Dresden Database
Systems Group

Master Thesis Defense

Systematic Analysis of Impact of Aggregation On Time Series Forecasting

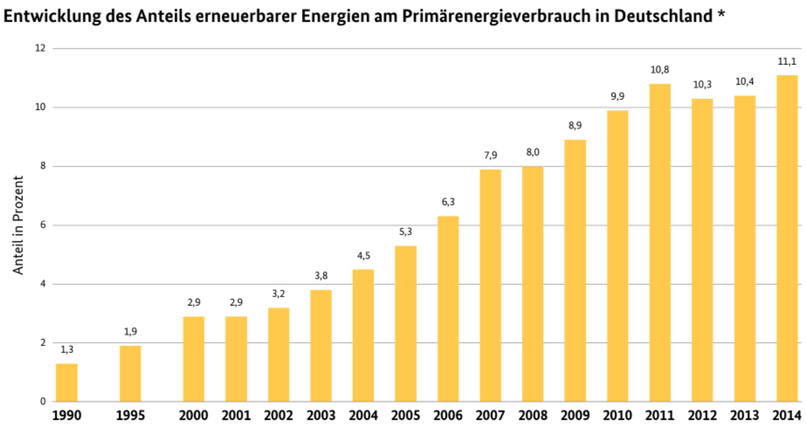
by:
Sachin Vittal

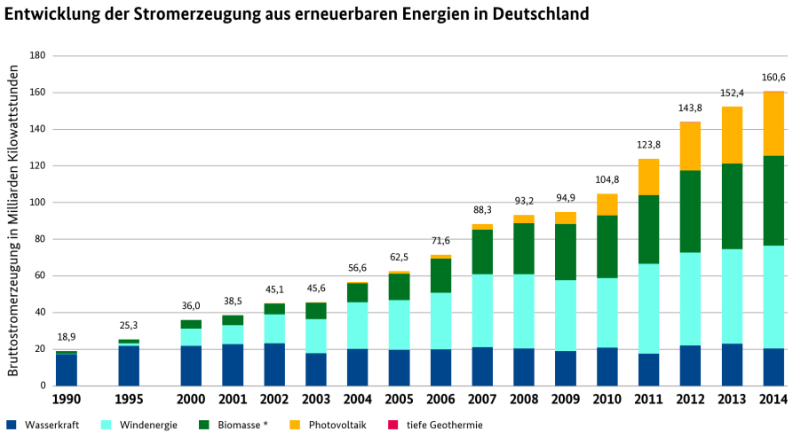
Supervised by:
Prof. Dr.-Ing. Wolfgang Lehner

Motivation

Need For Renewable Energy Forecasting

- Renewable energy source as major source of energy
- Wind and Solar are major contributors
- Fluctuation in Production
- Need for Renewable Energy forecast model





Motivation

GOAL

- Development of a reusable forecast model
- Data Preprocessing and reduction of input space
- Improved Forecast Results and Performance

ISSUES

- Issues with Wind Energy Forecasting
  - Nonlinear relation b/n Power output and wind
  - Individual site forecast errors are amplifed
- Aggregation of output from Ensemble of sites

RESEARCH QUESTION

- Can Aggregation of Power Time Series with high similarity leads to better forecast result
  - If yes, what is the criteria for finding the Time Series with high similarity
- Clustering and Aggregation of Time Series lead to better forecast results along with better performance

Approach

Time Series Subset

AGGREGATION

- Analyze the behavior of Time Series Forecast results on Aggregation
- Choose a small subset of Time Series
- Aggregate them at different sizes from lowest to highest
- Find the *Similarity Measure* between Time Series in each aggregate
- Perform forecasting of each aggregate and calculate the error
- Analyze the relationship between the *Similarity Measure* and *Forecast Error*

CLUSTERING

- Analyze the behavior of a set of Time Series upon clustering
- Apply standard clustering algorithm on a small subset
- Aggregate Time Series in each cluster and perform forecasting.
- Compare it with the Forecast result of Individual Time Series in the Subset

Datasets & Tools

Dresden Database Systems Group

NREL Wind Integraion Datasets

- Eastern & Western Integration Dataset
- Eastern Dataset contains data from ~1300 sites
- Modelled using mesoscale Numerical Weather Prediction(NWP) Model
- Data of 3 years from 2004 to 2006 of 10 minutes
- Wind Speed(m/s) and Power(MW)
- Geographic Coordinates of each site given

