MESSI Cluster Analysis

19.01.2023

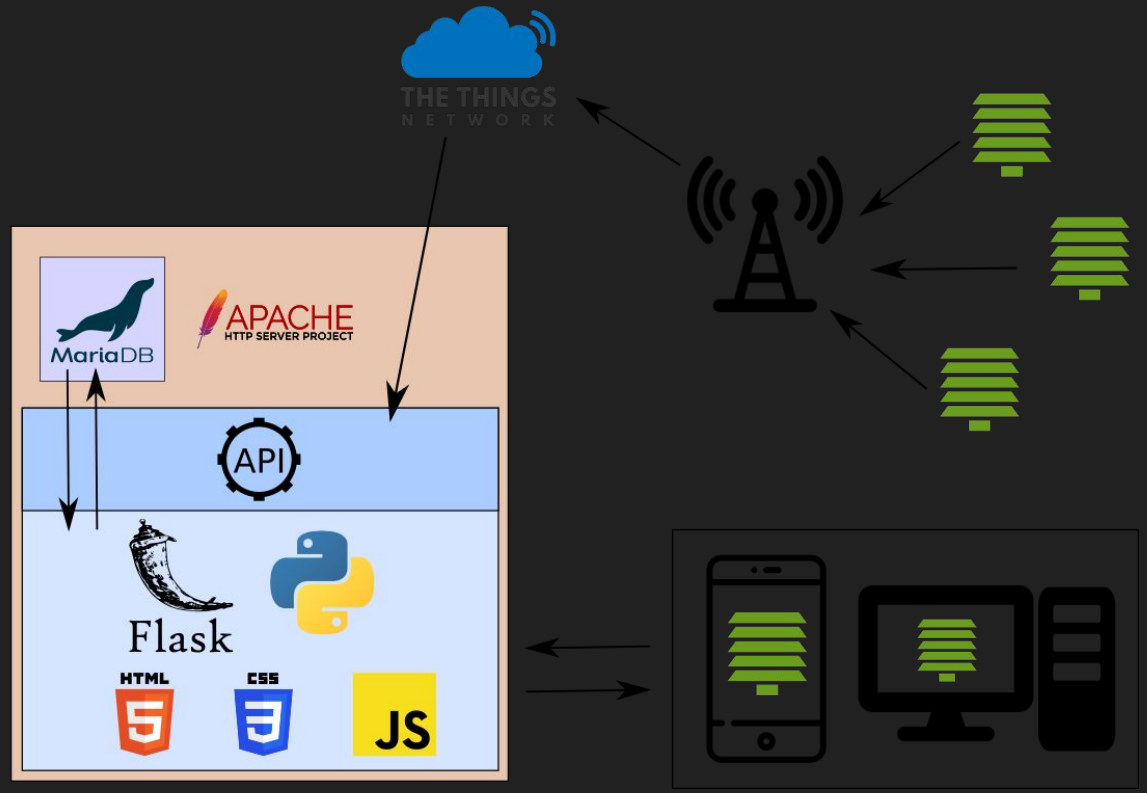
Robert Wright

Weather & Climate Diagnosis

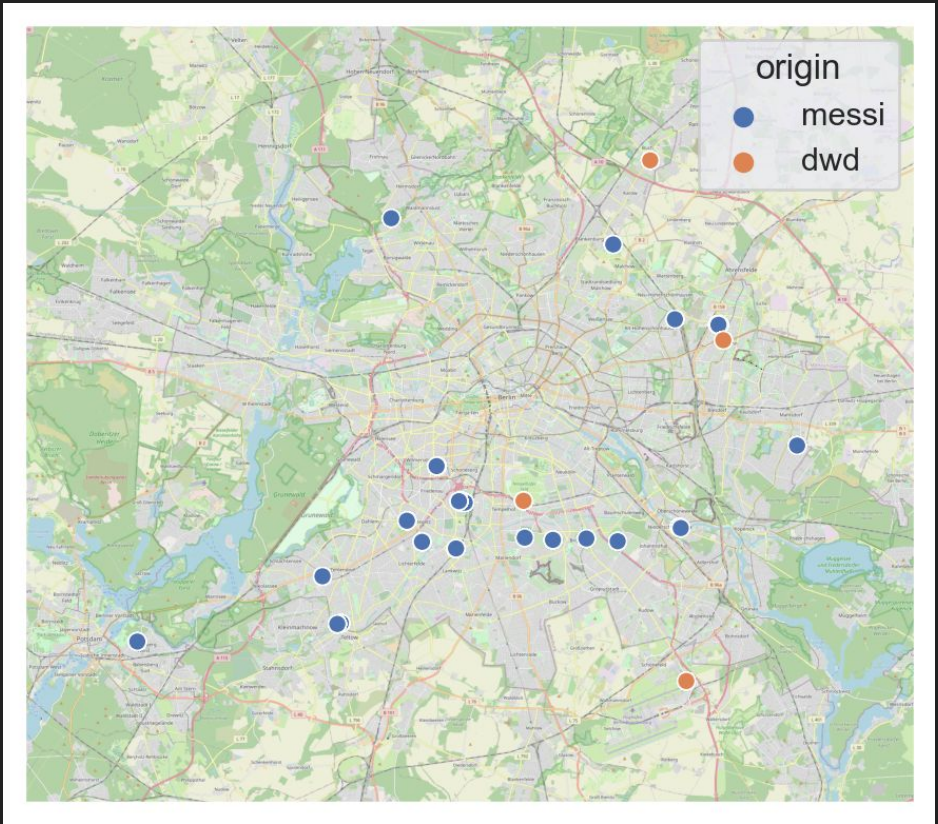
MESSIs

- = **M**ein **E**igenes **S**ub-**S**kalen **I**nstrument
- Designed by FU & TU
- Citizen Science Project: OpenUCO

