**Problem Statement**

PLE has collected large amounts of data that serve a multitude of purposes. They want this data to support and guide their operations for the foreseeable future. The principal questions under discussion here are determining which aspects of PLE's operations are important to their per-user market share and their customer satisfaction metrics, and the direction in which their various markets are expected to go. Each of these will be dealt with in turn below.

**Descriptive Analysis**

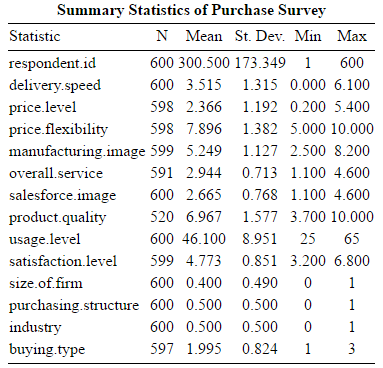
The first step in this process is summarizing the data and verifying its integrity. Most of the survey components have all of their data. Product quality, however, is missing 80 observations, or roughly 13%.

As for data quality, a quick check to see if there were any duplicate respondent ids revealed that there were none. There are some out of range values as well:

Overall service's minimum is .7, which is below the scale minimum of 1.

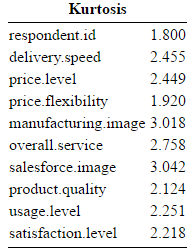
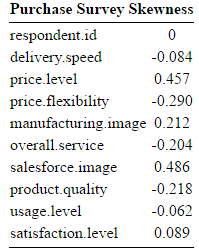
Buying type has a maximum of 4, which is off the scale of 1-3.

Improper values were rendered as NAs and removed for the purpose of further analysis. There were three bad values for buying type, and four for overall service. The table below is the corrected table without these bad values.



The tables below are for the survey skewness and kurtosis. Standard skewness is 0 and standard kurtosis is 3 (for the normal distribution). The levels of skew are all low enough to be negligible. Price flexibility has low-moderate kurtosis, as does product quality. The others are within acceptable bounds. Assumptions of normality are therefore assumed to not be violated.

Note that for categorical variables, namely purchasing structure, industry, buying type, and size of firm, there is no point in measuring deviation from normality for these, as they are, by definition, not going to have normal distributions.



Examinations of the histograms of the various survey responses yielded no surprises that were not revealed by the tables above, except that usage level has more of a uniform distribution, but since this is a dependent variable, it is not of much concern.



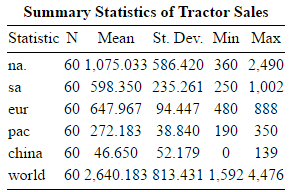
The boxplots reveal that manufacturing image and salesforce image have possible outliers. During the regression analysis, the Cook's distance measure will be used to determine if these values had a problematic influence.



The 13% missing data from product quality is worth individual attention. It needs to be determined if the data are missing at random (MAR), not missing at random (NMAR), or not completely missing at random (NCMAR). A logistic regression was done (missing values were imputed as column means for this only, as if they were omitted the logistic regression would not account for them and that defeats the purpose). Missing data were classified as either a 0 or 1 (1=missing) and then the regression was performed. Specifics follow, but **the effect was determined to be NMAR, yet negligible.**

The chi square derived p-value for this model satisfies a 90% confidence interval but not 95%, so it has some validity. It was determined that several variables may be partially responsible for the missing data, namely usage level, salesforce image, and buying type, when buying type is 2 or 3. However, the usage level only has a very slight push towards data being missing. Buying types of 2 or 3 are more likely to push the data to be present. A positive sales force image also tends to push the data to be present.

Having properly analyzed the survey data, the tractor data is up next. Since it is not being put through regression, the same level of treatment does not need to be undertaken. That being said, all observations are present, and the means and standard deviations do not seem suspect from a data quality perspective, nor do the mins and maxes.



**Regression**

Checking regression assumptions is an important process, and considerable attention was paid to this matter. Below are four plots for the **usage level** regression. The plot of residuals looks good, there isn't much shift off the line so the residuals have an even distribution. The QQ Norm plot is likewise acceptable, the tails being off is not of concern as most real data has this. This means the errors are approximately normally distributed. There is, however, some heteroskedasticity as shown by the third plot.

