances for $\eta_{i,t}$ and $\zeta_{i,t}$ and for the different knots. The restriction

$$\mu_{i,t} = \mu_{j,t}, \qquad 1 \le i, j \le 5, \quad i \ne j,$$

can be enforced.

The trigonometric seasonal process associated with a knot in the middle of the month may not be pronounced or at least less pronounced compared to the seasonal process of the last bank day. This is the basic motivation of introducing the flexibility of a unique trend and seasonal component for each knot and it also appeals to the periodic nature of the daily Tax Revenues as explored in section 2. Nevertheless, the restrictions of $\gamma_{i,t} = 0$ (no seasonality for ith knot), $\sigma_i^2 = 0$ (fixed seasonal for ith knot) or

$$\gamma_{i,t} = \gamma_{j,t}, \qquad 1 \le i, j \le 5, \quad i \ne j,$$

may be enforced. Further, different variances can be estimated for different frequencies of the sine-cosine terms of $\gamma_{i,t}$ or some sine-cosine terms may be dropped from $\gamma_{i,t}$.

The state space representation of the full model, where we assume that one year consists of six months, is given by

$$y_t = Z_t \alpha_t + G_t \varepsilon_t,$$

where

$$Z_t = \left[w_t' \otimes (1,0,1,0,1,0,1) \, , x_t' \right],$$

and

$$\alpha_t = \begin{bmatrix} \lambda_{1,t}^{\dagger} \\ \vdots \\ \lambda_{5,t}^{\dagger} \\ \delta \end{bmatrix}, \qquad \lambda_{i,t}^{\dagger} = \begin{bmatrix} \mu_{i,t} \\ \beta_{i,t} \\ \gamma_{i,1,t}^{+} \\ \gamma_{i,1,t}^{*} \\ \gamma_{i,2,t}^{+} \\ \gamma_{i,2,t}^{*} \\ \gamma_{i,3,t}^{+} \end{bmatrix}, \qquad i = 1, \ldots, 5.$$

The subscript i refers to the ith knot. In a real analysis, one year consists of twelve months and we need to include $(\gamma_{i,3,t}^{*}, \gamma_{i,4,t}^{+}, \gamma_{i,4,t}^{*}, \gamma_{i,5,t}^{+}, \gamma_{i,5,t}^{*}, \gamma_{i,6,t}^{+})'$ in $\lambda_{i,t}^{\dagger}$ for the full model.

This does entails a dimension for the state of $m = 13 \cdot 5 = 65$ for data with 12 months. This should be compared with the 20×13 regression parameters in a deterministic periodic regression model, of the kind presented in the previous section. With 19 knots and and zero disturbance vectors for the transition equations one obtains the periodic regression model in state space form.

Further, the transition matrix T is a time-invariant block diagonal matrix as given by

$$T = \mathrm{diag}(T_\lambda, T_\lambda, T_\lambda, T_\lambda, I),$$

where

$$T_\lambda = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \cos \pi_1 & \sin \pi_1 & 0 & 0 & 0 \\ 0 & 0 & -\sin \pi_1 & \cos \pi_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cos \pi_2 & \sin \pi_2 & 0 \\ 0 & 0 & 0 & 0 & -\sin \pi_2 & \cos \pi_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix},$$

and with a year equal to six months. The disturbance vector of the transition equation is given by

$$\varepsilon_t = \begin{bmatrix} \xi_t \\ \eta^\dagger_{1,t} \\ \vdots \\ \eta^\dagger_{5,t} \end{bmatrix}, \qquad \eta^\dagger_{i,t} = \begin{bmatrix} \eta_{i,t} \\ \zeta_{i,t} \\ \omega^+_{i,1,t} \\ \omega^*_{i,1,t} \\ \omega^+_{i,2,t} \\ \omega^*_{i,2,t} \\ \omega^+_{i,3,t} \end{bmatrix}, \qquad i = 1, \ldots, 5.$$