Data Assimilation

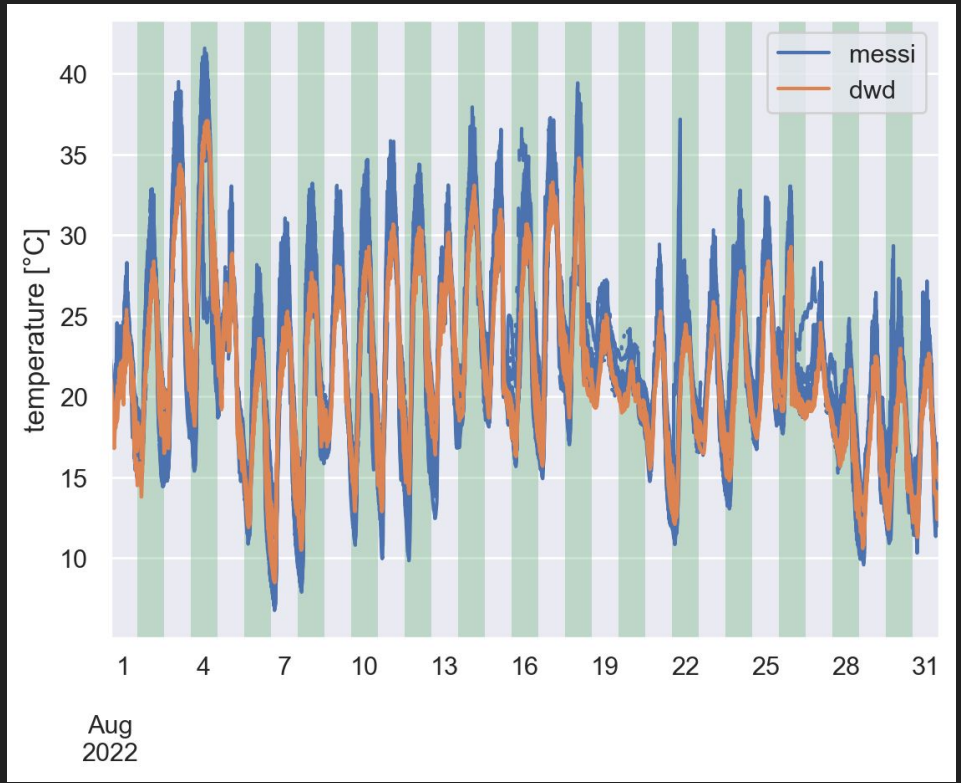


MESSI Locations



Data

- Temperature
- 10 min time interval
- August 2022

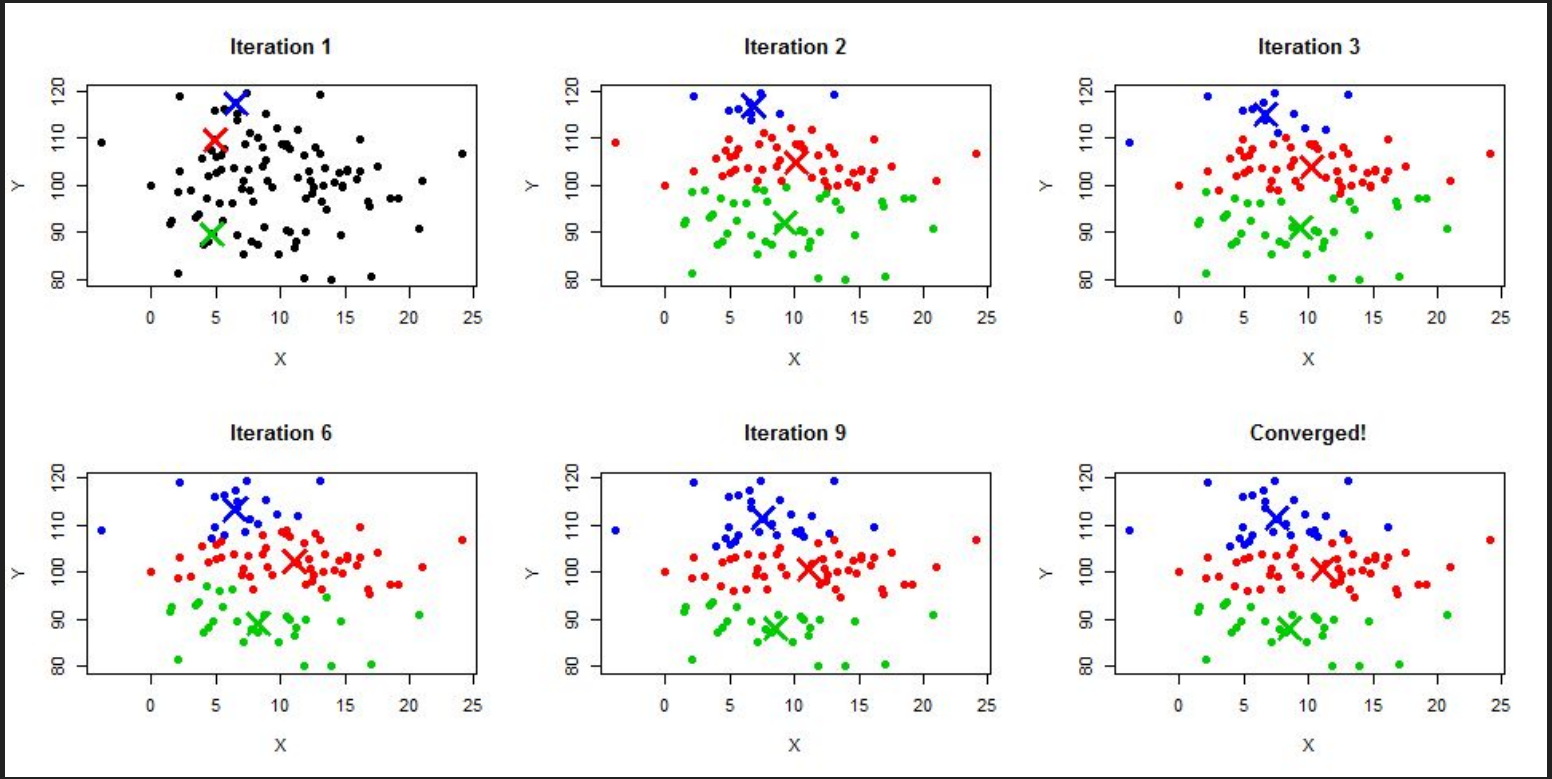


Clustering

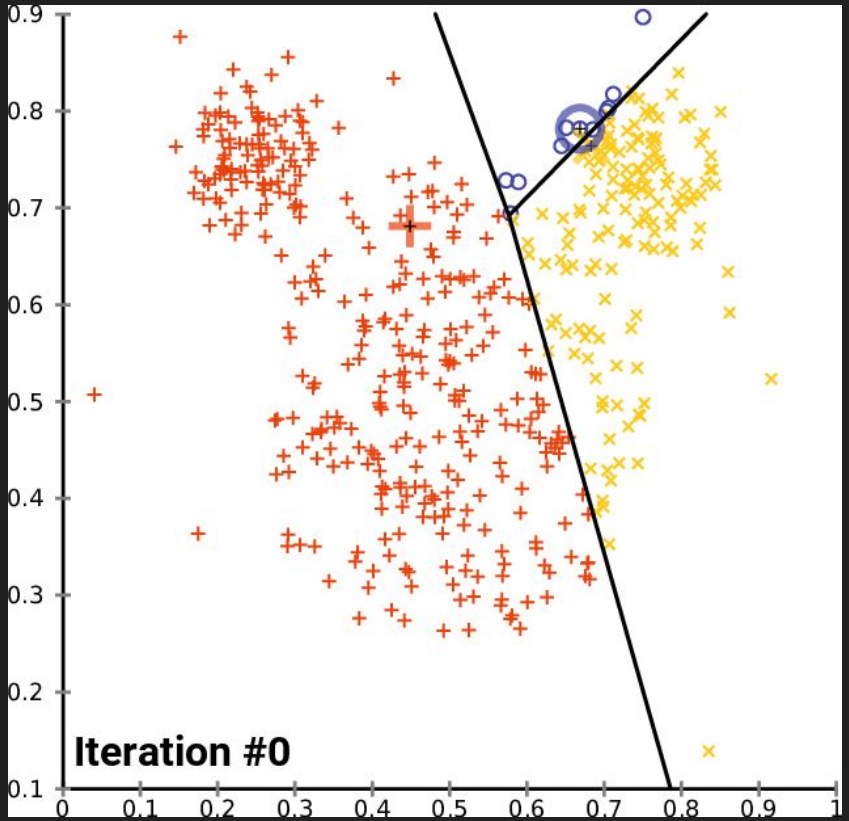
- kmeans
- Unsupervised learning = no target variable
- Separate samples into groups (*clusters*)
- # of clusters has to be specified
- Minimize within-cluster sum-of-squares (*inertia*):

