Journal of Applied and Natural Science
11(3): 661-665 (2019)
ISSN : 0974-9411 (Print), 2231-5209 (Online)
journals.ansfoundation.org

Unobserved components model for forecasting sugarcane yield in Haryana

Ekta Hooda*
Department of Mathematics and Statistics, CCS Haryana Agricultural University, Hisar-125004 (Haryana), India
Urmil Verma
Department of Mathematics and Statistics, CCS Haryana Agricultural University, Hisar-125004 (Haryana), India

*Corresponding author. E-mail: ektahooda@gmail.com

Article Info
https://doi.org/10.31018/jans.v11i3.2144
Received: July 10, 2019
Revised: August 14, 2019
Accepted: August 25, 2019

How to Cite
Hooda, E. and Verma, U. (2019). Unobserved components model for forecasting sugarcane yield in Haryana. *Journal of Applied and Natural Science*, 11(3): 661 - 665 https://doi.org/10.31018/jans.v11i3.2144

Abstract
Unlike classical regression analysis, the state space models have time-dependent parameters and provide a flexible class of dynamic and structural time series models. The unobserved component model (UCM) is a special type of state space models widely used to analyze and forecast time series. The present investigation has been carried out to study the trend of sugarcane(gur) yield in five districts (Ambala, Karnal, Panipat, Yamunanagar and Kurukshetra) of Haryana state using the unobserved component models with level, trend and irregular components. For this purpose, the time series data on sugarcane yield from 1966-67 to 2016-17 of Ambala and Karnal, 1971-72 to 2016-17 of Kurukshetra and 1980-81 to 2016-17 of Panipat and Yamunanagar districts have been used. For all the districts, the irregular component was found to be highly significant (p=0.01) while both level and trend component variances were observed non-significant. Significance analysis of the individual component(s) has also been performed for possible dropping of the level and trend components by setting their variances equal to zero. The state space models may be effectively used pertaining to Indian agriculture data, as it takes into account the time dependency of the underlying parameters which may further enhance the predictive accuracy of the most popularly used ARIMA models with parameter constancy. Moreover, the unobserved component model is capable of handling both stationary as well as non-stationary time series and thus found more suitable for sugarcane yield modeling which is a trended yield (i.e. non-stationary in nature).

Keywords: Forecast, Local level trend model, State space model, Sugarcane yield, Unobserved component model

INTRODUCTION

The Autoregressive Integrated Moving Average (ARIMA) models have been used to model agricultural time-series data related to yield and production of sugarcane(gur) and other crops in India, [Suresh and Priya (2011), Suman and Verma (2017)]. These models are suitable only for stationary time-series (Box and Jenkins, 1976). For the widely used ARIMA methodology, the rule of thumb is that one should have at least 50 observations but preferable more than 100 observations (Box and Tiao, 1975). This methodology could lead to inappropriate model specifications and prediction if the number of observations is less than 40.

State space models are widely used in time series analysis to deal with processes which gradually change over time. Expositions of the state space approach to multivariate forecasting can be found in Akaike (1976), Kitagawa and Gersh (1984) and Durbin and Koopman (2002). A good account on state space modelling is also given in the books by Aoki (1987) and Commandeur and Koopman (2007). Ravichandran and Prajneshu (2000) studied Box-Jenkins ARIMA and state space modelling approach using Kalman filtering technique for analyzing all-India marine products export data. The goodness of fit statistics viz., AIC, SBC and RMSE favoured the use of state space model as compared to ARIMA model. Rajarathinam *et al.* (2016) studied the trends in area, production and productivity of wheat in India during 1950 to 2014 using the unobserved component model.

Unobserved Component Modeling is a promising alternative approach to model time series data (Harvey, 2001). It is a flexible class of structural time-series models and decomposes a given time series into latent components such as trend, cyclical, seasonality, linear and non-linear regression

effects. The main feature of UCM is the latent components, which follows suitable stochastic models and provides a suitable set of patterns to capture the outstanding actions of the response series. UCM assumes that the latent components are stochastically independent of each other and allows for inclusion of explanatory variables. All the component models in UCM can be thought of as stochastic generalization of the corresponding deterministic time series patterns.