In the residuals vs leverage plot, we see that while two points towards the end do pull the relationship off, they do not have an exceedingly large Cook's distance, and so do not affect the data disproportionately. These are most likely the few outliers noted earlier in the initial data exploration. Aside from a bit of heteroskedasticity, **the assumptions for this regression (usage level) appear to be in order**.

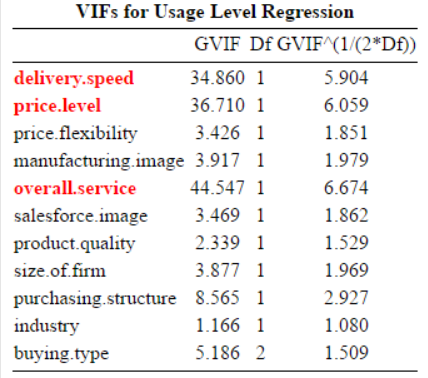
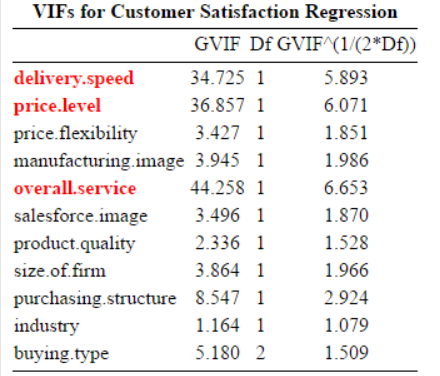


The next four graphs are the same ones for the **customer satisfaction** regression. There is some minor concern with the residuals vs fitted values, specifically the pattern, but it is not pronounced enough of an effect to affect the integrity of the results. The QQ Norm is acceptable and the scale-location is likewise fine. Curiously, it behaves differently in this as opposed to the other regression. Nothing in residuals vs leverage indicates that any one point is altering the regression line too much. **Ergo, the assumptions have not been violated and the analysis can continue.**

As for why the scale-location graph bears such a difference from the first graph, this arises from a different pattern of errors in the fitted values. This means that the usage level error increases as the fitted value goes up, while it remains roughly the same for customer satisfaction. This implies that higher levels of PLE usage do not follow the same pattern as low ones, but the effect is very slight. **The regression tends to underestimate the usage level, and likewise underestimate customer satisfaction at either end of the spectrum.**



Below are the VIF tables for testing for multicollinearity between the variables. Variables above the cutoff of 3 are noted in bold red (corresponding to an r squared of about seventy with the other variables), using the rightmost column. If the VIFs were above ten, then the regression results would be highly suspect.

An examination of the correlation matrix shows which variables these three are highly correlated with, and the following insights hold for both regressions, whether it be for PLE usage level or customer satisfaction.

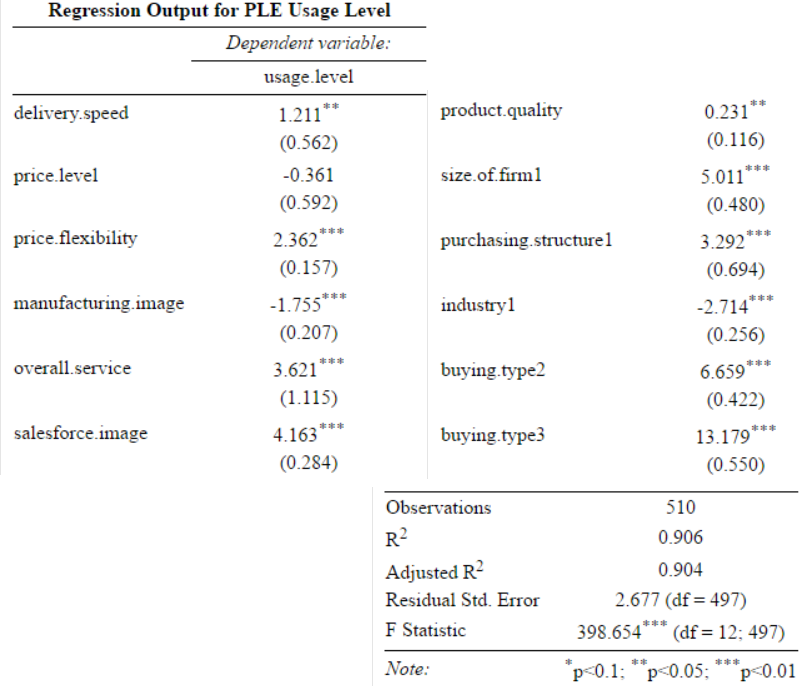
Delivery speed is highly positively correlated with buying type and highly negatively correlated with purchasing structure (both over or close to .7). Purchasing structure makes sense since it is easier to get product to the customer faster with more centralized operations, whereas just like with all things, as it becomes more diffuse, the process is inherently harder to control and maintain. Buying type likewise makes sense, since it is probable that the various levels of used equipment are stored in different manners and such.

