Cluster-based Forecasting for Laboratory samples

Research paper Business Analytics

Manoj Ashvin Jayaraj

Vrije Universiteit Amsterdam
Faculty of Science
Business Analytics
De Boelelaan 1081a
1081 HV Amsterdam

Host Organization:
IDEXX Laboratories
Scorpius 60
2132 LR Hoofddorp
Netherlands

May 2017

Abstract

Forecasting a future event is important for companies for a better planning in their daily operations, also it helps every company to understand their customers better. Companies use forecast methods to predict the future events on a daily, weekly, and monthly basis. Forecasting requires data from the past and present to predict the future.

This report addresses the methods that are used to forecast the number of samples to be received in 2017 for IDEXX Laboratories in the United Kingdom. The data used for this research paper contains the number of samples received from the veterinary practitioners at IDEXX Laboratories from January 2010 until December 2016. Because of the challenges in the data, customers are first clustered using three distance measures. The forecast is done on the clustered data using the Holt Winters model and the ARIMA model. The best forecasting model is then chosen with the lowest Root Mean Squared Error [RMSE].

Contents

1. Introduction **1**
2. Data Description and Pre-processing **2**
3. Approach **3**
4. Clustering **3**
4.1 Distance based Clustering 3
4.1.1 K-means Clustering 4
4.1.2 Partition Around Medoids (PAM) 4
4.1.3 Hierarchical Clustering 5
4.2 Determine the Optimal Cluster 5
4.2.1 Elbow Method 6
4.2.2 Silhouette Method 6
5. Time Series **9**
6 Forecasting **11**
6.1 Holt Winters Forecasting 11
6.2 ARIMA Forecasting 12
7 Evaluation **12**
7.1 ARIMA vs Holt Winter's evaluation 14
7.2 Results 14
8 Conclusion **17**
References **19**

1. Introduction

At IDEXX, the collection of samples in the UK has been outsourced to external courier companies and the route planning is done by IDEXX. Couriers are charged based on the distance they travel every day. Customers at IDEXX are grouped into two categories, namely regular and on-call customers. In the case of regular customers, an assigned courier will visit the clinic every day during the week. In the case of on-call customers, an assigned courier will plan a visit only when there is a sample available for pickup. Most of the regular customers do not have a sample to be picked every day and the distance traveled to such customers is considered as a loss in revenue. Numbers indicate that nearly 50% of regular customers fall into this category. Predicting the number of samples per day for each customer will help IDEXX to optimize their route planning operations and thereby reducing the loss of revenue.

The initial goal of this research paper was to predict the number of samples to be delivered per day by each customer to IDEXX Laboratories in the UK for 2017. To do the forecast will use the data from January 2010 until December 2016. Because of the structure of the data and the absence of additional information like the reason for variations in the number of samples and when a customer left the company, forecasting on a daily basis becomes very challenging. So, the goal of this research paper has been modified to forecast the number of samples to be delivered on a monthly basis. Not all the customers delivered samples every month of the year, so we decided to first cluster the customers by K-means, Hierarchy and PAM clustering methods using different distance measures after which the forecast is done for each cluster. Holt Winters and ARIMA models are used for forecasting and the best model is chosen with the least error value. Because of the structure of the data, assumptions had to be made to fit the model.

Finally, the residuals of the model were verified if they satisfy the stationary property, which means residuals should have zero mean, a constant variance and having a normal distribution.

2. Data Description and Pre-processing

At IDEXX, the customer data are retrieved from their LYNX system in Excel format, from which the customer code and the number of samples delivered per day from January 2010 till December 2016 were extracted. Not all customers were present from January 2010 and sometimes the customer codes of the customers who left the company were assigned to new customers. Since the information about such customers is not available they are treated as a single customer in this research paper. Also, to finalize the list of customers for further processing, we extracted the list of customers who were present in December 2016 and only considered the data of those customers from January 2010 till December 2016. So customers who were present only until November 2016 will be excluded in this research work.

In R-studio the data is viewed as a matrix and appears like in Fig-1 which only shows the entries of 9 customers with their first 12 months of data only. As we can see, the entries are having numeric values greater than zero, some entries with zeros that represent that the customer did not deliver any samples in those months and the NAs correspond to no information available for those customers which can be for two reasons, customers were registered at IDEXX from January 2010 but no information was present for that time or customers were registered at IDEXX somewhere between January 2010- December 2016 and the NAs correspond to the months when they were not registered. We also decided to replace the customer entries with NAs to zero, as NAs does not have a value.

	Jan-10	Feb-10	Mar-10	Apr-10	May-10	Jun-10	Jul-10	Aug-10	Sep-10	Oct-10	Nov-10	Dec-10
BB38	0	56	53	47	52	65	49	45	51	57	63	54
BB46	0	11	19	21	13	18	17	12	15	24	17	8
BB47	0	0	1	3	3	2	3	1	1	3	2	2
C0025	0	0	0	0	0	0	0	0	0	18	24	17
C01	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
C012	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
C02	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
C03	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
C04	0	56	70	45	47	61	56	36	35	37	36	44

Fig 1- Raw customer data

Before beginning with forecasting, we need to split the data into a training and test dataset. Since each observation has 84 variables (each month from January 2010 till December 2016 is represented as a variable), we split the data into 66 observations of training data and 24 observations of test data. Then the forecasting models are applied to the training set and the forecasted data will be validated against the test set from which we will determine the best forecasting model with the smallest error value.

3. Approach

In the pre-processing section we replaced NAs as zeros, after which we noticed that there were instances where the customers delivered samples only for two years out of seven years (example: 2013 and 2016 only, and some others in 2015 and 2016 only). Since not all customers delivered samples every month of the year, we decided to first cluster the customers by three distance-based clustering techniques, before we introduce it to the forecasting models. By clustering we can determine the similarities in customers based on the number of samples delivered. Once the optimal number of clusters is determined by each clustering method, the forecast is done for each cluster and the error from each cluster for each clustering method are summed up by assigning weights for each error measure. The weight assigned is the number of customers in each cluster, considering that the cluster with more customers is more important than the ones with fewer customers. Both Holt Winter and ARIMA models were trained on the data starting from January 2010 until December 2014 named as training set and the remaining data will be the test set. Since the Holt Winter and the ARIMA model use different error measures, the best model is chosen with the lowest Root Mean Squared Error [RMSE] value from the test set. The residuals of the model are finally verified to confirm that the errors are normally distributed with mean zero and constant variance with the help of a histogram plot.

