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

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 modeling 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 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 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 sources. The state space models, the unknown parameters include the observation and the state disturbance variances, i.e. $\sigma_{\epsilon}^2$, $\sigma_{\eta}^2$ and $\sigma_{\xi}^2$. These parameters are also called the hyper parameters.

If $\sigma_{\eta}^2 = 0$, 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

$$y_t = \mu + \epsilon_t \quad \epsilon_t \sim \text{NID}(0, \sigma_{\epsilon}^2)$$

$$\mu_{t+1} = \mu_t + \xi_t \quad \xi_t \sim \text{NID}(0, \sigma_{\xi}^2)$$

Eq. (3)

If both of the state disturbance variances $\sigma_{\epsilon}^2$ and $\sigma_{\eta}^2$ are zero then model given in equation (1) reduces to the classical regression model. In this case the linear trend models simplifies to

$$y_t = \mu + \nu \cdot g_t \quad \epsilon_t \sim \text{NID}(0, \sigma_{\epsilon}^2)$$

Eq. (4)

For $t = 1, 2, \ldots, n$, where, the predictor variable $g_t = t-1$ for $t = 1, 2, \ldots, n$ is time effective and $\mu$ and $\nu$ are the initial values of the level and slope.

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)

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (y_t - \hat{y}_t)^2}$$

- Mean Absolute Prediction Error (MAPE)

$$\text{MAPE} = \frac{1}{n} \sum_{t=1}^{n} |y_t - \hat{y}_t|$$

- Relative Deviation (%) (RD)

$$\text{RD} = \frac{(\text{Observed} - \text{Forecast})}{\text{Observed}} \times 100$$

- Akaike Information Criteria (AIC)

$$\text{AIC} = \frac{1}{n} \left[ -2n \log(L_{\eta}) + 2(q + \omega) \right]$$

where $\mu_t$ 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, $\epsilon_t$ is the overall error or irregular component at time $t$, which is assumed to be Gaussian white noise with variance $\sigma_{\epsilon}^2$. 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:

$$y_t = \mu_t + \epsilon_t \quad \epsilon_t \sim \text{NID}(0, \sigma_{\epsilon}^2)$$

$$\mu_{t+1} = \mu_t + \nu \cdot g_t \quad \xi_t \sim \text{NID}(0, \sigma_{\xi}^2)$$

Eq. (2)

for $t = 1, 2, \ldots, n$. This model contains two state equations, one each for modeling the level, and the slope. The stochastic slope $\nu_t$ in equation (2) is equivalent to regression coefficient in classical regression model and $\mu_t$ is the unobserved level at time $t$ which is equivalent to the intercept in the classical regression model. $\epsilon_t$ is the observation disturbance at time $t$, $\xi_t$ and $\eta_t$ 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. $\sigma_{\epsilon}^2$, $\sigma_{\eta}^2$ and $\sigma_{\xi}^2$. These parameters are also called the hyper parameters.

If $\sigma_{\eta}^2 = 0$, the model in (2) have stochastic level and deterministic slope and is known as Local Linear Model (LLM) or the random walk model.

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

$$y_t = \mu + s_t + c_t + \epsilon_t$$

Eq. (1)

where $\mu_t$ 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, $\epsilon_t$ is the overall error or irregular component at time $t$, which is assumed to be Gaussian white noise with variance $\sigma_{\epsilon}^2$. 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:

$$y_t = \mu_t + \epsilon_t \quad \epsilon_t \sim \text{NID}(0, \sigma_{\epsilon}^2)$$

$$\mu_{t+1} = \mu_t + \nu \cdot g_t \quad \xi_t \sim \text{NID}(0, \sigma_{\xi}^2)$$

Eq. (2)

for $t = 1, 2, \ldots, n$. This model contains two state equations, one each for modeling the level, and the slope. The stochastic slope $\nu_t$ in equation (2) is equivalent to regression coefficient in classical regression model and $\mu_t$ is the unobserved level at time $t$ which is equivalent to the intercept in the classical regression model. $\epsilon_t$ is the observation disturbance at time $t$, $\xi_t$ and $\eta_t$ 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. $\sigma_{\epsilon}^2$, $\sigma_{\eta}^2$ and $\sigma_{\xi}^2$. These parameters are also called the hyper parameters.

If $\sigma_{\eta}^2 = 0$, 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

$$y_t = \mu + \epsilon_t \quad \epsilon_t \sim \text{NID}(0, \sigma_{\epsilon}^2)$$

$$\mu_{t+1} = \mu_t + \nu \cdot g_t \quad \xi_t \sim \text{NID}(0, \sigma_{\xi}^2)$$

Eq. (3)

If both of the state disturbance variances $\sigma_{\epsilon}^2$ and $\sigma_{\eta}^2$ are zero then model given in equation (1) reduces to the classical regression model. In this case the linear trend models simplifies to

$$y_t = \mu + \nu \cdot g_t \quad \epsilon_t \sim \text{NID}(0, \sigma_{\epsilon}^2)$$

