Aasness, Jørgen; Belsby, Liv

Working Paper

Estimation of Time Series of Latent Variables in an Accounting System Petrol Consumption of Norwegian Households 1973-1995

Discussion Papers, No. 203

Provided in Cooperation with:
Research Department, Statistics Norway, Oslo

Suggested Citation: Aasness, Jørgen; Belsby, Liv (1997) : Estimation of Time Series of Latent Variables in an Accounting System Petrol Consumption of Norwegian Households 1973-1995, Discussion Papers, No. 203, Statistics Norway, Research Department, Oslo

Discussion Papers No. 203, October 1997
Statistics Norway, Research Department

Jørgen Aasness and Liv Belsby

Estimation of Time Series of Latent Variables in an Accounting System

Petrol Consumption of Norwegian Households 1973-1995

Abstract:
We present an approach for estimating time series of a set of latent variables satisfying accounting identities. We concentrate on a simple case study and comment on possible generalizations. The model consists of three main parts: (i) A system of accounting identities, e.g., a subsystem of the national accounts, which variables are considered latent. (ii) A measurement model connecting the latent variables to indicators from different sources, including micro and macro data. (iii) Stochastic processes of a subset of the latent variables in the accounting system, with stochastic trend and random walk as alternative models. The model is given a state space formulation and the Kalman filter and EM algorithms implemented in the software STAMP, are used to estimate the parameters and the time series of the latent variables. The approach is applied to estimate petrol consumption of the household and nonhousehold sectors in Norway 1973-1995, from observation of macro data on total petrol consumption and survey data of household expenditures for petrol. Satisfactory model properties are obtained. The stochastic trend model gives smooth and plausible estimates of the time series of latent petrol consumption of the household and nonhousehold sectors.

Keywords: National accounts, latent variables, stochastic trends, state space models.

JEL classification: C82.

Acknowledgement: We thank John, Dagsvik, Nils Henrik Risebro, Kristin Rypdal, Terje Skjerpen and Ib Thomsen for useful comments.

Address: Jørgen Aasness, Statistics Norway, Division of Microeconometrics,
P.O.Box 8131 Dep, N-0033 Oslo. E-mail: j2a@ssb.no

Liv Belsby, Statistics Norway, Division of Methods and Standards,
P.O.Box 8131 Dep, N-0033 Oslo. E-mail: lbe@ssb.no

Discussion Papers comprises research papers intended for international journals or books. As a preprint a Discussion Paper can be longer and more elaborated than a usual article by including intermediate calculation and background material etc.

Abstracts with downloadable postscript files of
Discussion Papers are available on the Internet: http://www.ssb.no

For printed Discussion Papers contact:

Statistics Norway
Sales- and subscription service
P.O. Box 8131 Dep
N-0033 Oslo

Telephone: +47 22 00 44 80
Telefax: +47 22 86 49 76
E-mail: Salg-abonnement@ssb.no

1 Introduction

A national accounting system can be considered as an economic interpretation of a system of definitional relations between a set of variables, cf. e.g. Aukrust (1966, 1994). These variables are latent (not directly observed) although they may be given precise

economic meaning. The main problem for the national accounting practitioners is to estimate these latent variables from a large set of different data sources, under the restriction that all the definitional relations between the latent variables shall hold exactly. This is a formidable task, and in practice these variables are estimated by quite informal methods. For example, if there are several indicators of one variable one may choose the variable one ”most trustworthy” and disregard the information in the other observables. If one variable has no directly observable indicator one may assume that it develops proportionately with some other observable, obtaining the factor of proportionality from an observation in a single period. If we try to formalize these procedures one may prove that the estimators are inefficient (not exploiting the available information). They may also be strongly biased if the assumptions of e.g. constant factors of proportionality are not appropriate approximations. As an alternative, we suggest to use formal statistical methods, exploiting time series of a set of indicators to simultaneously estimate time series of a system of latent variables, using soft assumptions and optimal estimation methods. We present our approach in terms of a simple empirical case study and point out possible generalizations.

The approach can be applied to a tiny subsystem of the national accounts, to several subsystems, and perhaps some day to the whole system by means of a hierarchy or a network of subsystems. The approach can, of course, also be applied to other similar, and simpler, statistical systems.

A basic advantage of the proposed approach is that we obtain a framework for testing the assumptions made, and we can accumulate experience and improve the models within a progressive research program connected to the daily accounting work. Furthermore, we obtain standard errors of the estimators of the latent variables, model diagnostics and goodness of fit measures for the models. It is interesting to note that many of the pioneers of national accounting used reliability measures, both in the publications and as a tool in the production process towards final figures, cf. Aukrust (1994, sec 4.10). This practice was stopped, probably because no formal methods were available. The standard errors, diagnostics, and the goodness of fit measures may be used as such formal reliability measures.

If one uses formal methods in some of the subsystems, and use more traditional informal methods when “balancing” the subsystems into a comprehensive set of national accounts figures, the reliability measures in the subsystems can be useful indicators when deciding how much the different variables

should be adjusted in order to obtain a fully consistent accounting system of estimated latent variables.

There has been an increased interest in methods which combine data taken from both macro and micro level. See e.g., Adler and Wolfson (1988) and Gorter and Laan (1992), where macro data and household survey data on consumption, income and wealth are used to estimate socio-economic accounts. Our case study, combining macro and micro data for the household sector, have particular relevance to this area of national accounting. Both these works emphasize the development of a consistent and appropriate accounting system. The present work, however, puts focus on formulation of appropriate stochastic models and applying modern statistical methods for testing and estimating.