Tools

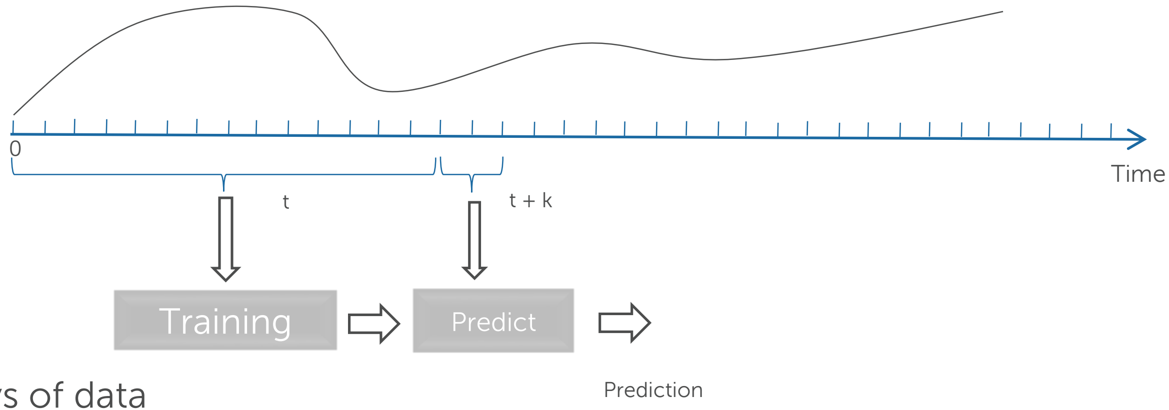
- R statistical environment
- Models were built using standard R packages
- MySQL used for data storage

Forecasting Method

Dresden Database Systems Group

Methodology

- Sliding window approach
  - Window Size = 24 hours
- Very Short forecasting
  - Model Training
  - Prediction
- Prediction in steps of 1 hour
- Total Length of 10 days and 50 days of data



Models used

- Gradient Boosted Model (gbm)
- Multivariate Adaptive Regression Splines (MARS)
- Multi Layered Feed Forward Neural Network (MLP)
- Bayesian Regularization of Neural Network (BRNN)

Regeression Based

Neural Network Based

Experiments Overview

Dresden Database Systems Group

Aggregation Experiments

- Aggregation of Subset of Time Series
- Deduce the Correlation between Similarity Measure and Forecast error

Clustering experiments

- Hierarchical Clustering
- Clustering & Aggregation at different levels
- Application on datasets of different sizes
- Performance Analysis

