Feasibility of Automation in Price Setting for Tea Auction

Diganta Mukherjee, Abhinandan Dalal, Subhrajyoty Roy
Indian Statistical Institute, Kolkata, India
digantam@hotmail.com, msdalal18@gmail.com, roysubhra98@gmail.com

Abstract

Tea Auctions across India occur as an ascending open auction, conducted online. Before the auction, a sample of the tea lot is sent to potential bidders, and a group of tea testers. The seller's reserve price is a confidential function of the tea-tester's valuation, which also acts as a signal to the bidders. In this paper, we work with the dataset from a single tea auction house, J Thomas, of tea dust category, on 49 weeks in the time span of 2018-2019, with the following objectives in mind:

- Objective classification of the various categories of tea dust (25) into a more manageable, and robust classification of the tea dust, based on source and grades.
- Predict which tea lots would be sold in the auction market, and a model for the final price conditioned on sale.
- To study the distribution of price and ratio of the sold tea auction lots.
- Discussion on the possibility of automation of the process without human intervention.

Keywords: Tea Auctions, Applied Econometrics, Model prediction, Price Setting, Agriculture, Business economics

1 Introduction

Tea (Camellia sinensis), a widely popular manufactured drink consumed throughout the world, requires specific tropical and sub-tropical conditions to grow. Following China, India is the second largest manufacturer of tea in the world, producing about 1.2 million tonnes in 2014, which is about 30% of the world.

Tea industry in India is one of the oldest organized industries, with a large network of tea producers, retailers, distributors, auctioneers, exporters and packers; and a large demand for it, both domestic (India largest consumer of black tea consuming 1000 million kg in 2016) and in foreign markets , thus driving exports (worth annual 800 million USD, fourth largest exporter). Tea industry being very labor intensive also provides a key source of employment generation.

Tea is heterogeneous both over season and region, its demand varying over price and income, demographics such as age, education, cultural background; while supply heterogeneous with respect to quality, and not all quality grades available throughout the year. It's demand is very price sensitive with elasticities for black tea vary between -0.32 and -0.80. Although seemingly volatile, prices seem to depict specific patterns, peaking with arrival of new crop, and going down with increase in supplies till end of season. Price may alter eve due to producer brand and quality.

Overview of e-auctions of tea: An e-auction is a primary marketing channel for selling tea to the highest bidder. This (i) facilitates price discovery by bringing the buyers and sellers to a common platform (ii) provides a guaranteed transaction protocol for the transaction (including delivery of tea to warehouse, storing, sampling, bidding and payment).

Since September 2016, the auctions are pan India, allowing any registered tea-trader to participate. Manufactured tea is dispatched from various gardens/ estates to the auction centres for sale through the appointed auctioneers, where it is catalogued on the basis of their arrival dates within the framework of the respective Tea Trade Associations. Registered buyers, representing both the domestic trade and exporters receive samples of each lot of teas catalogued, generally a week ahead of each sale enabling the buyers to taste, inform their principals and receive their orders well in time for sale.

The auctioneers taste and value the tea for sale and these valuations are released to the traders. This acts as a base price for the bids. Given the above complexities, the aim in this report is to evaluate the feasibility of automating the pricing process to the extent of dispensing with the manual testing and valuation steps. We first discuss the clustering exercise according to Grade and source; followed by the pattern of saleability of different tea lots, the Price to value ratio to gain some insight into the pricing pattern, finally to be used in pricing models developed in. The final section deals with some comments on the feasibility of automation.

2. Data Description:

In this report, we have used J-Thomas data-sets on the weekly tea details for Kolkata Dust tea, Orthodox tea details, CTC (Cut, Tear, Curl) tea details and Darjeeling tea details. Moreover, we have the e-auction statistics as a part and parcel of the data-set. We have used the data in 2018 (38 weeks of data available, from Jan to Dec) for modelling, as the training data, and the 2019 data (first 12 weeks) has been used for cross-validation. We assume that our model is not varied by the effect of time in such a short span, and the cross-validation supports our assumption The e-auction statistics (2018-19) consists of the name of tea leaf type, total lots offered in auctions, total sold in packets and in quantity, and average price. For each such tea leaf type, detailed info on the weekly sale, total sold in packets and in quantity, and average price has also been provided. In the J-Thomas datasets, we find lot numbers (hence the difference between the maximum and minimum lot number would give us the number of lots offered), the categorical variable- the grade of the tea, number of packages offered, the valuation given by the agency, and finally the auction selling price. The 25 available grades are available in Appendix A, where some of the grades are clubbed together due to very few data points to finally have 14
grades, named in the same appendix.

3. Classification and Clustering:

3.1 Clustering Based on Grade:

The number of clusters so formed in the previous subsection is still quite large and cumbersome, and hence we aim to combine them maintaining distinctive characteristics. A crude measure could be based on correlation based on Volume Weighted median Prices, using the metric **Dissimilarity**$(d) = 2(1 - \rho^2)$ where ρ is the correlation coefficient, on the basis of which a hierarchical clustering approach may be appointed. However, the use of the median ignores most of the dataset, and the correlation being insensititive to the change of scale and origin, we use an alternative approach based on the following idea:

- First we use **Bayesian Information Criteria** to figure out the number of clusters so that the model has the largest information. This figure came out to be 6.

- Then, we used a Gaussian Mixture Model and ran the Expectation-Maximization Algorithm (often known in the literature as the EM-GMM clustering). We performed clustering both based on median as well as the mean, these two resulting in slightly different clusterings. We stick with the median based clustering due to its robustness.
- Both the price and the valuation was considered when applying EM-GMM clustering, in order to effectively capture the whole pattern present in the data.
- Then, we define a new correlation structure. We see that in the clusters we have formed, in how many weeks do the i-th and the j-th grade have occurred in the same grade.

Thus, we form a similarity matrix characterized by the following: The (*i*, *j*)th entry of the similarity matrix contains the number of weeks when the Grade *i* and Grade *j* falls into the same cluster as obtained from the previous step. The Mosaic plot corresponding to this similarity matrix is given in Figure 1. Finally based on this similarity matrix, we conduct a hierarchical clustering to obtain the clusters, shown in Figure 2.