4. Clustering

4.1Distance based Clustering

Clustering is the process of grouping customers in such a way that the customers in a group have some similarity between them. In this paper, we discuss distance-based clustering which means that customers with the shortest distance between each other are grouped to form a cluster. Here we only focus on three types of clusters namely *K-means clustering, Hierarchy clustering, and Partition Around Medoids clustering*. Each clustering method is applied using three distance measures namely the *Euclidean distance, the Manhattan distance and the Maximum distance.* The smallest values of each distance measure are considered to form a cluster.

The distances are calculated using the formula as below.

Euclidean distance	$\sqrt{\sum_{i=1}^{N}\left(X_i - Y_i\right)^2}$,	--- [1]
Manhattan distance	$\sum_{i=1}^{N} \mid X_i - Y_i \mid$,	--- [2]
Maximum distance	$\max_i \mid X_i - Y_i \mid$.	--- [3]

4.1.1 K-means Clustering

The first step in the K-means clustering method is to specify the number of clusters (K). Choosing an optimal number of clusters will be discussed later in this paper after we explain the concepts behind each clustering method. The process begins by partitioning the 2,041 observations- with each observation as a customer, into K ($\leq$ 2,041) sets of clusters so as to minimize the within-cluster sum of squares (i.e., variance). The K-means method can be explained in a 4 step process

Step1 - The initial K means are randomly generated.

Step2 - The distance from each observation to the initial K means are calculated, observations that are closer to each other are grouped to form a cluster, and thus K clusters are created initially.

Step3 - The means are then calculated for each of the K clusters, which then become the new centroid[a].

Step4 - Step 2 and 3 are repeated until convergence is reached.

The distance in step 2 is calculated using the distance measure, by means of Equations [1], [2] or [3], and as a result we have formed K clusters for each distance measure.

4.1.2 Partition Around Medoids (PAM)

PAM clustering is somewhat similar to K-means and the number of clusters (K) should also be specified initially as in K-means. In contrast to K-means, PAM chooses one of the observations as the initial medoids[b], and then associates each of the data points to the closest medoids using the distance measure as per Equations [1], [2] and [3]. The total cost is then calculated by summing the cost for each cluster, with the cost being the sum of distances of points to their medoids. The process is then repeated by changing the initial medoids each time with one of the non-medoids until the total cost of the K cluster configuration decreases. The *cluster* package should be loaded in R-studio before one can use the PAM clustering technique.

a - Centroid is the mean position of all the points in an n-dimensional space
b- Medoids are objects of a cluster whose average dissimilarity to all the other objects in the cluster is minimal

4.1.3 Hierarchical Clustering

In this paper we will discuss only Agglomerative clustering- a bottom up approach where each observation starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy. The result of the clustering is usually represented in a dendrogram. A dendrogram is formed based on the distance measure as per Equations [1], [2], and [3]. Pairs of clusters are formed based on the maximum possible distance between points belonging to two different clusters. A representation of a dendrogram for some random set of customers is shown in Fig-2 with the height being the distance at which the observations are merged. Once the dendrogram construction is complete, customers can be grouped into K clusters starting from the top of the tree. For example, if we need 3 clusters then starting from the top of the tree, at height 1,500 if we draw a horizontal line, we will have 3 groups of customers with GODD9 being alone in one cluster. Fig 2 is only an example, as a representation for 2,041 customers will not be clearly visible.

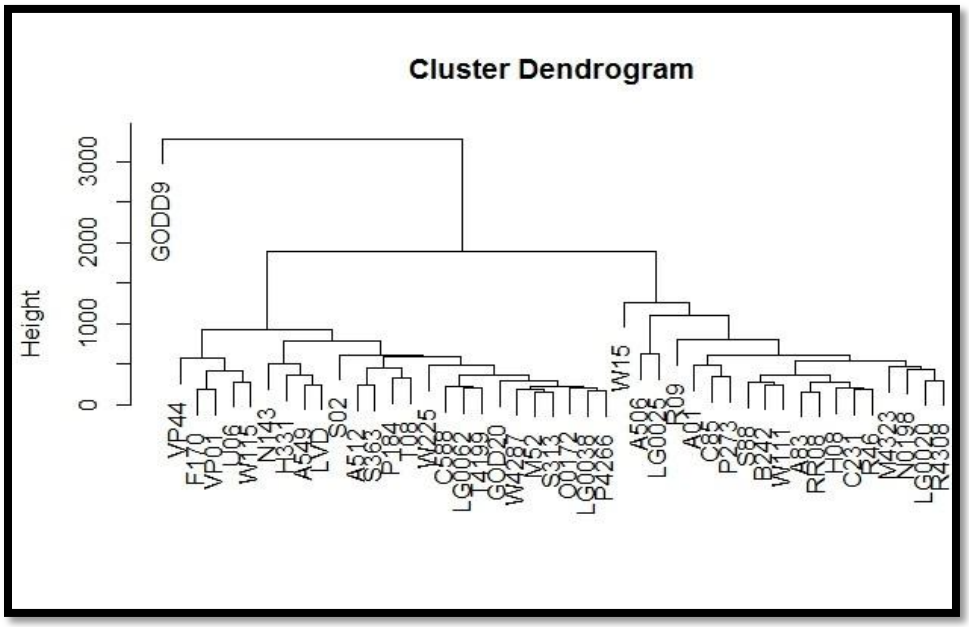


Fig 2- Dendrogram

4.2Determine the Optimal Cluster

In this section, we will discuss two ways of determining the optimal number of clusters that will be applicable for all the three clustering methods discussed in the previous section. Those two methods are,

1. The elbow method;
2. The silhouette method.

4.2.1 Elbow Method

The elbow method looks at the variation within the cluster and chooses the optimal number of clusters where adding more clusters will not give any improvement to the data. Plotting the *within the sum of squares* against the increasing number of clusters shows a graph as in Fig-3. As we can see in Fig-3 the variation is higher for one cluster and it gradually reduces as we keep increasing the number of clusters for the 2,041 observations. As the name of the method suggests, the point where the plot has a bend (knee) is generally considered as the optimal number of clusters.

