

2012

Forecasting fruit demand: Intelligent Procurement

FCAS Final Project Report

Predict fruit sales for a 2 day horizon to efficiently manage procurement logistics.

Dinesh Ganti(61310071) Rachna Lalwani(61310845), Ravi Shankar(61310210),
Shouri Kamtala(61310215), Supreet Kaur(61310595)
Section A Group 2
12/26/2012

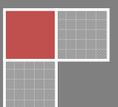


Table of Contents

- Executive Summary** 2
 - Problem description/Business goal 2
 - Forecasting goal..... 2
 - Data 2
 - Forecasting Methods and Performance Metrics..... 2
 - Conclusions/Recommendations..... 2
- Major Stakeholders and Benefits**..... 3
- Technical Summary**..... 3
 - Apple red delicious 3
 - Pineapples 4
 - Bananas 5
 - Watermelons 5
 - Pacham Pear..... 6
- Conclusions** 7
- Appendix** 8

EXECUTIVE SUMMARY

Problem Description/Business Goal - As the fruit supplier to the hypermarket, we wish to match our procurement with the fruit demand. This is important owing to the perishable nature of fruits. This reduces over/under stocking. Also, by matching the fruit demand accurately, we can provide value-add to the hypermarket and stay ahead of the competition.

Forecasting goal - To forecast the demand for five chosen fruit SKUs over a forecast period of 2 days; the chosen SKUs are Pineapple cuts (Kg) (Mb), Apple red delicious, Watermelon striped, Packham pear and Premium banana. The criterion for fruit selection was: High volume of transactions.

Data - Some main features to note within the data file are: (a) it has only 13 months of data which meant that any annual seasonality or monthly predictions were out of question (b) it had no information about the reason for zero demand (no demand versus stock out). Quantity sold for each SKU was aggregated at the daily level. The missing values were filled with seasonal naïve or zeros on a case-by-case basis depending on whether it was a stock out or zero demand. Please refer to exhibits for charts showing actual and predicted values for each SKU.

Forecasting methods and performance metrics - The primary performance metric used to evaluate the performance of different methods was the combined cost of under-stocking + over-stocking over validation period. As the fruit vendor, the cost of over stocking was wastage of items and was equal to the cost of the fruit. The cost of under stocking was lost sales and lost reputation with hypermarket. To capture this difference, we considered a 3:1 weight-age which means: $\text{Cost (Under Stocking)} = 3 * \text{Cost (Over Stocking)}$. In all the cases the seasonal naïve was our bench mark. Below is a consolidated summary of all the methods tried and the final performance as per the chosen cost metric:

	Holt Winters & AR	Naïve	Linear Regression	Trailing MA & AR	Holt Winters & AR	Double differencing & AR
PineApple	267.00	3,124.00				
Apple		1,060.00	654.00			
Pear		13,512.00	6,475.00		4,878.00	12,233.00
Watermelon		946.00		315.00		
Banana	601.00	75.00				

Conclusions/Recommendations – Different models work best for different fruits. Replicating the forecasting exercise for each SKU when there are 100s of them would be a costly affair. Automated data-driven methods are preferable. Data quality and consistency must be ensured for data driven models to work. One should also keep in mind the fact that data from the hypermarket will be available to the vendor only after a certain lag of at least a day. When predicting for a very short period, it is not possible to come up with prediction intervals around forecasts.

MAJOR STAKEHOLDERS AND BENEFITS

Fruit Supplier

- Efficient Procurement Management
- Reduce in under/over stocking
- Increase in revenues

Hypermarket

- Reduce in wastage
- Efficient inventory management

TECHNICAL SUMMARY

General approach across all SKUs is to first capture trend and seasonality through double differencing or regression or Holt-Winters and then captures any remaining signal in the model through AR.