Row vector G_t selects ξ_t from ε_t for the measurement equation and the appropriate disturbances for the transition equation are selected by the matrix H_t. In particular we have,

$$G_t = (\sigma_\xi, 0, \ldots, 0), \qquad H_t = \begin{bmatrix} 0_1 & \mathrm{diag}(H_{\lambda,1}, \ldots, H_{\lambda,5}) \\ 0_1 & 0_\delta \end{bmatrix},$$

where 0_1 are column vectors of zeroes and 0_δ is a zero matrix with number of rows equal to the dimension of the regression vector δ. Further,

$$H_{\lambda,i} = \mathrm{diag}(\sigma_{\eta_i}, \sigma_{\zeta_i}, \sigma_{\omega_{i,1}}, \sigma_{\omega_{i,1}}, \sigma_{\omega_{i,2}}, \sigma_{\omega_{i,2}}, \sigma_{\omega_{i,3}}), \qquad i = 1, \ldots, 5,$$

where σ_{η_i} and σ_{ζ_i} refer to the level and slope disturbance variances associated with the ith knot and $\sigma_{\omega_{i,j}}$ is the trigonometric seasonal disturbance variance associated with the ith knot and the jth seasonal frequency (in this section, $i = 1, \ldots, 5$ and $j = 1, 2, 3$).

4.2 Irregular number of bank days in a month

The number of bank days in a month is of course varying among different months. We therefore introduce a scale of P entries in a month where P is some moderately large number, say $P = 100$. Assume that the number of bank days in a particular month is M, then we can distribute the observations at equal intervals of m entries within the month where m is the integer closest to P/M. The Kalman filter will go through to each entry and treat the entries which are not used as missing values. It is shown in section x.x that the Kalman filter and related algorithms can handle missing observations straightforwardly.

However, some consequences of lengthening the scale by P should be taken into account. For example, the seasonal frequencies of the trigonometric seasonal are defined as $\pi_j = 2\pi j/s$ for $j = 1, \ldots, [s/2]$ but they should be modified to $\pi_j = 2\pi j/(Ps)$. Also the variances should be scaled by $1/P$ although this is not really necessary as long as the results are

interpreted correctly. Also, the monthly increment of the trend at time t is not the slope β_t but it is $P\beta_t$.

It is argued in section x.x that extending the sample with missing values, the Kalman filter will automatically produce forecasts of the state vector together with the mean square forecast error matrices. This strategy can still be pursued without further modifications. Of course, forecasting one month ahead now requires P forecasts instead of 1.

4.3 Forecasting monthly totals

to be added later

4.4 Model building and testing

For the exposition of our model and its statistical treatment given in this section, we have assumed that we need five knots within one month of daily observations. Whether this is appropriate in our case of Tax Revenues of the Dutch Ministry of Finance is to be investigated in section 4. Here we want to emphasize that decreasing or increasing the number of knots will not change our approach. Finding the appropriate model for each knot might be a formidable task. However, we may start with a regression model which is the model discussed in this section but with all variances set equal to zero except σ_ξ^2. In this case the Kalman filter provides the estimator for σ_ξ^2, that is

$$\hat{\sigma}_\xi^2 = n^{-1} \sum_{t=1}^{n} F_t^{-1} v_t^2,$$

see Harvey (1993). The final state estimator $a_{n|n}$ is also provided by the Kalman filter and it contains the least squares estimators of the trend, seasonal and regression parameters. The mean square error matrix of the regression estimates is the matrix $P_{n|n}$ from the Kalman filter. From this analysis we can use the usual t-tests to determine the individual statistical contribution of each state element. This may be a first step to model building. In the same way we can include or exclude components from knots or exclude knots completely.

When an optimal model is found in terms of a regression model, we may introduce time-variation of the regression parameter starting with the trend component.

5 Application and implementation of the model

Structural periodic models for daily data offer a wide range of possibilities for the online-modeller and forecaster. Each specification will correspond to different forecasts and forecast intervals. The identification of the model, i.e. the choice for specification for a particular appliation, 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.

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 needed 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 discusses the implementation of the developer environment for this task.

5.1 Identification, estimation and diagnostics

Sections 2, 3 and 4 described several aspects of structural time series modeling of daily time series. Here we recapitulate the modelling menu. At each stage the forecaster has to choose from several options.

5.1.1 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 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. Subsection 2.3 described some options.

5.1.2 Model mean 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. For a periodic structural model one may specify different seasonal and trend components for the different periodic indices, see §4. The seasonal and periodic components can be modelled using stochastic dummies, splines or trigonometric terms. The stochastic trends can have a fixed or a varying slope.

5.1.3 Knot positions 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.

5.1.4 Variances and autocovariances 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 few parameters modelling these variances can be estimated simultaneously. In the end it may be necessary to specify the "irregular" as an AR-process to whiten the innovations of the measurement equation.

5.1.5 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 modeled using single dummies.