Price level is highly positively correlated with purchasing structure. This could potentially mean that the more centralized operation means that a client is purchasing more of PLEs products and required access to a larger PLE distribution center.

Overall service, as one may expect, has a spate of middling strength correlations (.5-.6 range, all the large ones are notably positive correlations). This makes sense as a factor such as overall service is likely to be capturing many effects that these other variables nail down more specifically.

From the scatterplots, it is worth noting that the relationship between satisfaction level and either product quality or salesforce image appear to be curvilinear. This may affect the results. None of the scatterplots for PLE usage level appear to exhibit this behavior.

Having handled the various necessary tasks for showing that the regression is a valid model and useful, the actual outputs for the two are given below. The first is PLE usage level, and the second is for Customer Satisfaction.



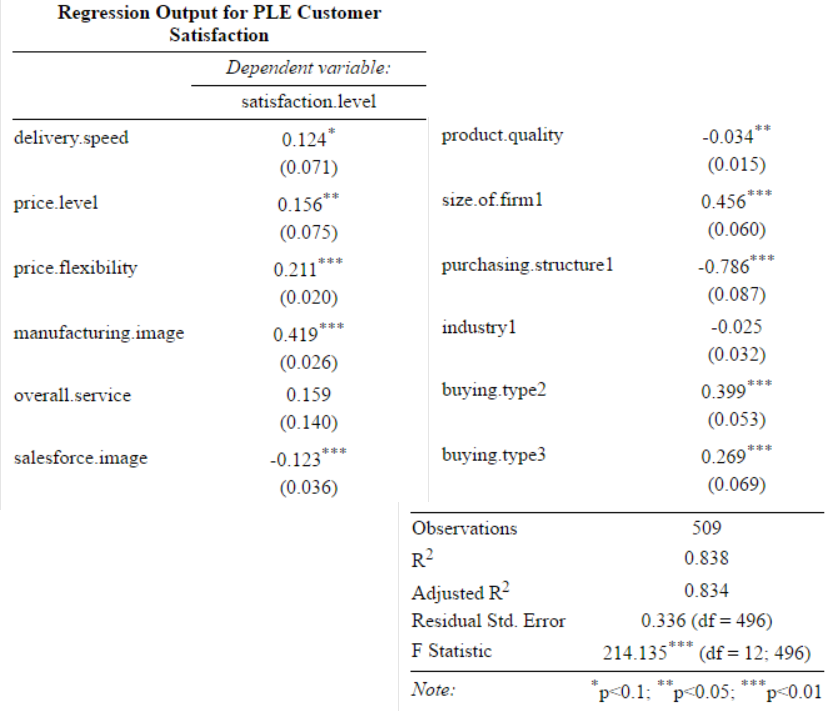
As seen above, many of the variables are significant at the chosen level for determining the amount of PLE products a customer will use. Curiously enough, the price level is not significant. One would tend to assume that the value ratio one thought they were getting for a product would influence how much of said product they buy, but that, from this alone, appears to not be the case. This is further supported by the relative insignificance of product quality compared to the other variables, but this may be at least slightly due to the curvilinear nature of the relationship between this variable and usage level.