Aggregation Experiments

Details

Dresden Database Systems Group

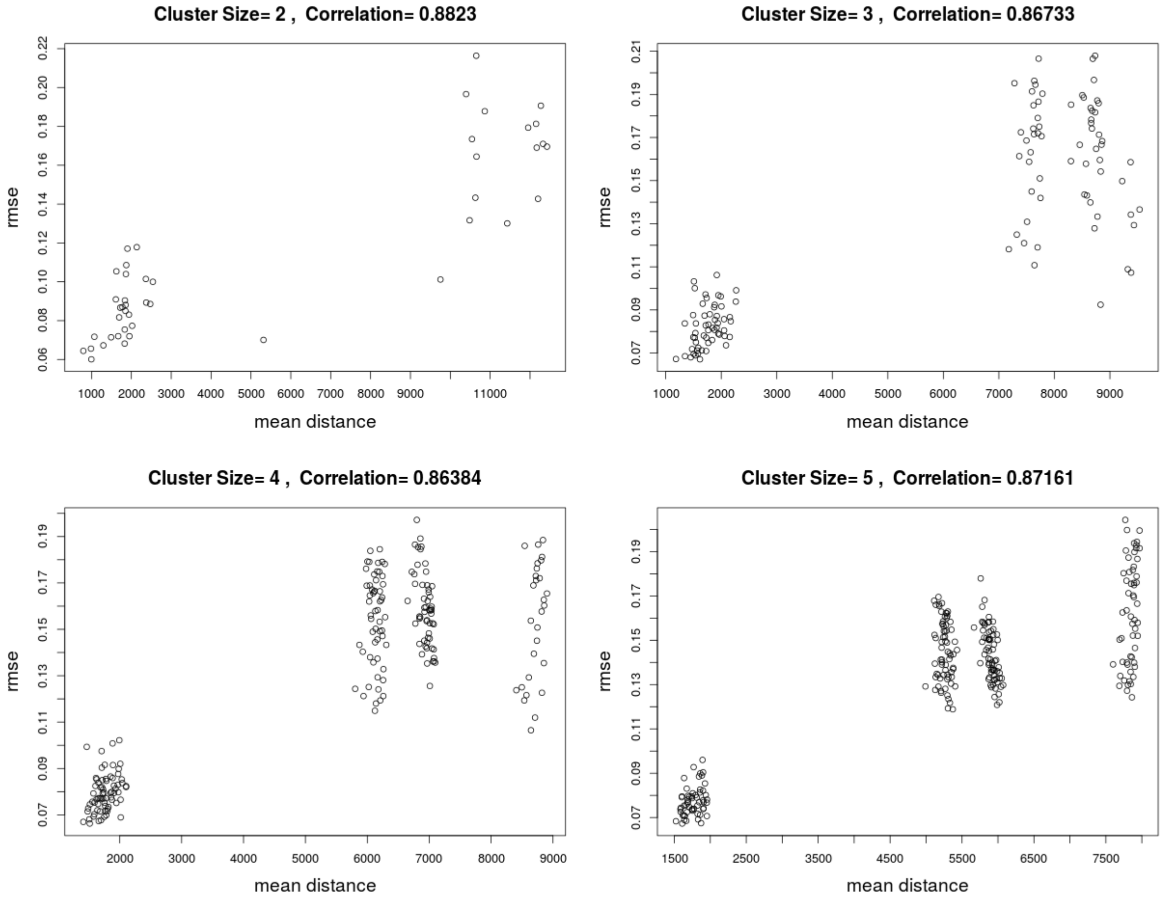
Goal

- Analyze the relationship between Time Series *Similarity Measure* and *Forecast Error*
- Idea: Lower the Similarity Measure then lower the Forecast Error
- Analyze this behavior using different Standard Similarity Measures