5.1.6 Relevant sample

The relevant sample for estimation, diagnostics and forecasting has to be chosen. Note that this sample need not be continuous: patches of additive outlying observations may simply be treated as missing. One may also delete (or select) the observations for particular days of the week, month or year. In this way one can specify a periodic model, say for weekdays, if data for weekends are considered irrelevant for forecasting weekdays.

5.1.7 Estimation

Estimation of the so-called hyperparameters, i.e. the free variances of the innovations of the different components, is performed by maximizing 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.

5.1.8 Diagnostics

See §3.4 above. One can check for nonnormality, heteroskedasticity and serial correlation for the innovations, or for the auxiliary residuals of the different components, both intra-monthly, intermonthly and interyearly. All other familiar and newly developed regression diagnostics can easily be programmed in Ox using a few lines of code.

5.2 Online Developer environments

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 (1995), which is optimized for "standard" structural modeling 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 forecast using an existing model. The user only has to record the observed revenue, put it in an easily accessible data base specified in calendar time and update the forecasts and confidence intervals for the next few days. The forecasts should be 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 labeled this program `ETE`, Econometric Tax Estimator.

At the second level one may want to see more diagnostics, perform sensitivity ananlysis, 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 econometrics and time series analysis should suffice to operate at this level. We labeled 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 heteroskasticity, 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 econometric knowledge and programming experience.

At the highest level we use Visual C++, Ox, GiveWin and `SsfPack`, a visual object oriented programming language, an advanded object oriented matrix language, a front-end for (two-index) data manipulations and visualisation, and a package of procedures for State Space modelling with a strong link with Ox, respectively. Of course, one can program everything directly in VC++, but this would require an unacceptable amount of programming time. Both Ox, GiveWin and SsfPack are written in (MS Visual)C(++), so only functions that combine or extend the functionality of these three programs have to be programmed in VC++. The programs at the highest level are combined to produce a second level program for the actual day-to-day modeling, forecasting and testing with the basic (periodic) structural time series model. They are also used for the production of the lowest level program for data input and short term online forecasting. Although these lower level programs will not be changed on a daily or weekly basis, we still consider the updating of these programs an online activity. Small inconveniences can lead to ineffective or improper use of the programs, so the programs need regular updates, both in functionality, layout and documentation.

Visual C++ is used to set up a database class in model time with three indices. This class is used to retrieve the data, stored in a text-file, including the names of the regressors. Other Visual C++ functions are used to customize the import and export of the model specification and the test specification in point-and-click menus. Finally there are Visual C++ functions for the generation of graphs. These graphs are made in GiveWin via calls of the basic graphical functions of Ox. The graphs can subsequently be edited in GiveWin if necessary.

Ox is used to perform the econometric computing: interaction with the databases (in our custom format, in GiveWin format, and in popular spreadheet formats like Excel) deriving test statistics using matrix notation, computing p-values from statistical functions, setting up the data in state space form required by `SsfPack`, and maximizing and analyzing the

likelihood produced by `SsfPack`. We have written functions for the interaction with VC++, to get the data and the model in matrix form. There are functions to interact with `SsfPack` and there are functions to present the results in graphs in GiveWin.

`Ssfpack` is used to set up the state space form for the basic structural components, but its main task is Kalman filtering, smoothing and the computation of the likelihood. A very wide class of models can be put into `Ssfpack`'s format, see Koopman, Shephard, and Doornik (1999), who also present and explain simple Ox sample programs to generate relevant applications.

Figures 8–10 give an idea of the look of `STSM`. Figure 8 shows the intial tasks of the modeller after loading the data: specification of the model and sample selection. Figure 9 shows the specification of the basic components, in this case a deterministic level and slope and a stochastic intramonthly spline and periodic heteroskedasticity. The position (and number) of knots can be specified in other windows. Figure 10 shows the window where diagnostics for the different components can be selected following Harvey and Koopman (1992).

5.3 Current model and results

The project we describe in this paper initially focussed on modelling and forecasting only the last day of each month, since that is 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 main method currently in use. That method is based on the distribution of the (remaining part) of a predicted value of the monthly total over the (remaining) days of the month. The monthly aggregate predictions are based on projections for the growth of the economy and (changes) in the different tax rates and collection policies. The parameters describing the current distribution are 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, $Y_\tau - 1, Y_\tau - 2, Y_\tau - 2$.

