Internship Interview

Minh-Khoi Pham

HUAWEI | IRELAND RESEARCH CENTRE

Introduction

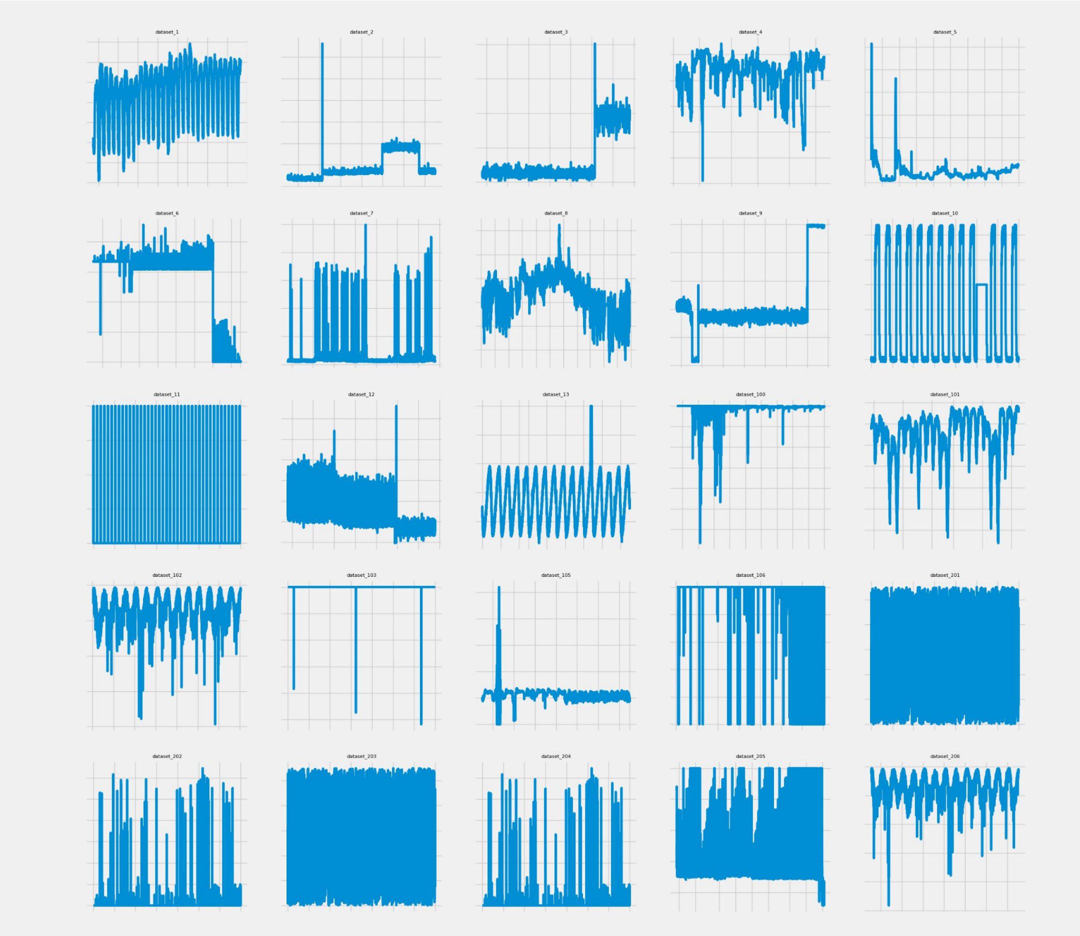
- Minh-Khoi Pham, ***2nd year Ph.D student*** in Computer Application, Dublin City University. Supervised by Prof. Martin Crane and Dr. Marija Bezbradica
- My topic: **Multimodal AI in studying healthcare process**
    - Working on a collaboration project with St James Hospital on using **ML/DL in studying patients' electronic health records to forecast patients' outcome.**
    - **Big tabular data** covers 4 years of **hospital bed days** recording 50,000 patients with nearly 1 million rows. **Diverse data types**: demographics, clinical codes, treatment pathways, admission history.
    - Apart from Ph.D works, do tasks requested by the hospital including **study the movement of an antimicrobial bacteria inside the hospital.**
    - Multimodal approach: from **traditional ML** on numeric features to **deep NLP models** on clinical codes, and **process mining** to study pathways with concentration on **models' interpretability**.
    - Just finished **2 papers** and are currently in **internal reviewing process**.