4.2.2 Silhouette Method

The quality of the cluster is determined using the silhouette method. This method determines how close each object lies within a cluster in comparison to other clusters

Silhouette Method $$S_{(i)} = \frac{|b_i - a_i|}{\max\{a_{(i)}, b_{(i)}\}}$$ --- [4]

where a_i = avg (dissimilarity within the cluster) and b_i = the lowest average dissimilarity to other clusters. The silhouette value usually lies in the range $-1 \leq S_{(i)} \leq 1$. The higher the silhouette value, the better the clustering. Like in the case of the elbow method, we plot the Silhouette value against the increasing number of clusters and the optimal number of clusters is chosen with the best silhouette value. Fig-4 represents an example of the silhouette value against the increasing number of clusters.

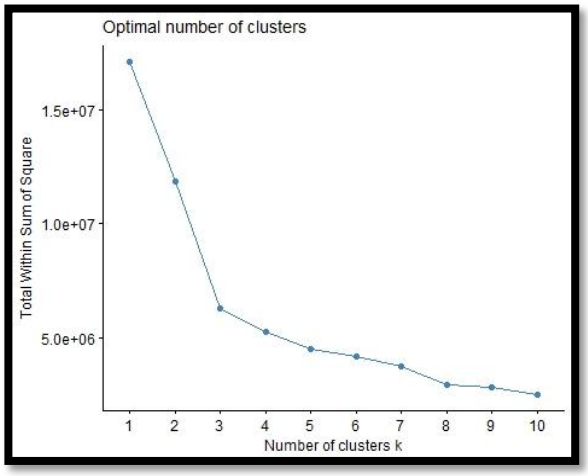


Fig-3 Optimal cluster using elbow method, K-means

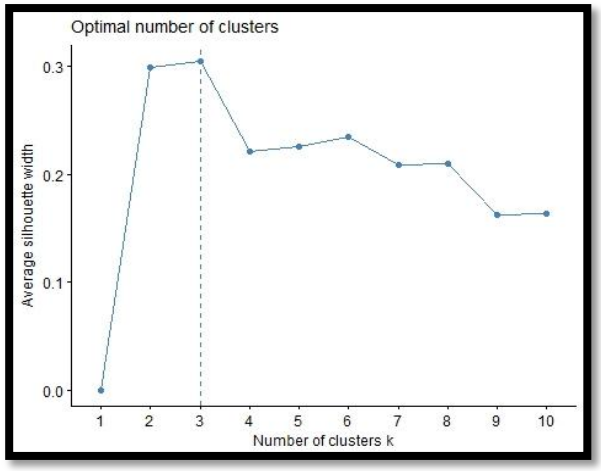


Fig-4 Optimal cluster using Silhouette method, k-means

Fig-3 and Fig-4 are plotted for the data with 2,041 observations and the clustering technique used is K-means. The result shows that 3 clusters will be optimal. Similarly, the optimal number of clusters for PAM and Hierarchical clustering methods can also be determined.Fig-5 is an example of how customers are clustered by Hierarchical clustering using the maximum distance. Also clustering is performed for other clustering methods using the three distance measure chosen as per Equations [1], [2], and [3]. As we

can see in Fig-5, the majority of the 2,041 customers are grouped in the 1st cluster where the 2nd cluster has 4 customers and the 3rdcluster has only one customer. Examining the average number of samples per month, shows that that the customers in the 2nd and 3rd cluster were delivering a higher number of samples per month, when compared to the customers in the 1st cluster, this analysis triggered two questions:

1. *Can we split the customers into n groups, based on the number of samples and then apply the clustering techniques for individual groups?*
2. *What would be the threshold number for splitting the customers into n groups?*

Since cluster 1 has more than 95% of the customers, we decided to split the customers into two groups and then apply the clustering technique to each group. The reason to split the customers into two groups was to distinguish the high revenue generating customers from others. However, there are several ways of grouping the customers depending on the business requirements, for example we can also group the customers based on the route they were allocated to (every customer at IDEXX is assigned to an individual route based on which courier agent will collect the samples). In this research paper, we will proceed with the idea of grouping the customers into two groups, with group 1 having customers delivering fewer than 150 samples per month and group 2 having customers delivering more than 150 samples per month.

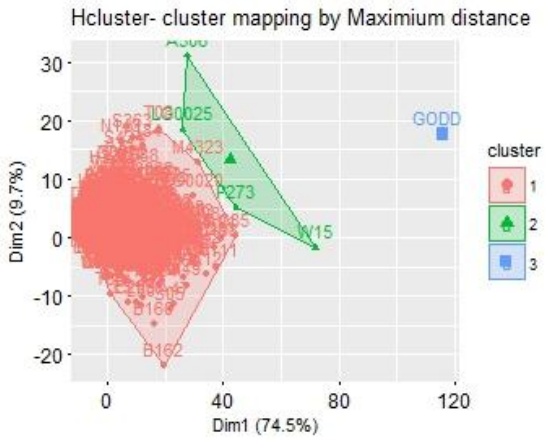


Fig 5 Hierarchical clustering of 2,041 customers

Once the customers are divided into two groups, the optimal number of clusters is again determined for each group using the elbow method and the silhouette method for all the three clustering methods. Fig-6 and Fig-7 are examples of PAM clustering from two groups using the Euclidean distance. These plots can be generated using the *fviz_nbclust()* function from the *factoextra* package in R studio.