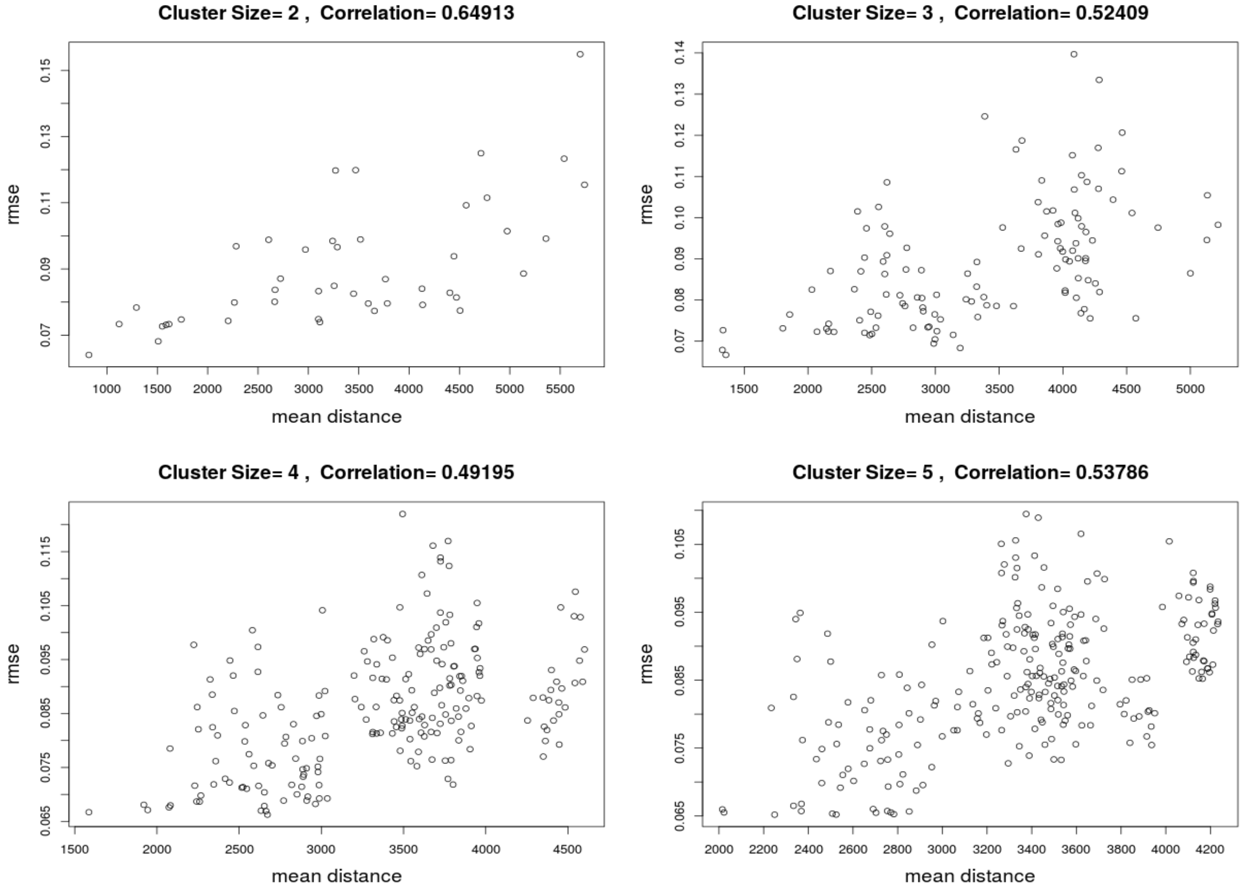
Methodology

- 10 different subset of Time Series of size 10 is chosen
  - 5 subsets are randomly chosen
  - 5 subsets belong to a specific geographic location
- Each subset is aggregated at different sizes from 1 to 10
- Forecasting is done on all the *combinations* of aggregates of each size
  - RMSE is calculated
  - Lengths = 10 days & 50 days
- Different standard Similarity Measures are used
  - Like Euclidean Distance, Fourier Distance and Principal Components
  - Weighted mean of Similarity Measure