Case study

Received 2 time-series datasets and 3 tasks in total to complete:

1. Anomaly detection
2. Future forecasting
3. Clustering

[Huawei Time-Series.ipynb - Colaboratory (google.com)]



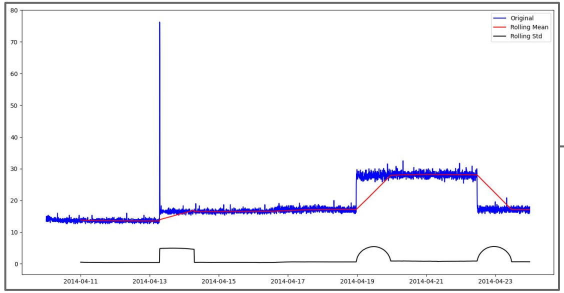
Anomaly Detection

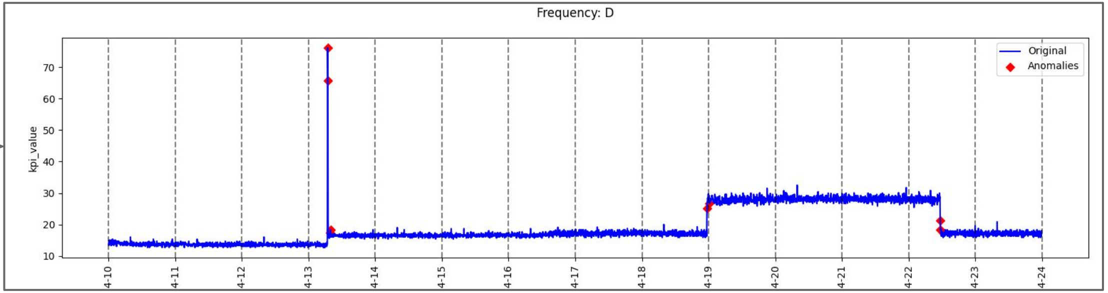
- Use 2 different techniques: rolling standard deviation and Seasonal-Trend Loess decomposition
- Rolling standard deviation
  - Suitable for time series with simple patterns, detecting short-term variability or fluctuation.
  - Easy to implement and interpret.
  - The choice of the rolling window size impacts the results
- STL decomposition
  - Applicable to time series showcasing complex patterns such as trends and seasonality.
  - May pose computational concerns, especially with larger datasets.
  - Have to identify the seasonal period

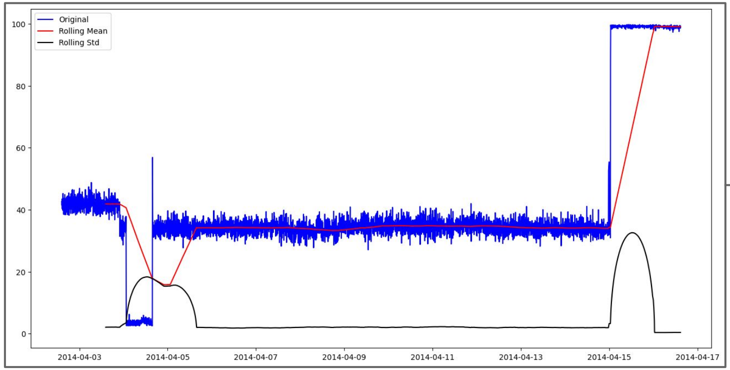
Rolling standard deviation

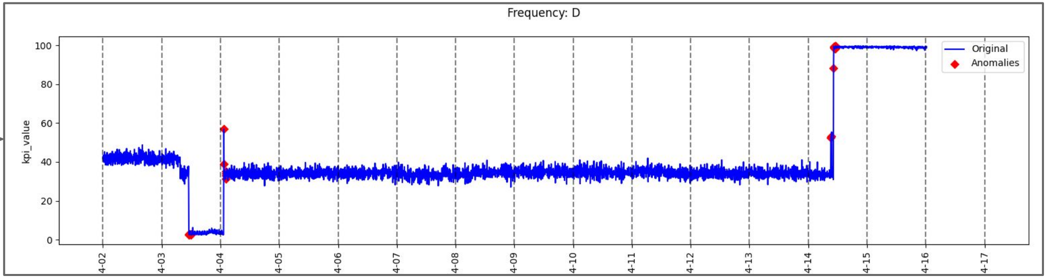
Calculate standard deviation for each time window

