## Appendix 1

### *Distance metrics for k-means clustering*

The choice of distance measures is a critical step in clustering. It defines how the similarity of 2elements (x, y) is calculated and will influence the shape of the clusters. In this study, the impact of 4 different distance metrics on clustering results was evaluated. The following table contains the description of different distance metrics and their formulas.

| Distance metric | Description | Formula |
|---|---|---|
| Sqeuclidean | Squared Euclidean distance. Each centroid is the mean of the points in that cluster. | $\mathrm{d}(\mathrm{x,y})=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$ (1) |
| Cityblock | Sum of absolute differences. Each centroid is the component-wise median of the points in that cluster. | $\mathrm{d}(\mathrm{x,y})=\sum_{i=1}^{n}\lvert(x_i-y_i)\rvert$ (2) |
| Cosine | 1 minus the cosine of the included angle between points. Each centroid is the mean of the points in that cluster, after normalizing those points to unit Euclidean length. | $\mathrm{d}(\mathrm{x,y})=1-\frac{\lvert\sum_{i=1}^{n}x_iy_i\rvert}{\sqrt{\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i^2}}$ (3) |
| Correlation | 1 minus the sample correlation between points (treated as sequences of values). Each centroid is the component-wise mean of the points in that cluster, after centering and normalizing those points to 0 mean and unit standard deviation. | $\mathrm{d}(\mathrm{x,y})=1-\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$ (4) |

### *Centroid initialization methods for k-means clustering*

K-means clustering aims to converge on an optimal set of cluster centers (centroids) and cluster membership based on distance from these centroids via successive iterations. It is intuitive in that the more optimal the positioning of these initial centroids, the fewer iterations of the k-means clustering algorithms will be required for convergence. This suggests that some strategic consideration to the initialization of these initial centroids could prove useful. Four methods for centroid initialization were tested in this study, as follows:

| Method | Description |
|---|---|
| Plus | First, a data point is randomly selected from the input data set, and its distance from the nearest cluster center is then calculated. A new data point with larger distance is then selected as the new cluster center. This procedure is repeated until k cluster centers are selected. The idea is that the initial seeds should be as far away from each other as possible. |
| Sample | The k cluster centers are selected from the data set randomly. |
| Cluster | Perform a preliminary clustering on a subset of 10% data points. This preliminary clustering is initialized using the 'sample' method. If the number of data points in the subset is less than k, then k data points are selected from the data set randomly. |
| Uniform | Select k data points uniformly and randomly from the range of data set. |

https://dx.doi.org/10.21037/qims-22-825