An Evaluation of Forecasting Methods and Forecast Combination Methods in Goods Management Systems

Carsten Schneider[1], Matthias Klapper[2], and Thomas Wenzel[3]

Abstract: In this paper we use 4 different time series models to forecast sales in a goods management system. We use a variety of forecast combining techniques and measure the forecast quality by applying symmetric and asymmetric forecast quality measures. Simple, rank-, and criteria-based combining methods lead to an improvement of the individual time series models.

Keywords: ARX forecasts, dynamic linear model, combination of forecasts, goods management system, asymmetric loss function.

Acknowledgement: Supported by Sonderforschungsbereich 475 and Deutsche Forschungsgemeinschaft, Graduiertenkolleg ”Angewandte Statistik”.

AMS 1991 Subject Classification: 62M10

1 Introduction

In this paper we will propose various methods for predicting sales figures in goods management systems and analyze the results. We use simple time series methods like autoregressive models and dynamic linear models and also model the seasonal behaviour of the time series. In addition to these simple univariate methods we focus on combinations of these models which may lead to better forecasts. The forecast quality may increase especially if the combined simple forecasts use different information bases in the sense of usual measures like

[1] Department of Economics, University of Wuppertal, 42097 Wuppertal, Germany, schneider@wwst09.wiwi.uni-wuppertal.de
[2] Department of Statistics, University of Dortmund, 44221 Dortmund, Germany, klapper@amadeus.statistik.uni-dortmund.de
[3] Department of Statistics, University of Dortmund, 44221 Dortmund, Germany, wenzel@amadeus.statistik.uni-dortmund.de

mean squared error or mean absolute deviation.

We will start in Section 2 by giving information about the data and the data quality which has great influence on the forecast quality. Section 3 describes the simple forecast strategies and the combination methods. In Section 4 we will derive suitable performance measures and use them to compare the quality of the various forecast models. The last section (5) will close with summaries and conclusions.

2 The data

There are 630 time series of sales items available for analysis with average weekly sales of 10 or more. The exact number of sales is registered by an automatic scanner cash desk and we also have calendar information available which is important information for many items. The 630 time series are rather short. There are 106 observations from week 51/1995 to week 52/1997, therefore estimating a seasonal behaviour may cause some problems.

Items 2170, 2347, and 40606 have missing values for the first 3-11 weeks. None of the regarded time series has any obvious wrong data points. Items 981, 1201, 1203, 1684, 2170, 3318, 9424, 10777, 11905, 14374, 15307, 15378, 22155, 24532, 27361, 34873, 35980, 36292 and 40132 have zero sales observation, zero prediction for one of the models, zero prediction error for several periods, and therefore no value for some combining methods. This causes some combining methods to show zero denominators and therefore to have no measure for these sales items. All 21 items listed were excluded from the analysis, and we will therefore deal with 609 sales items from here on.

We use the first twenty observations to generate the models. This leaves us with 84 time points that we use to analyse the forecast quality. Due to practical considerations two-step-ahead forecasts are more important than one-step-ahead forecasts and therefore we only look at two-step-ahead forecasts in this paper.

3 The forecasting models

3.1 Autoregressive Models

This is one of the most popular forecast strategies for the prediction of sales figures in goods management systems. In order to be able to forecast the sales figures automatically, we predict all time series with the same kind of AR model and only estimate the parameters individually. The model that fits most time series best is an AR(2) model of the kind:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \varepsilon_t \quad .$$

This model takes only the past of the time series into account and may therefore not be optimal if further information is available. In this case the model can be extended to an ARX model that contains the calendar information. Due to the short time series, a complete seasonal figure of length 52 can not be estimated, therefore only special seasonal influences are regarded. In this case we only consider the Christmas-effect. This effect usually starts to influence the time series some weeks before Christmas and increases up to a peak during Christmas-week. Some items only show a single peak during Christmas-week. Both kinds of effects are usually followed by one or a few weeks of very low sales figures. We model this effect relative to the correlation between subsequent observations in an ARX process. Therefore, we choose a somewhat complicated model of the following form:

$$y_t = \beta_0 + \beta_1\, y_{t-1} + \beta_2\, y_{t-2} + \beta_3\, c_t + \beta_4\, d_t + \varepsilon_t\ .$$

The model is of an ARX(2)-type with c_t as a dummy variable for Christmas and d_t as a dummy variable for the weeks after Christmas. We will code it as [AR(2)-2 Seas.]-model.

