4.1 Correlation parameters

In the Indian GST system, three types of taxes are collected, *viz*., CGST, SGST, and IGST.

- *CGST:* Central Goods and Services Tax is levied on intrastate transactions and collected by the central Government of India.
- *SGST:* State/Union Territory Goods and Services Tax, which is levied on intrastate transactions and collected by the state or union territory Government.
- *IGST:* Integrated Goods and Services Tax is levied on interstate sales and collected by the Central Government. Central Government takes half of this amount and passes the rest of the amount to the state, where corresponding goods or services are consumed.

The six correlation values used in clustering are derived from month-wise data in Table 1.

1. Correlation of *Total Sales Amount* and *Total GST Liability*: This is correlation coefficient between month-wise total sales value and month-wise total liability, which is equal to the sum of CGST, SGST and, IGST liabilities. We can derive this value by taking month-wise returns of taxpayers.

2. Correlation of *Total GST Liability* and *Total SGST Liability* : This is correlation coefficient between month-wise total liability and month-wise SGST liability.

3. Correlation of *Total SGST Liability* and *Total SGST paid in cash*: This is correlation coefficient between *SGST paid in cash*, which is equal to SGST Liability - (SGST ITC used to setoff SGST liability + IGST ITC used to setoff SGST liability), and SGST liability.

4. Correlation of *Total Sales Amount* and *Total SGST paid in cash*.

5. Correlation of *Total Tax Liability* and *Total ITC*: Total ITC is equal to the sum of SGST ITC, CGST ITC, and IGST ITC.

6. Correlation of *Total ITC* and *IGST ITC*.

4.2 Ratio parameters

These three ratio values used in clustering are derived from month-wise data in Table 1.

1. Ratio of *Total Sales* vs. *Total Purchases*: This ratio captures the value addition.

2. Ratio of $IGST\,ITC$ vs. $Total\,ITC$: This ratio captures how much purchase is shown as interstate or imports compared to total purchases.

3. Ratio of $Total\,Tax\,Liability$ vs $IGST\,ITC$.

4.3 TrustRank

Network of taxpayers: One of the independent variables in clustering is the *TrustRank*. To compute this, we created a weighted, directed graph(social network). Each vertex(node) in this directed graph corresponds to a taxpayer. We placed a weighted directed edge from taxpayer a to taxpayer b, where edge weight is the amount of sales done by the taxpayer a to taxpayer b during the period July 2017 to April 2019. Then the min-max normalization of edge weights is performed. For the same, we used the sales data explained in Table 1. This graph will capture the scale of interaction and(or) the exchange of money between taxpayers.

Computing Trust Rank: The TrustRank algorithm is a procedure designed to assign rate the quality to web pages [9]. The idea behind this method is almost similar to the PageRank algorithm. This method also takes the linking structure among web pages to generate a measure for the quality of a page. In this algorithm, we need to select some genuine web pages(seed set) as sources of trust initially. This is a manual process. This method then propagates the trust from seed pages to other pages based on the linking structure between pages. Trust is propagated similar to PageRank propagation.

We use the TrustRank [9] algorithm to assign weights to the dealers such that higher weights are assigned to the plausible genuine ones and lesser weights to the plausible fraudulent ones. We used the graph defined above towards this. This score is the 10^{th} and the last parameter used to performing cluster analysis.

4.4 Clustering Dealers

For obtaining the desired clusters, we have used different clustering techniques, and it is observed that spectral clustering works the best. Here we discuss more about the same.

Spectral clustering is a widely used technique for finding cliques in a weighted connected graph. It works well for sparse graphs. Spectral clustering uses eigenvalues of the similarity matrix to perform clustering in fewer dimensions after performing dimensionality reduction. In [16], Andrew, N *et.al.* discuss the scenarios and reasons for the better working of spectral clustering in spite of the fact that it uses K-means clustering algorithm.

4.5 Identifying Suspicious Taxpayers

We used kernel density estimation [17], which is a common technique in non-parametric statistics for the approximation of the unknown probability distribution of taxpayers within a cluster. We used Gaussian distribution as the kernel