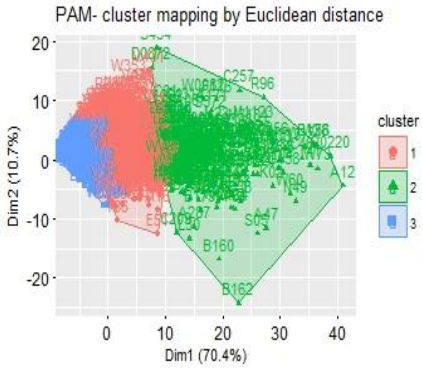


Fig 6 Clustering – customers with samples <150

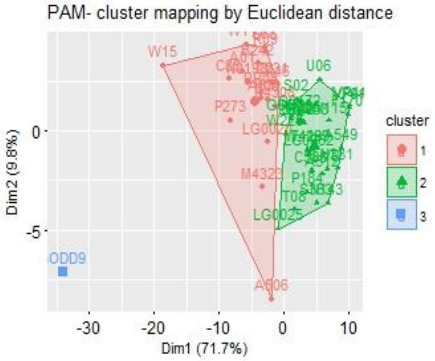


Fig 7 Clustering – customers with samples > 150

After applying the elbow and the silhouette method to both groups we found that 3 clusters are optimal for all the three clustering types. Fig-8 represents the division of customers from each cluster based on the optimal clusters suggested by the elbow and the silhouette method. The column represents the clustering method and each row represents one cluster forming 3 clusters in total. Thus, aggregating the samples from each customer will reduce the number of observations from 2,041 to 3 under each group.

By Euclidean distance		
Hclust	**PAM**	**Kmeans**
1946	639	518
37	229	128
13	1128	1350

By Maximum distance		
Hclust	**PAM**	**Kmeans**
1977	437	1248
14	915	216
5	644	532

By Manhatten distance		
Hclust	**PAM**	**Kmeans**
1925	567	1356
41	214	511
30	1215	129

Customers with samples below 150

By Euclidean distance		
Hclust	**PAM**	**Kmeans**
19	18	1
25	26	19
1	1	25

By Maximum distance		
Hclust	**PAM**	**Kmeans**
9	26	12
35	18	1
1	1	32

By Manhatten distance		
Hclust	**PAM**	**Kmeans**
20	17	25
24	27	1
1	1	19

Customers with samples above 150

Fig-8 Grouping of customers in each cluster via the 3 distance measures

5. Time Series

A time series is a sequence of data points listed in the order of time; usually the data will be starting from the past and not necessarily needs to end with the present day. In this paper we are going to forecast the number of samples for 2017.We use the available data from the past 7 calendar years. One of the most important concepts in the context of modeling time series is *stationarity*, which occurs in two forms: *strictly stationary* and *weakly stationary*. A time series is *strictly stationary* if, for any value of k, the joint distribution of $(X_{t+1},\ldots, X_{t+k})$ does not depend on t. Strict stationarity is a strong and, in practice, uncheckable assumption. For most practical purpose, including this research paper the assumption of weak stationarity is sufficient. A time series is second order stationary or weakly stationary if, for any value of *k*, EX_t and EX_{t+k} exist and do not depend on t.

The time series in our paper is the clustered data from Jan-2010 till Dec-2016 and the data exhibits some structure as seen in Fig-9 and it does not satisfy the weakly stationary properly as the mean increases over time. So our first step is to convert the non-stationary time series into a stationary time series in order to do a forecast. A plot in Fig-9 represents the data from the 1st cluster of the K-means using the Euclidean distance belonging to the group of customers with a number of samples fewer than 150.

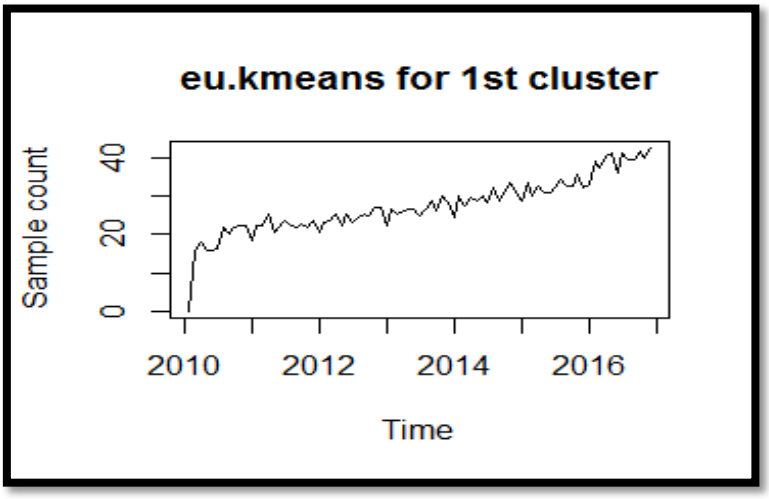


Fig-9 – Samples from the 1st cluster of K-means using Euclidean distance

Fig-9 represents the data from a cluster we obtained from the previous section, as we can see that the data exhibits some trend over time and there can be possibilities of having seasonal variation. The corresponding time series model for the data in Fig-9 can be represented by Equation [5].

$$X_t = m_t + s_t + Y_t \quad \text{--- [5]}$$

where m_t is a deterministic trend component, s_t is a deterministic seasonal component with known period d (in our case its 12 months) and Y_t is the random noise component which is assumed to be stationary so that $EY_t = 0$. So eliminating the trend and the seasonal component from the data can make the data stationary for which we need to first estimate the trend and the seasonal component in the given data. The decomposition function in R helps us to estimate the trend, seasonality in the model using moving averages. Fig-10 represents the decomposed plot for the data used to represent Fig-9, with the 1st plot being the plot of original data, 2nd being the estimated trend, 3rd being the estimated seasonal component and the last one is the estimated randomness in the data whose behavior is similar to the white noise.

