**Clustering Approach**

Clustering approach is employed to group cities with similar characteristics together for the purpose of trip distribution data assignment which is required for each city in the UAM demand forecasting methodology. The trip distribution data provides information about the number of vehicles travelling a certain distance over a particular time period. Four parameters, namely land area, population density, GDP of current year (2019) and forecasted year (2050) are used for the clustering analysis. A statistical computer program, JMP, is used to find the optimal number of clusters needed to best represent a list of 544 cities using these four attributes. The optimal number of clusters is found to be 28 through the JMP analysis. Among the 28 clusters, 13 clusters are identified as single-city clusters and 15 clusters are multi-city clusters.

A representative city for all the clusters is identified. For single-city clusters the city itself would be a representative city while for multi-city clusters a representative city is identified by looking at city closest to the statistical centroid of the cluster (assuming equal weighting of all attributes). DLR guidance was also incorporated to refine the selected representative city in multi-city clusters.

OASyS pursued trip distribution data for the 28 cluster cities from commercial data supplier INRIX. There was np data coverage gap in some Asian countries like China, Japan, India and South Korea. This led to a data gap for six cluster cities from these Asian countries. Out of the six cluster cities, five of them are single-city clusters while one of them is a multi-city cluster containing three cities in it; which means that we have to re-assign eight cities to other clusters due to this lack of data availability. A statistical approach is used to merge these eight Asian cities into other clusters.

All four clustering parameters [land area, population density, GDP of current year (2019) and forecasted year (2050)] were normalised to adjust values on different scale to notionally a common scale. Mathematically, normalisation equation is represented as,

Xnormalised = (X – Xminimum ) / (Xmaximum – Xminimum)

Where, X: of the observed values present in X

Xmaximum: maximum values in X

Xminimum: minimum Values in X

After normalisation, the averaged mean of all the clustering parameters is computed by taking a ratio of summation of all clustering parameters to number of cities in that cluster. As an example, if a cluster K = 12 have 10 cities we simply add the normalised land area of all cities within cluster 12 and dividing it by 10 to get a normalised mean for cluster 12, and, similar approach is followed to get normalised population density, GDP of current (2019) and forecasted year (2050) for all the 28 clusters. It should be kept in mind that all the clustering parameters are equally weighted.

Once we compute the mean of all normalised clustering parameters, error sum of squares (SSE) method is used to identify that which cluster out of 28 does the city of interest (Tokyo, Seoul, Beijing, Shanghai, Guangzhou, Osaka, Delhi and Mumbai) best resemble. Further, city of interest is merged in that cluster which means that newly merged city will have same trips distribution data as other cities in that cluster. As an example, the normalised clustering parameters of city of interest are compared to the sum squared error of the other 27 clusters and where the sum squared error is minimum that cluster was selected as a best match to merge the city of interest.

Mathematically, sum squared error is computed as,

SSE = [{ αcoi - αCluster}2 + {βcoi - βCluster}2 + {µcoi - µCluster}2 + {Δcoi - ΔCluster}2]

Where, α: Normalized Land Area

β: Normalized Population Density

µ: Normalized GDP 2050

Δ: Normalized GDP 2019

coi: City of interest

After execution of this step, the eight cities are merged into their statistically most-similar neighboring cluster, the results are indicated in the table below. This yields a final set of 22 representative cluster cities. It should be noted that there is no inherent hierarchy among the clusters, they are simply delineated by their properties in the four clustering parameters mentioned earlier and indexed for convenience.

**Representation of Merged Asian Cities using SSE Method**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 10** | **Cluster 11** |
| New York | Jakarta | Los Angeles | Hong Kong | Istanbul |
| Tokyo | Shanghai | Seoul | Bogota | Bangkok |
|  |  |  | Ahmadabad | Mexico City |
|  |  |  | Sura | Sao Paulo |
|  | Representative City |  | Aleppo | Beijing |
|  | Merged Asian City |  | Vijayawada | Osaka |
|  |  |  | Chittagong | Delhi |
|  |  |  | Asansol | Guangzhou |
|  |  |  | Mumbai |  |