$$SSE = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

kmeans


Iteration 1
Iteration 2
Iteration 3
Iteration 6
Iteration 9
Converged!
X
Y

kmeans



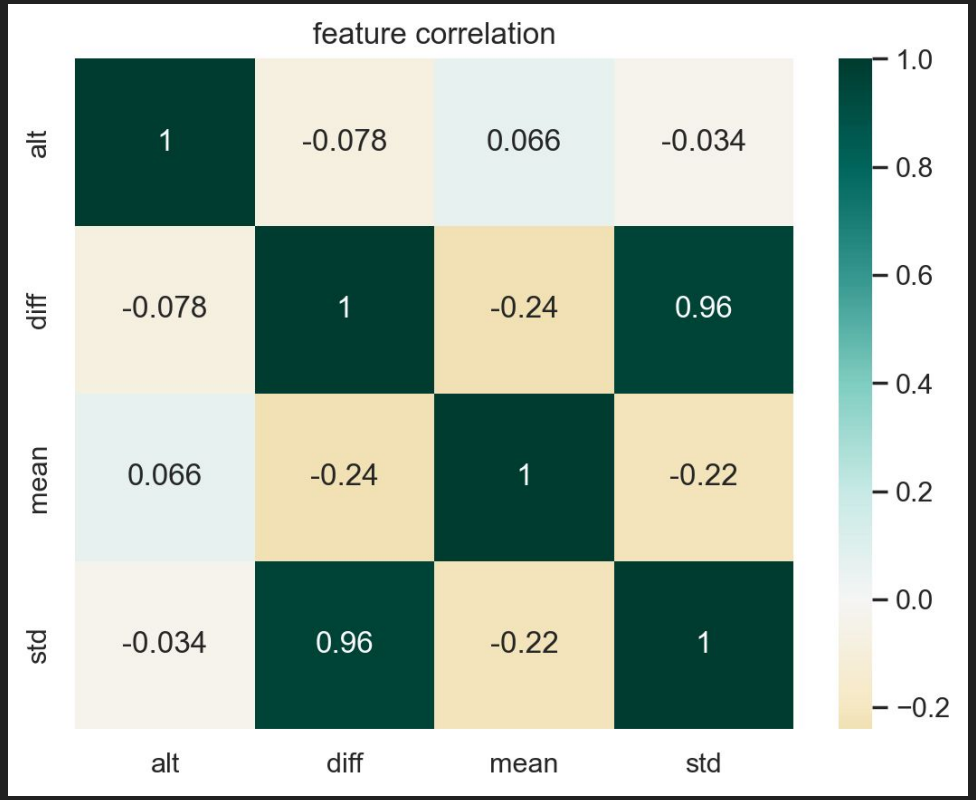
wikipedia.org

Feature Engineering

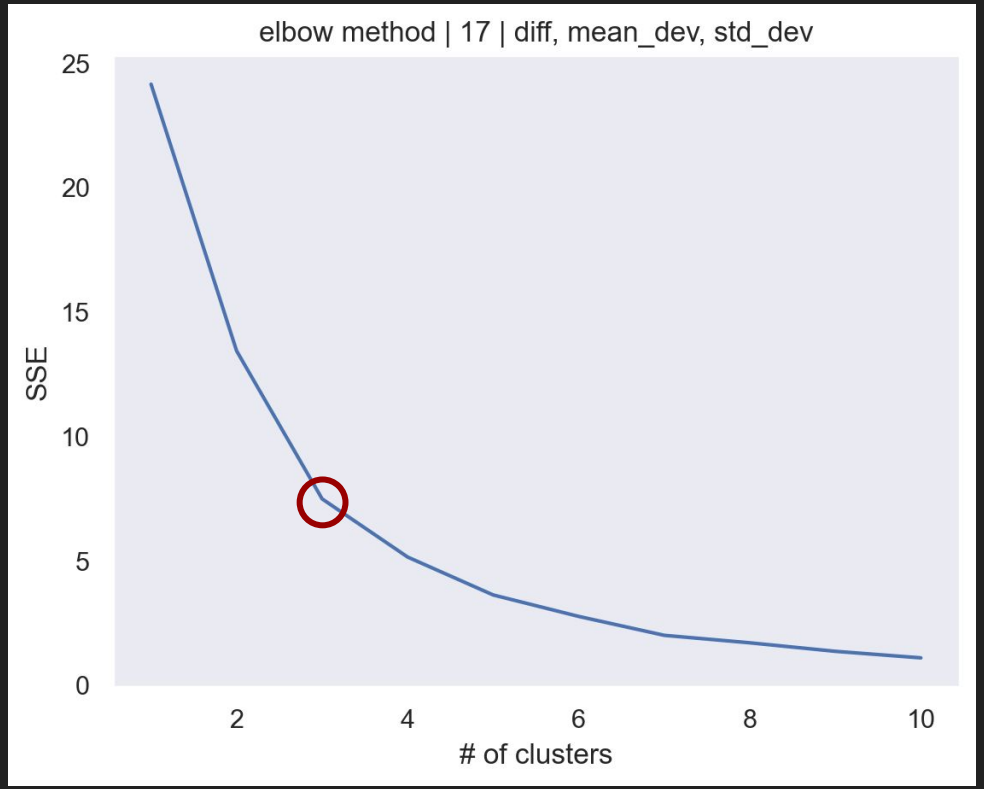
- Deviation from *dwd* daily mean: ΔT