A basic idea in our paper is to formulate the accounting identities in latent variables, as an important part of a fully specified statistical model. We are not aware of literature using this idea for producing statistics, such as national accounts, but this idea has a long tradition in consumer econometrics. A classic article in this respect is Liviatan (1961), where the accounting identity says that the consumers expenditures on different goods add up to total expenditure. Liviatan used a static model estimated by cross section data on households expenditures. Aasness, Biørn and Skjerpen (1993) follow up this idea, using a more refined and dynamic model estimated by panel data.

Often an unknown variable is observed through several kinds of data. Generally estimates based on data taken from several sources are more accurate than estimates made using only one source. One example of this is the household petrol consumption in Norway. The Energy Statistics (ES) for Norway includes accurate sales statistics on total petrol consumption. But there are no direct macro information on the fraction sold to the households. The Household (or Consumer) Expenditure Survey (HES), however, measures the petrol consumed exclusively by the households. Since this measure is based on a small sample, it will of course contain statistical errors. We combine these two data sources to obtain a more accurate estimate of the household consumption. In addition we obtain an estimate of the petrol consumption in the nonhousehold sector. We make use of the full time series of both observables to simultaneously estimate the full time series of the two latent variables.

Our model consists of three parts. (i) An accounting identity with two latent variables (petrol consumption of households and petrol consumption of nonhouseholds). (ii) Measurement relations connecting the observed variables to the latent variables and measurement errors. (iii) Stochastic processes of the two basic latent consumption variables. Our primary aim is to produce more reliable estimates of the petrol consumption by the households and nonhouseholds. The estimates should be less influenced by measurement errors than the yearly estimates from the household expenditure survey. The petrol consumption of the households constitute the main part of the total petrol consumption. Yearly fluctu-

ations measured exclusively by the household expenditure survey are likely to be measurement errors. Thus we will choose a model which smoothes yearly fluctuations measured exclusively by the household expenditure survey. However, the model should not have a particular direction. We will consider the Gaussian random walk and the stochastic trend model. The flexibility of the models are influenced by parameters, which are estimated from the data. The three parts of the model is described in section 2.1-2.3. In section 2.4. we express the model in state space form.

State space models provide a flexible representation of linear models, with rich opportunities for direct interpretation of the components, and *a priori* knowledge of the phenomenon can more easily be formulated than with ARIMA models. The state space models can be connected to powerful algorithms. The latent variables are estimated by the Kalman filter algorithm. This is optimal in a least square sense. The maximum likelihood estimates for the parameters are found by the EM–algorithm, see section 3. Two textbooks with introduction to state space modeling, the Kalman filter, and the EM–algorithm are Harvey (1989) and Shumway (1988). It is technically straight forward to apply this approach on complex models with many accounting identities, latent variables and indicators. Therefore these algorithms are very useful for combining several data sources observed over a time period. We use the computer program STAMP, see Koopmans et al (1995), in our empirical analysis.

During the nineties, state space models have increasingly been used both to smooth univariate time series and to combine more data sources for more accurate estimates. Patterson (1995) and Skjerpen and Swensen(1997) use state space models to combine preliminary estimates on consumer expenditure and manufacturing investment respectively. Furthermore, Gonzales and Moral (1995) analyze international tourism demand in Spain by combing income index, two price indexes and tourist tastes using state space models. Another example is given by Rahiala and Teräsvirta's (1993) approach. They combine autoprojective information from the Swedish and Finnish metal and engineering industries with information from business surveys to forecast the output from these industries.

A specific purpose of the present paper is to provide an approach, illustrated by a simple case study, to be considered for application in the production of parts of the Norwegian national accounts, and in particular for exploiting the yearly HES, cf. Andersen et al (1991) and Hansen et al (1992).

The rest of this paper is organized as follows: In section two we present the class of models we use in the empirical analysis, with comments on possible generalizations. The estimation procedure is described in section three, the empirical results are presented in section four, and we make some concluding remarks in section five.

2 The model

2.1 The accounting identity in latent consumption

Consider the following simple accounting identity,

$$z(t) = x_1(t) + x_2(t), \tag{1}$$

where the variables can be interpreted as follows, $z(t)$ is the total petrol consumption in Norway in year t, $x_1(t)$ is the petrol consumption of Norwegian households, and $x_2(t)$ is the petrol consumption in other sectors of the Norwegian economy. These variables are considered latent, and a basic problem is to estimate the time series of these latent variables from observed indicators which may contain random and systematic measurement errors.

Relation (1) is the simplest case of an accounting system. We may easily generalize by using a more detailed sector classification (including e.g. public sector and different industries), and several commodity groups (e.g. petrol, autodiesel, heating kerosene, and fuel oil).

In principle one could extend (1) to the whole set definitional relationships in a system of national accounts. However, it would be practically infeasible to carry out our approach directly on this system as a whole, but for separate subsystems it should be possible, and perhaps some day for the full system by means of a hierarchy or a network of subsystems.

2.2 Measurement relations

The estimation is based on two data sources, the Household Expenditure Survey (HES), and the Energy Statistics (ES) in the time period 1973–1995. The HES gives an estimate of the total consumption of petrol and autodiesel by the households in Norway. We have reduced the data from HES according to estimated autodiesel consumption. See the appendix for details on the data. Denote consumption derived from HES by $y_1(t)$. Because it is an estimate, it can be interpreted as an indirect measurement of the true petrol consumption by the households; $x_1(t)$. We can express this as

$$y_1(t) = x_1(t) + v(t), \tag{2}$$

where $v(t)$ is the measurement error. In order to make our first empirical exercise as simple as possible, we assume $v(t)$ to be Gaussian white noise, i.e., Gaussian uncorrelated variables with zero mean and constant variance[1].