3.2 Dynamic linear models

These models are based on the ideas of Kalman-filtering and are useful to predict time series if the generating process changes over time. In an AR model it would be necessary to make new identifications and estimations of the model if something may have changed whereas in a dynamic linear model these changes are included automatically. In this context we use two kinds of dynamic linear models, one that interprets the process to be mean stationary [2-DLM] and one with trend adjustment [2-DLM, TR].

The dynamic linear models are defined by two equations:

- Observation equation:

$$y_t = \boldsymbol{F_t}'\boldsymbol{\Theta_t} + \nu_t \qquad \nu_t \sim \boldsymbol{N}[0,\mathrm{V}_t]$$

- System equation:

$$\boldsymbol{\Theta_t} = \boldsymbol{G_t}\, \boldsymbol{\Theta_{t-1}} + \omega_t \qquad \omega_t \sim \boldsymbol{N}[0,\mathrm{W}_t]$$

The observation equation contains a parameter matrix $\boldsymbol{\Theta_t}$ which can be viewed as a usual parameter matrix in a least squares model. The difference is that the parameter matrix is treated as normal distributed random variable following the system equation. Therefore, the model is dynamic in the sense of possibly changing parameters through time. A dynamic linear model can by defined depending on the four variables $\boldsymbol{F_t}$, the regression vector, the evolution transfer matrix $\boldsymbol{G_t}$,, and the variance matrices V_t and W_t. For reference see West and Harrison (1997).

3.3 Combination of forecasts

These methods use several forecasts of the same event to generate a combined forecast. We will use several forecast combining techniques that are presented in detail in Russell and Adam (1987), Klapper (1998), and Granger and Ramanathan (1984), and therefore only give brief descriptions of the methods.

Simple combining methods:

- The simple arithmetic mean (SA) takes the arithmetic mean of the 4 forecasts in each time period. It does not depend on the past forecast errors and has shown good results in the literature.
- The simple median (SM) works like the SA but takes the median of the forecasts instead of the mean.
- The simple range method (SR) works like the SA but takes the midpoint of the range of the forecasts instead of the mean.

Rank based and criteria based methods:

- The method called Rank uses the inverse of the sum of the ranks of the last 10 time points for each forecaster divided by the sum of the inverses of all forecasters. This results in coefficients for each forecaster that are greater than zero and add up to one, which is the same for all other rank based methods.
- The RQua technique does the same as Rank but takes the quadrupled ranks instead of the simple ranks.
- RHist works like Rank but includes all past time points instead of the last 10.
- R0.5 averages the coefficients calculated like Rank and the coefficients of the previous time period.
- The cmse technique as explained in Russell and Adam (1987) takes the inverse of the MSE of the past 10 performance periods for each forecaster and puts it into relation to the sum of the inverse MSEs of all forecasters.
- The cmad technique works like cmse but uses the MAD instead of the MSE.

L1- and OLS-based methods:

- The method OLS does an OLS-regression of the individual forecasts on the realization using the last 10 time points at each step.
- OLS/1 works like OLS but restricts the coefficients to add up to 1.
- OLS/I works like OLS but also uses an intercept.
- The method OLS/H is an OLS regression of the individual forecasts on the realization using all available history.
- OLS/1H works like OLS/H but restricts the coefficients to add up to 1.
- OLS/IH works like OLS/H but also uses an intercept.
- The method L1 does an L1-regression of the individual forecasts on the realization using the last 10 time points at each step.
- L1/1 works like L1 but restricts the coefficients to add up to 1.
- L1/I works like L1 but also uses an intercept.
- The method L1/H is an L1 regression of the individual forecasts on the realization using all available history.
- L1/1H works like L1/H but restricts the coefficients to add up to 1.
- L1/IH works like L1/H but also uses an intercept.

Since OLS- and L1-regression based methods could result in negative weights for some individual forecasts and therefore generate a negative combined value, we correct these cases to zero because negative stock orders are not possible.