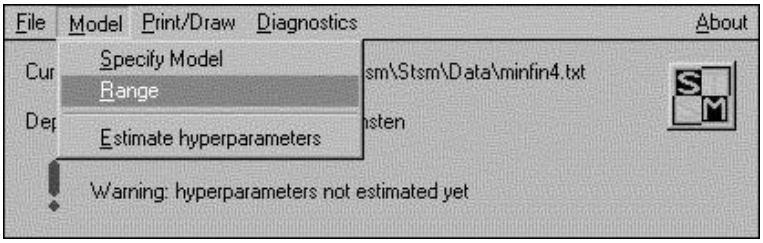

Figure 8: Main window of `STSM` after loading data

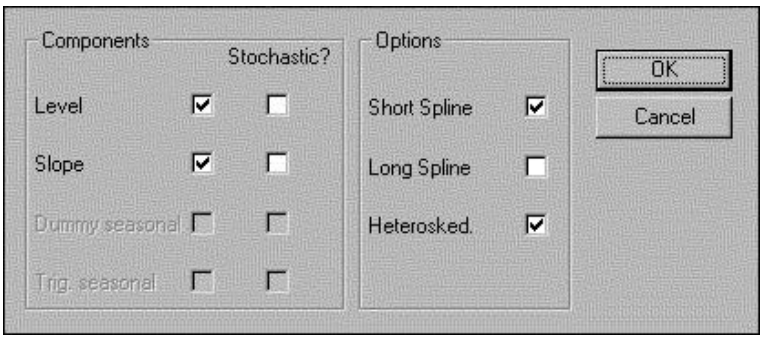

Figure 9: Model window of `STSM` to select basic components

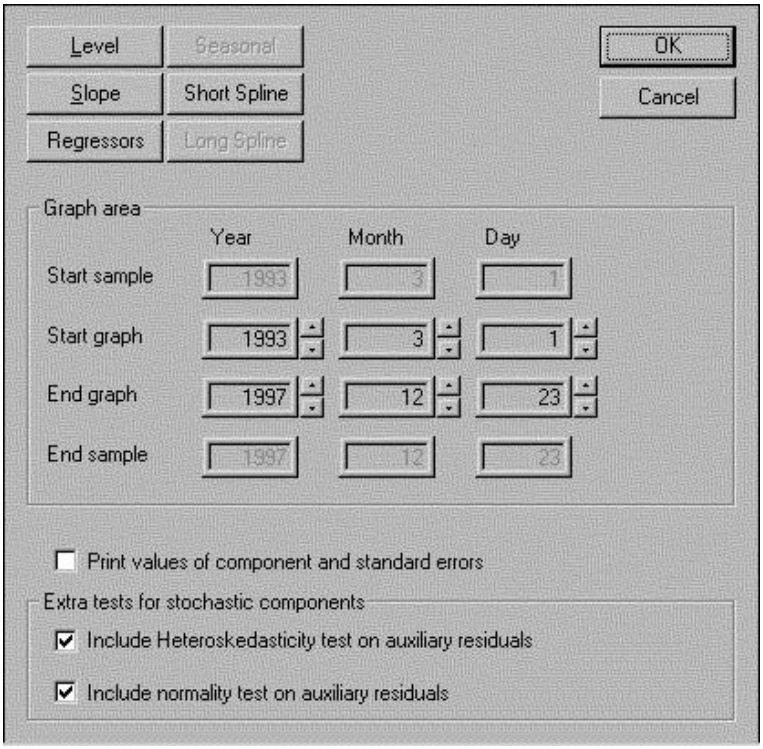

Figure 10: Diagnostics window for components `STSM`

The development of STSM has now reached a stage where it is much easier to adapt the model. Following the scheme of §5.1, we have made the following 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 §2.3. Second, we chose the following components. For the periodic intramonthly mean we chose a time-varying spline. For the 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, i.e. depending on $P * (s(t) - 1) + p(t)$ turned out to be insignificant. The stochastic trend is nonperiodic, i.e. it 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 make 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 $w_\tau = 3, 4, 5$, c.f. § 2 and a dummy measuring the length (in bank days) of the previous month, $M(t - P)$, c.f. § 2.3. 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 Table 3. Eigth, we present the following diagnostics: time series plots of in-sample and out-of-estimation-sample one-step ahead forecasts errors, $\widehat{v}_t$, and standardized forecast errors, $\widehat{F}_t^{-1/2}\widehat{v}_t$, 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 11–13. Except for a single outlier in June 1998, our model fared very well upto the middle of October 1998, when an unexpected change in the pattern around the centers 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. Finally Figure 14, illustrates the most important aspect of component-wise analysis: a plot of the component at the end of 1997. This is the type of spline we were looking for when started to look at time series plots like Figure 1.