[1] In future work we may model $v(t)$ more carefully by taking into account variation in sample size and specific features of the sample design, cf. Andersen et al (1991, p. 34). About 20% of the households take part in the household survey two

The ES measures the total amount of petrol sold by the oil companies to the petrol stations in Norway measured in liters ($y_2(t)$). These statistics are reliable and can be considered to give the correct total amount, i.e, $y_2(t) = z(t)$, thus

$$y_2(t) = x_1(t) + x_2(t), \tag{3}$$

where we have exploited the accounting identity (1).

These measurement relations may of course be extended in various ways, including random measurement errors in (3) and systematic measurement errors in (2). However, this may lead to identification problems, depending on the specification of the complete model and on the possibility for including other observables or other types of information.

2.3 Stochastic processes of latent consumption

Random walk

Our aim is to develop a model which can produce more reliable statistics, and not primarily to model and analyze petrol consumption. Thus we do not want to put strong restrictions on the model. The model for the latent process should therefore be quite flexible, and must allow non-stationarity. Because of the short time series available, the model should be simple, i.e., not have many parameters. The random walk model fulfills this demand. It has an appealing simplicity. But the forecast made by the random walk model is not very interesting, since it will merely be the estimated value for the last observed time period.

Thus, one option is to model the latent process as random walk, both because this is in itself interesting, and also to compare this model's properties with those of a more complicated model. The random walk model is written

$$x_i(t) = x_i(t-1) + w_i(t) \text{ for } i = 1, 2, \tag{4}$$

where $w_i(t)$ is centered, Gaussian, white noise.

Stochastic trend

From experience we know that there are often trends in consumption, thus we will also consider a more stable model with an explicit trend. The stochastic trend model is written

consecutive years. Thus $v(t)$ is expected to have an autocorrelation less than 20% with lag one, and zero with lags greater than 1. A more appropriate model for $v(t)$ could therefore be a moving average model of order one. Dagsvik (1978) gives an example of a model, with some similarities to ours, where the sample design, with rotating panels, is taken into account.

(5)
$$\begin{aligned} x_i(t) &= x_i(t-1) + \beta_i(t-1) + w_i(t) \\ \beta_i(t) &= \beta_i(t-1) + w_{\beta,i} \qquad \text{for } i = 1, 2, \end{aligned}$$

where $\beta_i(t)$ is the slope term, which is allowed to change over time, and $w_i(t)$ and $w_{\beta,i}(t)$ are Gaussian, centered white noise. Positive values for β_i signify increasing consumption, and negative values decreasing consumption. The forecast produced by this model equals the estimated value at the last time, plus the estimated slope. This seems a more appealing forecast than the forecast by the random walk model, that is simply estimated value at the last observed time.

2.4 State space formulation

Both the models for the observations, and for the latent processes, can easily be formulated in the state space model class, which covers a wide range of linear time series models. Generally state space models consist of a transition equation and an measurement equation. The transition equation models a latent, and indirectly observed, vectorial stochastic process $X(t)$ by a linear Markov model as

$$X(t+1) = F\,X(t) + W(t+1).$$

Here F is a deterministic transition matrix, and $W(t)$ is centered white noise. The transition matrix F may also depend on t deterministically. The measurement equation connects the latent process to an observable vector $Y(t)$ through

$$Y(t) = H\,X(t) + V(t),$$

where H is a non-stochastic matrix, which may also depend on t deterministically, and $V(t)$ centered white noise uncorrelated with $W(t)$. Let Q denote the covariance matrix of the white noise process $W(t)$.

When the latent processes are modeled as random walk, denoted by model A, the matrices defining the measurement link and the transition system are

(model A)
$$H = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ F = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The covariance matrix for then noise terms in the latent process is

(model A)
$$Q = \begin{pmatrix} \text{var } w_1 & \text{cov } (w_1, w_2) \\ \text{cov } (w_1, w_2) & \text{var } w_2 \end{pmatrix}.$$

Alternatively the latent processes are modeled as stochastic trends, denoted by model B. Then the X-vector consist of $(x_1, \beta_1, x_2, \beta_2)'$, and the matrices that defines the measurement link and the latent processes are

$$\text{(model B)} \qquad H = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \; F = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Furthermore, the covariance matrix for the noise terms in the latent processes reads

$$\text{(model B)} \qquad Q = \begin{pmatrix} \text{var } w_1 & 0 & \text{cov } (w_1, w_2) & 0 \\ 0 & \text{var } w_{\beta_1} & 0 & 0 \\ \text{cov } (w_1, w_2) & 0 & \text{var } w_2 & 0 \\ 0 & \text{cov } (w_{\beta_1}, w_{\beta_2}) & 0 & \text{var } w_{\beta_2} \end{pmatrix}.$$

The zeros in the matrix above illustrate that the slope term and x_i-value are assumed to be uncorrelated. However, the two slope terms may be correlated. The same applies to the two latent processes themselves.

The white noise in the observation is $(v_t, 0)'$. The last term is zero because there is no measurement noise in y_2. The covariance matrix is denoted by R and reads

$$\text{R} = \begin{pmatrix} r & 0 \\ 0 & 0 \end{pmatrix}.$$

Note that this apply to both models.