4 Results of forecasting and combining

To compare the performance of the models presented in Section 3, we have to find appropriate forecast quality measures that fit the economic needs behind the data. The MAD seems to be a more appropriate measure than the RMSE, since a linear loss makes more economic sense. Interest rates and stocking costs for overstocking and costs for suppressed sales due to stockouts usually grow linear and depend (as long as the differences are not too high) on the difference between forecast and sales and constants. We will also look at the average number of stockouts (MSO), the average number of suppressed sales (MSS), the average number of overstockings (MOS), and the average number of overstocked units (MOU), all described in detail in Arminger and Götz (1999).

Arminger and Schneider (1999) mention that asymmetric loss functions are more appropriate to be used in goods management systems. Their motivation is that the lost profit margin due to suppressed sales is usually higher than the interest and stocking costs for overstocked items. Arminger and Götz (1999) propose a lin-lin loss function in the following way:

$$MLL_{a,b} = \frac{1}{T-h}\sum_{t=h}^{T} L(y_t, f_t)$$

with

$$L(y_t,f_t) = \begin{cases} a(|y_t - f_t|) \; for \; y_t \le f_t \\ b(|y_t - f_t|) \; for \; y_t > f_t \end{cases}$$

where a and b are the penalties for overstocking and stockout, respectively, y_t and f_t are the realisation and the forecast at time t, respectively, and h is the index of the first observation included. To keep things simple, we do not consider a and b to be time dependent.

The results from combining the 4 individual forecasts for the first 10 sales items are displayed in Table 1 and 2. Table 1 shows the MAD of the combining methods in relation to the simple average's MAD. Numbers below 1 indicate that the particular combining method outperforms the simple average. The rank-based combining techniques like Rank score below 1 for items 186, 232, and 343. It turns out that usually all or none of the rank based techniques outperform the simple average and that the OLS- and L1- based techniques perform much worse. The best individual time series method is 2-DLM that scores 3 times below 1. This already indicates that the simple average may be a better choice than the individual methods.

Table 1. MAD relative to the simple average's MAD for the first 10 items

time series	186	195	211	229	232	237	298	309	342	343
individual TS methods:										
AR2-2	0,993	1,012	1,009	1,057	1,033	1,022	1,042	0,971	1,026	1,075
AR2-2, Seas	1,028	1,094	1,053	1,010	1,075	1,028	1,014	1,040	1,085	1,077
2-DLM	1,049	1,123	1,013	1,033	0,992	1,003	1,071	1,009	0,997	0,968
2-DLM, TR	1,122	1,052	1,079	1,085	1,032	1,081	1,068	1,104	1,008	1,008
simple combining methods:										
SA	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
SM	0,979	0,999	1,007	1,012	0,995	1,011	1,000	0,992	0,989	1,010
SR	1,023	1,021	0,994	0,996	1,011	0,998	1,006	1,015	1,014	0,997
rank/criteria-based methods:										
Rank	0,998	1,014	1,000	1,005	0,996	1,002	1,002	1,001	1,001	0,992
RQua	0,997	1,022	1,011	1,010	0,993	1,006	1,004	1,004	1,007	0,991
RHist	0,995	1,003	1,000	1,001	1,000	1,001	0,999	0,998	1,000	0,999
R0.5	0,997	1,012	1,000	1,004	0,998	1,002	1,002	1,001	1,001	0,993
cmse	1,001	1,009	1,009	1,007	0,994	1,004	1,001	1,002	1,008	0,993
cmad	0,999	1,013	1,002	1,005	0,995	1,002	1,001	1,002	1,003	0,992
L1- and OLS-based methods:										
OLS	4,349	2,337	1,777	4,256	3,640	3,869	2,025	5,301	2,683	1,704
OLS/1	2,047	2,542	1,640	1,751	1,471	1,853	1,614	3,126	1,591	1,772
OLS/I	4,579	2,256	2,219	4,879	2,484	2,072	2,875	6,344	3,647	2,113
L1	5,184	2,179	2,278	3,949	3,373	4,526	2,604	2,764	3,490	1,990
L1/1	2,162	1,995	1,670	2,122	1,703	1,577	1,636	2,919	1,774	1,690
L1/I	4,470	2,590	3,114	5,394	2,930	3,529	2,126	7,774	4,726	2,576
OLS/H	1,252	1,278	1,055	1,420	1,875	1,095	1,317	2,316	1,215	1,095
OLS/1H	1,223	1,337	1,089	1,372	1,099	1,164	1,272	2,543	1,051	1,102
OLS/IH	1,196	1,474	1,125	1,422	1,657	1,207	1,112	4,277	1,404	1,176
L1/H	1,332	1,350	1,082	1,518	1,936	1,232	1,412	1,167	1,365	1,493
L1/1H	1,277	1,402	1,105	1,525	1,246	1,235	1,280	1,935	1,173	1,143
L1/IH	1,236	1,505	1,119	1,935	1,889	1,406	1,330	3,504	1,865	1,496

