## Cluster models output

**DBSCAN**

*Best parameters: eps=500, min\_samples = 5*

*Model output: #-clusters: 6*

A colorful calendar with numbers

Description automatically generated with medium confidence

*External evaluation*:

Prediction accuracy MAE: 27.478

Prediction accuracy MAPE: 0.377

**GGM Model**

*Best parameters: n\_clusters = 3*

*Model output:* *A blue and yellow squares with black text

Description automatically generated*

*External evaluation*:

Prediction accuracy MAE: 31.468

Prediction accuracy MAPE: 0.410

**KMeans Model**

*Best parameters: 'n\_clusters': 5, 'init': 'random', 'n\_init': 'auto', 'algorithm': 'lloyd'*

*Model output:*

A colorful squares with months

Description automatically generated with medium confidence

*External evaluation*:

Prediction accuracy MAE: 28.394

Prediction accuracy MAPE: 0.298

**Agglomerative Model**

*Best parameters: 'n\_clusters': 5, 'metric': 'euclidean', 'linkage': 'ward'*

*Model output:*

**A colorful squares with numbers

Description automatically generated with medium confidence**

*External evaluation*:

Prediction accuracy MAE: 28.653

Prediction accuracy MAPE: 0.290

**Building models:**

To build the different models, the dataset was transformed to remove the missing values. After this step, the dataset has been implemented in the different model queries.

**Finding parameters**

Different methods have been used to determine the parameters of the models. For the DBSCAN, the min\_sample was set to 5, and a 5-NN distance graph has been used to determine the epsilon value. The value for the epsilon was around 750. However, when implementing this value, the model provided only 2 clusters, which did not align with literature. Therefore, a manual search has been started, which ended with an epsilon of 500 as a result.

For the GGM model, the BIC elbow method was used to determine the number of clusters. In this case, the number of clusters was 3. Other parameters were kept at their default value.

For both the KMeans and Agglomerative model, a gridsearch was used to determine the best parameters. The chose was based on the Calinski-Harabasz scores as these scores show that clusters are well dense and well separated from each other. Both models ended up with 5 clusters.

**Best model:**

According to the external evaluation, the KMeans and Agglomerative models perform equally good for this dataset. The models have similar outputs for the MAE and MAPE, and while not having the best value for the MAE, they do have the best value for the MAPE. Looking at the clusters with an expert view, the clusters in both models have logic separation, which can be substantiated with literature. For example, the models show good separation between the winter and summer seasons, weekdays and weekend days, and holidays.