1. Difference of maximum and minimum temperature: $\Delta T_{max} - \Delta T_{min}$
2. Mean: $\overline{\Delta T}$
3. Standard deviation: $\sigma \Delta T$

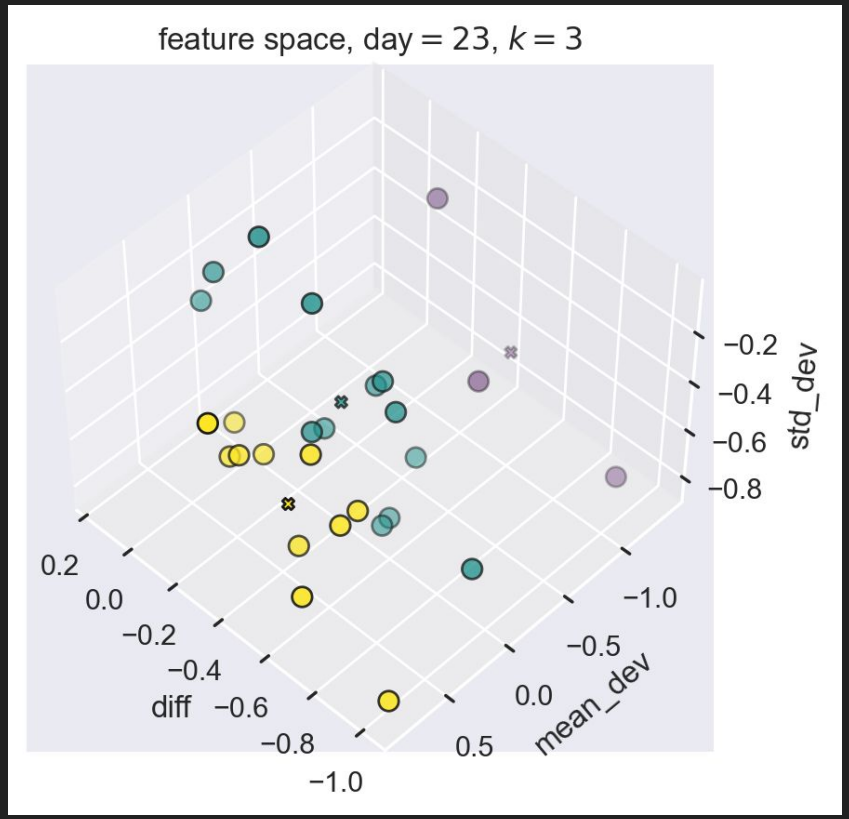
Feature Correlation



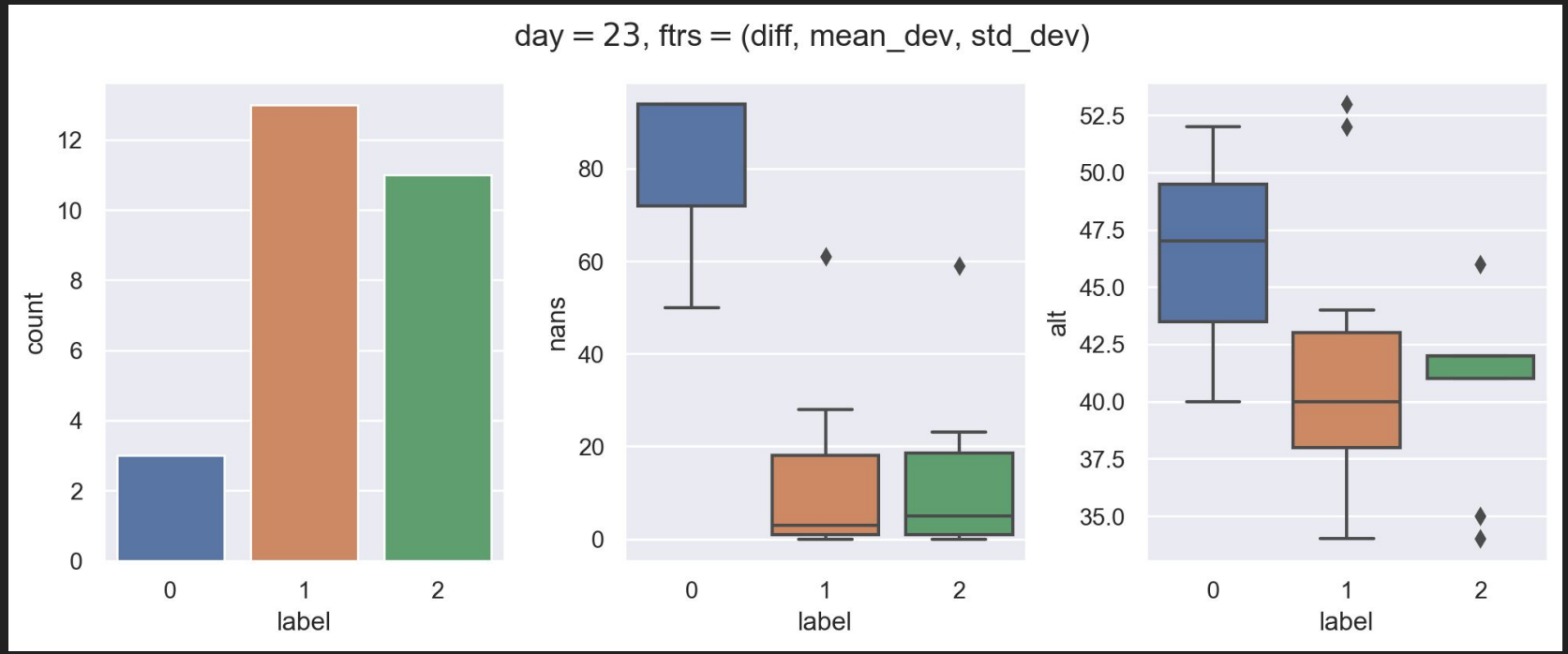
of Clusters


elbow method | 17 | diff, mean_dev, std_dev
SSE
25
20
15
10
5
0
2
4
6
8
10
# of clusters