Apples Red Delicious

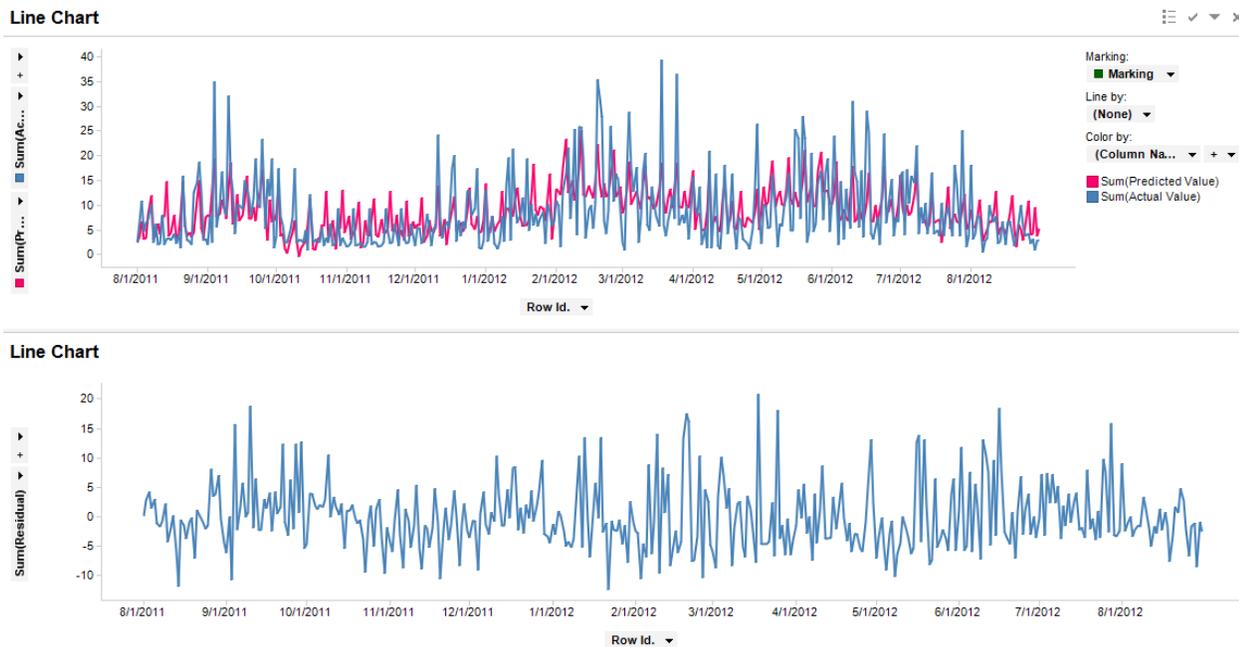
There are 5 apple varieties that were in reasonable demand over the year. Apples red delicious had a greater demand over the year and was therefore chosen for forecasting.

The plot of demand of apples red delicious Vs time revealed a weekly trend with demand peaking on Saturdays and Sundays, with a minor peak on Wednesday and almost constant sales on the rest of the weekdays.

The demand also showed a yearly cycle with the months of Jan, Apr & Dec having similar demands, Mar, May & Sep behaving similarly and the months of Feb and Oct each having a different demand value. Accordingly 3 dummy variables were chosen for weekly cycle. Similarly, 4 dummy variables representing the yearly cycle were also chosen.

Please refer to the plots of the Actual versus the Predicted daily demand and their residuals (**Exhibit 1A**) and ACF plot of the residuals (**Exhibit 1B**).

Exhibit 1A Plot of actual Vs predicted daily demand values for apple red delicious and its residuals