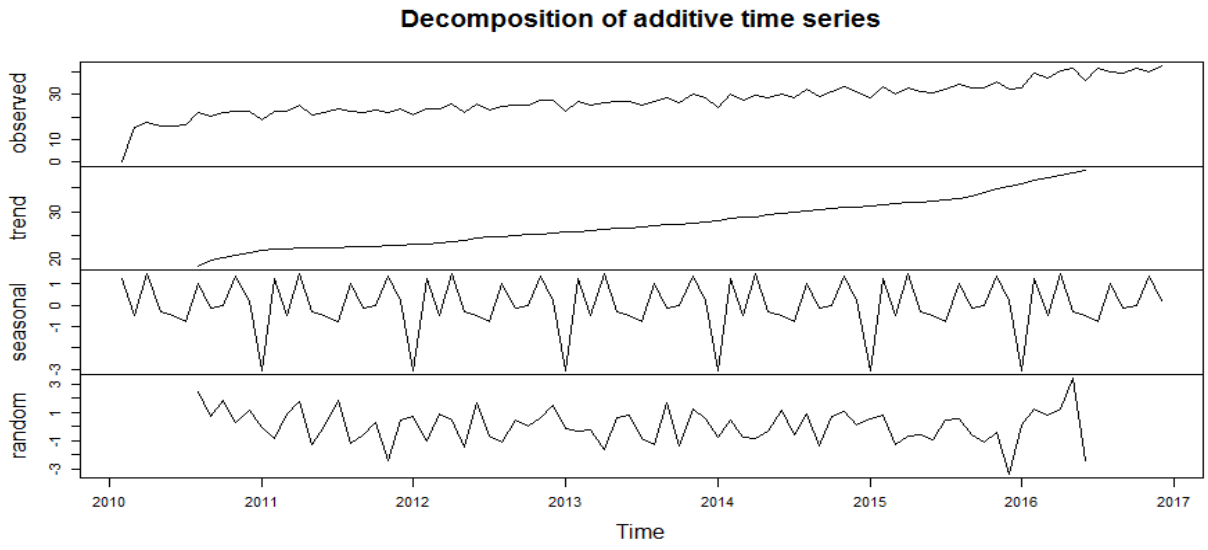


fig – 10 Decomposed time series data

Equation [5] is an example of *additive* time series where the trend and seasonal components are constant over time, however there can be situations where the trend and seasonal components can multiply over time. In such cases, a modified form of Equation [5] named as *multiplicative* time series will be applicable and can be represented by Equation [6].

$$X_t = m_t * s_t * Y_t \quad \text{--- [6]}$$

6 FORECASTING

6.1 Holt Winters Forecasting

The forecasting equation for an additive series, using the Holt Winters model can be expressed as Equation [7], made after observation at time t, where a_t is the estimated level, b_t is the estimated slope[1] and s_t is the estimated seasonal effect at time t.

Additive Holt Winters Forecast $\tilde{Y}_{t+k} = a_t + k * b_t + s_{t+k-p} \quad k \leq p,$ --- [7]

where $a_t = \alpha (x_t - s_{t-p}) + (1- \alpha) (\alpha_{t-1} + b_{t-1}),$

$b_t = \beta(a_t - a_{t-1}) + (1- \beta) b_{t-1},$

$s_t = \gamma(x_t - a_t) + (1- \gamma)s_{t-p}.$

α , β, and γ are the smoothing parameters and a_t, b_t and $s_t$are the one step ahead forecasts calculated using the estimates up to the time t, with p being the time series period length. Repeating this for each step ahead using the current estimates helps us to predict the result upto t+k periods.

In the case of multiplicative series, where the seasonal variations increase over time along with the trend forecast Equation [8] will be different from that of additive series forecasting.

Multiplicative Holt Winters Forecast $\tilde{Y}_{t+k} = (a_t + k * b_t) \ * s_{t+k-p} \quad k \leq p,$ --- [8]

where $a_t = \alpha \frac{x_n}{s_{t-p}} + (1- \alpha) (\alpha_{t-1} + b_{t-1}),$

$b_t = \beta(a_t - a_{t-1}) + (1- \beta) b_{t-1},$

$s_t = \gamma(\frac{x_t}{a_t}) + (1- \gamma)s_{t-p}.$

The value for the smoothing parameters α ,β and γ are usually estimated by minimizing the one-step-ahead prediction error (SS1PE). The smoothing parameter lies in the range between 0 and 1. A typical choice of the parameter value is 0.2 indicating that the forecast depends on the past data and the choice of smoothing parameters towards 1 indicates that more weight will be given to the recent data. The Holt Winters function in R studio automatically chooses the best smoothing parameters using the smallest SS1PE value. SS1PE value is calculated using the equation [9]

$$\text{SS1PE} = \sum_{t=2}^{n} (e_t)^2, \quad \text{--- [9]}$$

where $e_t^{\,2} = e_2^{\,2} + e_3^{\,2} + \ldots + e_n^{\,2};$

$e_t = \left(X_t - \hat{X}_{t|t-1}\right);$

X_t is the value of the variable to be predicted at time t;

$\hat{X}_t$ is the predicted value of X_t;

n is the number of observations.

6.2 ARIMA Forecasting

The ARIMA- Auto Regressive Integrated Moving Average, is a combination of an AR- autoregressive and an MA- moving average model. Since the timeseries can be non-stationary, the concept of differencing is used in the ARIMA model to make the time series stationary. If, X_t is a non-stationary time series, then $\nabla^d X_t$ will be the differenced time series which can be stationary, with d being the order of differencing. Such a model is represented as ARIMA(p,d,q). In the case of stationary time series, the differencing parameter will be of no use making the model as ARIMA(p,q) represented using Equation [10].

$$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p * X_{t-p} + Z_t + \beta_1 * Z_{t-1} + \ldots + \beta_q * Z_{t-q,} \quad \text{--- [10]}$$

where X_t is the stationary process;

p and q are the ARIMA parameters;

Z_t is the white noise- with mean zero and variance σ^2.

In R-studio the auto.arima() function helps to determine the best model based on the best *AIC* value. Given a collection of models for the data, the AIC estimates the quality of a model, relative to each of the other models. Hence, the AIC provides a means for model selection. Since the customer data is divided into clusters using three clustering methods, the best clustering method is chosen based on the ARIMA model with the least AIC value. AIC value is calculated using Equation [11].

$$\text{Akaike information criterion (AIC)} = 2k - 2\ln(\hat{L}) \quad \text{--- [11]}$$

where k is the number of estimated parameter in the model;

$\hat{L}$ is the maximum value of the likelihood function for the model; i.e. ($\hat{L} = P(x|\hat{\theta}, M)$;

M is the ARIMA forecasting model for the time series x;

$\hat{\theta}$ are the parameter values that maximize the likelihood function.