Similarly, extreme values in the rolling standard deviation often indicate anomalies

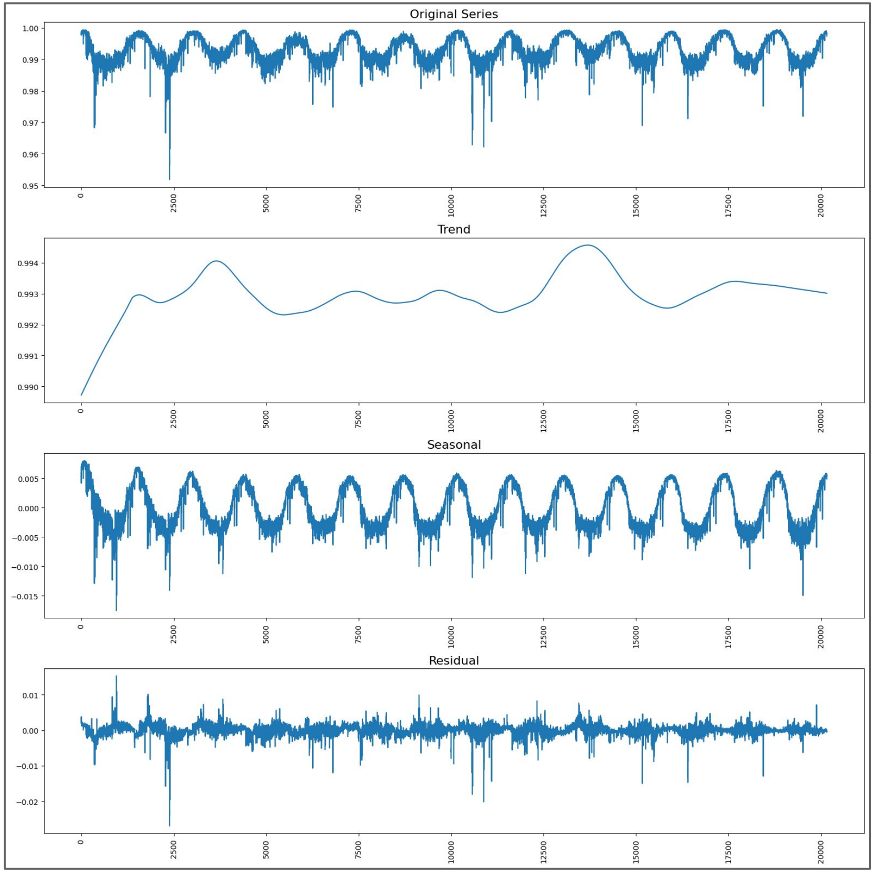