Eq. (4)

for $t = 1, 2, \ldots, n$, where, the predictor variable $g_t = t-1$ for $t = 1, 2, \ldots, n$ is time effective and $\mu$ and $\nu$ are the initial values of the level and slope.
Where $y_t$ is the actual or observed value and $\hat{y}_t$ is predicted/forecast value, $n$ is the number of observations in the time series, $\log(L_n)$ 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 Table 1. 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

| District  | Component | Parameter    | Estimate   | Approx. Std Error | $t$-value | $p$-value |
|-----------|-----------|--------------|------------|-------------------|-----------|-----------|
| Karnal    | Irregular | Error Variance | 23.05125000 | 5.36078          | 4.30      | <0.0001   |
|           | Level     | Error Variance | 0.00107000 | 0.93647          | 0.00      | 0.9991    |
|           | Slope     | Error Variance | 0.02300000 | 0.02911          | 0.79      | 0.4295    |
| Ambala    | Irregular | Error Variance | 21.41172000 | 5.14952          | 4.16      | <0.0001   |
|           | Level     | Error Variance | 0.53367000 | 1.04454          | 0.51      | 0.6093    |
|           | Slope     | Error Variance | 0.00000001 | 0.00008          | 0.00      | 0.9988    |
| Yamuna Nagar | Irregular | Error Variance | 30.00041000 | 8.27450          | 3.63      | 0.0003    |
|           | Level     | Error Variance | 0.00002000 | 0.01466          | 0.00      | 0.9987    |
|           | Slope     | Error Variance | 0.02767000 | 0.06047          | 0.46      | 0.6472    |
| Panipat   | Irregular | Error Variance | 22.18011000 | 5.98372          | 3.71      | 0.0002    |
|           | Level     | Error Variance | 0.00001000 | 0.00536          | 0.00      | 0.9987    |
|           | Slope     | Error Variance | 0.02552000 | 0.04356          | 0.59      | 0.5581    |
| Kurukshetra | Irregular | Error Variance | 33.46329000 | 7.78006          | 4.30      | <0.0001   |
|           | Level     | Error Variance | 0.00000020 | 0.00069          | 0.00      | 0.9997    |
|           | Slope     | Error Variance | 0.00000010 | 0.00003          | 0.00      | 0.9987    |

| District  | Component | DF | Chi-Square | $p$ > ChiSq |
|-----------|-----------|----|------------|-------------|
| Karnal    | Irregular | 1  | 4.37       | 0.0365      |
|           | Level     | 1  | 1097.24    | <0.0001     |
|           | Slope     | 1  | 9.07       | 0.0026      |
| Ambala    | Irregular | 1  | 0.07       | 0.7875      |
|           | Level     | 1  | 1243.98    | <0.0001     |
|           | Slope     | 1  | 34.42      | <0.0001     |
| Yamuna Nagar | Irregular | 1  | 11.30      | 0.0008      |
|           | Level     | 1  | 619.13     | <0.0001     |
|           | Slope     | 1  | 0.92       | 0.3377      |
| Panipat   | Irregular | 1  | 1.11       | 0.2927      |
|           | Level     | 1  | 1063.50    | <0.0001     |
|           | Slope     | 1  | 8.49       | 0.0036      |
| Kurukshetra | Irregular | 1  | 0.02       | 0.8844      |
|           | Level     | 1  | 1676.72    | <0.0001     |
|           | Slope     | 1  | 121.84     | <0.0001     |

| Component | District | Karnal Estimate | Karnal SE | Ambala Estimate | Ambala SE | Yamuna Nagar Estimate | Yamuna Nagar SE | Panipat Estimate | Panipat SE | Kurukshetra Estimate | Kurukshetra SE |
|-----------|----------|-----------------|-----------|-----------------|-----------|----------------------|----------------|-------------------|-------------|----------------------|----------------|
| Level     |          | 75.03           | 1.29      | 76.73           | 0.75      | 63.75                | 0.13           | 73.61             | 0.456       | 74.42                | 0.47           |
| Slope     |          | 22.26           | 0.43      | 1.92            | 0.75      | 2.56                 | 0.13           | 2.26              | 0.47        | 1.29                 | 0.44           |
Table 4 presents the Goodness of Fit criterion values for LLTM and LLM based on residuals and likelihood.

The table shows the MSE, RMSE, and MAPE values for each district, along with the AIC values. The data is presented in a tabular format with columns for district, year, actual yield, forecast yield, and relative deviation. For example, in the Karnal district for the year 2011-12, the actual yield was 79.68 kg/ha, the forecast yield was 72.18 kg/ha, and the relative deviation was 13.09%.

Table 5 provides the post-sample prediction performance of UCMs for sugarcane yield. It includes columns for district, year, actual yield, forecast yield, and relative deviation. The table shows the performance of the model for the years 2011-2016, with the actual and forecasted yields for each year.

The text discusses the results of the analysis, focusing on the significance of level and slope components across different districts. For Yamunanagar, the slope component is not significant. For all other districts, the level and slope components are significant. The AIC values for different models are presented, with the LLT model having lower values for Karnal and Ambala, and the LLM model for Yamunanagar and Kurukshetra.

The text concludes by noting that the LLT model is a better fit for Karnal, Ambala, and Yamunanagar, while the LLM model is preferred for Kurukshetra. The results are supported by the goodness of fit criterion values, with the LLT model consistently providing lower AIC values across all districts.
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 sugar-cane 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/10.1080/01621459.1984.10478060
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 Praneshu (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