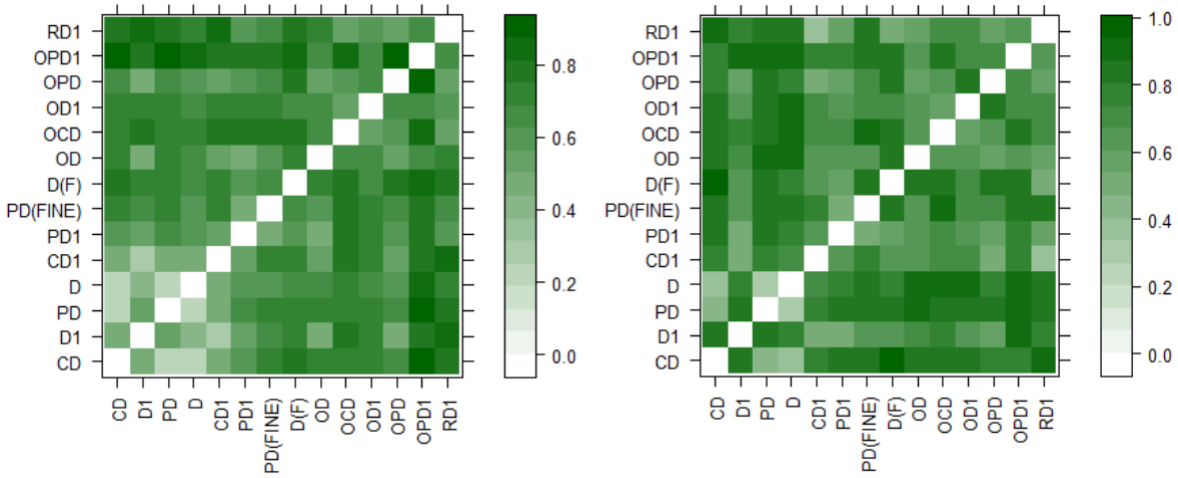


Figure 1: Mosaic plot for similarity matrix based on EM-GMM using Volume weighted Mean (left) and Median(right) Price & Valuations (Darker shades of green indicates higher degree of dissimilarity)

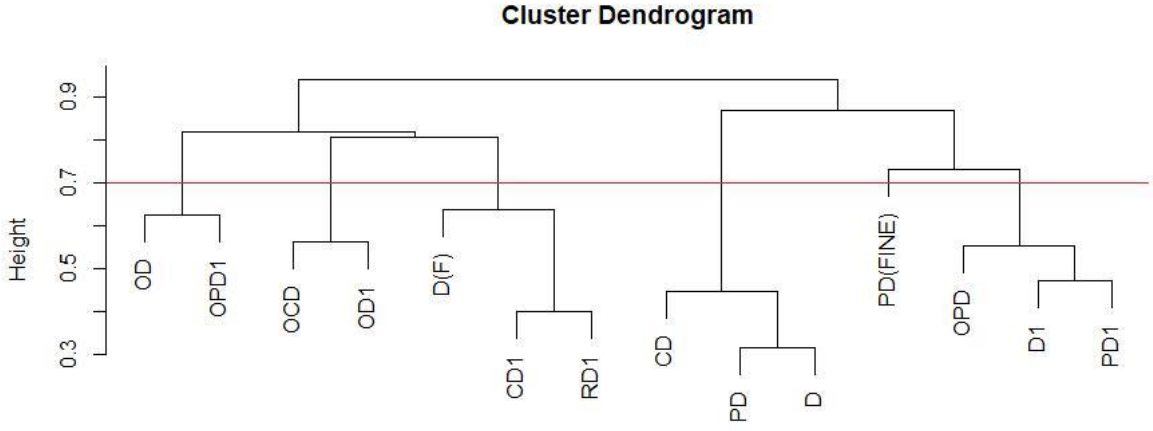


Figure 2: Dendogram for tea grades for volume weighted medians by EM-GMM

Thus we obtain the following **6 clusters based on grade** (while the GT Dust category has been left out due to lack of sufficient data points):

- **Cluster 1**: OD, OD-Special, OPD1
- **Cluster 2**: OCD, OCD1, OD1
- **Cluster 3**: D-Fine, CD1, CHD1, RD1
- **Cluster 4**: D, D-Special, CD, CHD, CHU, PD, PD-Special
- **Cluster 5**: PD-Fine
- **Cluster 6**: OPD, OPD-Clonal, ORD, D1, D1-Special, PD1, PD1-Special

Wikipedia [2] gives a different type of classification, of 8 categories, which differs slightly with the classification suggested by our data. We have dealt with this using which classification better explains the variance, in Appendix B, to finally stick with our original clustering.

3.2 Clustering by Source:

The source of the tea packets, although useful, is very cumbersome due to their large variety. There are 238 tea gardens (293 including their Clonal, Royal, Gold and Special variants) from which the tea has originated, keeping track of which is intractable and probably redundant; as the tea also show similarity in characteristics in terms of market behavior. Hence we undergo clustering based on volume weighted median to obtain dendogram given in Figure 3. These dendogram has again been clustered using the EM-GMM algorithm and then the time based similarity matrix as in the preceding section.

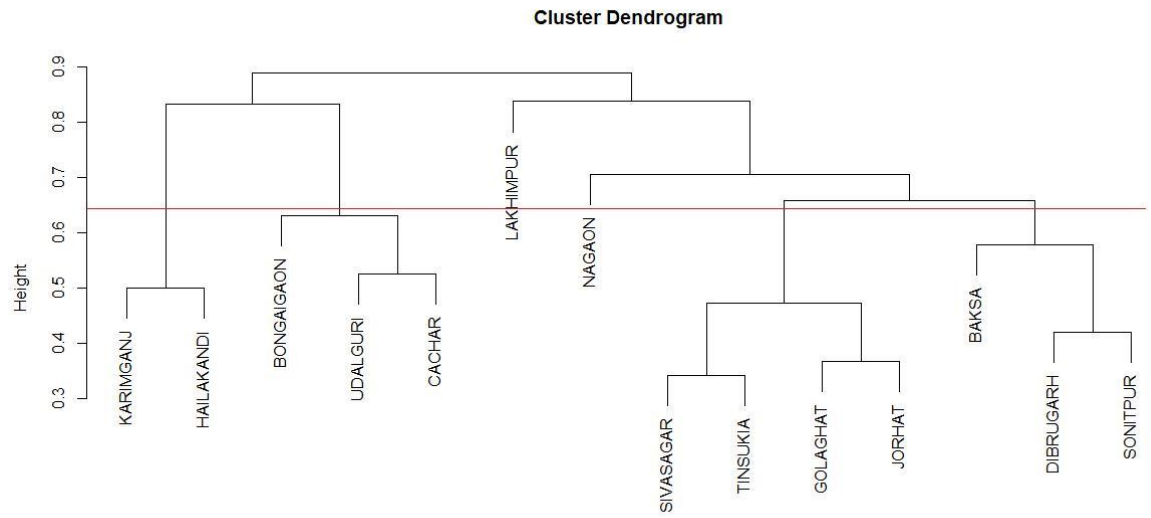


Figure 3: Dendogram for volume-weighted medians by source

Before going into the final source based clusterings, we need to keep the following things in mind as well:

- Labor factors and socio-economic factors of West Bengal and Assam vary distinctly.
- Geographical proximity of districts that belong to the same cluster makes more sense.
- Topographical, soil structural and water source of the tea producing regions.