Apart from the forecast, structural time series models give estimates of these unobserved components. In many time series the adjacent observations are more closely correlated with each other than observations those are far apart. The UCMs are local in nature and give higher weights to the recent observations than observations in the distant past. These models tend to predict better than models that treat time-series data globally as in the deterministic time trend model. Keeping in view the above points, UCMs have been developed to fit the trend in sugarcane yield of five districts (Ambala, Karnal, Panipat, Yamunanagar and Kurukshetra) in Haryana assuming the level and trend components to be locally linear as well as when level and trend components remain constant without any persistent upward or downward drift.

MATERIALS AND METHODS

The Haryana state comprised of 22 districts is situated between 74° 28` to 77° 36` E longitude and 27° 37` to 30° 35` N latitude. The time series data on sugarcane yield from 1966-67 to 2016-17 of Ambala and Karnal, 1971-72 to 2016-17 of Kurukshetra and 1980-81 to 2016-17 of Panipat and Yamunanagar districts compiled from statistical abstracts of Haryana have been used for the present study. The data for the last six years i.e., 2011-12 to 2016-17 have been used to check the validity of the developed models for district-level sugarcane yield prediction. The PROCUCM procedure available in SAS have been used for data analysis.

Unobserved component model: The unobserved component model can be considered as a multiple regression model with time-varying coefficients. It is based on the principles that a time series can be decomposed into trend, seasonal and cycle components and that in many time series the adjacent observations are more closely correlated with each other than observations those are far apart.

The UCM consists of trend, cycle, seasonal and irregular components and is expressed as

$y_t = \mu_t + s_t + c_t + \varepsilon_t$ Eq….(1)

Where μ_t denotes the stochastic trend in the time series y_t at time t, s_t the stochastic seasonal effect at time t and c_t the cyclical effect at time t. Here, ε_t is the overall error or irregular component at time t, which is assumed to be Gaussian white noise with variance σ^2_ε. In case of annual time series, the seasonal and cyclic effects cannot be identified and the UCM also called the Local Linear Trend Model (LLTM) is formulated as:

$y_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \sim NID(0, \sigma^2_\varepsilon)$

$\mu_{t+1} = \mu_t + v_t + \xi_t, \quad \xi_t \sim NID(0, \sigma^2_\xi)$ Eq…..(2)

$v_{t+1} = v_t + \eta_t, \quad \eta_t \sim NID(0, \sigma^2_\eta)$

for t = 1,2 …..n. This model contains two state equations, one each for modeling the level, and the slope. The stochastic slope v_t in equation (2) is equivalent to regression coefficient in classical regression model and μ_t is the unobserved level at time t which is equivalent to the intercept in the classical regression model, ε_t, is the observation disturbance at time t, ξ_t and η_t are the level and slope disturbances respectively. For the LLTM, the slope also determines the angle of the line with the time axis. The important difference is that the regression coefficient is fixed in classical regression model, whereas, the model in equation (2) allows both the level and slope to vary over time. In LLTM, the slope is also referred to as the drift. In state space models, the unknown parameters include the observation and the state disturbance variances, i.e. σ^2_ε, σ^2_ξ and σ^2_η. These parameters are also called the hyper parameters.

If $\sigma^2_\eta = 0$, the model in (2) have stochastic level and deterministic slope and is known as Local Linear Model (LLM) or the random walk model. This model can be written as

$y_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \sim NID(0, \sigma^2_\varepsilon)$

$\mu_{t+1} = \mu_t + v_1 + \xi_t, \quad \xi_t \sim NID(0, \sigma^2_\xi)$ Eq…..(3)

If both of the state disturbance variances σ^2_ξ and σ^2_η are zero then model given in equation (1) reduces to the classical repression model. In this case the linear trend models simplifies to

$y_t = \mu_1 + v_1 g_t + \varepsilon_t, \quad \varepsilon_t \sim NID(0, \sigma^2_\varepsilon)$ Eq…...(4)

for t = 1,2, ……, n, where, the predictor variable g_t = t-1 for t = 1,2, …,n is time effective and μ_1 and v_1 are the initial values of the level and slope (Commandeur and Koopman, 2007).