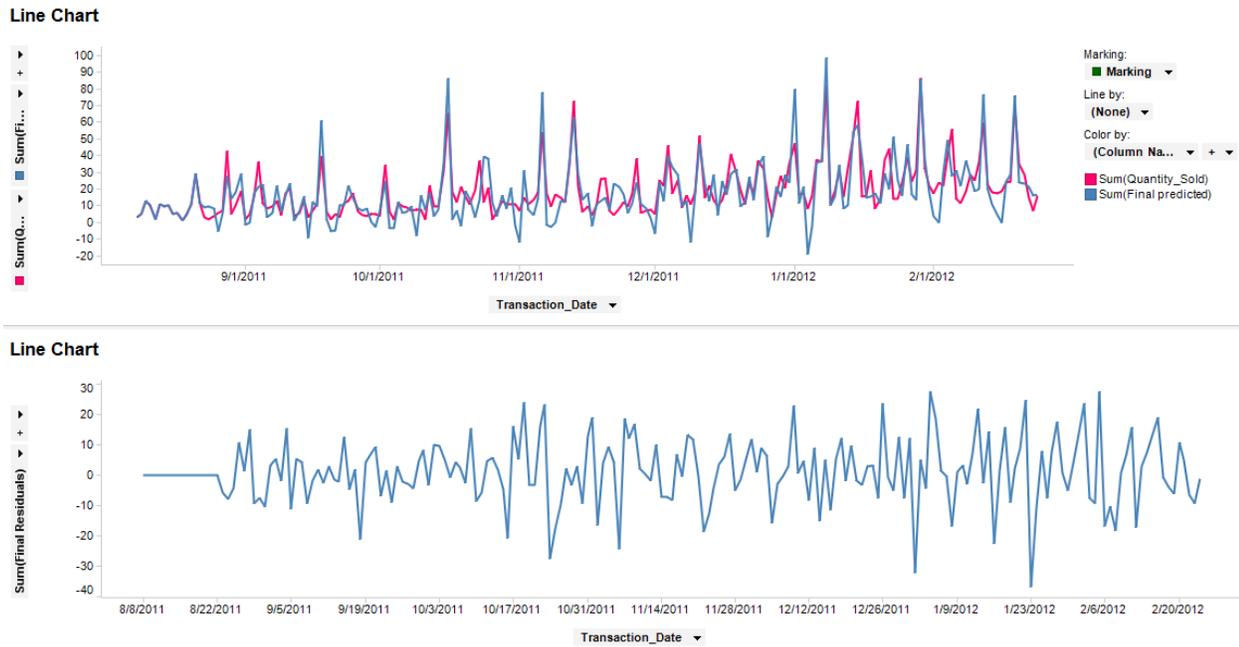
Pineapples

The most suitable SKU within the pineapple family was chosen to be Pine Apple Cuts KG. Initially, the data was explored for the presence of any day of month or day of week trends. We identified that Sunday was a special case of high demand and tried a linear regression with dummy variables. However, the regression failed as it could not capture the weekly seasonality and the high peaks and low troughs. The over forecasting and under forecasting errors were very high. Our next method was Holt Winters. Holt Winters additive suited the data and we tweaked the gamma parameter from the default 0.05 to 0.3 to achieve satisfactory coverage of changing seasonal pattern. No automatic optimization of alpha, beta or gamma was performed to avoid over fitting.

After Holt winters, the residuals were still see to have some seasonality left. See Exhibits 2A, 2B and 2C for the holt-winters performance and the left over residuals. None of the AR methods or linear regression with lags did a good job in capturing the left over signal in residuals. ARIMA worked relatively well to reduce the residuals. So after Holt-winters additive followed by ARIMA, the predicted values were back calculated.

The traditional error metrics on the validation periods chosen were not conclusive but the cost measure favored the chosen model.

Exhibit 2A: Pineapple – Holt-winters + ARIMA – Plots of Actual Vs Forecasted and Errors below



Bananas

Banana is known to be a fruit available throughout the year. In the given data, there was no seasonality evident in the ACF plot even though the time series plot looks like there is some seasonality. Refer to **Exhibit 3A**.

We ran multiple models and finally chose Holt-Winters without trend to model what seems like seasonality from the graph (**Exhibit 3B**). Finally, an ACF was re-run on the residuals to check if there is any seasonality left. Refer to **Exhibit 3C**. After running ARIMA, added the forecasted residual values to forecasted Holt-winters, and plotted the final residuals.

Watermelons

The most suitable SKU within the watermelons family was chosen to be Water Melon Striped. Initially, the data was explored for the presence of seasonality and trend. We ran ACF on the values of daily demand of watermelon striped SKU and found that there is a strong auto correlation in lag-1 (Refer to **Exhibit 4A**). So we used a trailing moving average method with a window size of 2 days (Refer to **Exhibit 4B**). We then plotted ACF for the residuals. The plot showed strong correlation in lag2 (Refer to **Exhibit 4C**). We then took lag-2 error values as a predictor, predicted the error values and added them