- Tea dust of West Bengal is prominently dominated by Jalpaiguri, which suggests some succinct features different from rest of West Bengal.

The map of tea producing regions of West Bengal and Assam, in Figure 10 and Figure 9 in the Appendix H helps to visualize the proximity of the regions. Thus we finally use the following 7 clusters based on the source of the tea dust.

- **Cluster 1**: Darjeeling, Cooch Behar, Uttar Dinajpur, Jalpaiguri (4 West Bengal districts)
- **Cluster 2**: Karimganj, Hailakandi (2 districts)
- **Cluster 3**: Bongaigaon, Cachar, Udalguri, Darrang, Dima Hasao (5 districts)
- **Cluster 4**: Lakhimpur
- **Cluster 5**: Nagaon
- **Cluster 6**: Sivasagar, Tinsukia, Golaghat, Jorhat
- **Cluster 7**: Baksa, Dibrugarh, Sonitpur

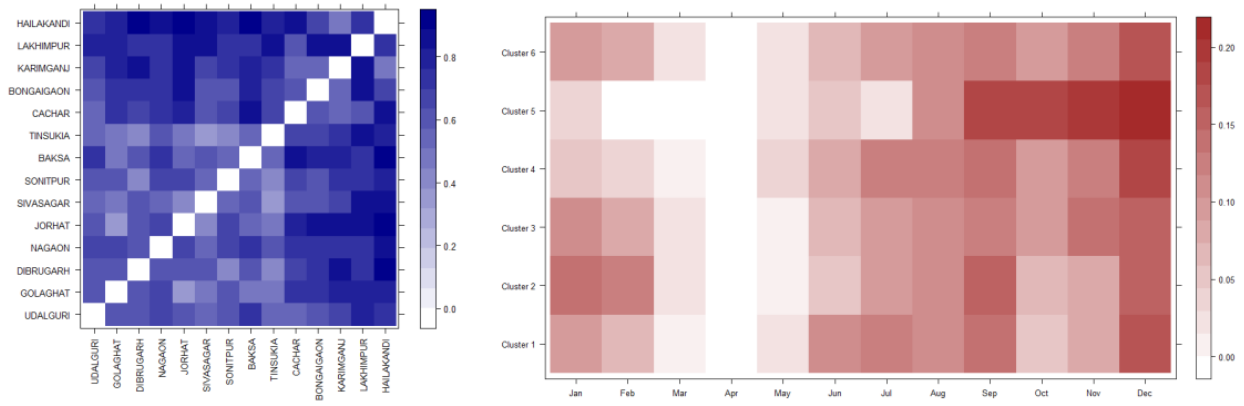


Figure 4: (Left) Mosaic plot for dissimilarity based on EM-GMM clustering for source (Darker shades for more dissimilarity) (Right) Mosaic plot for proportion of Volume of tea occurring for the grade clusters across months (Darker shades suggest larger proportions)

4. Predicting Saleability of Offered Tea Lots

4.1 Time Dependence of Sale:

Before moving on to predict saleability, a glimpse into the distribution of volume over different months may be useful, the values can be obtained in Appendix C and the plot in Figure 4. For analyzing the auction of tea packets, we should concern ourselves with the proportion of tea packets to be sold, and relate its valuation and several other characteristics to it. A successful attempt at predicting the probability of being sold (or being unsold) of an incoming tea packet based on its Grade, Source, Valuation and the current month, would give an insight to automate the auction process, alongside enabling an opportunity to study the effect of Valuation on determining the market characteristics of tea grades.

4.2 A primary inspection:

A primary inspection is made to see whether any particular type of tea grades are more likely to be sold at the auction than other type of tea grades. Appendix D shows the proportion of tea lots being sold and the total number of tea lots offered across different grades. The primary

inspection uses separate One-Way ANOVA models for both based on grade and based on gardens. The results show that Fine variant of a tea grade is offered more rarely than its original variant, and its probability of getting sold at the auction also increases. Also, the Special variant of any tea grade is more rare to be offered than its Fine variant, and hence, there is not number of observations is too small to be conclusive. From this, we note that if the tea packet has come from a Clonal tea garden, its selling probability is expected to be higher than Regular ones by 0.044, and the smaller value of p-value indicates evidence to support this claim. Similarly, the tea packets produced from Gold type variant of Garden is expected to be 15% less probable to be sold at the auction. Also, with 95% confidence, we can say that the tea packets produced from Royal type variant of Garden is 22% less probable to be sold.

4.3 Model for Prediction

In order to build a predictive model to predict whether a tea packet will be sold at the auction or not, based on its Valuation, Grade, Source and the current time, we use a mixture of logistic regression, which we present here. We have also tried by using simple logistic regression and generalized linear models, but the former shows the most promising results. The remaining results may be obtained from the authors on request.

We divide the total dataset of year 2018 into training and cross validation sets, with training set containing 70% of the samples. The cross validation set is used to select the model. The dataset of year 2019 is kept as the final test set, which is used to evaluate the performance of the finally selected model for prediction. The percentage of sold tea lots is kept almost similar for both the training and testing sets about 81%. For each cell, the proportion of sold tea lots is compared and assessed to obtain the final model, which turned out to be mixture of logistic. For analytical comparison, we use three measures as follows- **Null and residual deviance, RMSE** and **MAE**.