7 Evaluation

In the previous section, we trained the two groups of data using the Holt Winters and the ARIMA model and obtained the forecasting result from both the models. Now we will explain how to determine the best model based on their performance. Since the Holt Winters model works on the data by removing trend and seasonality, we have 3 observations for each cluster which are Holt winters with the removal of trend only, Holt winters with additive seasonality, and Holt Winters with multiplicative seasonality. Since three clustering methods are applied using three distance measures and the optimal number of clusters is three

in each method, we have 27 observations under ARIMA and 27 observations under each of the Holt Winters models making it challenging to choose the best model as visible in Fig-11. The best model is chosen based on the smallest error value from their forecast. As we can see from Fig 8 that each clustering method differs in their division of customers. Only adding the errors from each clustering method will not lead to a good choice for the best forecasting model. Instead, we assign weights to the error from each cluster and then sum the weighted error under each clustering method. This reduces the 27 results to 9 under the Holt Winters and the ARIMA model. The weights assigned are the number of customers in each cluster, because the clusters with more customers are more important than clusters with fewer customers.

Clusters	SSE_trend_alone	SSE_additive	SSE_multiplicative	ARIMA_AIC
eu.kmeans_cluster_1	888.117115	126.1722328	145.7832714	266.13134
eu.kmeans_cluster_2	28.368453	4.3426922	5.2509043	77.58303
eu.kmeans_cluster_3	6151.236867	968.725781	1110.528471	440.60158
max.kmeans_cluster_1	693.366489	80.5266584	92.0482924	252.08261
max.kmeans_cluster_2	2808.867758	512.0763889	594.6087365	389.01345
max.kmeans_cluster_3	32.735332	4.582028	5.4996654	82.72976
man.kmeans_cluster_1	969.223724	139.6115068	161.724435	259.65004
man.kmeans_cluster_2	6328.691275	976.2132593	1114.220623	442.82078
man.kmeans_cluster_3	23.072786	3.8540286	4.5564485	70.45854
eu.hclust_cluster_1	195.718201	27.6110021	32.6122042	182.15261
eu.hclust_cluster_2	10279.71494	1282.061808	1432.729813	469.17139
eu.hclust_cluster_3	14361.56432	2382.779042	2213.044631	504.41844
max.hclust_cluster_1	242.745328	32.7468703	38.6723678	193.35315
max.hclust_cluster_2	13120.23061	2049.96666	1957.103255	500.93661
max.hclust_cluster_3	236.661421	169.6479656	148.0326786	204.18071
man.hclust_cluster_1	184.891996	24.3203915	28.4860723	180.80775
man.hclust_cluster_2	9388.614396	1250.379738	1425.378298	383.78838
man.hclust_cluster_3	6950.616106	2145.348326	2339.028526	453.4199
eu.pam_cluster_1	512.398703	58.3111638	67.2847874	230.21705
eu.pam_cluster_2	4115.570821	702.9978758	810.657939	415.40514
eu.pam_cluster_3	9.805479	2.1730006	2.6651355	24.44343
max.pam_cluster_1	1877.463579	285.6784186	333.7879923	297.98719
max.pam_cluster_2	192.014694	22.4864804	25.9139116	177.81718
max.pam_cluster_3	1.141046	0.7010614	0.8665233	-56.347
man.pam_cluster_1	646.657211	76.6742667	89.9517766	234.94609
man.pam_cluster_2	4355.18051	726.5426209	831.0069884	419.05347
man.pam_cluster_3	11.983306	3.0940094	3.5493999	44.22671

Fig 11* –SSE value from Holt Winters with trend alone, additive seasonality, multiplicative seasonality and AIC from ARIMA

where, eu- Euclidean distance, man- Manhattan distance, max- Maximum distance

kmeans, pam, and hclust are the three clustering methods

***** Fig 11 represents the group of customers with samples less than 150

The model with the smallest SSE value is then chosen as the best from Holt Winters, first among the different methods (Trend only, additive seasonal effect, multiplicative seasonal effect) and then between the different clustering methods. Similarly, the best ARIMA model is chosen from the results with the smallest AIC value. Fig-12 represents the errors from Holt Winters and ARIMA after assigning weights to the error values.

Cluster	ARIMA_RME	HW_RMSE
Euclidean_kmeans	298990.1	195216.8
Maximum_kmeans	321381.6	159167.1
Manhattan_kmeans	285346.8	202499.1
Euclidean_hclust	378385.8	129936.9
Maximum_hclust	390293.2	92880.17
Manhattan_hclust	377392.8	162442.8
Euclidean_pam	269808.7	200698.5
Maximum_pam	256635.7	145868.1
Manhattan_pam	276627.3	202713.7

Fig 12*- SSE and AIC value after assigning weights

7.1 ARIMA vs Holt Winter's evaluation

The ARIMA and the Holt Winters model have different error measures and so we cannot compare the SSE and AIC values to choose the best between them. To solve this, we calculated the Root Mean Squared Error (RMSE) for both the ARIMA and the Holt Winters model and the model with the smallest RMSE for the test data is chosen as the final model and the forecast from that model will be the best fit. The RMSE is calculated for individual clusters and weights are assigned to each RSME value. Evaluation is done to both the group of customers, customers with samples fewer than 150 per month and above 150 per month andthe best model is chosen in both cases.

7.2 Results

Computations show that in the case of customers delivering samples below 150 per month, the Holt Winters model using the Hierarchical clustering method by means of the maximum distance performs better among the other models. For the Holt Winters model, the 1st cluster follows a time series model with additive seasonality, where as the 2nd and 3rd cluster follows a time series model with multiplicative seasonality. In the case of customers with samples above 150 per month, the ARIMA model using the PAM clustering method by means of the maximum distance performs better among the other ARIMA models. A plot of the forecast for 2017 can be viewed in Fig-13. The training set (64 observations – 2010 January

* Fig 12 represents the group of customers with samples less than 150

till 2014 December) is indicated in black and the test set (24 observations – 2015 January till 2016 December) in blue along with the forecast for 2017.