3 Estimation

The Kalman filter algorithm, see e.g. Shumway (1988) or Harvey(1989), is an efficient recursive algorithm for estimating the latent process in the state space models. For linear models the estimates are generally optimal in a least square sense, and for Gaussian variables the estimates are generally optimal. We assume that the random variables are Gaussian.

The algorithm can conveniently cope with missing data, which is a difficult problem for ARIMA models. Another advantage is that the Kalman filter can easily handle non-uniformly spaced sample points, and has no restriction on stationarity. We assume that the random variables are observed through

$Y(t)$ at discrete sample points $t = 1, 2, \ldots T$. The Kalman filter, *forward recursion*, calculates the *Kalman filter* estimators

$$E(X(t)|Y(1)\cdots Y(t)),$$

i.e., the expectation of $X(t)$ given the observations $Y(1), \cdots Y(t)$. For each time t the algorithm first calculates

$$E(X(t)|Y(1)\cdots Y(t-1)),$$

and then uses the new observation $Y(t)$ to correct the "preliminary" estimate to get

$$E(X(t)|Y(1)\cdots Y(t)).$$

The "preliminary" estimate can also be used as a forecast. This is the reason why it is technically trivial to make forecasts using the Kalman filter. The *forward recursion* equations are

$$\begin{aligned} \hat{x}(t|t-1) &= F\hat{x}(t-1|t-1)) \\ \hat{x}(t|t) &= \hat{x}(t|t-1) + K(t)\left(Y(t) - H\hat{x}(t|t-1)\right), \end{aligned}$$

where $\hat{x}(t|t-1)$ and $\hat{x}(t|t)$ are the estimators for the latent process at time t, given $Y(1)\cdots Y(t-1)$ and $Y(1)\cdots Y(t)$ respectively. The *Kalman gain matrix* K, Kalman (1960) and Kalman and Bucy (1961), is defined by

$$K(t) = P(t|t-1)H'\left(HP(t|t-1)H' + R\right)^{-1},$$

where $'$ means transposed, and the matrix $P(t|t-1)$ is updated recursively by

$$\begin{aligned} P(t|t-1) &= FP(t-1|t-1)F' + Q \\ P(t|t) &= P(t|t-1) - K(t)HP(t|t-1). \end{aligned}$$

The matrices $P(t|t-1)$ and $P(t|t)$ are the estimated covariance matrices for $x(t)$ given respectively the data sets $Y(1)\cdots Y(t-1)$ and $Y(1)\cdots Y(t)$.

If the parameters are known, and one is primarily interested in making forecasts, or in estimating the latent process at the time of the last observation, the *forward recursion* is all one needs.

Often, as in our case, one is interested in estimating the latent process for the whole time period. This is done by calculating the expectation conditioned on the whole data set,

$$E(X(t)|Y(1)\cdots Y(T)),$$

the so called *Kalman smoother* estimators. These are calculated by the *backward recursion*

$$\hat{x}(t-1|T) = \hat{x}(t-1|t-1) + J(t-1)(\hat{x}(t|T) - \hat{x}(t|t-1)),$$

where

$$J(t-1) = P(t-1|t-1)F'P(t|t-1)^{-1}.$$

The matrix J is the backwards analogue of the Kalman gain matrix K in the forward recursion equation.

The estimated covariance matrix for the smoothed estimators satisfies the recursion equation

$$P(t-1|T) = P(t-1|t-1) + J(t-1)(P(t|T) - P(t|t-1))J(t-1)'.$$

The maximum likelihood estimates for the general set parameters can be found by the EM–algorithm by Shumway and Stoffer (1982). However, the models in the present paper contains only unknown parameters by the covariance matrix for the latent process and by variance for the measurement error. In this case the EM–algorithm can be simplified, Koopman (1993). The original EM–algorithm iterates between estimates for the parameters and the latent process, while the EM–algorithm by Kooper requires no iteration, and thus it is less computationally intensive.

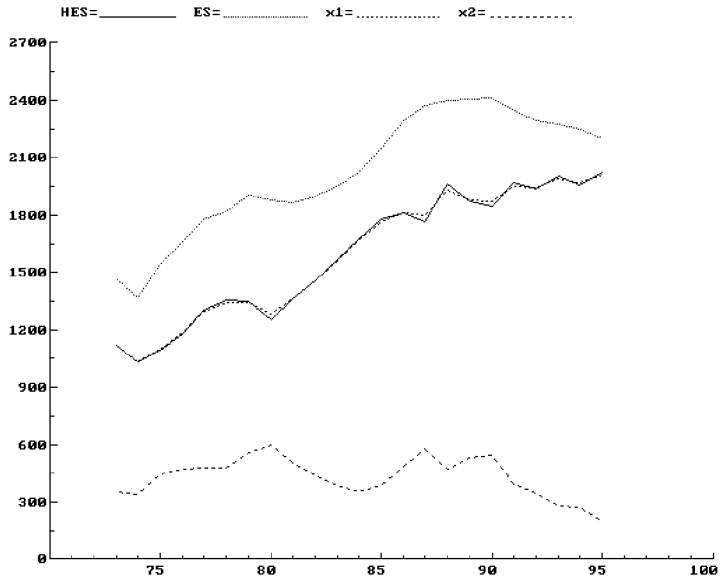
The Kalman filter combined with the EM–algorithm by Koopman is implemented in the STAMP program, see Koopman et al (1995). The estimated latent processes and the model diagnostic in the next section are estimated using STAMP.