Model selection criteria: The following criteria have been used for comparing the performance of LLM and LLTM models developed for sugarcane yields of various districts:

Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Mean Absolute Prediction Error

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Relative Deviation (%)

$$RD(\%) = \left(\frac{Observed - Forecast}{Observed}\right) x\ 100$$

Akaike Information Criteria (AIC)

For state space models, the AIC takes the form

$$AIC = \frac{1}{n}[-2n\log(L_d) + 2(q + w)]$$ Eq…..(5)

Where y_i is the actual or observed value and $\hat{y}_i$ is predicted/forecast value, n is the number of observations in the time series, log(L_d) is maximized diffuse log-likelihood function, q is the diffuse initial values in the state and w is the total number of error variances estimated in the analysis.

RESULTS AND DISCUSSION

Unobserved Component Modeling approach was used to fit the trend in sugarcane yield of five districts i.e., Karnal, Ambala, Yamunanagar, Panipat and Kurukshetra in Haryana. Initially, all possible components viz., level, trend and irregular were estimated and tested using the UCM or local linear trend model given in Equation 2. In the initial stage, the analysis aimed to identify the existing stochastic components in the model. Error variances of irregular, level and slope components, also known as free parameters of the model were estimated and are given in Table1. The estimates along with their corresponding t-values and the associated p-values have also been given for testing the stochastic nature of the components.

The results of LLTM shown in Table 1 reveal that the error variance of irregular component is highly significant for all the districts under consideration. However, the disturbance variances of level and slope components are found non-significant for all the five districts. It indicates that level and trend components can be treated as constant as they have near zero estimated variances for the five districts. Therefore, it might be useful to determine whether, they could be dropped from the model by examining the significance analysis of the components.

The significance analysis of components is shown in Table 2. The table indicates that the slope and

Table 1. Final estimates of free parameters for sugarcane yield.

District	Component	Parameter	Estimate	Approx Std Error	t-value	Approx Pr> |t|
Karnal	Irregular	Error Variance	23.0512500	5.36078	4.30	<.0001
	Level	Error Variance	0.0010700	0.93647	0.00	0.9991
	Slope	Error Variance	0.0230000	0.02911	0.79	0.4295
Ambala	Irregular	Error Variance	21.4117200	5.14952	4.16	<.0001
	Level	Error Variance	0.5338700	1.04454	0.51	0.6093
	Slope	Error Variance	0.0000001	0.00008	0.00	0.9988
Yamuna	Irregular	Error Variance	30.0004100	8.27450	3.63	0.0003
Nagar	Level	Error Variance	0.0000200	0.01466	0.00	0.9987
	Slope	Error Variance	0.0276700	0.06047	0.46	0.6472
Panipat	Irregular	Error Variance	22.1801100	5.98372	3.71	0.0002
	Level	Error Variance	0.0000100	0.00536	0.00	0.9987
	Slope	Error Variance	0.0255200	0.04356	0.59	0.5581
Kurukshet-	Irregular	Error Variance	33.4632900	7.78006	4.30	<.0001
ra	Level	Error Variance	0.0000002	0.00069	0.00	0.9997
	Slope	Error Variance	0.0000001	0.00003	0.00	0.9987

Table 2. Significance Analysis of components (based on the final state) of sugarcane yield.