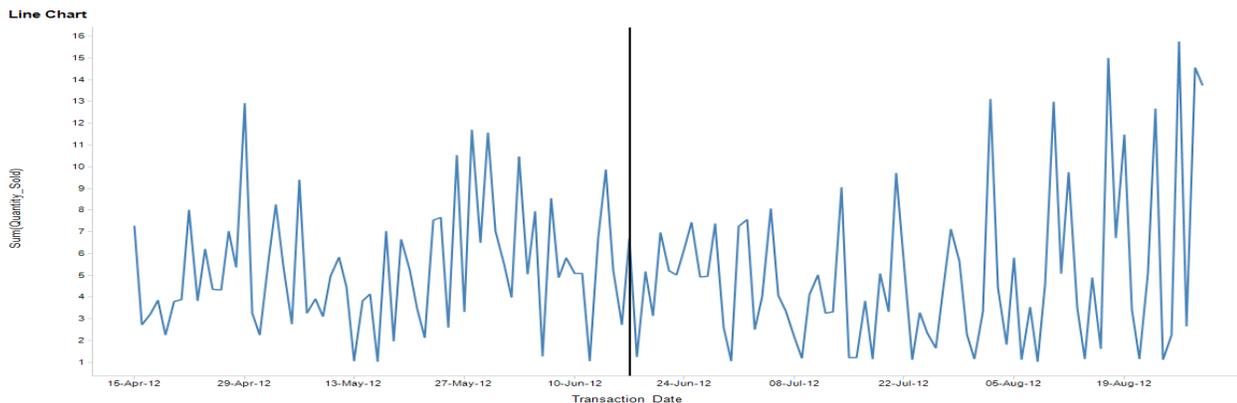
to the forecasted values of the moving average method. Please refer to **Exhibit 4D** for the final forecasted value Vs actual value and residuals after applying this second-level forecasting.

Exhibit 4D: Forecasted Vs Actual values for water melon striped and plot of residuals below



Packham Pear

Imported fruits were one category we chose to look into because of the high volume of transaction. Among these fruits, we chose Packham Pear because it had the highest number of data points. Data before 15th April has been ignored due to too many stock outs. 15th April onwards till 17th June, the data seems to follow a different pattern as compared to data after 17th June. The latter has been picked up for forecasting as that is the most recent. Below diagram shows the split.



Following three methods have been tried:

1. Double differencing (Lag 1 and Lag 7) followed by AR model (Lag 1 and Lag 4)
2. Multiple Linear Regression
3. Holt-Winters multiplicative seasonality (Smoothing parameters = 0.25/0.25/0.55) followed by AR

The cost metric over validation period for each of the methods above along with seasonal naïve has been included in the executive summary. Please refer to the **Exhibits 5** for Actual Vs Forecasted graphs and the forecast errors for each of the above methods.

CONCLUSIONS

We have the following conclusions after the entire modeling exercise:

- **LAG:** There could be a certain lag in getting the data as the demand for a day would be known only at the end of business day, however, we would need to forecast and supply demand for the next day by the end of previous business day
- **Data-driven:** As a fruit vendor, we would be dealing with a lot more variety of fruits than just 5 SKUs. It will be difficult to build manual, model-based models for each SKU. Automatic, data-driven methods are the way to go and for this there must be adequate data collection mechanisms
- **Adjustments:** Care should be taken that forecasts once given out shouldn't be adjusted to suit personal agendas. A record of the forecasts must be kept so that model performance can be evaluated over time
- **Review:** The models should be periodically reviewed and revised. We suggest that models be revised every 2 to 3 months.
- **Importance of Naïve:** We are doing better than the naïve for all fruits except banana. For banana, it is possible that external factors have a higher predictive power than the series itself
- **Improvement over Naïve:** Any improvement in terms of cost over the naïve is preferable and considered worthy effort in building a model
- **Prediction intervals** would have been useful but we couldn't come up with the same as the validation period is only 2 days long.

APPENDIX

Exhibit 1B: ACF Plot of apples red delicious residuals after applying multiple linear regression

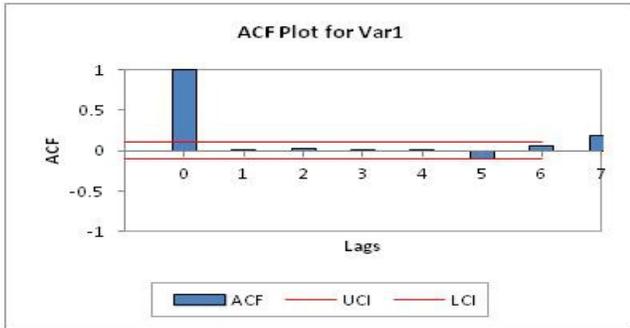


Exhibit 2B: Holt winter model on Pineapple SKU (Layer 1)