4.4 Results from Mixture of Logistic Regression

We try using a mixture of logistic regressions to predict the selling potential of the packets. Appendix E gives an idea of the model in use. Here, we have used our independent variable x to be the corresponding valuations, T being the number of packets arriving, and success denoting the event that the packet is sold. The concomitant variables are the source, grade cluster and month of the packets.

Here we consider 2 to 5 component mixture for this. Based on Bayesian Information Criterion, the 3 component mixture of logistic regression is chosen which yields a BIC value 6183.014. The following give the performance metrics of the chosen model.

- Updated model achieves RMSE of 0.214 and 0.2519 in the whole training set and testing set respectively.
- This achieves an MAE of 0.146 and 0.1684 in the whole training set and testing set respectively.
- Figure 5 provides the plots for fitted model for both training and testing set. Table 6 in Appendix E provides the summary of this model.

5. Distribution of price to valuation ratio for grade clusters

Valuation, by experts, does provide a significant predictor of what the final price of transaction would be. In fact, the entire process runs with the base price being set at some fixed

proportion of the valuations, and thus the following price of transaction revolves significantly across this measure. Now, we attempt to fit distributions over the ratio of price and valuations to account for its variability and shape of distribution curves. We have performed this exercise with the natural logarithm of the ratio of price and volume, to have a full support over the real numbers. The histograms of the ratio of price and valuations for various grade clusters as obtained from the data are given in Figure 5 and Figure 6. All but cluster 2 (OCD, OCD1, OD1) admit **a mixture of two log-normal distribution**, while the exception cited gives in to a **unimodal log-normal distribution**. The exact estimates of mixtures obtained can be found in Appendix F.

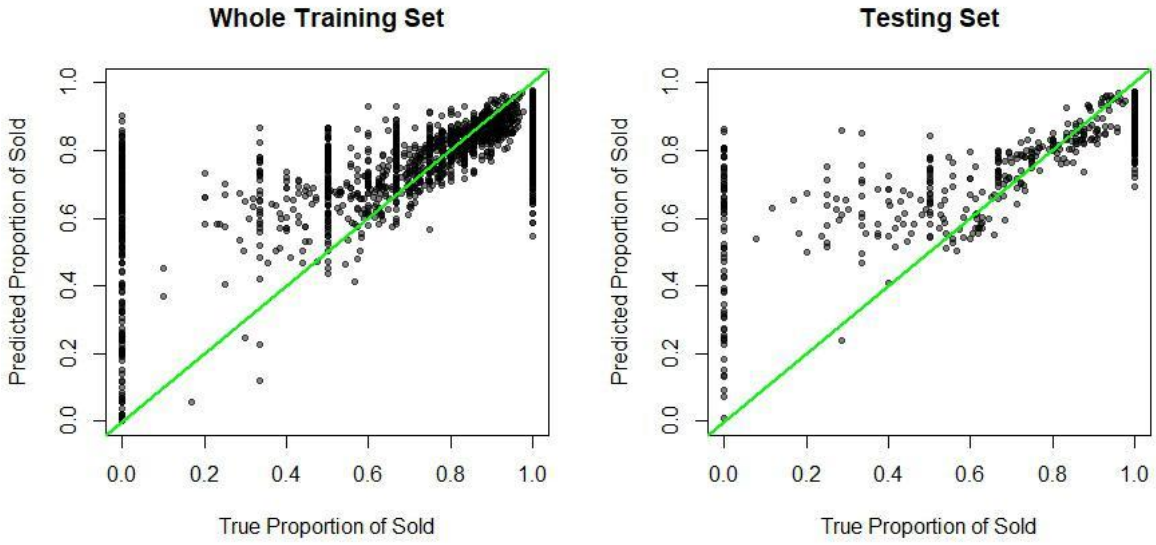


Figure 5: Fitted model of 3 component mixture of logistic regression for predicting sold grades for whole training and test sets

6. Analysis of the Price Model

To model the pricing system of the tea market, we consider modelling the demand side by consideration of Valuation of tea grades, its grade, source and the month in which the tea lots are available. On the other hand, to model the supply side of the market, we consider the volume of the tea lots as our main predictor. Therefore, our pricing model should include these variables.

To check whether the variant of the tea gardens (Clonal, Gold, Royal, Special etc.) should be included in the pricing model, we simply fit a one way Analysis of Variance model with Price as our response variable and the variant of tea garden as the possible treatment variable. It is found that this factor explains a sum of squares of 842405 on 4 degrees of freedom, yielding an F-statistic value of 102.41 and consequently extremely small p-value. Therefore, based on the data, we find a sufficient evidence to incorporate this factor into our pricing model in order to have better predictability.

For ease of interpret-ability, we start with a simple linear Analysis of Covariance model (ANCOVA) Ω_1 for pricing system with grade, source clusters, month of availability, variant of source garden, volume of the tea packet and valuation of the tea packets as our predictors. The model is given by the following, whose values of results are summarized in Table 7 in Appendix G. The takeaway is that all the factors turn out to be highly significant. On the other hand, to proceed with the aim of eliminating need for valuation, we use a competing Ω_2 model where we

model the logarithm of the price as a linear model of the remaining predictors. Note that we have avoided using a simple linear model in Ω_2 as the errors turn out to be heteroscedastic in that case. The results of Ω_2 are also provided in Appendix G Table 8.