However, in Fig-13 we can see that the forecasted result for the test set, indicated in blue does not follow the actual results indicated in red. The reason for the deviation is due to the variation in the training data. We can observe a higher difference in the case of the ARIMA forecast- cluster 3 (**customers with more than 150 samples*) and the reason is also clear from the same figure, there is a drastic increase in the number of samples delivered from January 2016 which was not considered in the training data. In case of Holt Winters forecast – cluster 2 (**customers delivering samples below 150)* the forecasted result is even lower than the actual which explains that the Holt Winters model forecasts are based on the observations from last few years but the test set does not follow the trend observed until June 2015 instead the number of samples increases starting from July 2015 and that increasing behaviour is not covered in the training set.

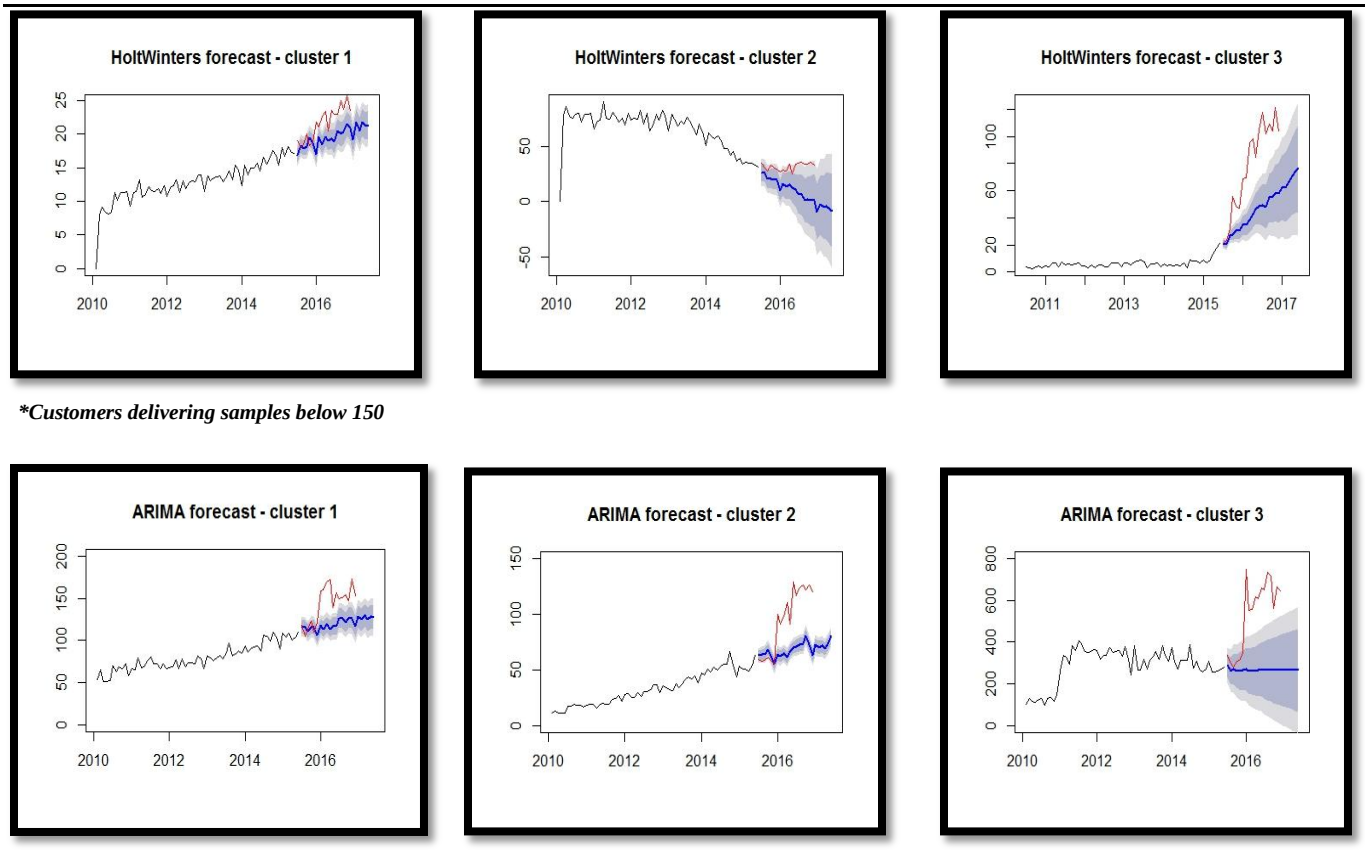
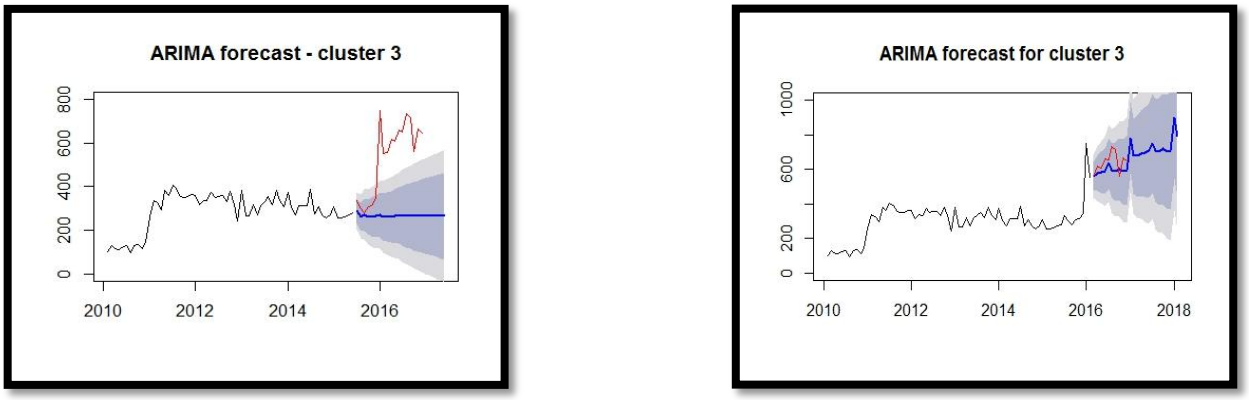