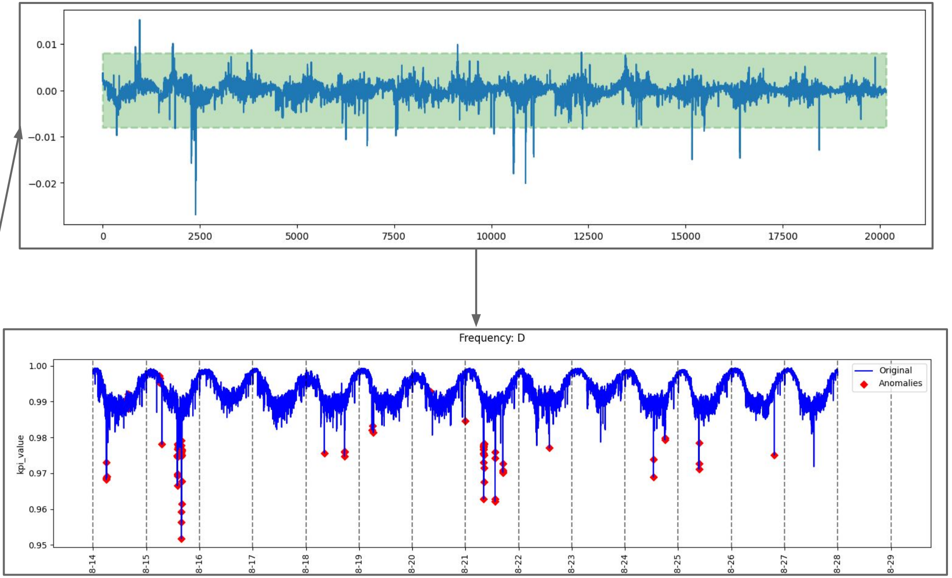




STL Decomposition



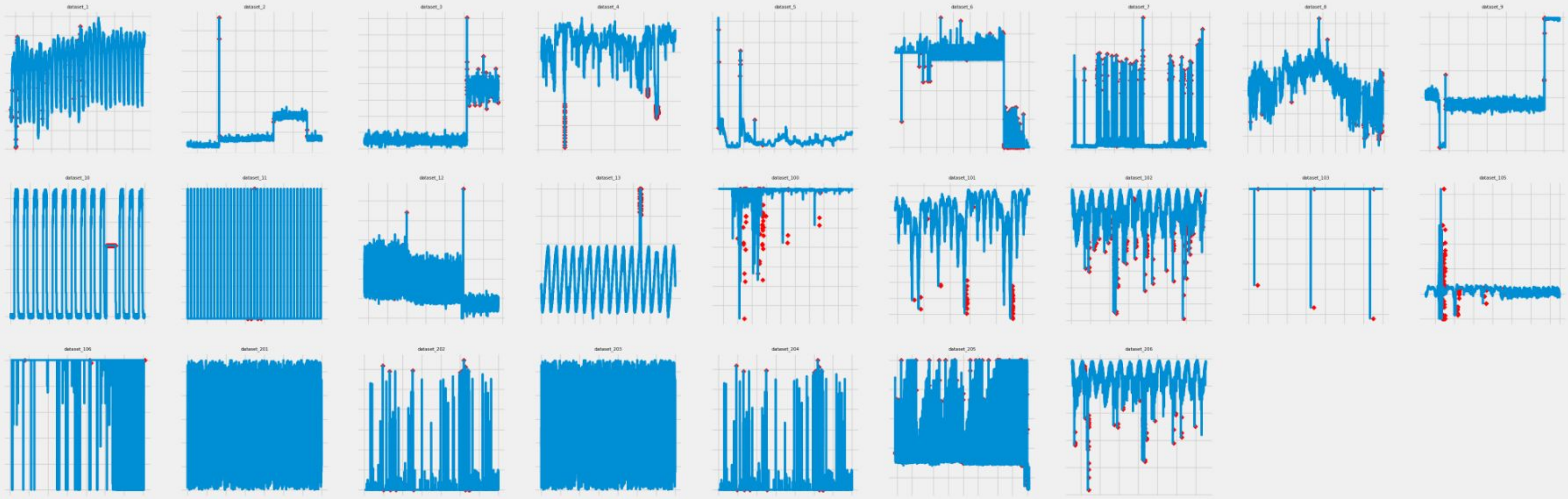
STL Decomposition into Trend, Seasonality and Residual