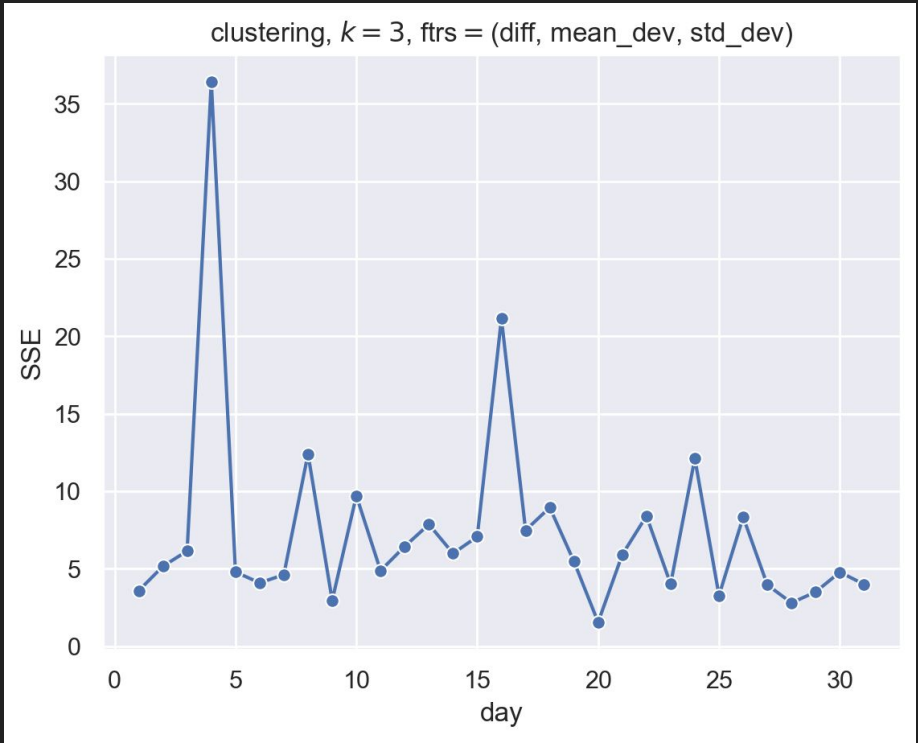
Clusters in Feature Space

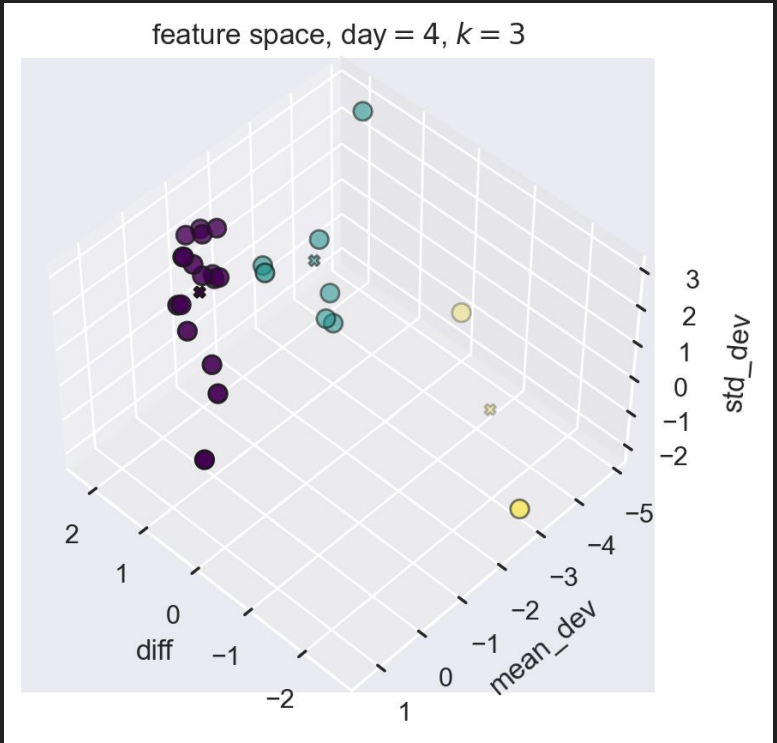


Characteristics of Clusters

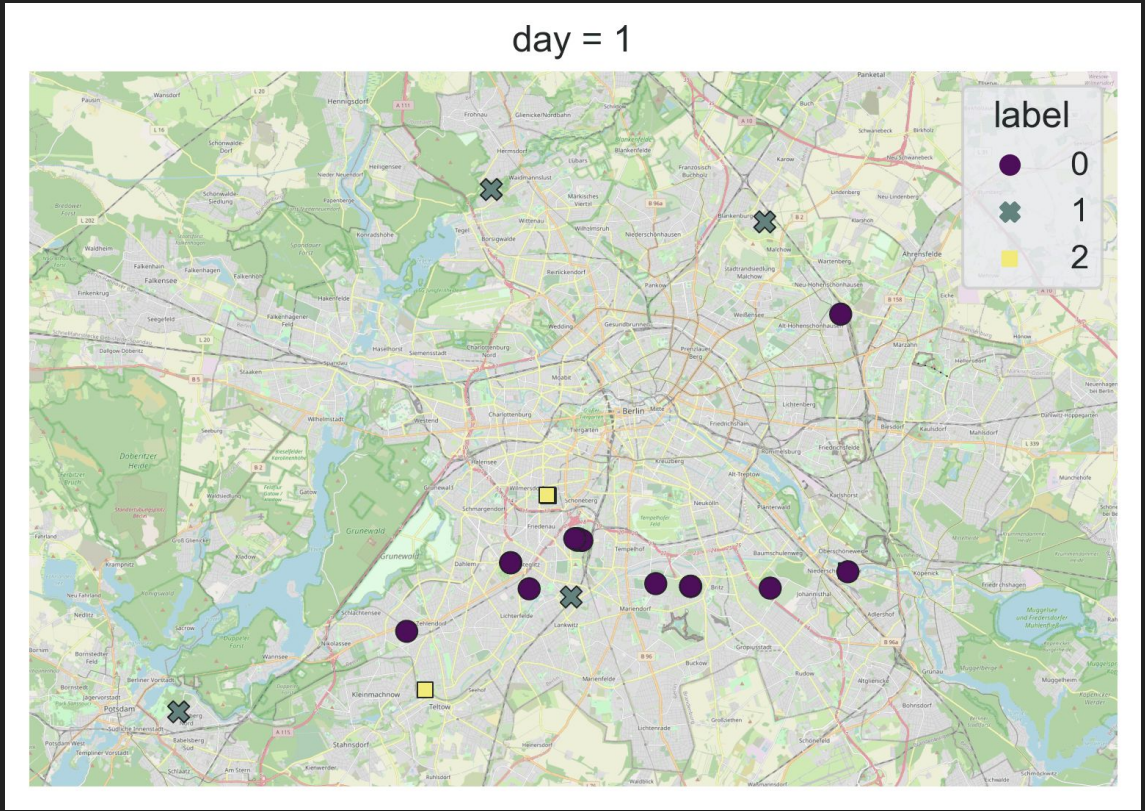


Inertia of Clusterings

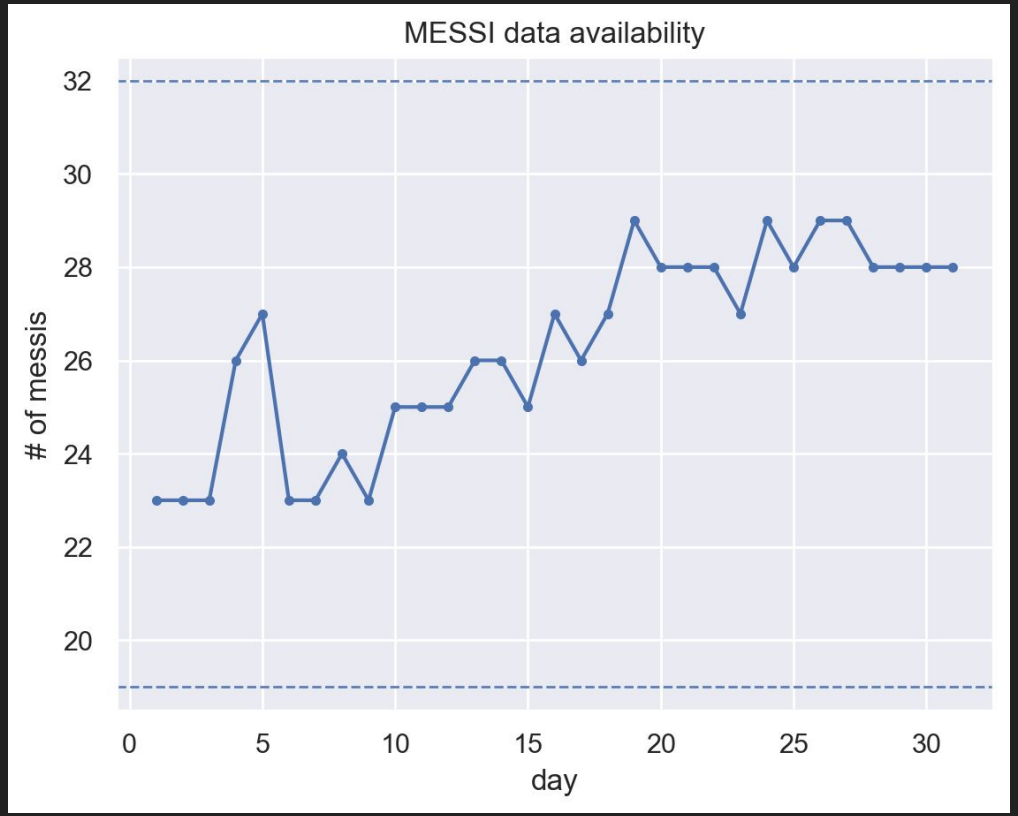




Clusters on Berlin map