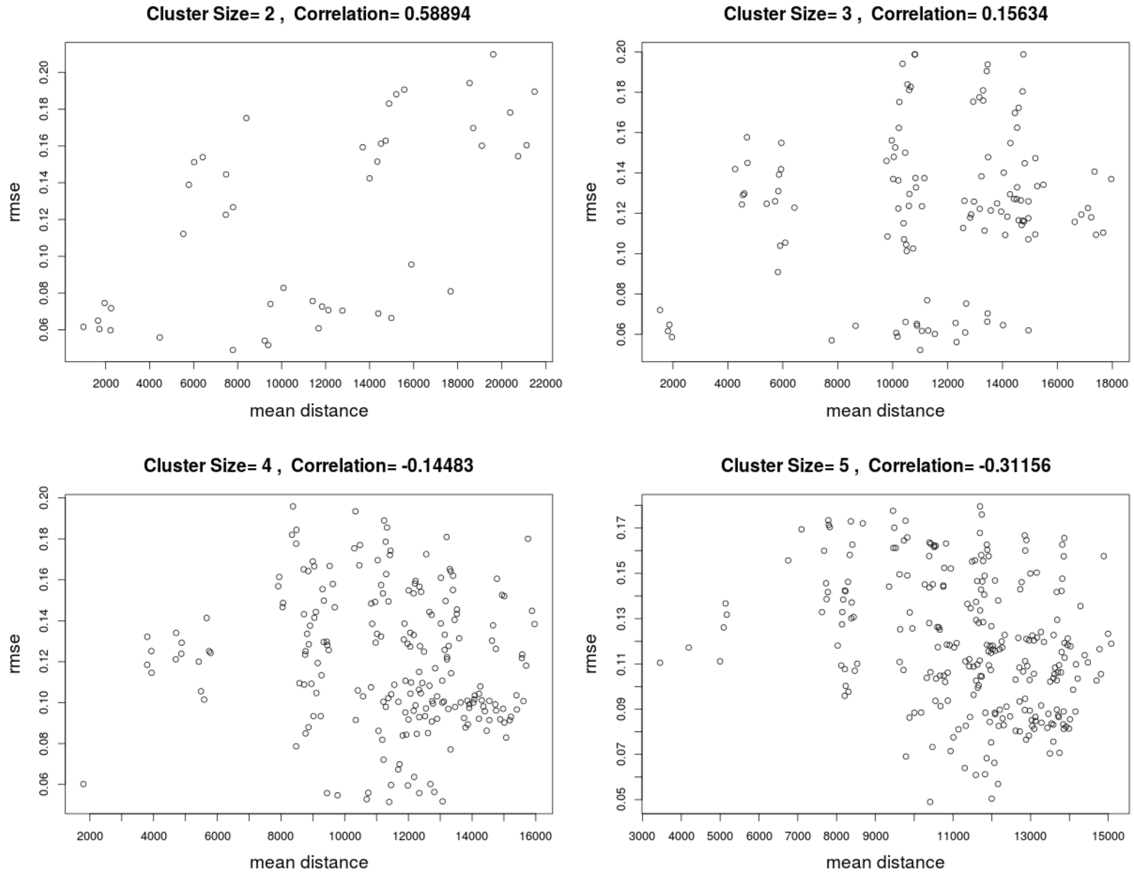
High Correlation



Diluting Correlation



Poor Correlation



Results

- High Correlation is obtained with average similarity measure < 2500
- Correlation starts diluting with average similarity measure ≈ 5000
- No correlation with average similarity measure > 10000
- Similar results were obtained by experiments with different History Lengths

➢ High Similarity $\Rightarrow$ Low Forecast error

➢ High Dissimilarity $\Rightarrow$ Ambiguous results

Clustering Experiments

Details

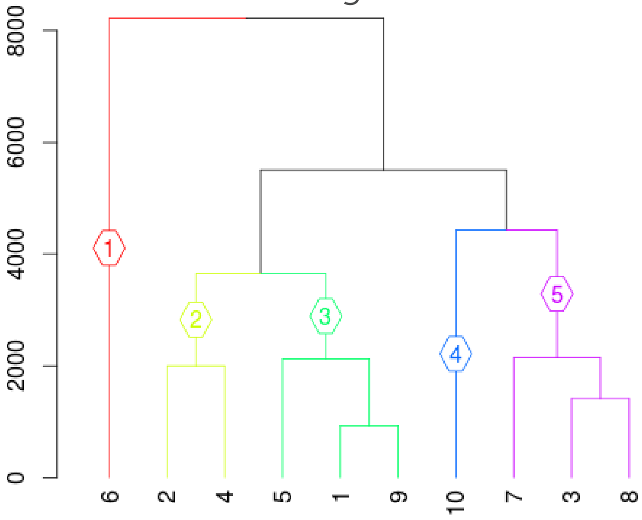
Goal

- Analyze the effect of Clustering and Aggregation on a small subset of Time Series
- Perform the analysis using different Standard Time Series Similarity Measures