The overall model explains roughly 90% of the variation in usage level.

Which buying type is used heavily alters the amount of PLE product the customer uses, even between level two or three. Using buying type 3, on average, results in 13% more product usage holding all other variables equal.

Larger firms have an average of 5% more PLE usage than others, and a centralized purchasing structure confers an additional 3.2%.

Resellers have a lower PLE percentage than non-reseller customers, but this is to be expected considering they are probably dealing with many different product types and companies. Salesforce image drives the effect up a fair amount, this either implies that that companies with more PLE use are more familiar with their sales staff and thus rate them better, or that a better sales staff develops better relationships and gets customers to use more of PLE.



An initial comparison of the F statistics and R squared values for the two models demonstrates that the data explain PLE usage level better than they do customer satisfaction. That being said, the model is still significant and useful. The following important variable changes occurred:

• Price level is now significant

• Manufacturing image now has a positive effect instead of negative

• Overall service is no longer significant

• Sales force image now has a much smaller, and negative, effect

• Purchasing structure experienced a large increase in effect and changed directions

• industry is no longer important

• buying type no longer has a large effect

Several of these make immediate sense; for instance, how innovative PLE appears to be is going to impact how a customer feels, if they believe they're getting a cutting-edge product, they are liable to feel more satisfied by it. Overall service's lack of significance is probably due to its effects being captured by other more specific measurements.

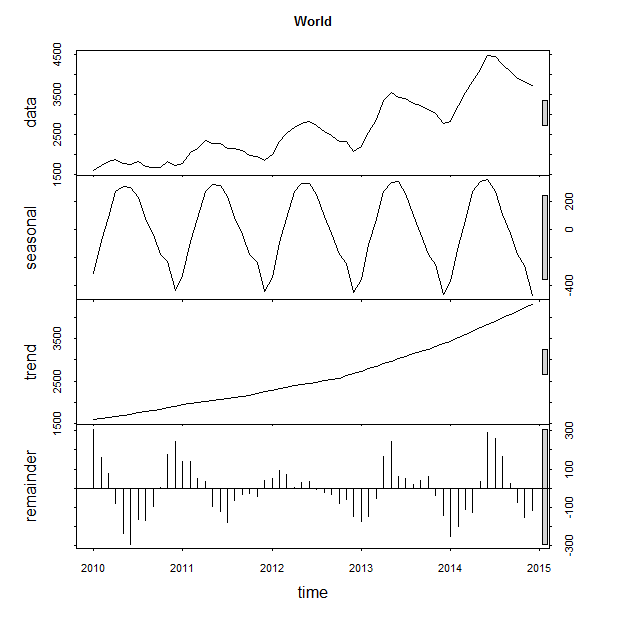
The effect of the sales force image being negative and small is notable, but its low size probably indicates that the survey respondents were thinking of the product itself, and not PLE, when indicating how satisfied they were (this also applies to overall service's lack of significance).

A centralized procurement area appears to negatively impact satisfaction a great deal. The loss of significance for industry is also readily explained, as the product will either perform or not regardless of whether or not they intend to resell. Buying type's huge loss of effect is also partially explained by this.

**Forecasting**



From these graphs, we can see how the various trends for the regions are playing out, and check to see if any of them are multiplicative or seasonal. However, some deeper digging was done on each region and the world to better understand what's going on. Below is an example of the charts generated for this process.



The above graph, as just mentioned, is one of six breakdown graphs generated to understand the tractor sales data before engaging in forecasting. This one represents the world. It shows that there is definitely a strong seasonal trend present in the data, and also shows the general trend of the data itself. The remainder graph shows how much irregular, or unexplained, variation is present in the data. It's good that this amount remains roughly the same over time, and likewise that it is not too high as compared to the actual data itself.

By region, the following effects were observed:

North America: a possibly multiplicative upward trend.

South America: a general upward trend.