4 Empirical results

Figure 1 shows the observed time series and the estimated latent petrol consumption by the households and the nonhouseholds for model A, where the petrol consumptions are modeled as random walks. The estimates follow the data almost completely, and do not fulfill our intention to smooth the radical changes. The household petrol consumption x_1 is a substantial part of y_2 - about 75%. Thus is seems reasonable that the estimates should at least smooth the changes which are exclusively measured by y_1, and not by the total y_2.

Figure 2 shows that modeling the petrol consumptions as stochastic trend gives estimates which smooth the data. Note that the two radical changes exclusively measured by y_1 in 1980 and 1987 are smoothed. The shape of the total petrol consumption y_2 seems to influence the estimate of the household petrol consumption strongly. However, the decrease in the total towards the end of the time series is not reflected in the estimated household consumption. This is because the model does not allow the estimates of to depart systematically from the data from the household expenditure over the years. The estimate of the household's petrol consumption increases with increasing values of y_1, while the estimated consumption by the nonhousehold decreases. This development is consistent with some

Figure 1: The data and the estimates based on model A (random walk)



estimates in million liters from the Energy accounts, cf. table 1.

Table 1: Alternative estimates of nonhousehold petrol consumption[a]

1976	1980	1985	1990	1991	1992
400	389	430	398	382	374

a) Computed and used by the Energy accounts at Statistics Norway. Million liters.

The estimated standard deviations and correlation of the noise in random walk model A and the stochastic trend model B are given in table 1 below.

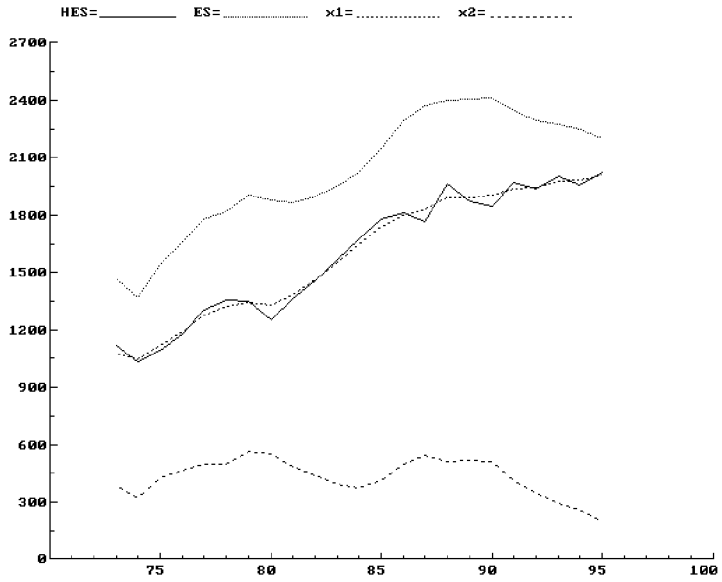
Table 2. Estimated standard deviations and correlations of the noise in model A and B

Model	s.d w_1	s.d w_2	corr(w_1, w_2)	s.d w_{β_1}	s.d. w_{β_2}	corr$(w_{\beta_1}, w_{\beta_2})$	s.d v
A	80.73	77.31	-0.48	-	-	-	30.62
B	44.77	52.84	-0.57	16.95	28.07	1.00	46.96

The ratio between estimated standard deviations, $\frac{\text{s.d}_v}{\text{s.d } w_i}$, reflects the difference in smoothness for the estimates by the two models. The higher the value, the more will the model smooth the data. For the random walk model is this ratio 0.38 for x_1 and 0.39 for x_2, while the ratios for the stochastic trend model are 1.05 and 0.89, respectively. This means that the stochastic trend model will smooth the data more than the random walk model. Furthermore, for both the models the estimated the standard deviations for the noise is bigger for $\hat{x}_2$ than for $\hat{x}_1$, reflecting that $\hat{x}_2$ varies more than $\hat{x}_1$.

For both models the correlations between w_1 and w_2 are negative. These negative correlations may indicate that the short term chances are in opposite directions. But it may also be explained by $y_2 = x_1 + x_2$, i.e., that the sum is observed without measurement error. A positive error term for x_1

Figure 2: The data and the estimates based on model B (stochastic trend)



may be linked with a negative error term for x_2 to satisfy this relation. Furthermore, the correlation between the noise terms for the slope equals 1. This indicates that long term trend for the households and the nonhouseholds are estimated to develop in the same direction. However, the interpretation of the correlation values should be taken with some care.

The standard errors of the estimates of latent consumption are equal for the household and non-household sector, since they add up to a variable measured without error, cf. relation (3) above. For the last year (T) the standard error is estimated to be 28.3 in model A and 36.5 in model B[2]. The estimated standard errors are thus lowest in the simple random walk model, which is an argument in favor of choosing this model. But one should not give this argument too much weight since the estimates of the standard error may be biased if the model is too simple.

Now we consider some model diagnostics to evaluate our models. A successful model will aim to capture the systematic movements in the data. If this goal is achieved, then what remains, namely the residuals, should be random. In other word they should contain no predictable system component. Hence, if the model is well-specified, the residuals should be random. The diagnostics are based on the residuals computed by the standardized one-step-ahead prediction errors, which are standardized estimates for the noise in the measurements. See Harvey (1990) and the STAMP manual by Koopman et al (1995) for precise definitions of the diagnostics.

[2]These standard errors are estimated by STAMP and correspond to the square root of the diagonal elements in the matrix $P(T/T)$ above. In the stochastic trend model we also obtain estimates for the standard errors of $\beta_1(T)$ and $\beta_2(T)$, which are 23.6 and 37.2 respectively.