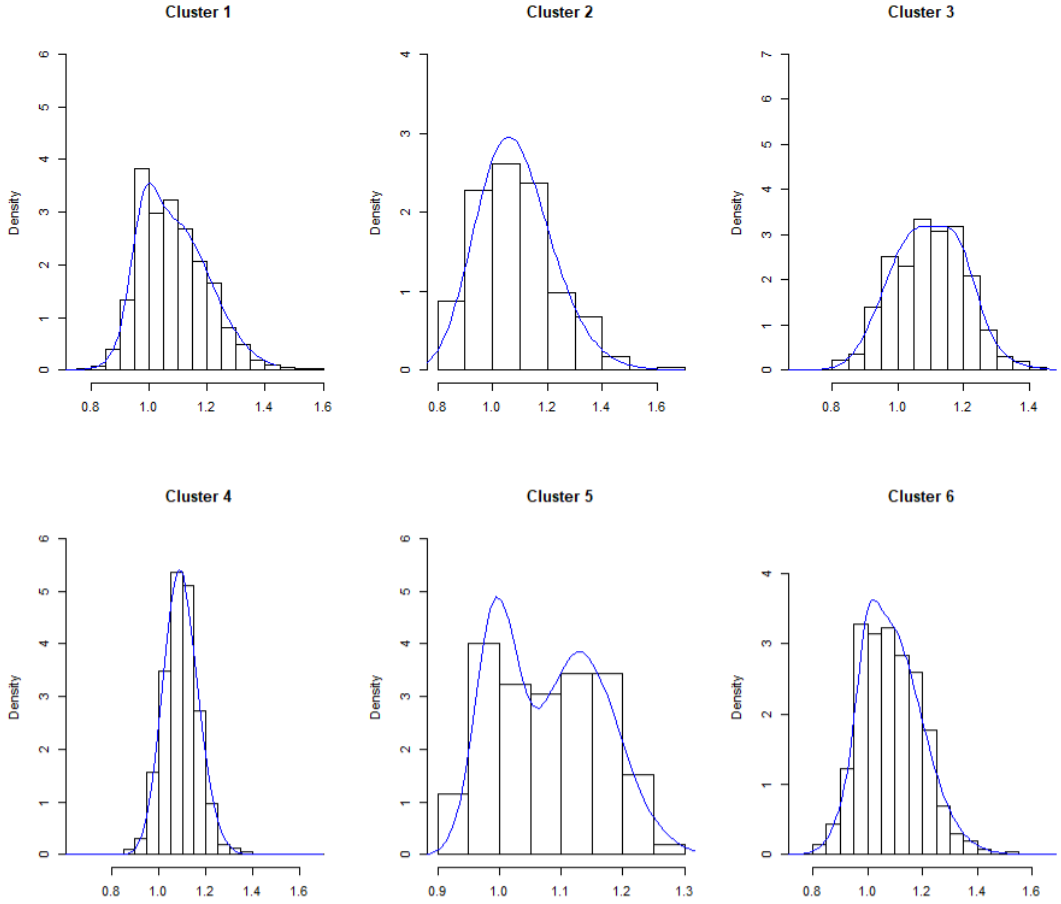


Figure 6: Histogram of Price/Value Ratio and Fitted Distributions for Clusters 1,2,3 (first row left to right) 4,5 and 6 (Second row left to right)

$$\Omega_1: Price = \beta_0 + \beta_{Grade} + \beta_{Source} + \beta_{Month} + \beta_1 Valuation + \beta_2 Volume + \varepsilon$$

$$\Omega_2: \log(Price) = \beta_0 + \beta_{Grade} + \beta_{Source} + \beta_{Month} + \beta_2 Volume + \varepsilon$$

where $\varepsilon \sim N(0,\sigma^2)$ independently and identically distributed in both the cases, and the variables with the variables in the suffix are categorical variables, taking a factor value for each of the possibilities. Note that both the garden and the source values are taken with respect to the classification obtained earlier. The following suggest the key features of the two models.

- Multiple R-squared for Ω1 is 0.9234 (excellent), while for Ω2 is 0.6431 (moderate). The drop in the R^2 is too high to be just attributed to 1 increase in the number of parameters.
- The ANOVA table finds all the predictors significant; however, the not so excellent fit suggests that the valuation as a factor is difficult to eliminate.

- Both the models find reasonably good qqplots and a few outliers towards the tail, in Figure 7 and in Figure 8
- Based on the evaluation of Ω_1 in testing dataset, the 2.5% quantile of the residuals is 8.59145, while the 97.5% quantile of the residuals is 23.43802. Hence, this pricing model approximately makes an error about 20 Rupees, to both positive and negative side, considering a robust measure of variation. A classical approach to measure the standard error, we find the interval, with predicted value as the center and an error of 24.3229 added (or subtracted) to both sides contains the true price for 95% of the time.
- Ω_2 on the other hand, based on a classical standard error estimation procedure, finds that the predicted prices lie between 71.09% and 140.65% of the true prices about 95% of the time.

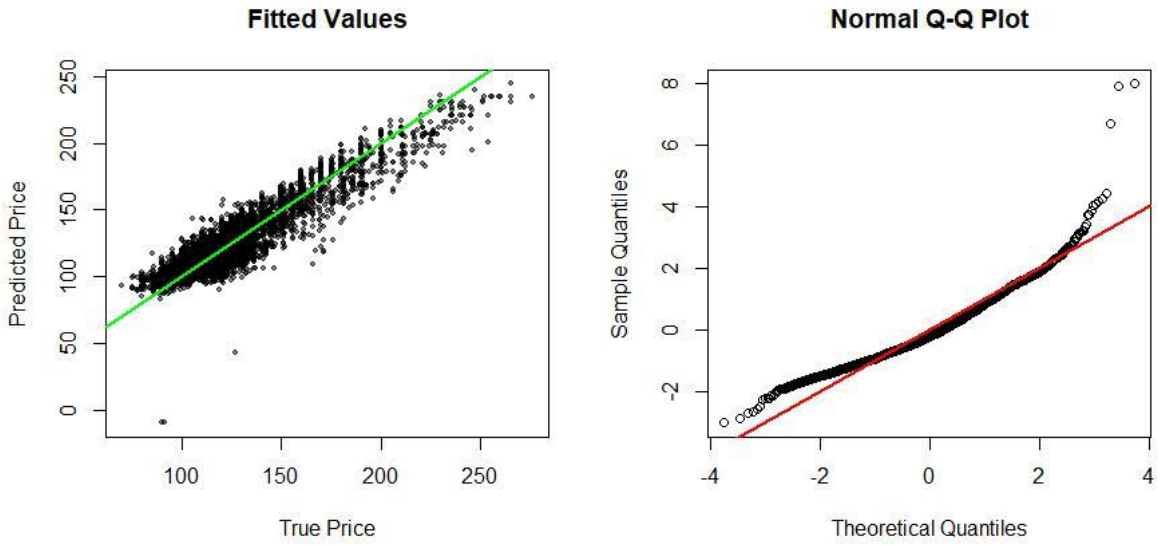
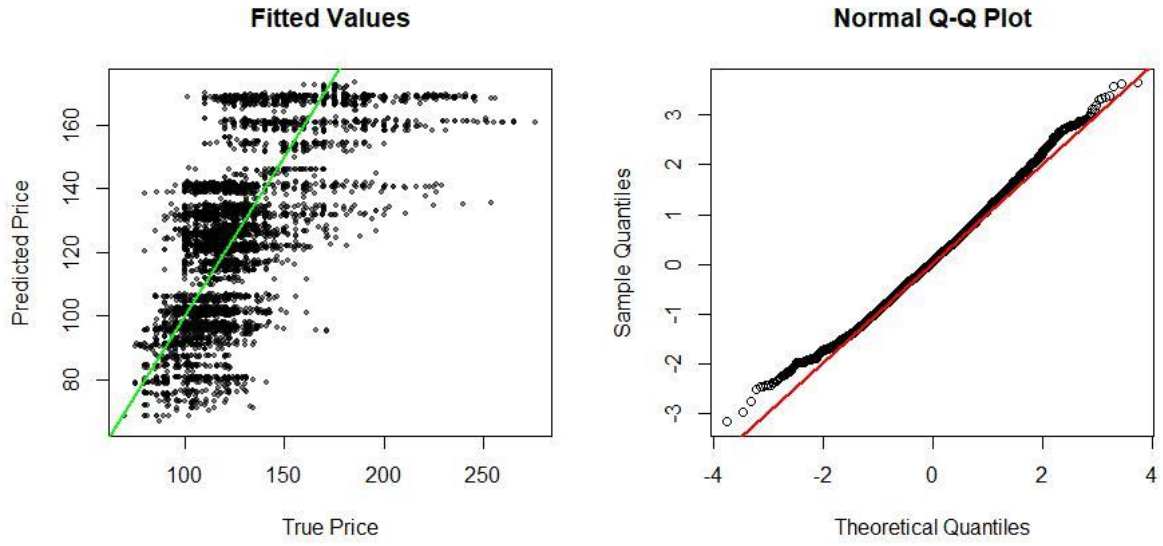