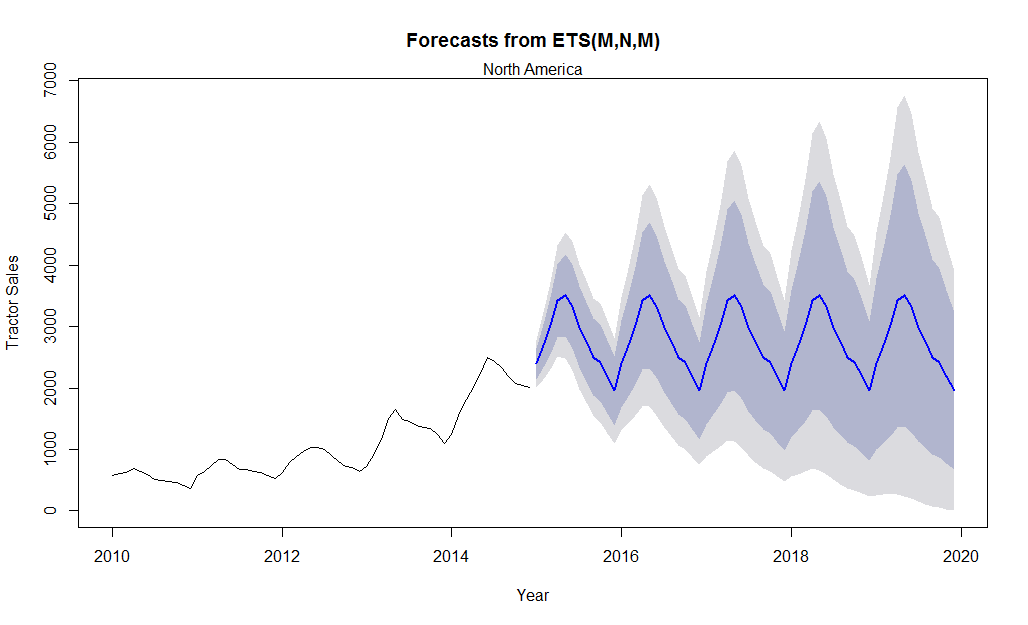
Europe: a downward trend that is levelling off.

Pacific: a more recent downward trend, but a sharper slope.

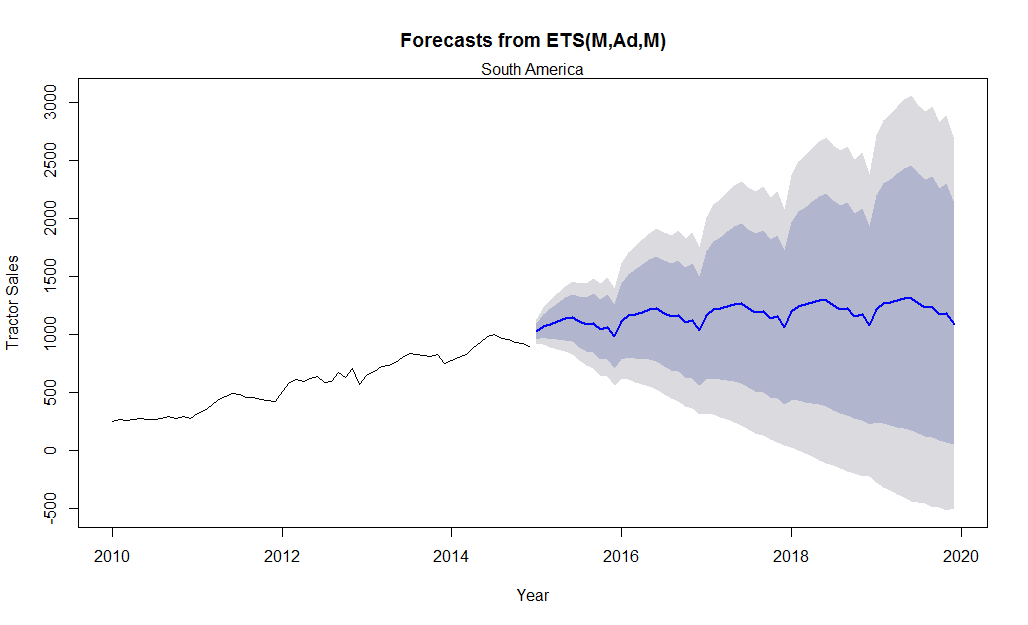
China: an upward trend that is levelling off.