Table 2 shows the average number of suppressed sales (MSS) which is a crucial component of the cost analysis. Customers may be dissatisfied since they do not find the product they are looking for and the company loses its profit margin that is usually higher than interest rates for overstocking. Here, the cmse beats the simple average 8 out of 10 times and the other rank- and criteria-based methods are also mostly below 1. The OLS- and L1-based methods perform much better for the MSS compared to their MAD performance. They beat the simple average by a higher margin of up to 22.0% for L1/IH for item 342.

Table 2. MSS for the first 10 items relative to the SA's MSS

time series	186	195	211	229	232	237	298	309	342	343
individual TS methods:										
AR2-2	1,085	1,023	1,205	1,160	1,182	1,161	1,083	0,905	1,100	0,899
AR2-2, Seas	1,016	1,190	1,225	1,159	1,188	1,187	1,145	1,004	1,224	0,928
2-DLM	1,034	1,128	0,882	0,928	0,874	0,876	1,053	1,016	0,925	1,066
2-DLM, TR	1,058	0,913	0,806	0,899	0,861	0,885	0,865	1,217	0,854	1,268
simple combining methods:										
SA	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
SM	0,985	0,979	1,025	1,021	1,004	0,996	1,018	0,989	0,999	0,993
SR	1,017	1,039	0,975	0,985	1,001	1,011	0,987	1,020	1,004	1,015
rank/criteria-based methods:										
Rank	0,999	1,022	0,998	1,002	0,991	0,998	1,004	1,001	0,996	1,027
RQua	0,987	0,993	0,993	0,993	0,976	1,000	0,999	1,002	1,001	1,046
RHist	0,999	0,999	1,002	1,003	0,998	1,003	0,993	0,995	1,001	1,008
R0.5	0,999	1,019	0,999	1,000	0,993	0,998	1,004	1,000	0,997	1,022
cmse	0,983	0,988	0,996	0,992	0,979	1,001	0,983	0,996	1,002	1,036
cmad	0,993	1,014	0,998	0,998	0,989	1,001	0,997	0,999	0,997	1,027
L1- and OLS-based methods:										
OLS	1,401	1,220	1,388	0,980	0,973	1,321	1,231	1,502	1,422	1,967
OLS/1	1,271	0,877	1,298	1,093	1,136	0,967	1,186	1,108	1,461	1,883
OLS/I	1,363	1,162	1,597	1,116	1,349	1,694	1,313	2,206	2,093	1,905
L1	1,506	1,332	1,725	1,131	1,079	1,380	1,114	1,738	1,163	2,016
L1/1	1,351	0,953	1,349	1,140	1,150	0,968	1,321	1,518	1,161	1,903
L1/I	1,416	1,155	1,952	1,263	1,454	1,601	1,560	2,084	1,818	2,147
OLS/H	1,190	1,042	1,019	0,975	0,975	0,872	0,847	1,098	0,898	1,718
OLS/1H	1,202	0,883	1,116	1,041	0,876	0,932	0,861	1,062	0,950	1,610
OLS/IH	1,214	1,329	1,083	1,006	1,031	1,199	0,984	1,090	0,839	1,534
L1/H	1,371	0,886	1,190	1,075	0,877	0,831	0,920	1,511	0,945	1,664
L1/1H	1,312	0,827	1,273	1,113	0,891	1,043	0,901	1,070	0,907	1,750
L1/IH	1,459	1,290	1,235	1,233	1,060	1,444	1,138	1,205	0,780	1,518