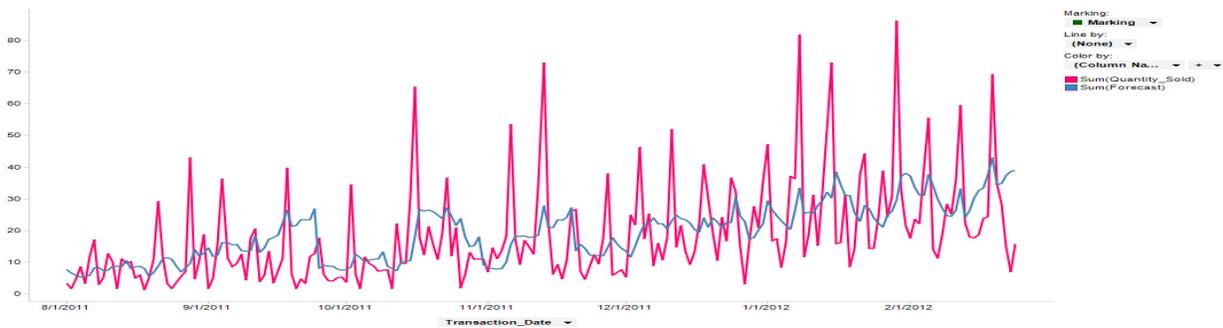


Exhibit 2C: Left over residual on pineapple SKU after Holt-winter additive

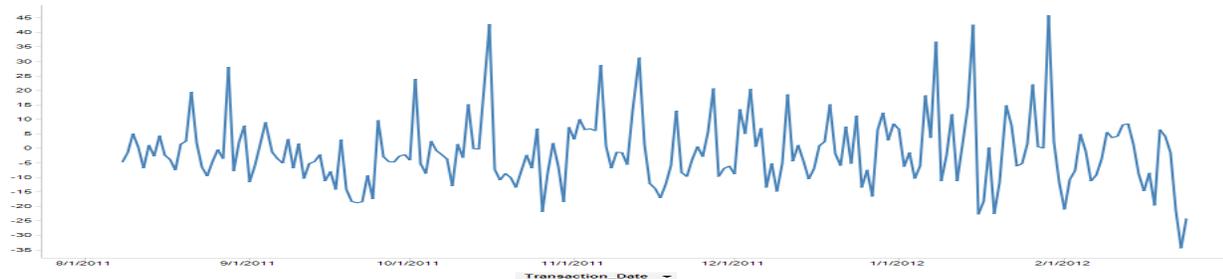


Exhibit 3A: ACF plot on banana series to detect any seasonality presence

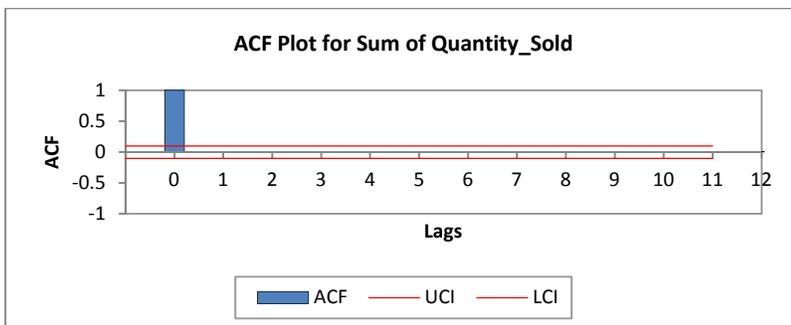


Exhibit 3B: Bananas - Holt-winters – Plots of Actual Vs Forecasted and Errors below



Exhibit 3C: ACF plot on banana residuals after applying holt-winters no trend model