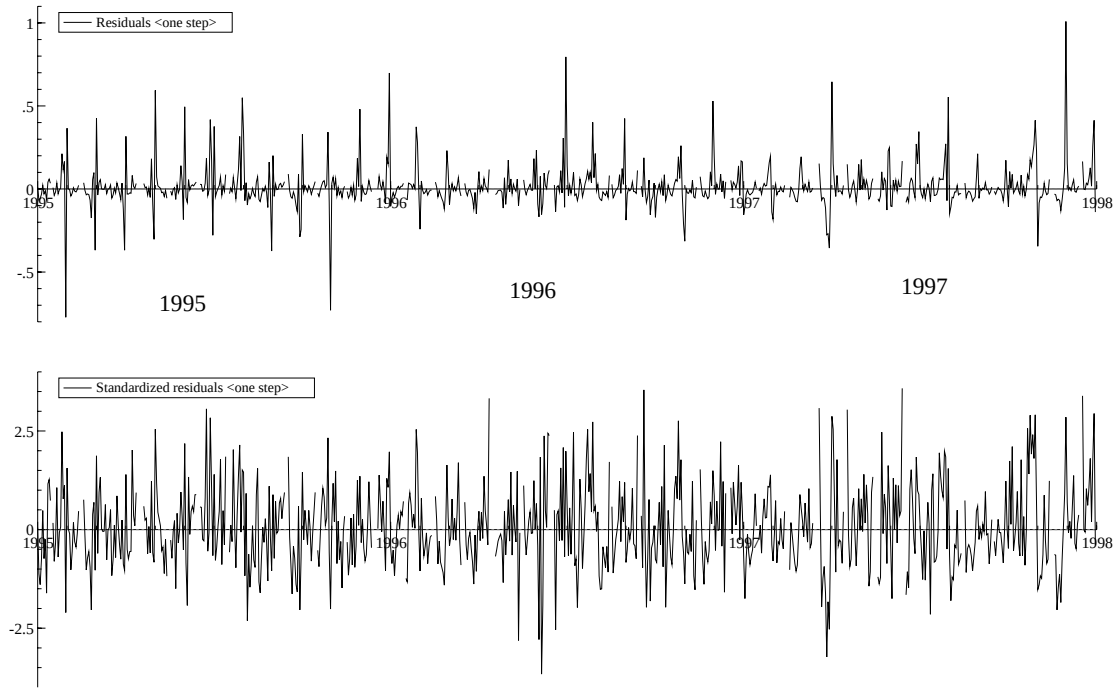

Figure 11: One-step (standardized) forecast errors in-sample 1993.1.1-1997.12.23: $(\widehat{F}_t^{-1/2})\widehat{v}_t$

Periodic and seasonal heteroskedasticity tests and normality tests for auxiliary residuals of level and irregular indicate that this basic model does not fit all days and all months equally well, see e.g. Table 5. Low probabilities in this table indicate rejection of the null hypothesis of equal variance for two days of the month in favour of the alternative where the variance on the day with "row index p" is higher. High probabilities reject the null in favor of the alternative where the variance on the day with "column index p" is higher. In the last column we see that the variances for $p = 4, 7$ are "significantly" lower than for $p = 23$, the variance for $p = 16$ is higher than for $p = 23$, but not significantly. Given the large number of tests and the stage of the modelling process one should not interpret the probabilities strictly as $p - values$ for classical hypothesis testing. This outcome should not come as a surprise since the periodic features of the model are limited in comparison with the evidence for periodicity presented in Table 1 above.