Detect anomalies based on extreme values of residuals

Anomaly Detection

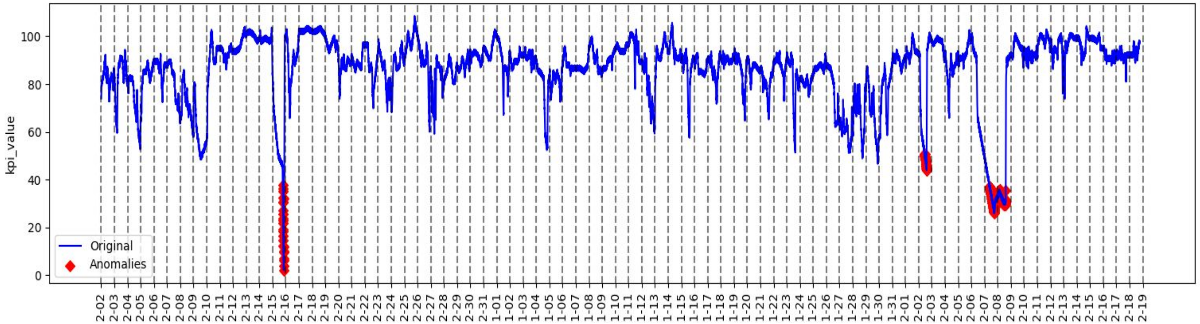
Visualization of anomalies detection results on 25 time series

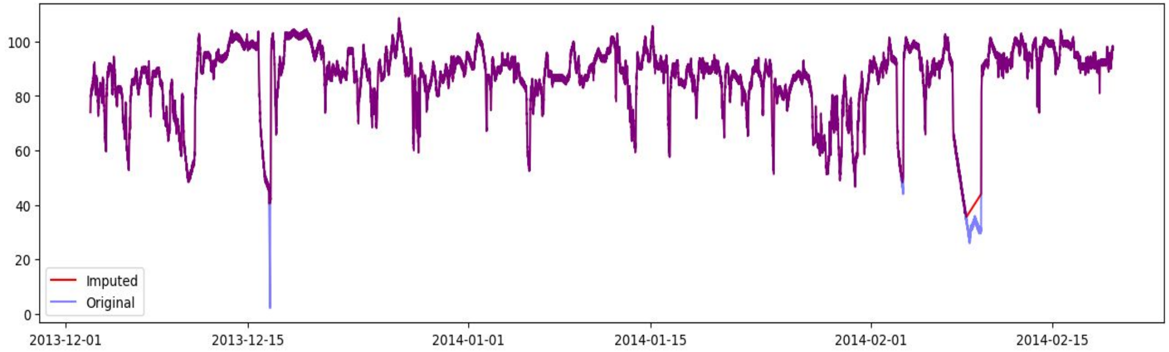


Forecasting

- Resampled the data, interpolated missing data, normalized data and remove outliers.
- Use ARIMA/SARIMA as main model for each time series. It is suitable for univariate time series exhibiting linear trend and seasonality, but requires stationary, so differencing is required.
- To check seasonality, autocorrelation functions are used
- Cross validation to prevent overfitting; use MAPE as optimization metric
- Hyperparameter tuning