Changing # of MESSIs throughout August

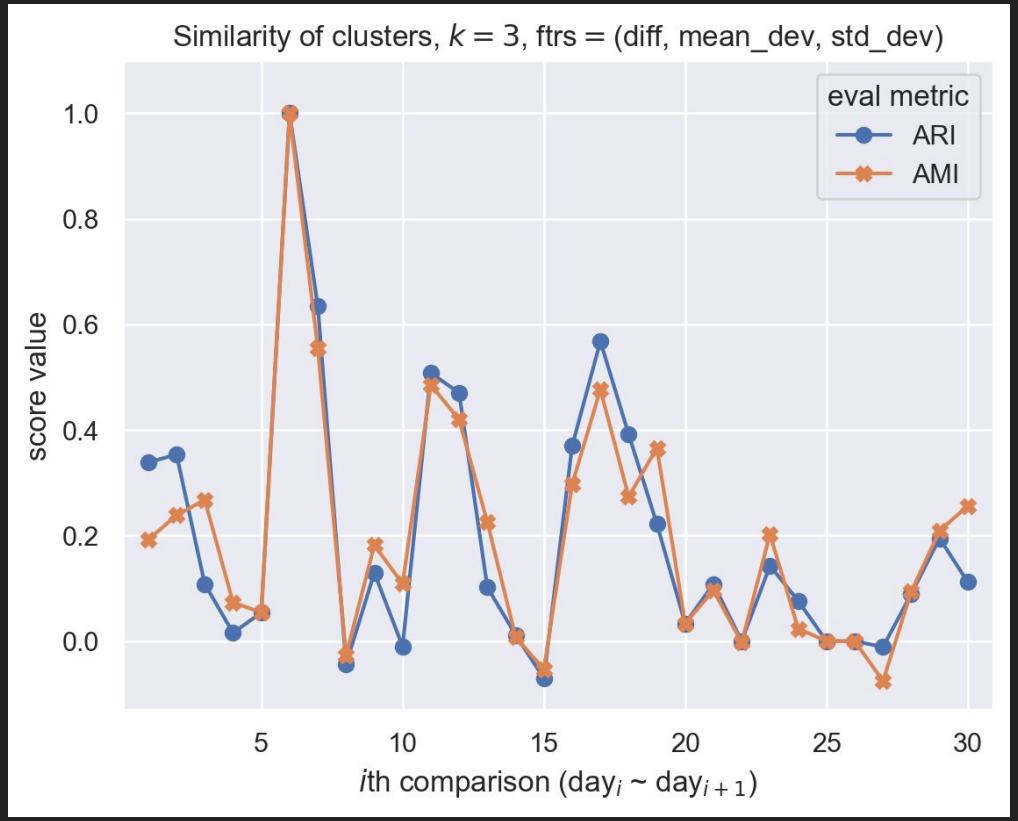


Rand Index

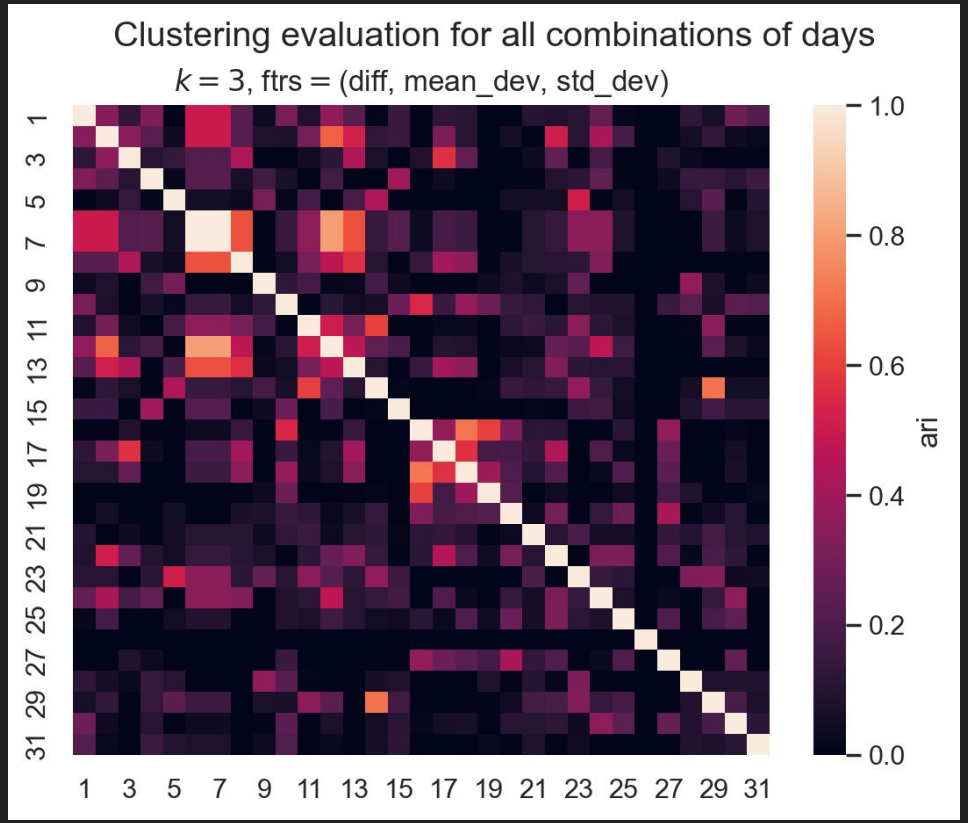
- Measures similarity of two cluster assignments
- Values between 0 (independent labellings) and 1 (perfect labelling)