The estimated residual autocorrelation for the random walk model at lag one and seven are

$$\hat{r}_{\hat{x}_1}(1) = -0.090 \;\; \hat{r}_{\hat{x}_1}(7) = 0.17 \;\; \hat{r}_{\hat{x}_2}(1) = 0.025 \;\; \hat{r}_{\hat{x}_2}(7) = 0.18,$$

while the estimated residual autocorrelations for the stochastic trend model at lag one and eight are

$$\hat{r}_{\hat{x}_1}(1) = 0.18 \;\; \hat{r}_{\hat{x}_1}(8) = 0.041 \;\; \hat{r}_{\hat{x}_2}(1) = -0.14 \;\; \hat{r}_{\hat{x}_2}(8) = -0.095.$$

Under the null hypothesis that the autocorrelation equals zero, the $\hat{r}$ is approximately N(0,$\frac{1}{23}$). The critical values for the null hypothesis at 5% are ± 9.40. Thus both the models have all the $\hat{r}$ values well within this interval.

The Durbin-Watson test was derived to have high power against an alternative of a first-order autoregressive residuals. However, a significant value may indicate a wide range of mispecifications, including incorrect functional form, Harvey (1981, p. 20). The general test for serial dependence is the portmanteau Box-Ljung Q-statistic, Harvey (1981, p. 148), which is based on a sum of residual autocorrelations. The Durbin-Watson statistics and the Q-statistics for the random walk, model A and the stochastic trend, model B are

$$\text{Model A: } \text{DW}_{\hat{x}_1} = 1.56 \;\; \text{DW}_{\hat{x}_2} = 1.87 \;\; \text{Q}_{\hat{x}_1} = 3.10 \;\; \text{Q}_{\hat{x}_2} = 4.92$$

$$\text{Model B: } \text{DW}_{\hat{x}_1} = 1.54 \;\; \text{DW}_{\hat{x}_2} = 2.07 \;\; \text{Q}_{\hat{x}_1} = 9.55 \;\; \text{Q}_{\hat{x}_2} = 6.60.$$

In a correctly specified model, the Durbin-Watson statistic is approximately N(2,$\frac{4}{23}$), see Koopman et al (1995, p. 231), while the Q-statistic are approximately distributed as χ^2_6 for both models, see Koopman et al (1995, p. 231). The acceptance intervals for the null hypothesis at 5% level are respectively (1.19, 2.81) and (1.24,14.45). Neither the Durbin-Watson nor the Q-statistics give any indication of serial correlation.

The statistics for testing heteroskedasticity and normality for the two models are

$$\text{Model A: } \text{H}_{\hat{x}_1} = 0.62 \;\; \text{H}_{\hat{x}_2} = 2.15 \;\; \text{N}_{\hat{x}_1} = 1.46 \;\; \text{N}_{\hat{x}_2} = 0.79$$

$$\text{Model B: } \text{H}_{\hat{x}_1} = 0.59 \;\; \text{H}_{\hat{x}_2} = 0.95 \;\; \text{N}_{\hat{x}_1} = 2.28 \;\; \text{N}_{\hat{x}_2} = 1.36.$$

The H-statistics should be tested against the $F_{7,7}$ see Koopman et al (1995, p. 232), and the N statistics should be tested against the χ^2_2, see Koopman et al (1995, p. 185). The acceptance intervals are (0.20, 4.99) and (0.051, 7.38). Thus neither of the tests give an indication of heteroskedasticity nor of non-normality. Though, the actual distributions of the tests may depart some from F and the χ^2 distributions respectively, due to small sample size.

A basic measure of goodness-of-fit is the prediction error variances, see Harvey (1989, p. 263) or Koopman et al (1995, p. 202), which is essentially the same as comparing the sums of squares of their one-step-ahead predictions errors. Their square roots, s, are

$$\text{Model A:} \quad s_{\hat{x}_1} = 88.86 \quad s_{\hat{x}_2} = 85.90$$

$$\text{Model B:} \quad s_{\hat{x}_1} = 74.29 \quad s_{\hat{x}_2} = 83.15,$$

and indicate that the stochastic trend model gives a better fit than the random walk model.

Since the models have different numbers of parameters, it is also appropriate to compare their AIC values, which gives models with more parameters a "handicap" , cf. Koopman et al (1995, p. 229). The calculated values are,

$$\text{Model A} \quad \text{AIC}_{\hat{x}_1} = 9.58 \quad \text{AIC}_{\hat{x}_2} = 9.52$$

$$\text{Model B} \quad \text{AIC}_{\hat{x}_1} = 9.66 \quad \text{AIC}_{\hat{x}_2} = 9.88.$$

The lower AIC value the better. Thus the AIC criterion slightly favors the random walk model.

None of the diagnostics indicate that the models are wrongly specified. The difference in terms of goodness-of-fit is not dramatic. The stochastic trend model results in a degree of smoothing, which corresponds more to what we think is reasonable than the random walk model do.

5 Conclusions

By means of a simple case study, we have presented an approach for estimating time series of latent variables in an accounting system. The approach exploits information from different sources, including macro and micro data. By combining different data on the same phenomena we can obtain better estimates (of e.g. household petrol consumption) than when relying on only one source. By exploiting the accounting identities we can estimate variables on which we have no direct measurements (e.g. nonhousehold petrol consumption). By formulating the accounting identities in latent variables and explicitly model the random and systematic measurement errors we obtain (i) efficient and consistent estimates given the model, (ii) possibilities for testing the model assumptions, (iii) a framework for systematic empirical research and progress in developing applied models for producing accounting statistics.