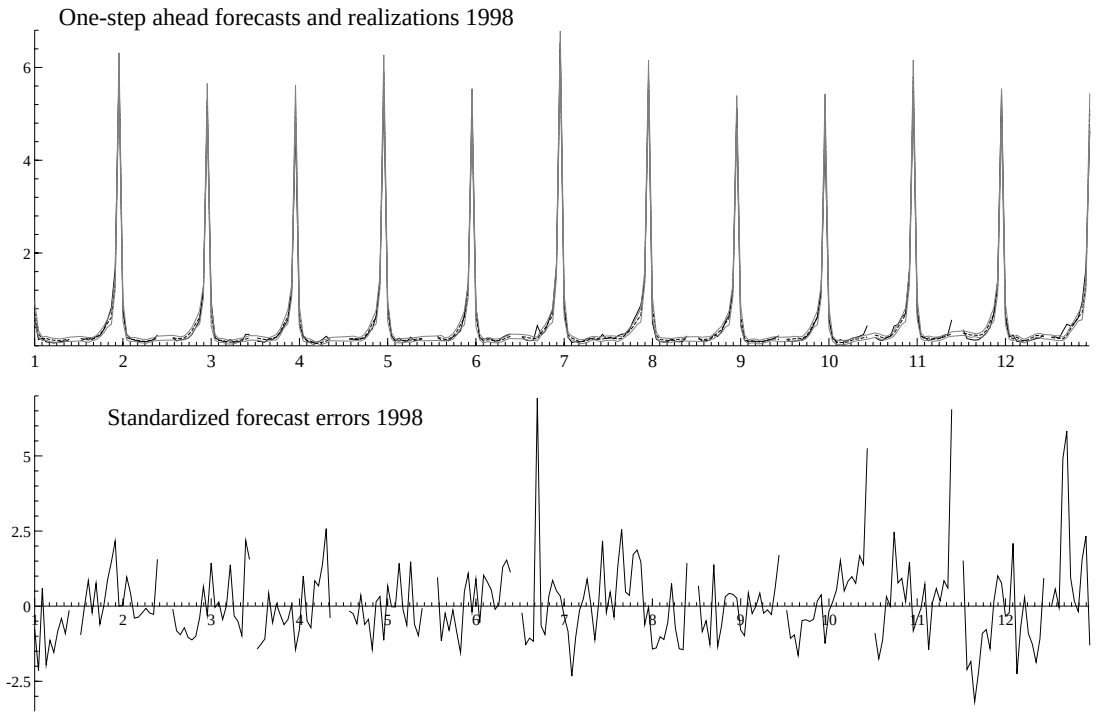

Figure 12: One-step ahead forecasts and standardized forecast errors outside estimation sample for hyperparameters of Table 3

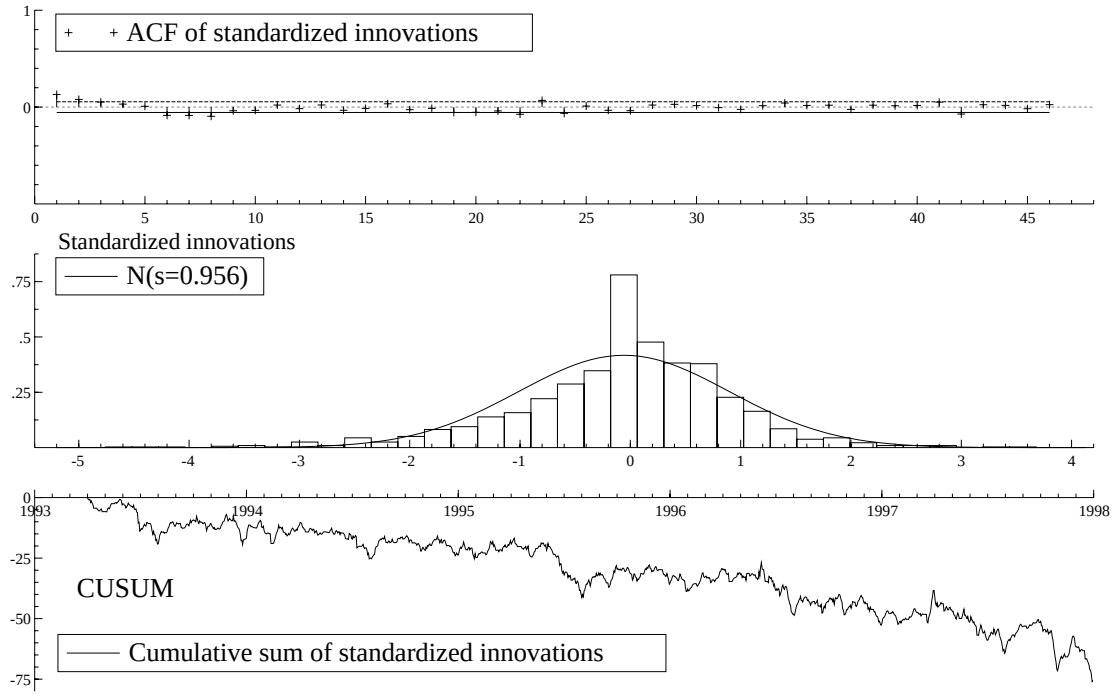

Figure 13: Diagnostics for standardized forecast errors in-sample $\widehat{F}_t^{-1/2}\widehat{v}_t$