Fig-13 Holt Winters and ARIMA forecast for 2017

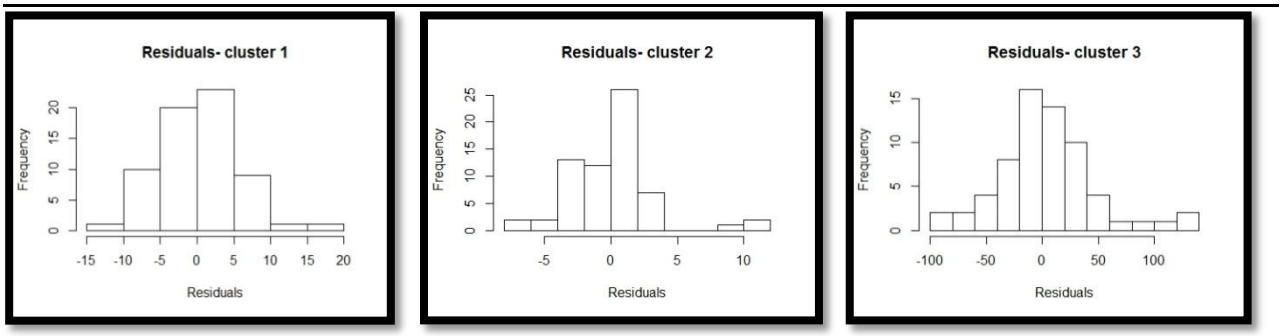
Because of the variations post-June 2015, we again trained the data by including the observations from July 2015 and tested them for different periods. We observed that the accuracy was much better when the data post-June 2015 was ìncluded in the training set.However in this report we will show the result from one observations only, but the results were also tested for all other observations and the accuracy is better when compared to the result from Fig-13. Fig-14 shows the results of ARIMA forecasting method for different lengths of training data and the model with training data extended until February 2016 performs far better than the earlier case.

To verify, if the forecasting errors are normally distributed we also plotted the residuals of the forecasting model using histogram as shown in Fig-15, we can see that the distribution of forecasting errors are centered on zero and is more or less normally distributed, although it seems to be slightly skewed compared to a normal curve. However, the skew is relatively small, and so it is plausible that the forecast errors are normally distributed with mean zero.

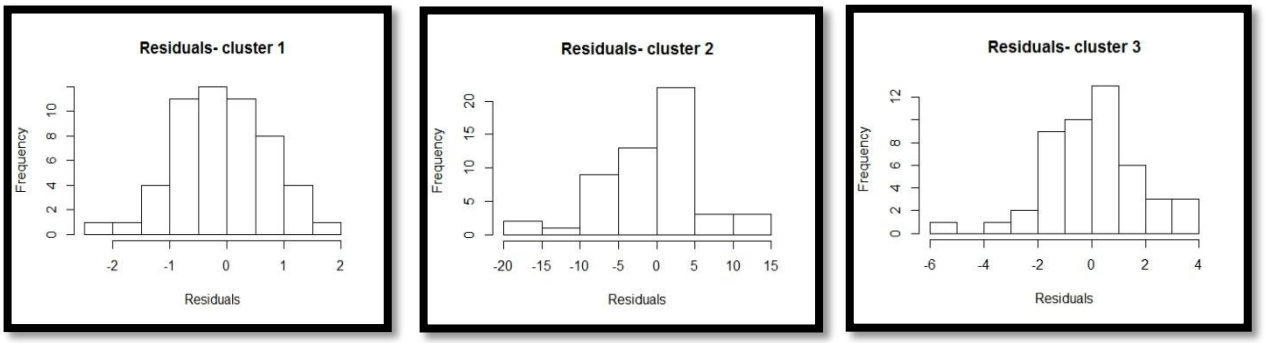


Result for Training data until Dec-2015 **Result for Training data until February-2016**

Fig-14 Results for ARIMA forecast with different lengths of training data



***Customers delivering samples below 150**



***Customers delivering samples above 150**

Fig 15- Residuals of forecasting model

8 Conclusion

In this research paper, we forecasted the number of samples to be delivered in 2017 for IDEXX laboratories in the UK using Holt Winters and ARIMA forecasting methods. Because of the variations in the number of samples, like customers only delivering samples for 1 or 2 years between January 2010 and December 2016 , we decided to cluster the customer using three distance measures. Based on the observations from Fig-5, 6 and 7 we see that the vast majority of customers are under one cluster. We also decided to divide the customers into two groups before clustering (customers delivering below 150 samples and above 150 samples). The threshold of 150 samples was considered based on the assumptions that the customers delivering more than 150 samples fall under the category of high revenue generating customers, however based on business requirement the grouping can be done in different ways.

We also discussed the possible ways to determine the optimal number of clusters. Finally, we trained the data using Holt Winters and ARIMA models for different lengths of training data and the best forecasting modelwas chosen based on the smallest Root Mean Square Error. We observed that the ARIMA model performs better in the case of customers delivering samples above 150 per month and the Holt Winters model performs better for the other group of customers. However, choosing different measures for the

customer grouping before clustering can give a different result in comparison to what we have achieved but we leave this discussion for future research.

References

[1] Paul S.P.Cowperwait and Andrew V.Metcalfe: *Introductory Time Series with R.* Springer, 2009.

[2] Arvil Coglan: *A little book of R for Time Series*- Release 0.2 –April 2017.

[3] https://www.stat.berkeley.edu/~s133/Cluster1.html and

[4] https://www.stat.berkeley.edu/~s133/Cluster2a.html

[5] http://www.sthda.com/english/wiki/determining-the-optimal-number-of-clusters-3-must-known-methods-unsupervised-machine-learning