World: a general upward trend that with an increasing slope.

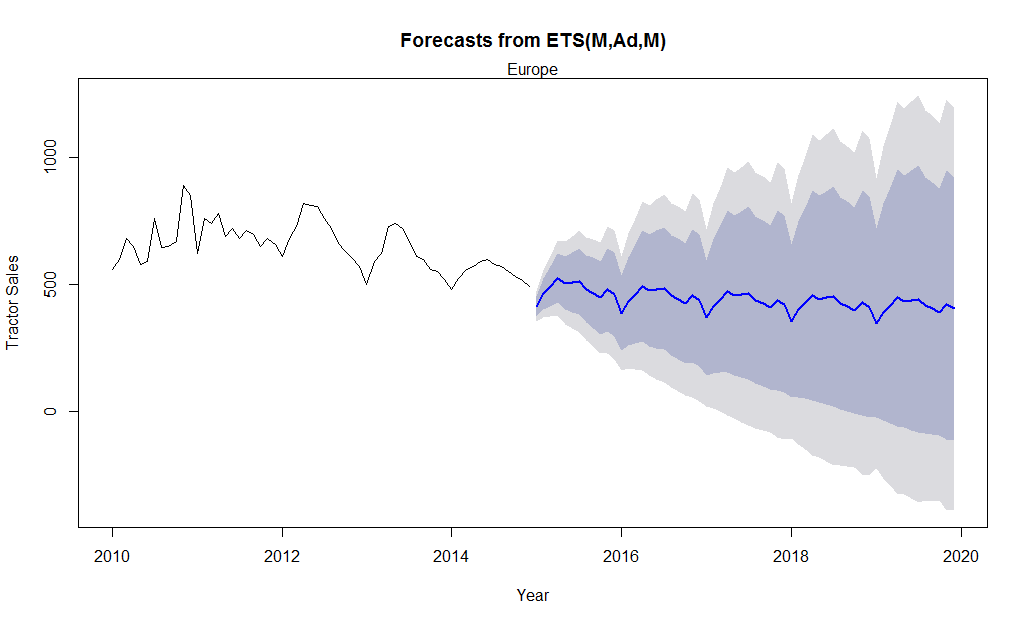
The forecasts themselves are given below. The blue line represents the given forecast, the dark blue shaded area represents a 95% confidence interval, and the light blue represents an 80%.



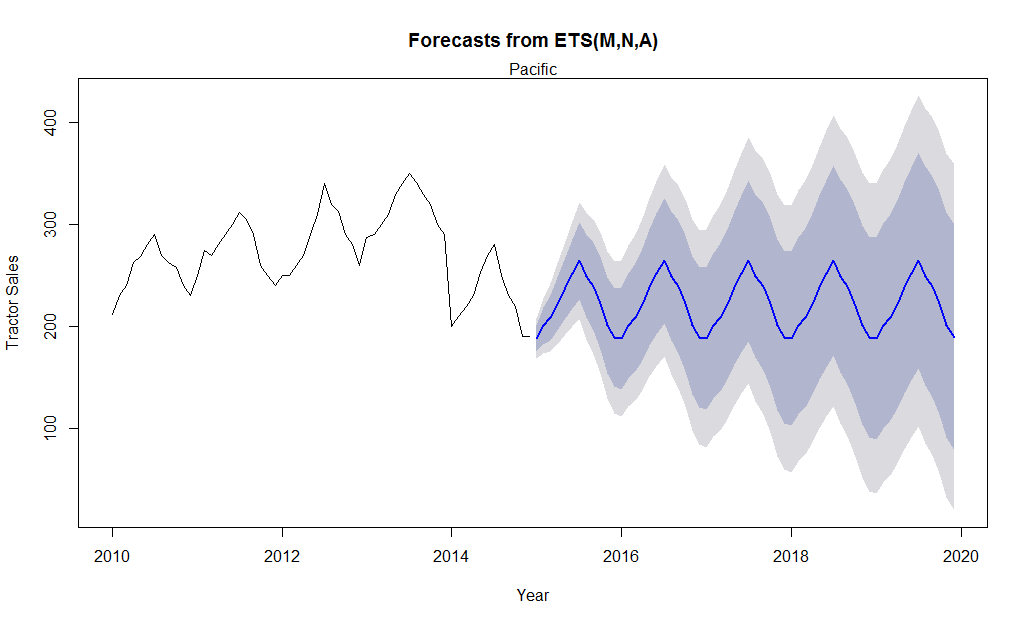
The first forecast performed was on North America. It yielded a forecast with a MAE of 71, signifying a forecast that is on average incorrect by 71 tractor-units, with a MAPE of 6%. The forecast implies that North American demand will taper off, despite a multiplicative model being used.