For the lin-lin loss based average cost of suppressed sales and overstocking per period and item (MLL), we use the parameters a=0.10 and b=0.20. This can be interpreted as 10% cost for storage and interest for overstocked items and 20% cost for suppressed sales due to the lost profit margin. In this case, b gets a higher weight than a as mentioned above. Since we do not have the individual profit margin and interest rates, we have to weight all products the same. Looking at all 609 items we can see from Table 3 that the average MADs for the individual methods are between 1.4% and 9.4% higher than average MADs of the simple average. This can even be improved by the rank- and criteria-based methods that improve this result by another 0.2%. R0.5 beats the simple average most often and RHist scores best versus the best individual forecasting method. In 333 out of 609 cases the 2-DLM is the best individual time series method. If we look at the distribution of the relative MADs in Figure 1, we can see that most of the values are greater than 1 for the individual methods. In 3 out of 4 cases even more than 75% of the time the simple average is better than the individual model. Table 3 shows a little less than half of the 2-DLM model beating the MAD of the simple average compared to 53% for the best combining method R0.5.

Other criteria deliver similar results. The simple average beats the 4 individual forecasting models by 3.4%-9.1% for the average number of supressed sales (MSS), and the average number of overstocked units (MOU) for the simple average is 2.7%-8.9% lower than for the individual forecasting models. For the overstockings the method RQua even outperforms the simple average by 0.9%. Many of the regression based combining techniques beat the MOS of the simple average by up to 6.1% for L1-regression with intercept. This measure, on the other hand, is not that crucial since it only counts the occurrences of overstocking but usually the actual severity measured by the MOU is more important, where the regression based methods perform very poorly.

The weighted measure of stockouts and overstockings MLL shows similar results as the MAD. If we weight the cost for suppressed sales twice the cost for overstockings, the simple average beats the individual models by 3.2%-9.0%. All rank/criteria based methods even beat the simple average by up to 0.2%.

Table 3. Performance of the individual and combined forecasts

	beat MAD of			stockout		overstocking		lin-lin-loss
method	SA	best indiv.	MAD	MSO	MSS	MOS	MOU	$MLL_{0.10,0.20}$
individual TS methods:								
AR2-2	126	160	1,045	1,002	1,050	0,998	1,043	1,048
AR2-2, Seas	45	59	1,094	1,007	1,091	0,993	1,089	1,090
2-DLM	278	333	1,014	0,998	1,034	1,002	1,027	1,032
2-DLM, TR	75	57	1,062	1,034	1,055	0,967	1,083	1,064
simple combining methods:								
SA	-	209	1,000	1,000	1,000	1,000	1,000	1,000
SM	251	194	1,004	0,997	1,002	1,003	1,007	1,004
SR	238	181	1,004	1,005	1,007	0,995	1,002	1,005
rank/criteria-based methods:								
Rank	314	211	0,998	1,004	1,002	0,996	0,993	0,999
RQua	298	199	0,998	1,008	1,003	0,992	0,991	0,999
RHist	323	213	0,999	1,000	1,000	1,000	0,998	1,000
R0.5	324	209	0,998	1,003	1,002	0,997	0,993	0,999
cmse	309	200	0,998	1,004	0,999	0,997	0,995	0,998
cmad	318	206	0,998	1,003	1,001	0,997	0,994	0,998
L1- and OLS-based methods:								
OLS	0	0	3,001	1,032	1,571	0,969	4,406	2,516
OLS/1	0	0	2,274	1,020	1,411	0,980	3,202	2,008
OLS/I	0	0	4,090	1,053	1,837	0,950	6,475	3,383
L1	0	0	3,270	1,050	1,673	0,952	4,719	2,688
L1/1	0	0	2,381	1,034	1,473	0,967	3,282	2,076
L1/I	0	0	4,478	1,064	1,963	0,939	7,030	3,652
OLS/H	11	0	1,604	1,001	1,100	0,998	2,058	1,419
OLS/1H	13	3	1,532	0,995	1,075	1,004	1,934	1,361
OLS/IH	6	1	1,793	0,994	1,132	1,005	2,480	1,581
L1/H	4	1	2,218	1,054	1,194	0,950	2,036	1,474
L1/1H	14	4	1,534	1,012	1,107	0,988	1,878	1,364
L1/IH	4	2	1,938	1,051	1,241	0,953	2,553	1,678

MAD, MSO, MSS, MOS, MOU, and MLL in relation to the simple average. For MLL a=0.10 and b=0.20.