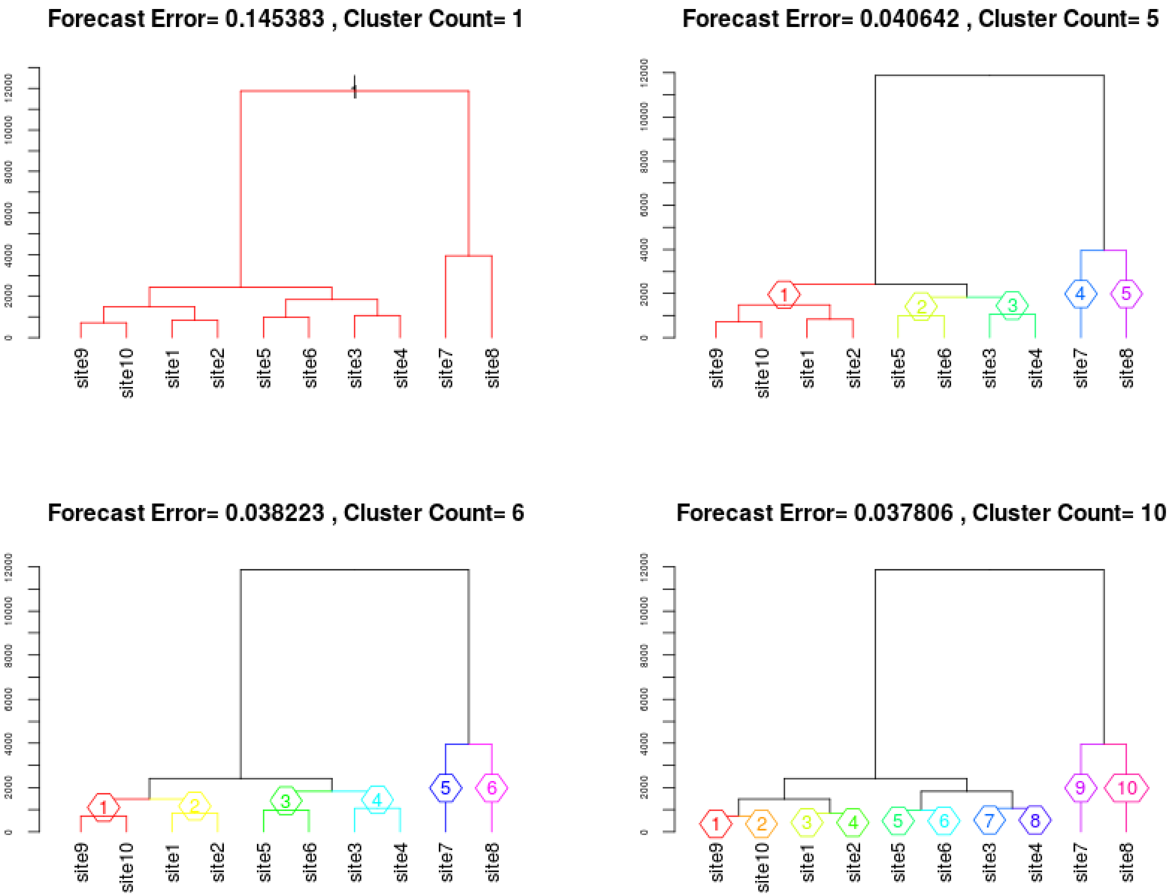
Methodology

- Apply Hierarchical Clustering on a smaller subset
- Cut the Dendrogram tree at different levels
  - At each level different number of clusters are formed
- Time Series in each cluster is aggregated and subjected to forecasting
- Overall error is calculated at each level

Dendrogram



Results

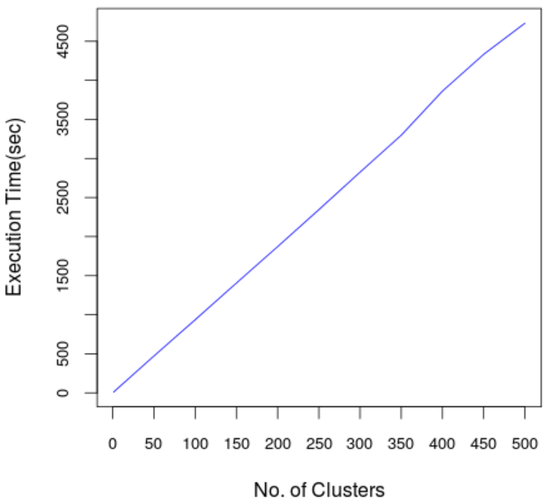

Forecast Error= 0.145383 , Cluster Count= 1
Forecast Error= 0.040642 , Cluster Count= 5
Forecast Error= 0.038223 , Cluster Count= 6
Forecast Error= 0.037806 , Cluster Count= 10
0
2000
4000
6000
8000
10000
12000
site9
site10
site1
site2
site5
site6
site3
site4
site7
site8

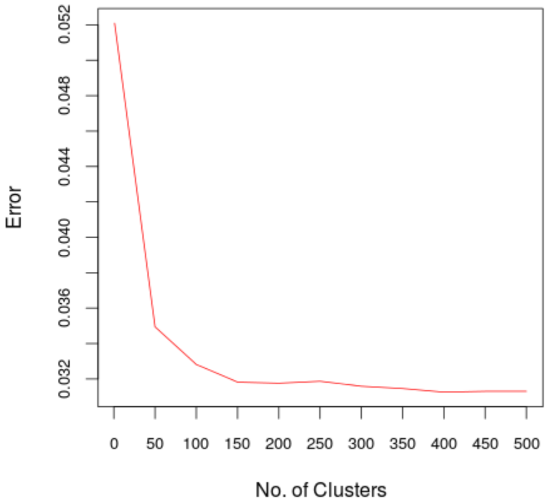
Further Analysis

Performance Measure

- Apply Hierarchical Clustering on a bigger data subset
- Similar procedure is followed
- Calculate the computation time as performance measure