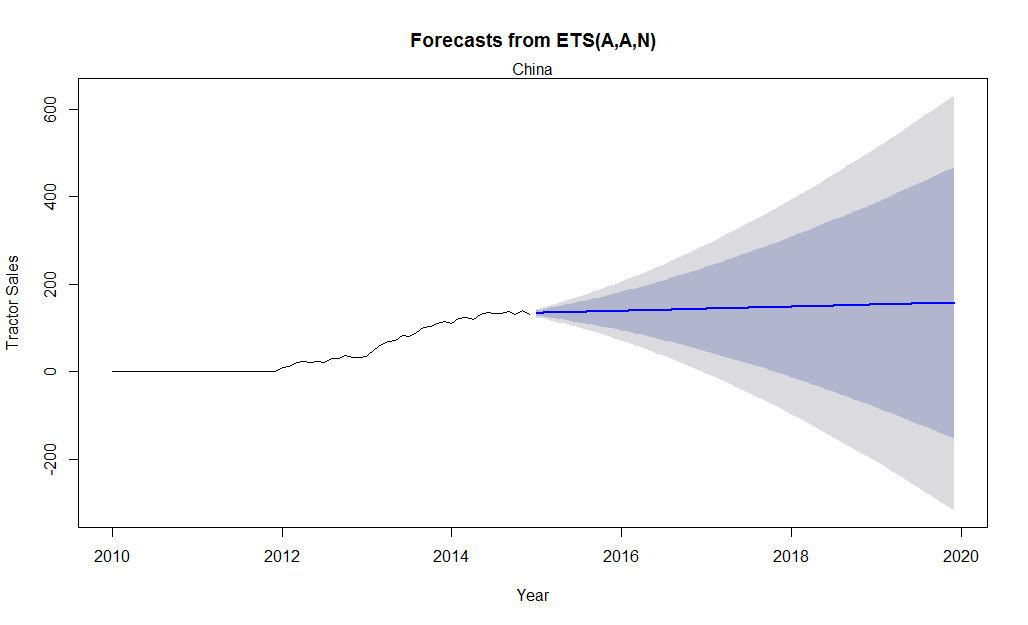
The next forecast was for South America. It has a MAE of 22, an improvement over the NA forecast, and a MAPE of 3.8%. It implies that demand in South America will steadily rise.