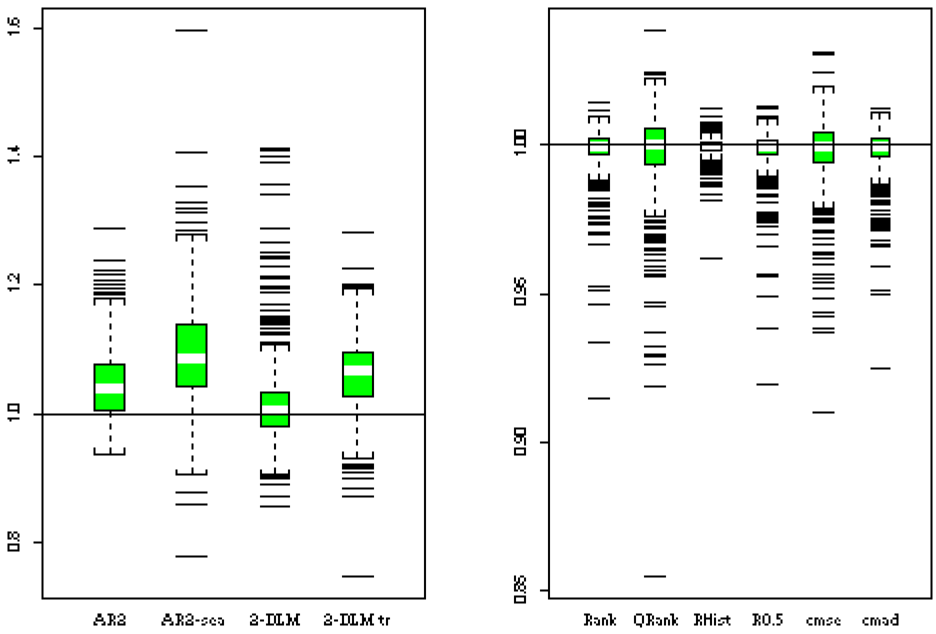

Figure 1: Distribution of the MADs relative to the Simple Average's MAD.

5 Summary

We use 4 different time series models to forecast sales for 609 sales items of a drug store chain in Germany. The dynamic linear model with no trend performs best among these methods, scoring the lowest MAD over all 609 items. Next, we use 21 forecast combining techniques to create combined forecasts. Our result: Combining forecasts in goods management systems turns out to be successful. The simple average that was used as our primary benchmark lowered the average MAD of the individual forecasts by 1.4%-9.4% and the economically more relevant MLL by 3.2%-9.0%. This is topped by the rank based methods that lower the MAD of the simple average by another 0.1%-0.2%.

The regression based methods do not perform well. They do not constrain their weights between 0 and 1, to sum up to 1, and are therefore not limited within the range of the individual forecasts. We have to remark here that further research is needed with an analysis of the influence of the data structure on the quality of the combination techniques. There are some cases where the OLS-based methods produce better results as described in Granger and Ramanathan (1984). This raises a general question in the theory of combining forecasts, whether one should use a combined forecast that is smaller than the minimum or larger than the maximum of the individual forecasts.

It makes sense to combine and now the user has to decide whether the small gains over the simple average justify the extra amount of time and costs that are necessary for calculating the combining weights of the rank based methods.

Bibliography

- **Arminger,** Gerhard and **Schneider,** Carsten (1999): Frequent problems of model specification and forecasting of time series in goods management systems, *Technical Report 21/1999, SFB 475, University of Dortmund.*
- **Arminger,** Gerhard and **Götz,** Norman (1999): Asymmetric loss functions for evaluating the quality of forecasts in time series for goods management systems, *Technical Report 22/1999, SFB 475, University of Dortmund.*
- **Klapper,** Matthias (1998): Combining German macro economic forecasts using rank-based techniques, *Technical Report 19/1998, SFB 475, University of Dortmund.*
- **Russell,** Thomas D. and **Adam,** Everett E. Jr. (1987): An empirical evaluation of alternative forecast combinations, *Management Science Vol. 33, 1267-1276*
- **Granger,** C.W.J and **Ramanathan,** R. (1984): Improved methods of combining forecasts, *Journal of Forecasting 3, 197-204*
- **West,** M. and **Harrison,** J. (1997): *Baysian forecasting and dynamic models,* 2nd Ed., Springer-Verlag, New York