Figure 7: [Left] True Price vs Predicted Price for the fitted linear pricing model; [Right] The qqplot of standardized residuals for residual diagnostics

7. Remarks on Automation of the Process and Conclusion

The preceding section shows us the utmost significance of the manual valuation of the tea packets that come in, in predicting the final price level. However, a possibility of automation of the entire procedure, without any human interference in the auction mechanism, cannot be yet eradicated. Since the packets of tea are sold off one unit at a time, and to only one bidder, the stage can be set up as a single unit auction, with reserve prices from the seller's perspectives. We can safely assume, keeping in mind the finally high demand of tea dust among the consumers, that the bidders need not worry about the tea packets being sold off in the market, provided they set them at the correct competitive price. Furthermore, the pool of sellers can be considered large enough to ensure a perfect competition approximation to the final market scenario. Thus the entire model for the auction boils down to that of a **Common Value** (CV) Model, as given in [6]. The essence of this model lies in the fact that the buyers in this auction are not the immediate consumers of the product, but rather intermediate retailers who aim to sell them in the market. Furthermore, the competitive market assumption leads them to have no control over the final market price, and thus

the ex-post value is the same for all the bidders, since they expect to capture the same market share at the same price. However, the seller may set a reserve price that is above the manufacturing costs, so as to increase her revenue, obviously with the chance of forfeiting possible transaction. The valuation, with the help of the confidential formula for converting valuation to base price for the auctions, provide essentially such a reserve price for the seller.



.

Figure 8: [Left] True Price vs Predicted Price for the fitted log-linear pricing model; [Right] The qqplot of standardized residuals for residual diagnostics

However, since this valuation is based on the inherent characteristics of the tea dust packets, and the volume of packets that arrive, we strongly believe that the practice of using valuations to set base prices can be done away with. George and Hui, in their paper [7], provide an ingenious way to estimate demand in the auction market, under the Independent Private Value Model second price auctions. A generalization of this method, to the CV model, may be helpful in our case. Then, knowing the distribution of the bidder values, an optimal reserve price may be set to maximize the ex-ante expected revenue, which is a function of this distribution.

However, Levin and Smith, in their paper [8], have shown that under non IPV case, the optimal reserve price for the seller converges to her true value, ie, here, her manufacturing costs. Hence if the pool of bidders grow largely, then it would be safe for the seller to set the reserve price to be her own manufacturing costs (or whatever minimum price at which she would be willing to sell than possessing the good).

Collusion among the bidders is often a very practical problem to ponder about, and most methodologies fail under scenarios not robust to such behaviour. This is even a possibility here, given that auctions occur often and sequentially. Our pricing model with valuation, due to its excellent fit, provides a way to detect such mischievous behaviour on the part of the bidders. As in the case of collusion, the final price of transaction would be very low than that of the expected transaction price, large deviation from the predictions would detect such anomalies.

Further research in the aforesaid aspects would be helpful in the path of automation.

References

[1] *www.indiatea.org*
[2] *https://en.wikipedia.org/wiki/Tea_leaf_grading*
[3] *https://assam.gov.in/*
[4] *https://www.mapsofindia.com/maps/westbengal/westbengal-district.htm*
[5] **Bettina Grun and Friedrich Leisch**: Applications of finite mixtures of regression models by Regression examples in *cran.r-project.org/web/packages/flexmix/vignettes/*
[6] **Vijay Krishna**: *Auction Theory*
[7] **George, Edward, and Sam Hui** (2012*), Optimal Pricing Using Online Auction Experiments: A Polya Tree Approach*, Annals of Applied Statistics, 6(1), 55-82.
[8] **Levin D. and J. Smith**, *Optimal Reservation Price in Auctions* 1996, Economic Journal, Vol. 106, 1271-1283

Author Information:

Diganta Mukherjee is an Associate Professor in the Sampling and Official Statistics Unit, at Indian Statistical Institute, Kolkata. On the other hand, Abhinandan Dalal and Subhrajyoty Roy are both students of Masters of Statistics, in their first year, at Indian Statistical Institute, Kolkata.

Appendices:

A. Names of Original tea Grades:

CD: Churamani Dust, **CD1**: Churamani Dust 1, **CHD**, **CHD1**, **CHU**, **D**: Dust, **D(F)**: Dust Fine, **D(SPL)**: Dust Special, **D1**: Dust 1, **D1(SPL)**: Dust 1 Special, **GTDUST**: Golden Tea Dust, **OCD**: Orthodox Churamani Dust, **OD**: Orthodox Dust, **OD(S)**: Orthodox Dust (S), **OD1:** Orthodox Dust 1, **OPD**: Orthodox Pekoe Dust, **OPD(Clonal)**: Orthodox Pekoe Dust (Clonal), **OPD1**: Orthodox Pekoe Dust 1, **ORD** : Orthodox Red Dust, **PD**: Pekoe Dust, **PD(FINE)**: Pekoe Dust (Fine), **PD(SPL)**: Pekoe Dust Special, **PD1**: Pekoe Dust 1, **PD1(SPL)**: Pekoe Dust Special, **RD1**: Red Dust 1

We finally consider 14 grades by clubbing, due to almost indistinguishable features within each clubbed group, or lack of large enough data points.