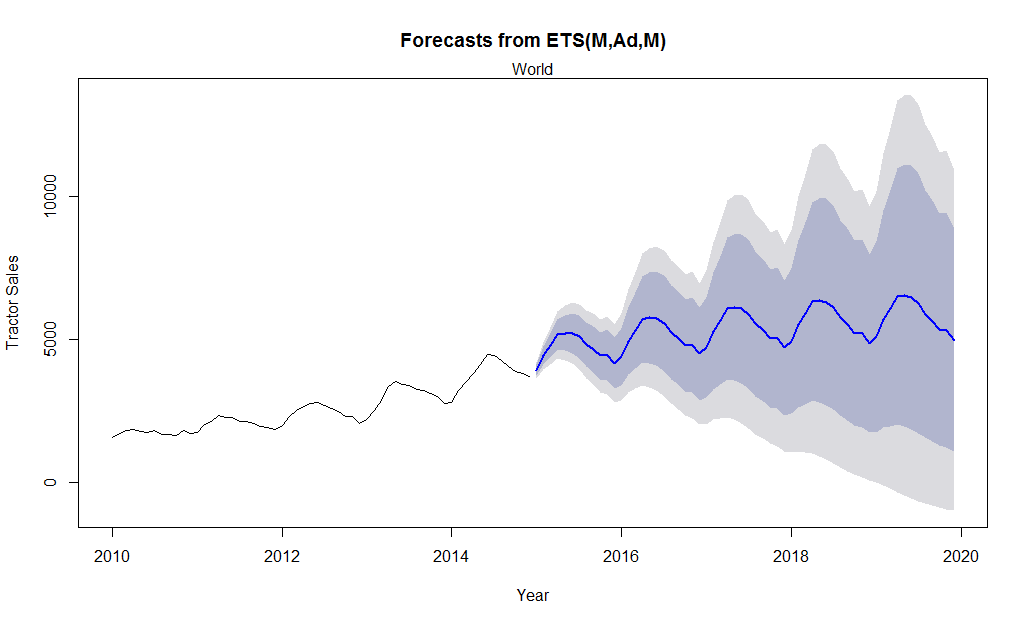
The third forecast is for Europe. It has a MAE of 35, which is acceptable, and a MAPE of 5.38%, high, with respect to previous forecasts. Europe's seasonality is expected to lose much of its variability, but still be slightly negative.



The forecast for the Pacific has an MAE of 8 and a MAPE of 3.33%, making is the most accurate thus far. A word of caution, the MAE should be taken in context with the volume of sales themselves, since it represents actual units. The trend for the Pacific is expected to stay fairly level and maintain somewhat normal seasonality.



The Chinese forecast is curious since it includes some zero-value data. Thus, the MAPE cannot be calculated. The MAE is 3, the lowest of the lot. An ARIMA model may be a better use for recent China data. However, it fails the stationary test for ARIMA, and the standard forecast performs adequately regardless of being able to determine the MAPE, at least sufficiently so to determine the trend of the data. That being said, very little can be properly determined judging by the intervals; outside data must be used to better determine the nature of the Chinese market.



The forecast for the world market in totality is given above. It resulted in an MAE of 73, with a MAPE of 2.85%. Considering the number of tractors in total represented, the MAE is largely meaningless if being compared to the other forecasts. The forecast appears to show a trend that maintains its seasonality. The slope of the trend is expected to decrease.

**Assumptions & Limitations**

For the regressions, the direction of causality (if indeed there is any causality) cannot be inferred, so any interpretations should be taken with a grain of salt.

Regression assumptions were rigorously checked, but it is possible errors were made. It is also worth remembering that the possibility of incorrectness exists despite statistical significance; p-values are only likelihoods, after all.

All the models, the regressions and forecasts, should be re-run with test and training sets to ensure no overfitting is present, if the values given are intended to be used in a more specific case as opposed to measuring general trends and such. They do work for general recommendations as to what areas to focus on.

The forecasts, since they are for periods so far out, are highly suspect in terms of actual values. The trends may be correct, but outside information should be considered as well.

**Recommendations**

1. Delivery practices from decentralized purchasing structures should be reviewed in order to improve them and thus improving PLE usage and customer satisfaction.
2. Larger customers should be encouraged to move to a more centralized purchasing structure if possible, as this raises PLE usage level and customer satisfaction.
3. If market share (i.e. PLE usage) is desired in an area, seeking out larger firms there may be better as these tend to use more PLE products. Also, establishing centralized procurement centers will help this as well. However, the sales force should be monitored, as an overly attached sales force appears to lower customer satisfaction, so they may drive customers off.
4. The large procurement centers need to be better managed in general to improve customer satisfaction.
5. Offers or sales to encourage purchases of tractors in North America, the Pacific, and Europe to reduce the sharp seasonality peaks by raising the troughs may be warranted to keep business more stable throughout the year in these areas.
6. Sales in the Pacific region and Europe are waning; something should be done to increase them. A strategy of appealing to large firms with effective sales staff that de-emphasize the cutting-edge qualities of PLE's products, and offering high price flexibility in these areas may stall this trend.