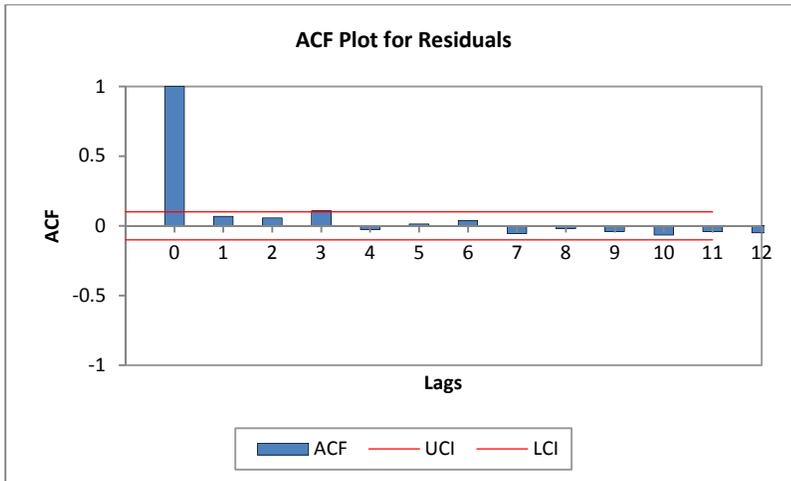


Exhibit 4A: ACF plot of quantity of watermelon striped sold on daily basis

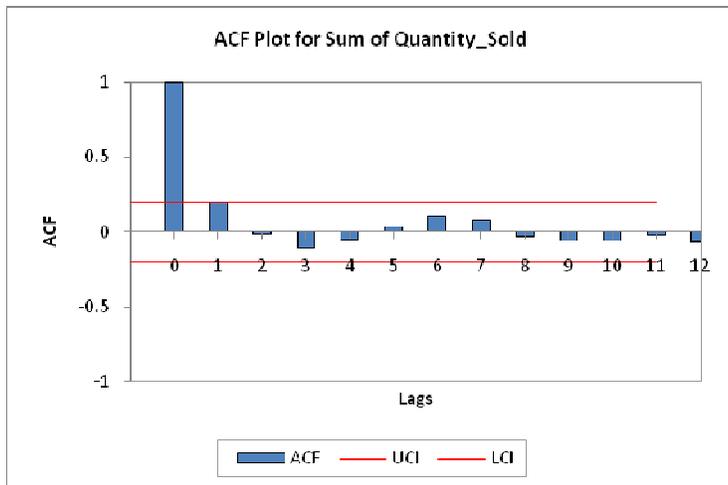


Exhibit 4B: Moving average forecast of watermelons striped sold on daily basis

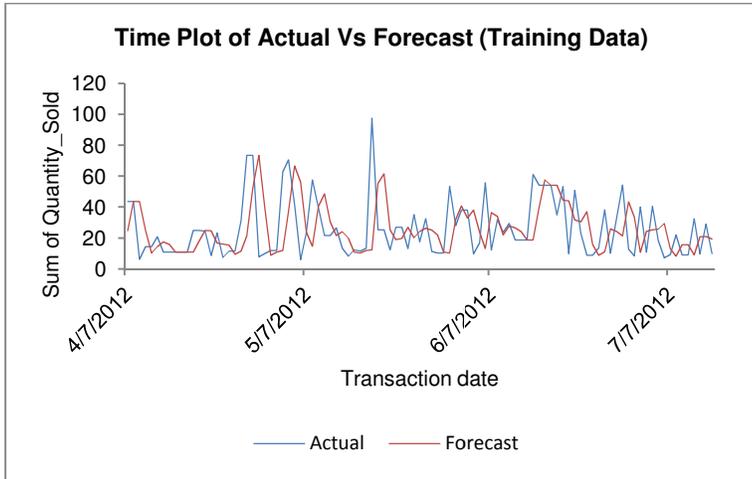


Exhibit 4C: Plot of residuals of Moving Average Forecast

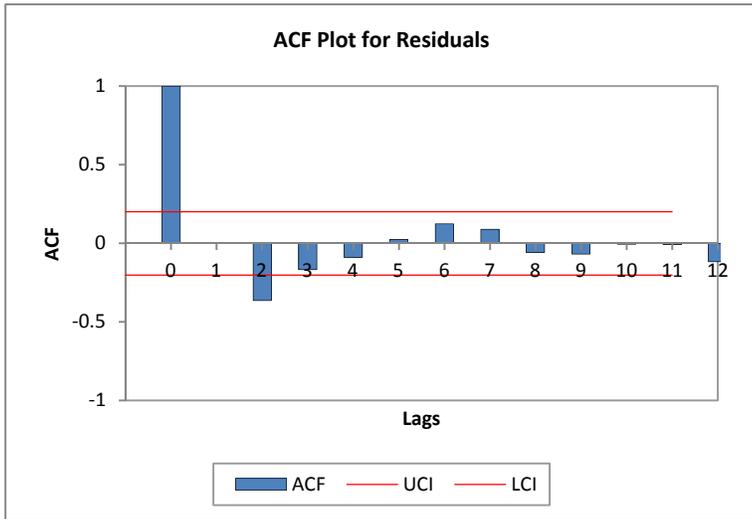