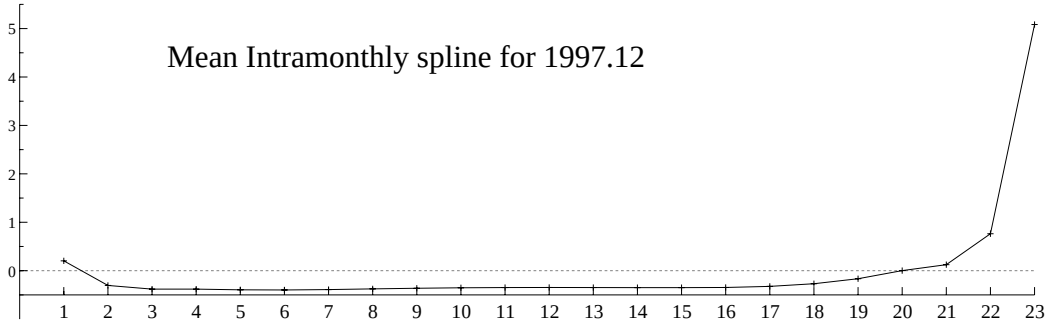

Figure 14: Conditional mean of intramonthly spline at the end of 1997

We can formulate the currently estimated model for y_t, $t = 1, \ldots, n$, 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 so as to have mean (very close to) zero over the span of a year, so that the level component can 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 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 normalized on the variance for $p(t) = 22$.

$$\begin{aligned}
G_t^{-2} = (&2.648, .147, .054, .067, .088, .059, .077, .070, .048, .102, .102,\\
&.102, .102, .102, .134, .037, .066, .169, .243, .369, .115, 1, 5.74)
\end{aligned}$$

The variances for the irregulary occurring $p(t) = 10, \ldots, 14$ were simply fixed.

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

Table 3: *Estimated hyperparameters*

σ_1^2	σ_2^2	σ_η^2	σ_ε^2
2.8193e-6	7.3364e-007	4.1839e-005	0.017979

Sample 1993.3.1 – 1997.12.23, measured in 10^9 Euro for the penultimate bank day of the month, $p(t) = 22$, $b_\tau = -1$

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 $.66 \cdot 10^9$ Euro per day is above the sample average, indicating an upward trend. This recursive estimate of this trend (not reproduced here, but naturally available in `STSM`) is relatively straight.

6 Conclusion

Although the model of the previous section gives reasonable forecasts, the diagnostics indicate that we can improve the model. We will do this by implementing the model of §4 first. We will also compare our forecasts with naive Holt-Winters type forecasts and see whether we outperform those significantly. We will also try to incorporate external information on (predictions for) monthly totals, first to adjust the forecasts over longer horizons and second to test the viability of these external forecasts online.

Table 4: *Estimated states at 1997.12.23*

State	mean	t-value
$\lambda(p=1)$	0.2033	4.99
$\lambda(p=2)$	-0.3046	-14.96
$\lambda(p=3)$	-0.3800	-23.96
$\lambda(p=5)$	-0.3491	-34.01
$\lambda(p=9)$	-0.36151	-35.04
$\lambda(p=20)$	0.0027	0.15
$\lambda(p=21)$	0.1246	5.21
$\lambda(p=22)$	0.7644	24.34
μ	0.6594	15.09
β	0.0001	0.58
Tuesday	-0.0190	-5.42
Wednesday	-0.0146	-4.33
Thursday	-0.0114	-3.32
$M(t-P)$	-0.0043	-2.34
$p(t)=1, s(t) \mod 3 = 1$	0.0964	1.54
$p(t)=22, s(t) \mod 3 = 1$	0.1844	4.77
$p(t)=23, s(t) \mod 3 = 1$	0.7065	7.61
$p(t)=1, s(t)=6$	0.1659	1.59
$p(t)=22, s(t)=6$	0.3649	5.64
$p(t)=23, s(t)=6$	1.2836	8.18