- **CD**: CD, CHD and CHU
- **CD1**: CD1 and CHD1
- **D**: D and D special
- **D(FINE)**
- **D1**: D1 and D1 Special
- **OD1**
- **OD**: OD and OD-special
- **OPD**: OPD, OPD-Clonal and ORD
- **OPD1**
- **PD**: PD and PD-Special
- **PD(FINE)**
- **PD1**: PD1 and PD1-Special
- **OCD**
- **RD1**

B. Diagnostics of Grade Clustering:

We compare, for the clustering based on grades, the following two classification to find out which clustering better explains variance: and hence obtain Table 1 to finally adhere to the authors' clustering.

- Method 1: The authors' 6 clusters classification.
- Method 2: 8 clusters, cluster 1, 2, 3, 5 remaining as it is, while cluster 3 gets broken into two separate clusters, one containing grades PD, PD-Special and another containing CD, CHD, CHU, D, D-Special. Similarly, we divide cluster 6 into two separate clusters, one containing OPD, OPD-Clonal, ORD, D1, D1-Special and another containing PD1, PD1-Special.

Table 1: Proportion of Explained Variation of Weekly Price and Valuation by Clusters

Week	Explained Proportion of Variance			
	For Valuation		For Price	
	With 6 clusters	With 8 clusters	With 6 clusters	With 8 clusters
2	54.33%	54.81%	44.1%	44.8%
3	56.79%	57.28%	44.2%	44.5%
4	58.26%	58.77%	48.2%	50.1%
		:		
		:		
45	63.93%	65.27%	60.2%	61.4%
46	56.84%	57.09%	51.4%	51.6%

C. Distribution of Volume over Different Months

The data for 2018, presented in the form of weeks, have which then are put into buckets of months. This gives us the Table 2. To obtain Table 2, we find out the number of tea packets offered in a lot, and multiply it with net average weight of the tea packets to obtain the total amount (or volume) of tea offered. Then, for each cluster of grade, we find the proportion of its total volume which is offered during a specified month.

Month → Cluster ↓	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	0.099	0.070	0.008	0.000	0.027	0.108	0.129	0.115	0.139	0.051	0.084	0.171
2	0.134	0.125	0.018	0.000	0.010	0.052	0.091	0.109	0.155	0.062	0.086	0.156
3	0.105	0.082	0.019	0.000	0.008	0.069	0.100	0.114	0.123	0.089	0.132	0.160
4	0.056	0.037	0.009	0.000	0.031	0.077	0.125	0.129	0.138	0.090	0.129	0.179
5	0.038	0.000	0.000	0.000	0.018	0.053	0.025	0.109	0.177	0.184	0.191	0.205
6	0.100	0.085	0.000	0.000	0.015	0.059	0.098	0.115	0.128	0.096	0.126	0.162

Table 2: Tea grade proportions by months : Figures denote ratio of volume of each cluster grade occurring in that month and the total volume of all packets of tea grade of that cluster in the dataset

D. Primary Inspection of Saleability of Packets:

Table 3: Proportion of being Sold across different Grades

Grade	Total lots	Sold Proportion.		Grade	Total lots	Sold Proportion.
CD	1720	0.768		OD(Spl)	3	0.333
CD1	1400	0.819		OD1	125	0.896
CHD	5	0.8		OPD	1064	0.822
CHD1	23	0.696		OPD(Clonal)	3	1
CHU	5	1		OPD1	61	0.754
D	7932	0.816		ORD	8	1
D(Fine)	216	0.842		PD	7119	0.771
D(Spl)	6	1		PD(Fine)	111	0.945
D1	3477	0.842		PD(Spl)	23	0.609
D1(Spl)	1	1		PD1	500	0.872
OCD	206	0.888		PD1(Spl)	5	0.8
OD	2229	0.796		RD1	77	0.987

Table 4: Output for Analysis of Variance of Saleability on variant of Tea Grades

Coefficients	Estimate	Standard Error	t-statistic	p-value

Regular(intercept)	0.808	0.002721	297.20	2e-16
Fine	0.077	0.025131	3.10	0.00194
Special	-0.098	0.063752	-1.54	0.12347

Table 5: Output for Analysis of Variance of Saleability on variant of Tea Gardens

Coefficients	Estimate	Standard Error	t-statistic	p-value
Regular(intercept)	0.808	0.002881	280.653	2e-16
Clonal	0.044	0.010272	4.259	2.06e-5
Gold	-0.151	0.032931	-4.588	4.05e-6
Royal	-0.225	0.113282	-1.987	0.0469
Special	-0.027	0.013983	-1.96	0.05

E. Mixture of Logistic Regression

The model, in general, for a mixture of S components, is given by [5]

$$H(y|T, \boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\Theta}) = \sum_{s=1}^{S} \boldsymbol{\pi_s}(\boldsymbol{w}, \alpha)\mathrm{Bi}(y|T, \theta_s(x))$$

where **w** stands for the concomitant variables, on which the mixing proportions $\boldsymbol{\pi}_s$ depend, Bi(y|T,θs(x)) is the binomial distribution with number of trials equal to T and success probability $\theta_s \in (0,1)$, given by $\mathrm{logit}(\theta_s(x)) = \boldsymbol{x}^T\boldsymbol{\beta}^s$. The concomitant variable is assumed to be a multinomial logit model, ie, of the form

$$\boldsymbol{\pi_s}(\boldsymbol{w}, \boldsymbol{\alpha}) = \frac{\exp(\boldsymbol{w^T\alpha_s})}{\sum_{u=1}^{s}\exp(\boldsymbol{w^T\alpha_u})} \ \forall \boldsymbol{s}$$

For our model the mixture estimates are given in the following table:

Table 6: Summary of Mixture of Logistic Regression Fitted Model

	Component 1		Component 2		Component 3	
	Intercept	Valuation	Intercept	Valuation	Intercept	Valuation
Estimate	0.5355	0.007586	-1.48	0.0177	7.2017	-0.03344
Std. Error	0.1622	0.00092	0.1819	0.00092	0.3329	0.001882
Z value	3.3020	8.2702	-8.1369	12.778	21.635	-17.772
P value	0.0009	2.2e-16	4.05e-16	2.2e-16	2.2e-16	2.2e-16

F. Price to Volume Distribution Estimates

Cluster	Distribution	Ratios	μ	σ	Expectation	Variance	Fit
		0.191	-0.01	0.045	0.985	0.002	

Cluster 1	Mixture of 2 lognormal	0.809	0.099	0.106	1.111	0.014	Chisq pval: 0.227 KS pval:0.159
Cluster 2	Unimodal lognormal	1	0.073	0.127	1.084	0.019	Chisq pval: .267 KS pval:0.691
Cluster 3	Mixture of 2 lognormal	0.856	0.073	0.102	1.081	0.012	Chisq pval: 0.385
		0.144	0.175	0.043	1.193	0.003	
Cluster 4	Mixture of 2 lognormal	0.894	0.089	0.059	1.095	0.004	Chisq pval: .500 KS pval:0.324
		0.106	0.088	0.116	1.099	0.016	
Cluster 5	Mixture of 2 lognormal	0.4	-0.004	0.036	0.996	0.001	Chisq pval: .823 KS pval:0.989
		0.6	0.125	0.055	1.136	0.004	
Cluster 6	Mixture of 2 lognormal	0.092	-0.001	0.036	0.996	0.001	Chisq pval: .062
		0.908	0.087	0.103	1.096	0.013	

where chisq pval stands for the p-value corresponding to the chi-squared goodness of fit statistic, and KS pval for the Kolmogorov Smirnov goodness of fits pvalue.

G. Pricing Model ANCOVA Estimates:

Table 7: Analysis of Variance table for fitted linear pricing model with Valuation

	Df	Sum Sq.	Mean Sq.	F-value	p-value
Grade	5	20140487	4028097	25153.19	2.2e-16
Source	6	1696940	282823	1766.07	2.2e-16
Month	10	5473468	547347	3417.87	2.2e-16
Volume	1	62247	62247	388.7	2.2e-16
Source Variant	4	186240	46560	290.74	2.2e-16
Valuation	1	13191583	1319583	82373.97	2.2e-16
Residuals	21106	3379970	160		

Table 8: Analysis of Variance table for fitted linear logarithmic pricing model without Valuation

	Df	Sum Sq.	Mean Sq.	F-value	p-value
Grade	5	835.15	167.030	5566.005	2.2e-16
Source	6	52.15	8.684	289.375	2.2e-16
Month	10	248.20	24.820	827.086	2.2e-16
Volume	1	0.564	0.564	18.779	1.475e-5
Source Variant	4	5.37	1.342	44.728	2.2e-16
Residuals	21107	633.40	0.030		

H. Maps of Tea Producing Regions of West Bengal and Assam:

Figure 9: Map of North of West Bengal

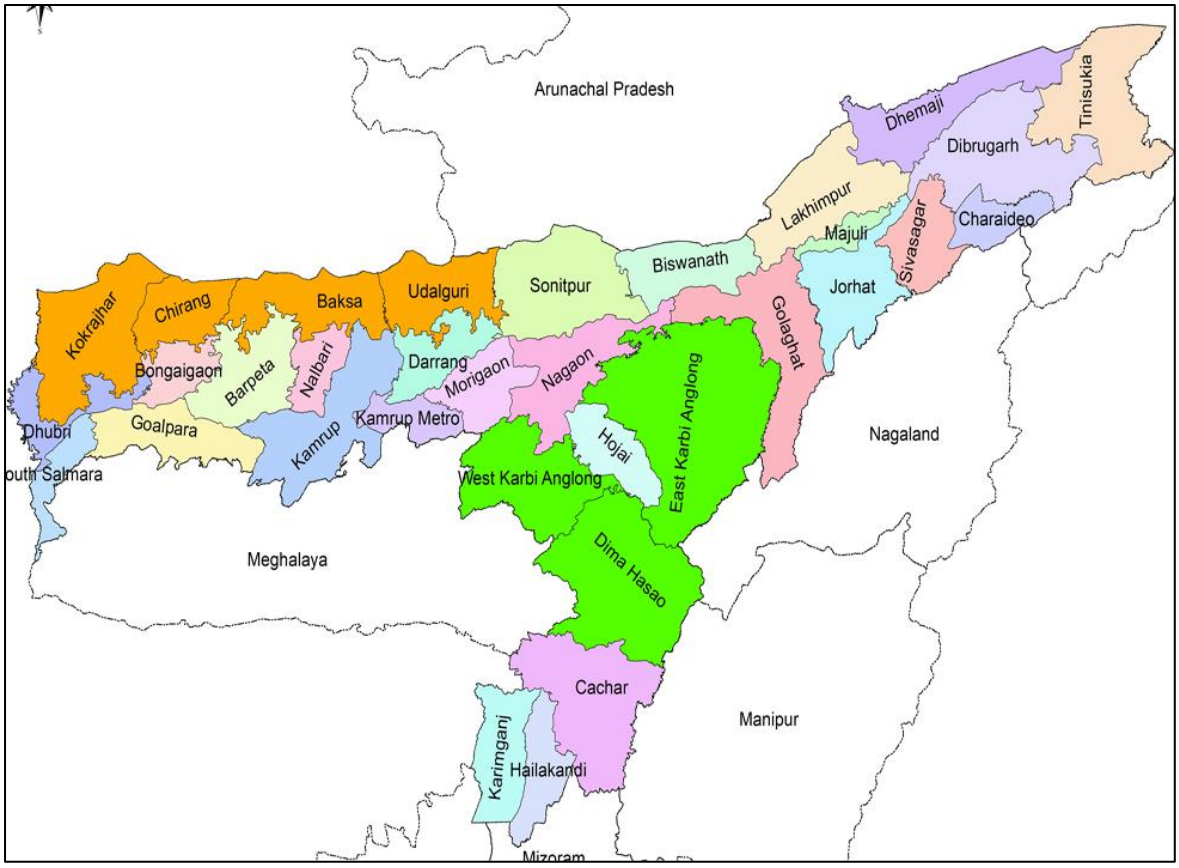


Figure 10: Map of Assam