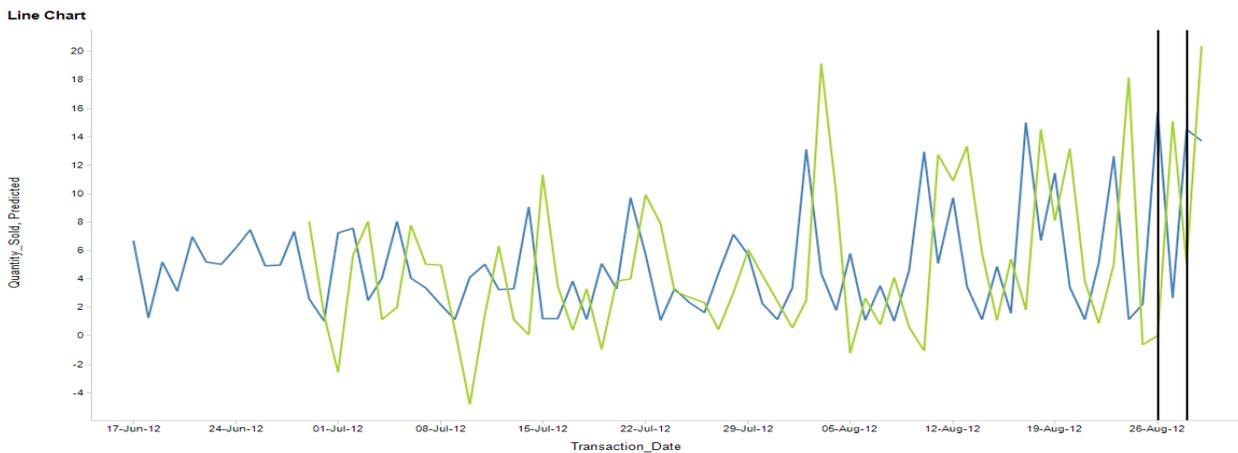


Exhibit 5A: Packham Pear - Double differencing + AR model – Actual Vs Forecast & Residuals (below)



(The shorter green series is the predicted series)



Exhibit 5B: Packham Pear - Regression – Actual (Pink) Vs Forecast & Residuals (below)

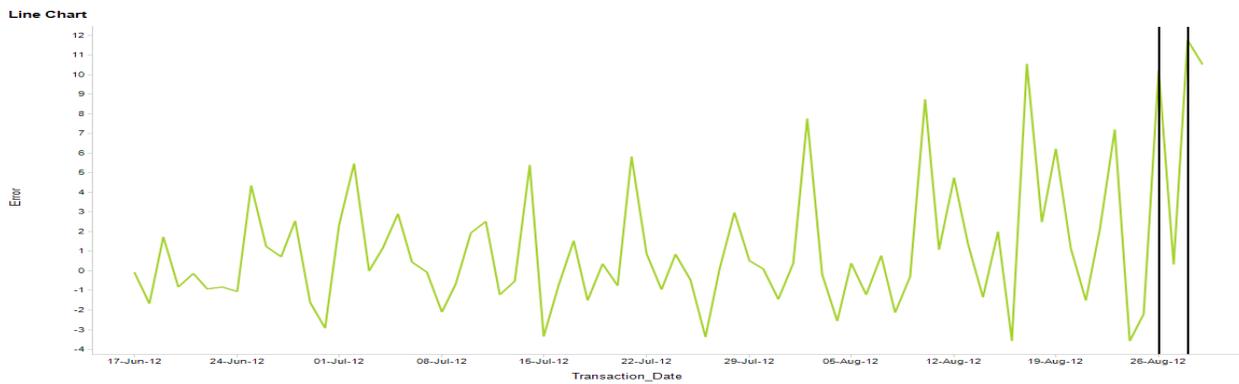
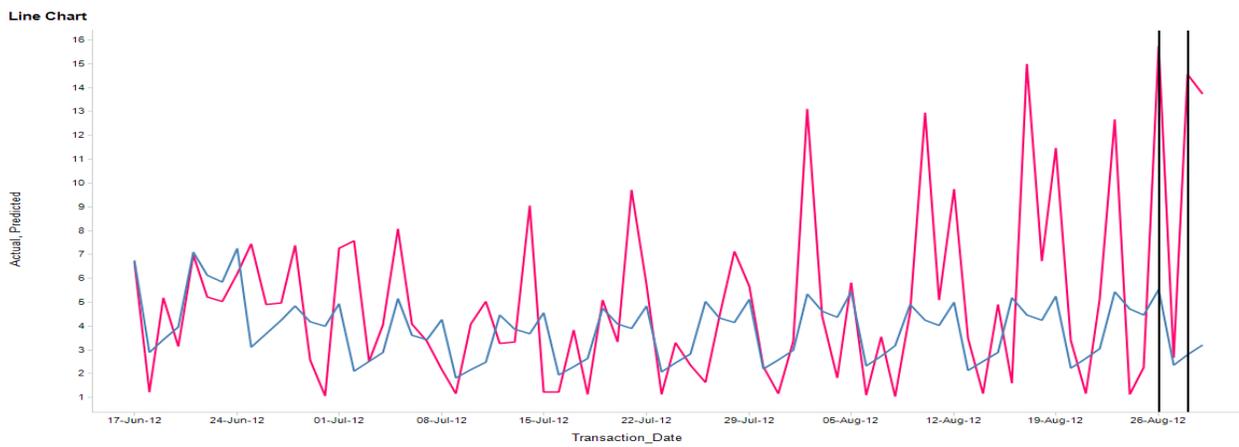