Estimation sample 1993.3.1 – 1997.12.23

λ: spline (see also Figure 14), μ: level, β: slope

Table 5: *Probabilities of F test for heteroscedasticity in third index p :*

p	2	3	4	5	6	7	8	9	10	14	15	16	17	18	19	20	21	22	23
1	.11	.08	.00	.07	.39	.01	.68	.24	.05	.64	.84	.89	.28	.11	.88	.49	.76	.30	.68
2	x	.50	.07	.72	.83	.14	.96	.68	.32	.93	.98	.99	.73	.56	.99	.90	.97	.74	.95
3	x	x	.06	.71	.82	.13	.96	.67	.40	.93	.99	.99	.72	.49	.99	.89	.97	.80	.96
4	x	x	x	.98	.99	.64	.99	.97	.86	.99	.99	1.0	.98	.92	.99	.99	.99	.98	.99
5	x	x	x	x	.64	.04	.88	.45	.50	.83	.99	.98	.50	.27	.96	.76	.90	.86	.97
6	x	x	x	x	x	.02	.79	.30	.08	.71	.88	.98	.35	.20	.92	.62	.83	.38	.76
7	x	x	x	x	x	x	.99	.93	.73	.99	.99	.99	.95	.88	.99	.99	.99	.95	.99
8	x	x	x	x	x	x	x	.09	.02	.40	.69	.91	.12	.05	.74	.32	.56	.15	.49
9	x	x	x	x	x	x	x	x	.21	.86	.96	.99	.55	.31	.97	.79	.92	.59	.87
10	x	x	x	x	x	x	x	x	x	.93	.99	.99	.85	.54	.99	.93	.99	.84	.97
11	x	x	x	x	x	x	x	x	x	.91	.99	.99	.79	.60	.99	.91	.98	.80	.96
12	x	x	x	x	x	x	x	x	x	.32	.56	.66	.08	.02	.64	.20	.46	.08	.36
13	x	x	x	x	x	x	x	x	x	.22	.51	.70	.04	.01	.52	.15	.34	.07	.30
14	x	x	x	x	x	x	x	x	x	x	.84	.90	.17	.06	.79	.40	.64	.31	.53
15	x	x	x	x	x	x	x	x	x	x	x	.60	.06	.01	.58	.16	.40	.06	.30
16	x	x	x	x	x	x	x	x	x	x	x	x	.01	.00	.22	.03	.11	.03	.21
17	x	x	x	x	x	x	x	x	x	x	x	x	x	.26	.96	.75	.90	.51	.85
18	x	x	x	x	x	x	x	x	x	x	x	x	x	x	.98	.88	.96	.73	.94
19	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	.13	.31	.04	.23
20	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	.73	.29	.68
21	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	.10	.39
22	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	.83
23	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Based on Goldfeld-Quandt-statistics for auxiliary residuals in different subsets p.

Low values reject residual homoskedasticity because of a higher variance for row index p,

Sample 1993.3.1 – 1997.12.23.

References

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

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

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

Doornik, J. A. and D. F. Hendry (1996). *GiveWin, An Interface to Empirical Modelling*. London: International Thomson Business Press.

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

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.

Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman Filter*. Cambridge, UK: Cambridge University Press.

Harvey, A. C. (1993). *Time Series Models* (2nd ed.). Hemel Hempstead: Harvester Wheatsheaf.

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.

Harvey, A. C., S. J. Koopman, and J. Penzer (1998). Messy time series. In T. B. Fomby and R. C. Hill (Eds.), *Advances in Econometrics, volume 13*. New York: JAI Press.

Jones, R. H. (1980). Maximum likelihood fitting of ARIMA models to time series with missing observations. *Technometrics 22*, 389–95.

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. *Biometrika 76*, 65–79.

Koopman, S. J. (1993). Disturbance smoother for state space models. *Biometrika 80*, 117–126.

Koopman, S. J. (1997). Exact initial Kalman filtering and smoothing for non-stationary time series models. *J. American Statistical Association 92*, 1630–1638.

Koopman, S. J., A. C. Harvey, J. A. Doornik, and N. G. Shephard (1995). *STAMP 5.0, Structural Time Series Analyser, Modeller and Predictor*. Chapman and Hall, Andover, England.

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.

Ooms, M. and P. H. B. F. Franses (1998). A seasonal periodic long memory model for monthly river flows. Discussion Paper Report 9842/A, forthcoming in Environmental Modeling and Software, Econometric Institute Erasmus University Rotterdam.

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

Schweppe, F. (1965). Evaluation of likelihood functions for Gaussian signals. *IEEE Transactions on Information Theory 11*, 61–70.

Young, P. (1984). *Recursive Estimation and Time Series Analysis*. New York: Springer-Verlag.