District	Component	DF	Chi-Square	Pr > ChiSq
Karnal	Irregular	1	4.37	0.0365
	Level	1	1097.24	<.0001
	Slope	1	9.07	0.0026
Ambala	Irregular	1	0.07	0.7875
	Level	1	1243.98	<.0001
	Slope	1	34.42	<.0001
Yamuna Nagar	Irregular	1	11.30	0.0008
	Level	1	619.13	<.0001
	Slope	1	0.92	0.3377
Panipat	Irregular	1	1.11	0.2927
	Level	1	1063.50	<.0001
	Slope	1	8.49	0.0036
Kurukshetra	Irregular	1	0.02	0.8844
	Level	1	1676.72	<.0001
	Slope	1	121.84	<.0001

Table 3. Trend information (based on the final state) for sugarcane yield.

Component	District									
	Karnal		Ambala		Yamuna Nagar		Panipat		Kurukshetra	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Level	75.03	2.26	67.73	1.92	63.75	2.56	73.61	2.26	74.42	1.82
Slope	1.29	0.43	0.75	0.13	0.456	0.47	1.29	0.44	0.91	0.08

Table 4. Goodness of Fit criterion values for LLTM and LLM based on residuals and likelihood.

Criterion	LLT Model					LL Model				
	Karnal	Ambala	Yamuna Nagar	Panipat	Kurukshetra	Karnal	Ambala	Yamuna Nagar	Panipat	Kurukshetra
MSE	38.22	36.52	43.92	39.38	53.47	39.09	36.52	43.89	40.16	53.47
RMSE	6.18	6.04	6.63	6.27	7.312	6.25	6.04	6.62	6.34	7.31
MAPE	9.02	10.34	9.62	8.63	10.27	9.30	10.34	9.61	8.71	10.27
AIC	279.1	275.19	198.88	190.94	253.06	278.29	273.19	197	189.8	251.06

Table 5. Post-sample prediction performance of UCMs for sugarcane yield.

District	Year	Actual yield (kg/ha)	Forecast yield (kg/ha)	Relative Deviation (%)
Karnal	2011-12	78.38	75.78	3.32
	2012-13	81.60	76.49	6.26
	2013-14	78.81	77.21	2.03
	2014-15	85.04	77.93	8.36
	2015-16	85.04	78.65	7.51
	2016-17	95.00	79.37	16.45
Ambala	2011-12	71.58	68.49	4.31
	2012-13	79.68	69.25	13.09
	2013-14	71.23	70.00	1.73
	2014-15	70.55	70.75	-0.29
	2015-16	70.55	71.51	-1.36
	2016-17	78.13	72.26	7.51
Yamunanagar	2011-12	66.02	66.68	-1.00
	2012-13	74.01	67.40	8.93
	2013-14	68.66	68.12	0.79
	2014-15	69.90	68.84	1.52
	2015-16	69.90	69.56	0.49
	2016-17	79.66	70.28	11.78
Panipat	2011-12	83.72	74.91	10.52
	2012-13	92.39	76.20	17.52
	2013-14	76.91	77.50	-0.76
	2014-15	83.56	78.79	5.71
	2015-16	83.55	80.08	4.15
	2016-17	83.30	81.38	2.31
Kurukshetra	2011-12	69.93	75.33	-7.73
	2012-13	77.09	76.24	1.10
	2013-14	75.47	77.15	-2.23
	2014-15	81.64	78.06	4.39
	2015-16	81.64	78.97	3.27
	2016-17	85.57	79.88	6.65

level components are significant for all districts except for Yamunanagar where slope component is non-significant. Thus, slope and level components cannot be dropped from the model but could be made deterministic by holding the value of its variance fixed at zero. The contribution of the irregular component is found to be non-significant for all districts except Karnal. But it being a stochastic component, cannot be dropped from the model. Thus, fixing the slope variance at zero, the free parameters were again obtained (Table 3).