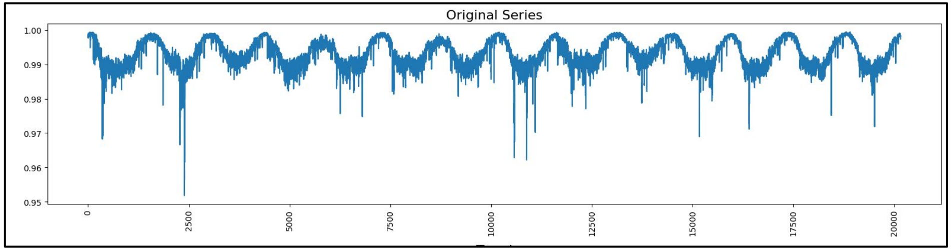
Anomalies Removal

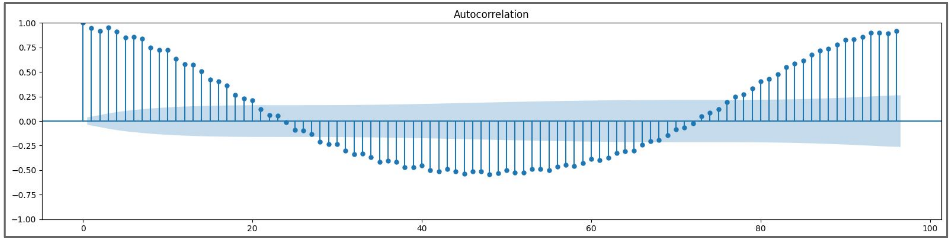




Anomalies are replaced by linearly interpolated values

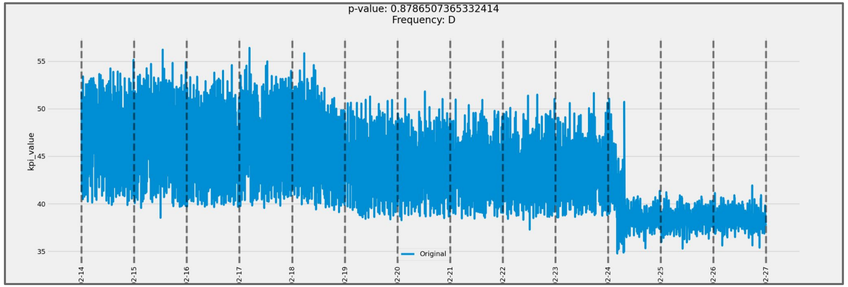
Seasonality/Stationary Check

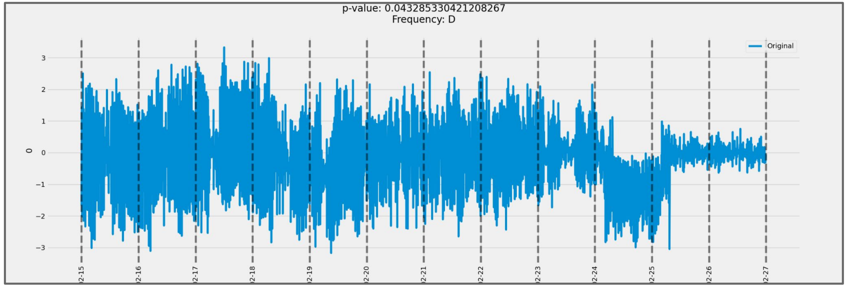




Autocorrelation Function helps detect recurring pattern at every 96th lag

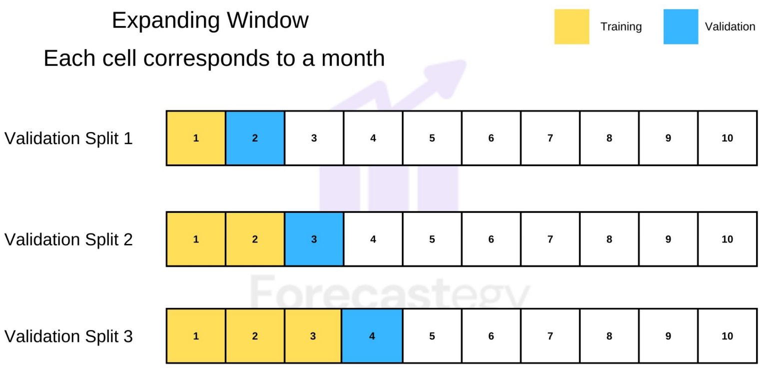
Non-stationary time series



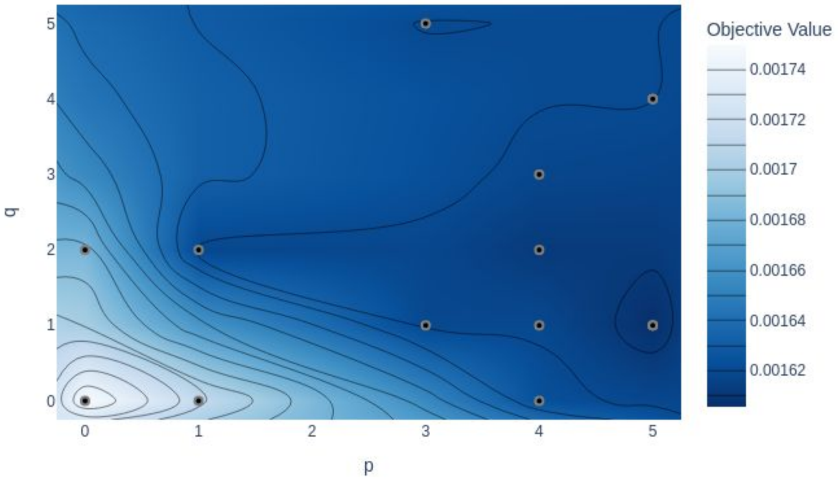


Stationary achieved after applying seasonal differencing

Hyperparameter Search And Cross-validation



Time series Cross validation



Contour plot of hyperparameter (p and q) searching to minimize MAPE

⇒ On a single step of searching, with a set of parameters of ARIMA, perform 5-fold cross validation training on the data; calculate the average score and compare with others