The stochastic processes of a set of latent variables in the accounting system is an important part of the model. The stochastic trend model seems to be good candidate model and gave plausible results

in our case study. The comparison between random walk and stochastic trend should be carried out in other models too, and by using simulation studies to compare how these model capture postulated true time series of the latent variables.

The state space formulation, Kalman filter and EM algorithm are convenient tools in our setting. Thus there is hope for successful use of much more complicated model specifications, for e.g. production of some parts (satellites) of a national account system.

Appendix: Data

The data and the calculations to achieve y_1 and y_2 are documented in table 3 below. Here we give a short documentation of the calculations. The micro statistics, y_1, is taken from the Norwegian Household Expenditure Survey (HES) in the period 1973-1995, cf. Statistics Norway (1996). The number of households interviewed in 1973 was 4707, while 3363 answered. For the rest of the period, the total sample size was around 2400 households and the with around 1500 answers, cf. Andersen et al (1991, p.34). The household consumption was measured in NOK per household.

We want to combine the data from HES with data on the total amount of petrol sold in Norway, measured by the Energy Statistics (ES). The data from ES is measured in million liters. We choose to transform the expenditure data from HES to consumption in million liters. There are mainly four types of petrol sold in the period we consider. We use the fractions sold of the different types in Norway and their prices to calculate a price index for petrol, transforming expenditure to liters, see table 3 for details. We base the calculation on the assumption that the distribution of the different kinds of petrol for the households is the same as for the total consumption, including the nonhousehold consumption.

By this calculation we have achieved an estimate for the average petrol consumption for the households measured in million *liters*. However, we need the estimate for the population. The consumption for the population at large is found using the average household size in the Norwegian survey of consumer expenditure, cf. Andersen et al (1991, p. 34), and the population statistics, cf. Statistics Norway (1996b, table 32,column 4). The calculation is displayed in the first term of y_1 displayed in the footnote a) for table 3.

The data from HES gives estimate for the total consumption of both petrol and autodiesel. The household consumption of the latter are probably small compared to the former, but may nonetheless influence systematically the empirical results. To reduce the bias, we have subtracted estimates of the consumption of autodiesel from the the estimates from the household survey, see again y_1 in footnote a) for table 3. The diesel estimates are derived from different sources used in the energy account at Statistics Norway. The estimates suggest that up to the mid 1970s the households' consumption of autodiesel was less than one percent of their petrol consumption. However, in the mid seventies the prices of petrol increased, as did the number of diesel cars and the consumption autodiesel. Autodiesel is cheaper than petrol and the car consume less autodiesel than petrol per kilometer. A change in the tax system from tax per kilometer to tax per car also contributed to the increase of the diesel consumption. Our numbers suggest that in 1995 the households consumed around five percent diesel compared to petrol.

The total amount of petrol sold in the period 1973–1995 is given in the Energy statistics, NOS C 347, table 3.11, column 3. We use only the petrol for cars, excluding petrol used by e.g. aircraft. This is the variable y_2.

Table 3: The time series for y_1 and y_2. The weights, the petrol prices, the population size in Norway and the estimated diesel consumption in million litres used to calculate y_1. The data covers the period 1973-1995

	Petrol cons. in mill. liters		expend.[c]	weights in price index[d]				prices[e]				diesel cons.[f] in mill. liters	household size[g]	population in Norway[h]
year	y_1[a]	y_2[b]	z	α_1	α_2	α_3	α_4	p_1	p_2	p_3	p_4	y_3	n	N
73	1122	1471	1310	.27	.73	0	0	157.1	161.8	-	-	0	2.88	3960613
74	1030	1369	1613	.28	.72	0	0	205.1	210.1	-	-	0	2.99	3985258
75	1091	1544	1577	.26	.74	0	0	204.6	208.6	-	-	0	2.79	4007313
76	1181	1659	1820	.25	.75	0	0	219.1	223.1	-	-	6	2.78	4026152
77	1304	1779	2157	.24	.76	0	0	232.3	236.9	-	-	7.1	2.82	4043205
78	1357	1822	2413	.24	.76	0	0	258.5	262.5	-	-	9.5	2.74	4058671
79	1348	1907	2509	.25	.75	0	0	277.4	281.7	-	-	10.7	2.68	4072517
80	1252	1880	2998	.30	.70	0	0	363.5	371.5	-	-	14.3	2.62	4085620
81	1365	1865	3870	.31	.69	0	0	427.0	435.0	-	-	19.0	2.65	4099702
82	1456	1899	4297	.30	.70	0	0	451.7	460.5	-	-	23.8	2.61	4114787
83	1567	1948	4969	.30	.70	0	0	480.2	492.5	-	-	28.6	2.63	4128432
84	1672	2021	5354	.29	.71	0	0	505.3	520.9	-	-	31.0	2.52	4140099
85	1780	2150	5629	.27	.73	0	0	501.8	512.8	-	-	32.14	2.53	4152516
86	1812	2297	5097	0	.82	.17	0	-	476.0	457.0	-	33.33	2.46	4167354
87	1768	2376	5421	0	.81	.19	0	-	510.0	489.0	-	33.33	2.49	4186905
88	1962	2402	6085	0	.77	.23	0	-	536.0	503.0	-	33.33	2.43	4209488
89	1870	2409	6287	0	.73	.27	0	-	578.5	540.5	-	54.4	2.43	4226901
90	1845	2413	6803	0	.64	.33	.03	-	642.8	594.0	622.1	67.4	2.41	4241473
91	1970	2346	7966	0	.53	.40	.070	-	741.0	677.0	705.0	73.7	2.33	4261732
92	1935	2292	8377	0	.45	.45	.10	-	795.0	717.0	747.0	71.5	2.37	4286401
93	2004	2274	8550	0	.31	.50	.19	-	836.2	756.0	787.1	69.0	2.26	4311991
94	1959	2247	8591	0	.078	.52	.40	-	851.1	760.5	791.4	77.3	2.35	4336613
95	2024	2204	9099	0	.071	.56	.37	-	893.0	807.0	838.0	83.9	2.28	4359184