Results





Further Experiments

Experiments with Solar

Dresden Database Systems Group

Solar Dataset

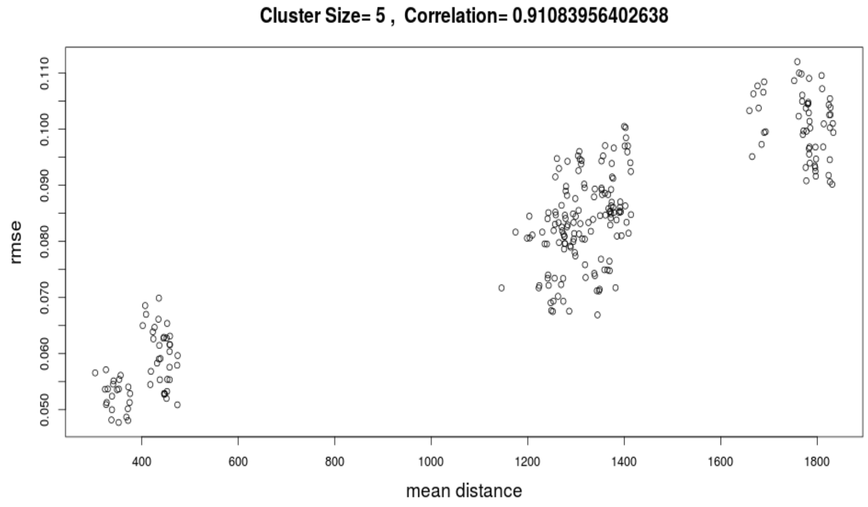
- NREL Solar Power Dataset
- Contains data from ~6000 PV plants
- Hourly data of Power(MW) generated from each plant
- Solar Radiation information obtained from SolarAnywhere
- Dataset is prepared by taking radiation information nearest to the site

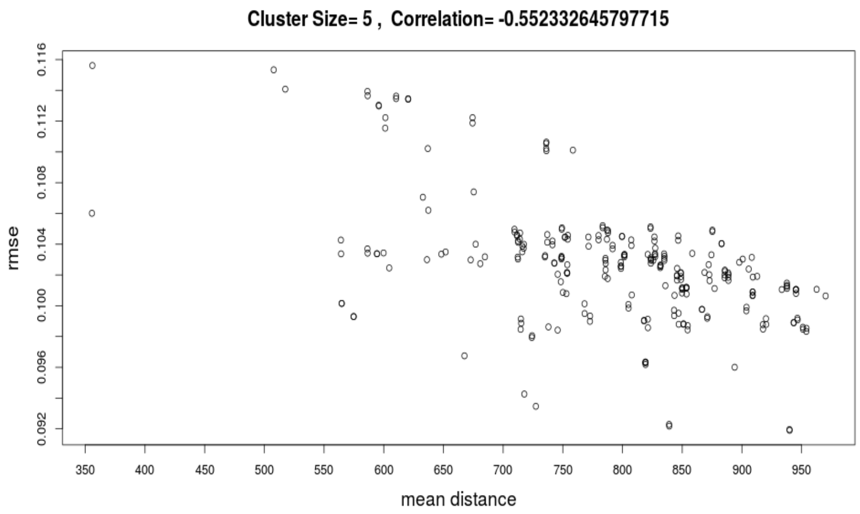
Methodology

- 6 different subset of Time Series of size 10 is chosen
  - 3 subsets are randomly chosen
  - 3 subsets belong to a specific geographic location
- Similar Forecasting method was followed
- History Length of whole year
  - Presence of Seasonality

Results

- Obtained results were not conclusive
- Difference between positive and negative results not clear
- Model Chosen was not complex
- Solar Data contains seasonality





Conclusion & Future Work

Conclusion

Dresden Database Systems Group

Aggregation experiments

- Lower Simialrity Measure leads to better Forecast results

Clustering Experiments

- Hierarchical clustering gives better results than simple aggregation
- Not better than Forecasting individual Time Series
- But, forecasting individual Time Series is slower in case of Large Dataset
  - Reduction of input space
  - Improves performance

Future Work

Dresden Database Systems Group

Clustering

- Deriving the Threshold for identifying best clusters
- Analyze the perfomance on very large datasets

Dataset

- Work on Solar Dataset by creating better forecast model
- Different datasets other than renewable energy

Ecast Framework

Questions

Thank You