1- Background:

1.1- Clustering by correlation

Clustering is the problem of dividing the data into groups based on a specific similarity factor while correlation is the definition of how similar these items are in a group or a cluster, so Correlation Clustering is to calculate the correlation between the items (cells in our problem) and associate the highly correlated cells into an optimum number of clusters.

Hierarchical Clustering initiates each object as a cluster. Then, it merges the most similar clusters together as one cluster, this process will loop until all the clusters are merged.

2- Design

2.1- Traffic Forecasting

After receiving the preprocessed data, we will feed it to the first step of the proposed model which is the Long-Short-Term-Memory (LSTM), this model will serve as a time series analysis model to forecast both of the independent features (SPEECH_TRAFFIC) and (DATA_TRAFFIC). These two features have no dependency on any other features however they vary through time. In our dataset, there are 30 records for every cell representing the monthly succession of the traffic on this cell, that’s why we will use the time series analysis model (LSTM) to forecast both of these features as we found that it will be the most suitable model based on its functionality that we defined previously.

Our challenge was the nature of the dataset which represents the data as repeated time stamps across various cells, the first 30 records represent the first cell the second 30 represent the second cell, and so on, this nature mismatches with the nature of the time series analysis problems so we tried to find a workaround. First of all, we tried to make one-cell models which means that we would train a model for every cell, which was a valid solution, but of course not practical at all since that we have over 19,000 cells which requires around 16,000 models. Subsequently, we handled this problem by using **Clustering Techniques**: These cells would be clustered to a rational number of clusters. First, we had to calculate the correlation between the SPEECH and DATA TRAFFIC for all cells, then highly correlated cells were segmented as one cluster using **SciPy Cluster Hierarchy Linkage**, to achieve this task we had to vectorize the cells into SPEECH_DATA_TRAFFIC vectors then every cell was represented as a vector, this vector represented the SPEECH traffic values followed by the DATA traffic values, finally, we had around only 16 clusters out of over 19 thousand of cells. Highly correlated cells are cells that have a similarity in distribution which means that all cells that have somehow similar behavior of ups, downs and, trends will be treated as correlated and will be segmented as one cluster like what is shown in the figure

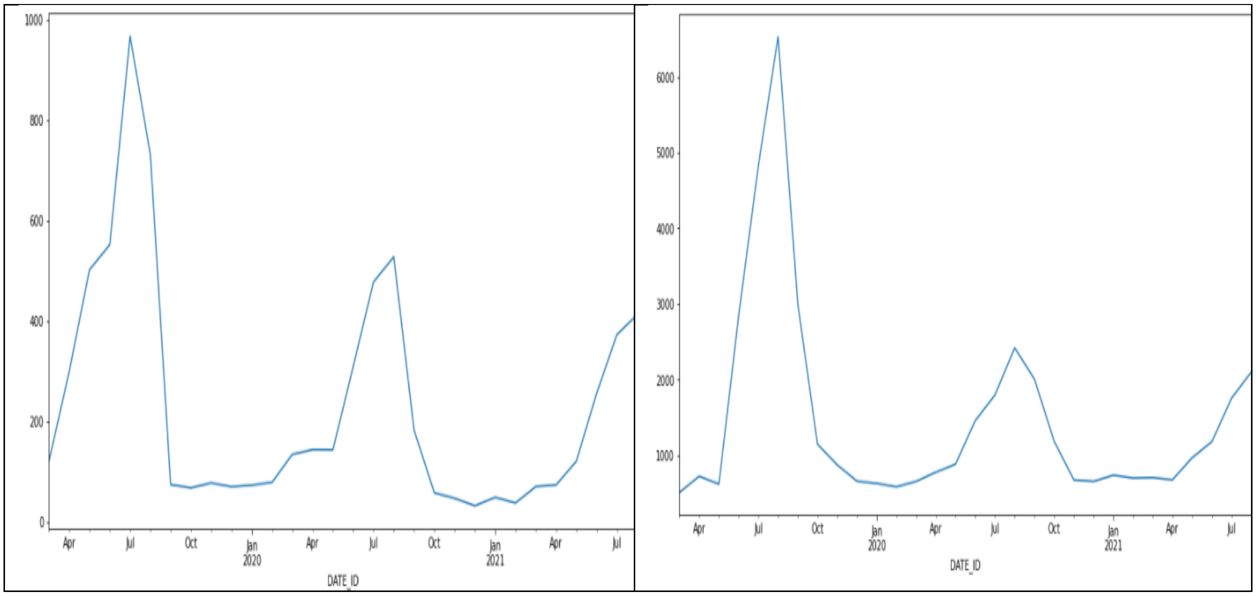

Figure 1SPEECH TRAFFIC for A1202 and A1218 Cells

After clustering cells to highly correlated cells, we made custom datasets to feed them to the comprehensive optimized general LSTM models. Simply, we merged the records of all cells of the same cluster which will give the model a better opportunity to be exposed to much more various records around this cluster's cells which will significantly improve the generalization of the model.

Time Series Analysis: Then, it was time to train a model for every bunch of clustered cells, and since we had the time series of the DATA_TRAFFIC and the SPEECH_TRAFFIC so, we had to train two models for every cluster, one model for each feature.

The output of this stage would be the input of the following stage, as we would feed the DATA and SPEECH traffic to the regression model that should predict the utilization.

3- Results

3.1- Traffic Forecasting

After exploring the results, we found that the integration of the cells had widened the range of the intersected values which caused a fine deduction of the performance, this was expected since previously we were dealing with every cell independently, but after the integration, we dealt with every bunch of clustered cells, so we can say that we sacrificed fine performance for the sake of the generalization and practicality.

As for the SPEECH_TRAFFIC, the results were satisfying enough despite the lack of records. Here are some of the SPEECH clusters' results:

Cluster \ Metric	Train RMSE	Test RMSE	Test MAE	Test MAPE
Speech Cluster 1	352.77	498.96	198.93	17.71 %
Speech Cluster 2	792.45	1103.88	551.91	31.72 %
Speech Cluster 3	296.41	681.87	359.05	36.84 %
Speech Cluster 4	750.56	1406.34	960.53	28.43 %

After discovering the DATA_TRAFFIC values we found a massive variance of values (from 0 to around 16 million) that led to model disturbance and bias. Subsequently, we tried to handle this problem by normalizing and standardizing the values, however, the variance was still huge. So as future work, we planned to perform additional value-based clustering for the DATA values so that we can slice every bunch of cells with similar values together to overcome this variance and enhance the results to reach the initial approach's results.

Here are the average results we reached using a model-for-every-cell approach:

MAE	MAPE	RMSE
230	20%	360

4- Evaluation

4. 1- Traffic Forecasting

We used the evaluation metrics of the Mean Absolute Error (MAE) and also the Root Mean Squared Error (RMSE) for both of the SPEECH_TRAFFIC and the DATA_TRAFFIC, but these metrics may be somehow misleading because of the variance between clusters values. This means that the same error may be acceptable for a cluster while it's not acceptable for another one. That's why we used the Mean Absolute Percentage Error (MAPE) to act as a normative metric.

The average of the MAPEs for the SPEECH_TRAFFIC is around 20% which is very satisfying for the operator.

References:

https://www.displayr.com/what-is-hierarchical-clustering/

https://en.wikipedia.org/wiki/Correlation_clustering

https://coderedirect.com/questions/207814/how-do-i-get-the-subtrees-of-dendrogram-made-by-scipy-cluster-hierarchy

https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html

https://en.wikipedia.org/wiki/Correlation_clustering#:~:text=Clustering%20is%20the%20problem%20of,specifying%20that%20number%20in%20advance.