a)Estimates based on Survey of household expenditure, $y_1 = \frac{N\,z}{n\sum_{i=1}^{4}\alpha_i p_i} - y_3$, cf. the other columns for definitions of symbols.

b)Energy statistics 1995, Statistics Norway, NOS C 347, table 3.11, column 3.

c)Mean household expenditure on 621 Gasoline and oil from the Norwegian survey of household expenditure, 1973–1995, cf. Statistics Norway (1996)

d)Energy statistics, Statistics Norway, 1:Regular gasoline, 2:Super gasoline (leaded),3:Unleaded gasoline 95,4:Unleaded gasoline 98.

e)Energy statistics, Statistics Norway; NOS A 977 table 37, row 2 and 4; NOS B 709 table 31, row 2 and 4; NOS C 347, table 4.9, row 2,4,6 and 8, Statistics Norway

f)Estimated approximately diesel consumption by the households from the energy account, Statistics Norway.

g)Average household size in the Norwegian survey of household expenditure, Statistics Norway.

h)Population size in Norway, Norwegian statistical yearbook 1996, table 32, column 4.

References

Aasness, J., E. Biørn and T. Skjerpen (1993): Engel functions, panel data and latent variables, *Econometrica* **61**, 1395-1422.

Adler, H.J. and M. Wolfson(1988): A prototype micro – macro link for the Canadian household sector, *The Review of Income and Wealth* **34.**

Andersen, A., S. Opdahl and J. Aasness (1991): Nytte og kostnader ved alternative opplegg for SSB's forbruksundersøkelser (Cost and benefit of different designs of SSB's household expenditure surveys), Interne Notater 91/22, Statistics Norway, Oslo.

Aukrust, O. (1966): An axiomatic approach to national accounting: An outline, *The Review of Income and Wealth* **12,** 179-190.

Aukrust, O. (1994): "The Scandinavian contribution to national accounting" in Z. Kenessey (ed): *The Accounts of Nations*, Amsterdam: IOS Press, 16-65.

Dagsvik, J. (1978): Filter estimation in repeated sample surveys, mimeo, Statistics Norway.

Gonzales, P. and P. Moral (1995): An analysis of international tourism demand in Spain, *International Journal of Forecasting* **11,** 233-251.

Gorter C. and P. Van der Laan (1992): An economic core system and the socio-economic accounts module for the Netherlands, *The Review of Income and Wealth* **38,** 347-392.

Hansen H., N. Langbraaten, and Røstadsand J.I. (1992): Sosio-økonomisk husholdningsregnskap, Interne notater 92/12, Statistics Norway.

Harvey, A.C. (1981): *Time series models,* Deddington, Oxford: Phillip Allan; Atlantic Highlands, New Jersey: Humanities Press.

Harvey, A.C. (1989): *Forecasting, the structural time series and the Kalman filter*, Cambridge U.K: Cambridge University Press.

Kalman R.E. (1960): A New Approach to Linear Filtering and Prediction Problems, *Journal of Basic Engineering, Transactions ASME*, Series D **82**, 35-45.

Kalman R.E and Bucy R.S. (1963): New Results in Linear Filtering and Prediction Theory, *Journal of Basic Engineering, Transactions ASME*, Series D **83,** 95-108.

Koopman S.J., A.C. Harvey, J.A. Doornik and N. Shephard (1995): *STAMP 5.0 - Structural Time Series Analyzer, Modeller and Predictor,* London: Chapman and Hall.

Liviatan, N. (1961): Errors in variables and Engel curve analysis, *Econometrica* **29**, 336-362.

Patterson, K.D. (1995): An integrated Model of the Data Measurement and Data Generation Processes

with Application to Consumer Expenditure, *The Economic Journal* **105,** 54-76.

Rahiala, M. T. Teräsvirta (1993): Business Survey Data in Forecasting the Output of Swedish and Finnish Metal and Engineering Industries: A Kalman Filter Approach, *Journal of Forecasting* **12,** 255-271.

Ripley B.(1987): *Stochastic simulation*, New York: Wiley.

Skjerpen T and A. R. Swensen (1997): *Forecasting Manufacturing Investment Using Survey Data*, Reports 97/3, Statistics Norway.

Shumway R.H. (1988): *Applied Statistical Time series Analysis*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Shumway R.H. and D.S. Stoffer (1982): An approach to time series smoothing and forecasting using the EM algorithm, *Journal of time series Analysis* 1982, 3, 253-264.

Statistics Norway (1996): *Survey of consumer expenditure 1992–1994*, Official Statistics of Norway NOS C 317.

Statistics Norway (1996b): *Statistisk arbok 1996*.

Statistics Norway (several years): *Energy statistics*, Official Statistics of Norway, NOS A977, B709, B995 and C 347.