Exhibit 5C: Packham Pear - Holt-Winters + AR – Actual (Pink) Vs Forecast & Residuals (below)

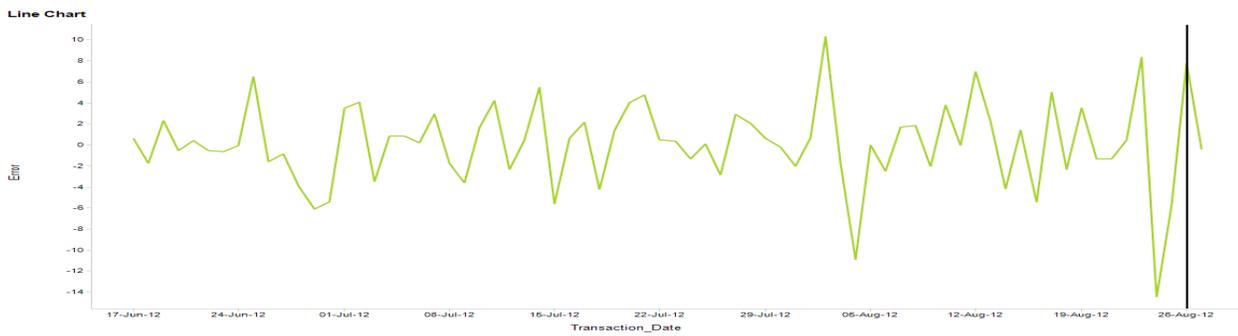
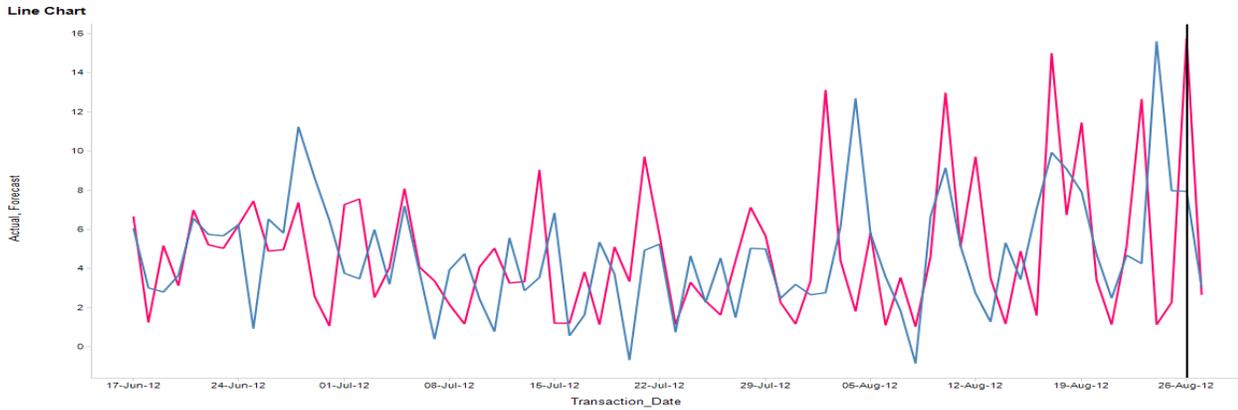


Exhibit 5D: Seasonal naïve – Actual (Pink) Vs Forecast & Residuals (below)