$$\mathrm{RI} = \frac{a + b}{C_2^{n_{samples}}}$$

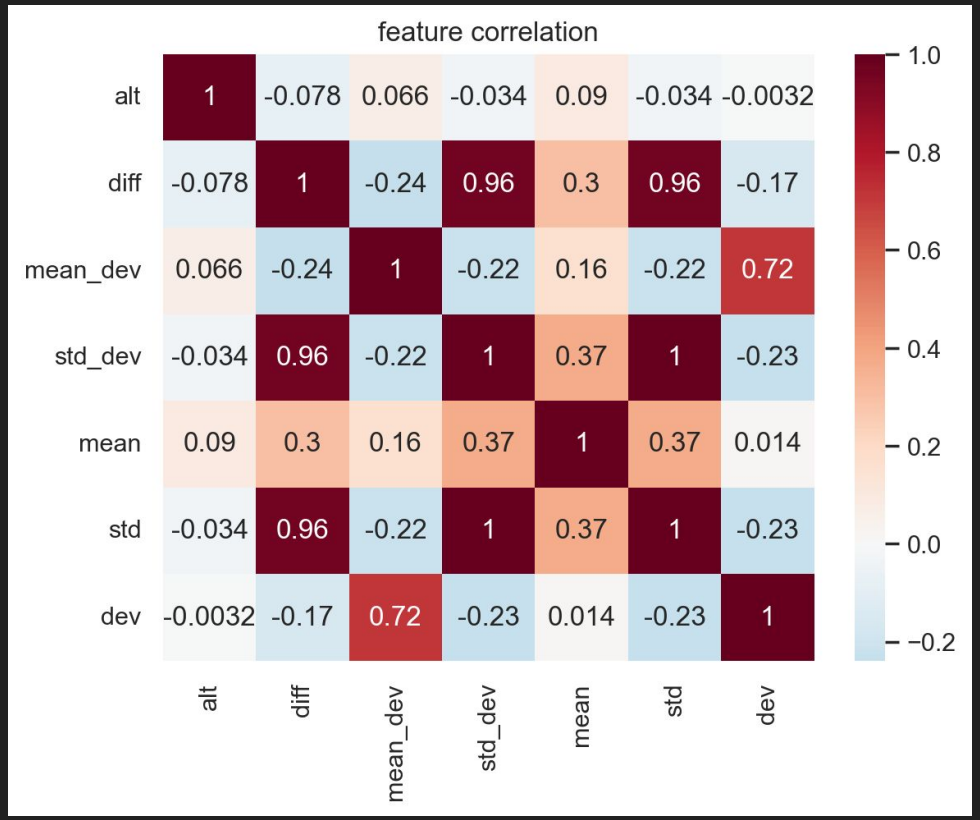
Cluster Similarity I



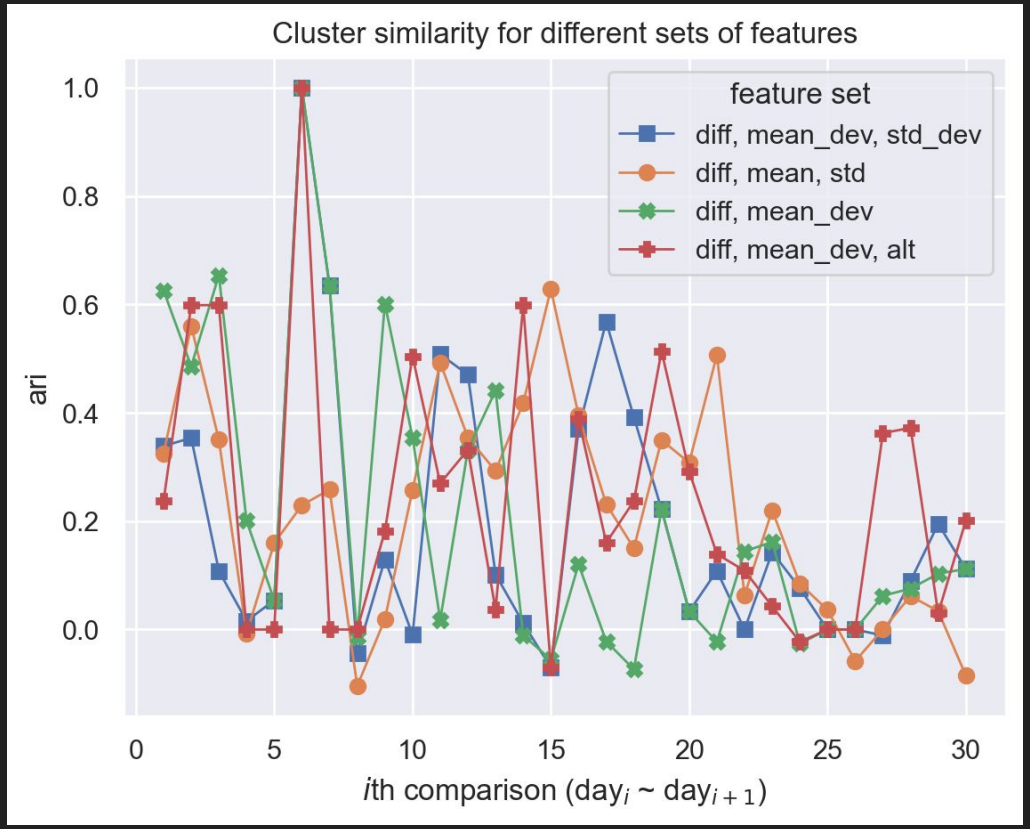
Cluster Similarity II



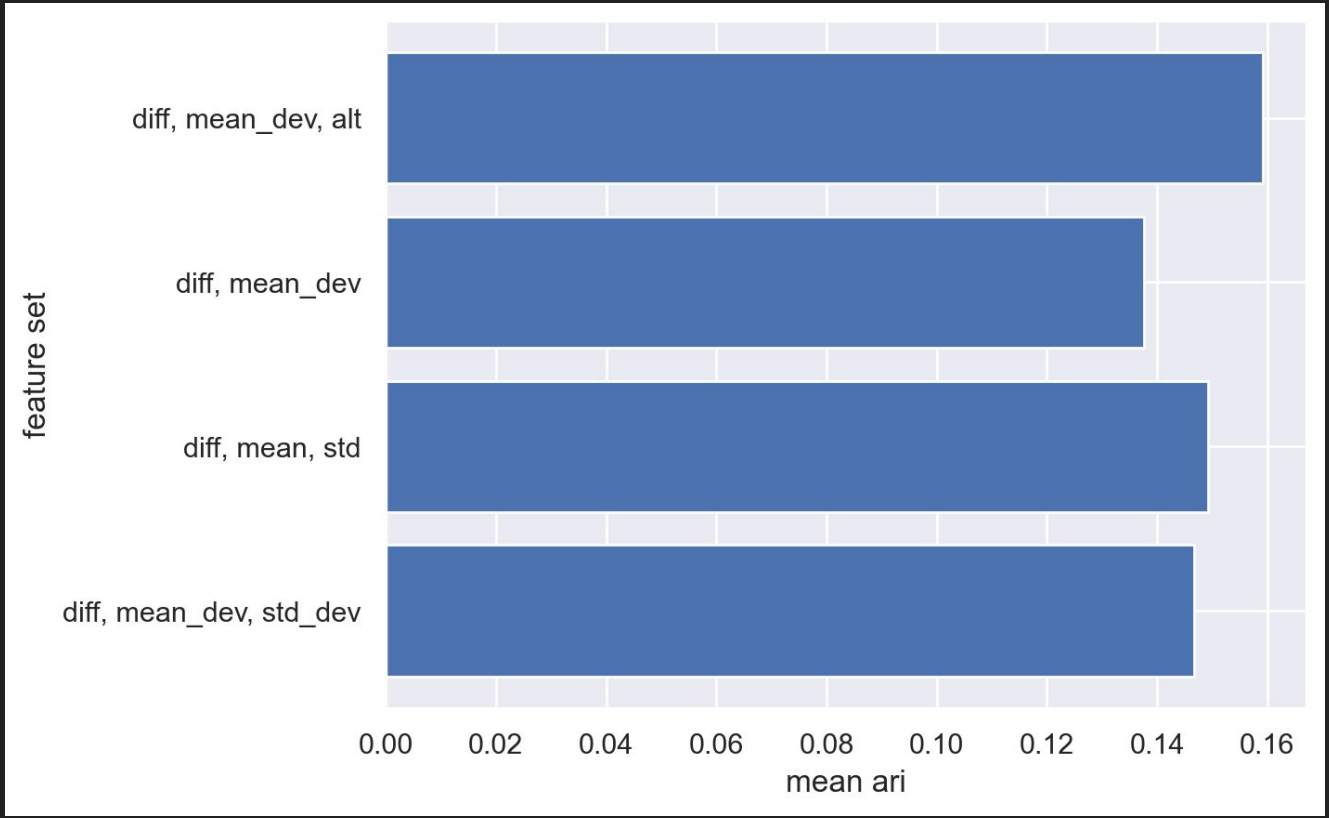
Different Sets of Features



Results I


Cluster similarity for different sets of features
feature set
diff, mean_dev, std_dev
diff, mean, std
diff, mean_dev
diff, mean_dev, alt
ari
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
ith comparison (day$_i$ ~ day$_{i+1}$)

Results II


diff, mean_dev, alt
diff, mean_dev
diff, mean, std
diff, mean_dev, std_dev
feature set
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
mean ari

Future Research

- Note: # of clusters is determined on a single day
- More MESSIs with continuous data / fill data gaps

  → Run clustering & performance evaluation on **same** data points

- Identify “meaning” of clusters with more meta-information

MESSI Cluster Analysis

19.01.2023

Robert Wright

Weather & Climate Diagnosis