Fit statistics based on residuals and likelihood are presented in Table 4. For Karnal district, by considering the model with all components i.e., irregular, slope and level, the AIC value came out to be 279.1. When the slope component was taken as constant because of its variance being approximately zero then for the modified form of model (random walk model/LLM), the AIC value was found to be 278.29. It indicated a relatively better fit model. Further, we also considered both the level and slope as constant components, but the AIC value jumped to 282.47 making it a poor fit model. Hence, the random walk model having only slope as constant component is found to be the best fit model for Karnal district. Using AIC values, similar results were also obtained for the remaining four districts. After fixing the slope component; the MSE, RMSE and MAPE values obtained are shown in Table 4. Based on AIC values, LLM was found better than LLTM for sugarcane yield prediction of Karnal, Ambala, Yamunanagar and Kurukshetra districts however, the LLTM was found to be better for Panipat district. The post-sample prediction(s) using the best fit models have been given in Table 5 along with the prediction error(s) for the period 2011-12 to 2016-17. The MAPE for Karnal, Ambala, Yamunanagar, Panipat and Kurukshetra were found to be 6.61, 3.63, 3.10, 5.96 and 3.31 respectively. MAPE values indicate that the local linear trend model has

relatively good post-sample forecast performance for Yamunanagar, Ambala and Kurukshetra districts. Sugarcane yield prediction in Haryana was also studied by Suman and Verma (2017) using ARIMA and state space models, however in terms of percent relative deviation, the unobserved component model (UCM) outperformed ARIMA and is found out to be at par with the state space models. Also, unobserved component model (UCM) provides an easy alternative to the state space models and is capable of modeling stationary as well as non-stationary times series.

Conclusion

The LLM was found better than LLTM for sugarcane yield prediction of Karnal, Ambala, Yamunanagar and Kurukshetra districts however, the LLTM was found to be better for Panipat district. The UCM performed well in capturing tolerable percent relative deviations for district-level sugarcane yield forecasts in all time regimes. The developed models are capable of providing the reliable estimates of sugarcane yield well in advance of the crop harvest while on the other hand, the real-time yield estimates from State Department of Agriculture are obtained quite late after the actual harvest of the crop.

REFERENCES

1. Akaike, H. (1976). Canonical correlations analysis of time series and the use of an information criterion in advances and case studies in system identification (R. Mehra and D.G. Lainiotis (Eds.)), Academic Press, New York. https://doi.org/10.1016/S0076-5392(08)60869-3
2. Aoki, M. (1987). State space modelling of time series, Springer, Berlin.
3. Box, G. E. P. and Jenkins, G. M. (1976). Time series analysis: Forecasting and Control, *Holden Day, San Francisco*.
4. Box, G. E. P. and Tiao, G. C. (1975). Intervention Analysis with Applications to Economic and Environmental Problems. *Journal of the American Statistical Association,* 70(349): 70-79.
5. Commandeur, J. J. F. and Koopman, S. (2007). An introduction to state space time series analysis. *Oxford: Oxford University Press*.
6. Durbin, J. and Koopman, S. J. (2002).Time series analysis by state space methods. *Oxford: Oxford University Press*.
7. Harvey, A.C. (2001) Forecasting, Structural Time series models and the Kalman filter. *Cambridge University Press, UK*.
8. Kitagawa, G. and Gersch, W. (1984). A smoothness priors-state space modelling of time series with trend and seasonality, *J. Amer. Statist. Assoc.*, 79: 378 - 389. https://doi.org/10.1080/0162 1459.1984.104 78060
9. Rajarathinam, A., Vetriselvi, R. and Balamurugan, D. (2016). Unobserved component model for forecasting wheat production. *Int. J. Agricult. Stat., Sci*., 12(1): 161-167.
10. Ravichandran, S. and Prajneshu (2000). State space modelling versus ARIMA time series modelling, *Journal of the Indian Society of Agricultural Statistics,* 54 (1), 43-51.
11. Suman and Verma, U. (2017). State space modeling and forecasting of sugarcane yield in Haryana, India. *Journal of Applied and Natural Science*, 9(4): 2036-242. https://doi.org/10.31018/jans.v9i4.1485
12. Suresh, K. K. and Krishna Priya, S. R. (2011) Forecasting sugarcane yield of Tamil Nadu using ARIMA models, *Sugar Tech*, 13(1):23-26. https://doi.org/10.1007/s12355-011-0071-7