Forecasting

MAPE: Mean Absolute Percentage Error

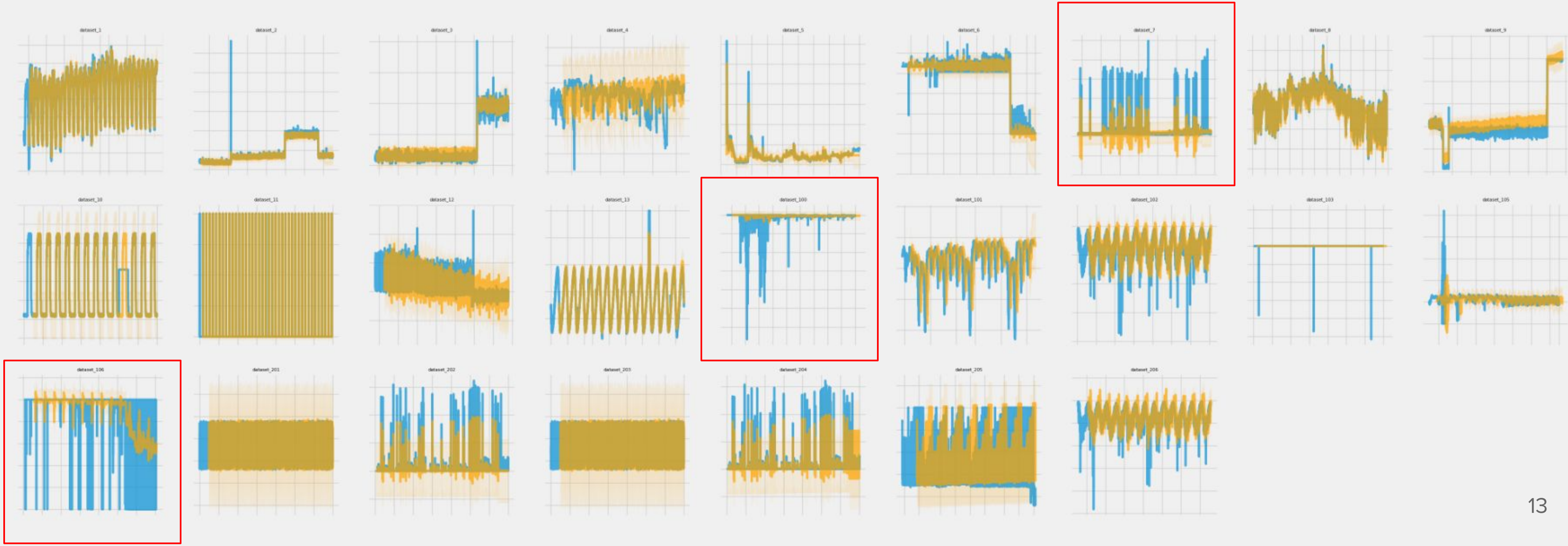
$$\mathrm{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

id	train_mape
dataset_1	0.004890
dataset_2	0.033688
dataset_3	0.051049
dataset_4	0.168646
dataset_5	0.261502
dataset_6	0.082886
dataset_7	0.403302
dataset_8	0.013893
dataset_9	0.265624
dataset_10	0.221111
dataset_11	0.001319
dataset_12	0.065029
dataset_13	0.042679

id	train_mape
dataset_100	0.54587
dataset_101	0.000138
dataset_102	0.000658
dataset_103	0.000000
dataset_105	4.177103
dataset_106	0.018822
dataset_201	0.265633
dataset_202	0.050794
dataset_203	0.265633
dataset_204	0.079665
dataset_205	0.110656
dataset_206	0.000705

Forecasting

- Most time series have seasonal pattern, allowing SARIMA to fit easily; meanwhile the model struggle with some time series without any clear patterns

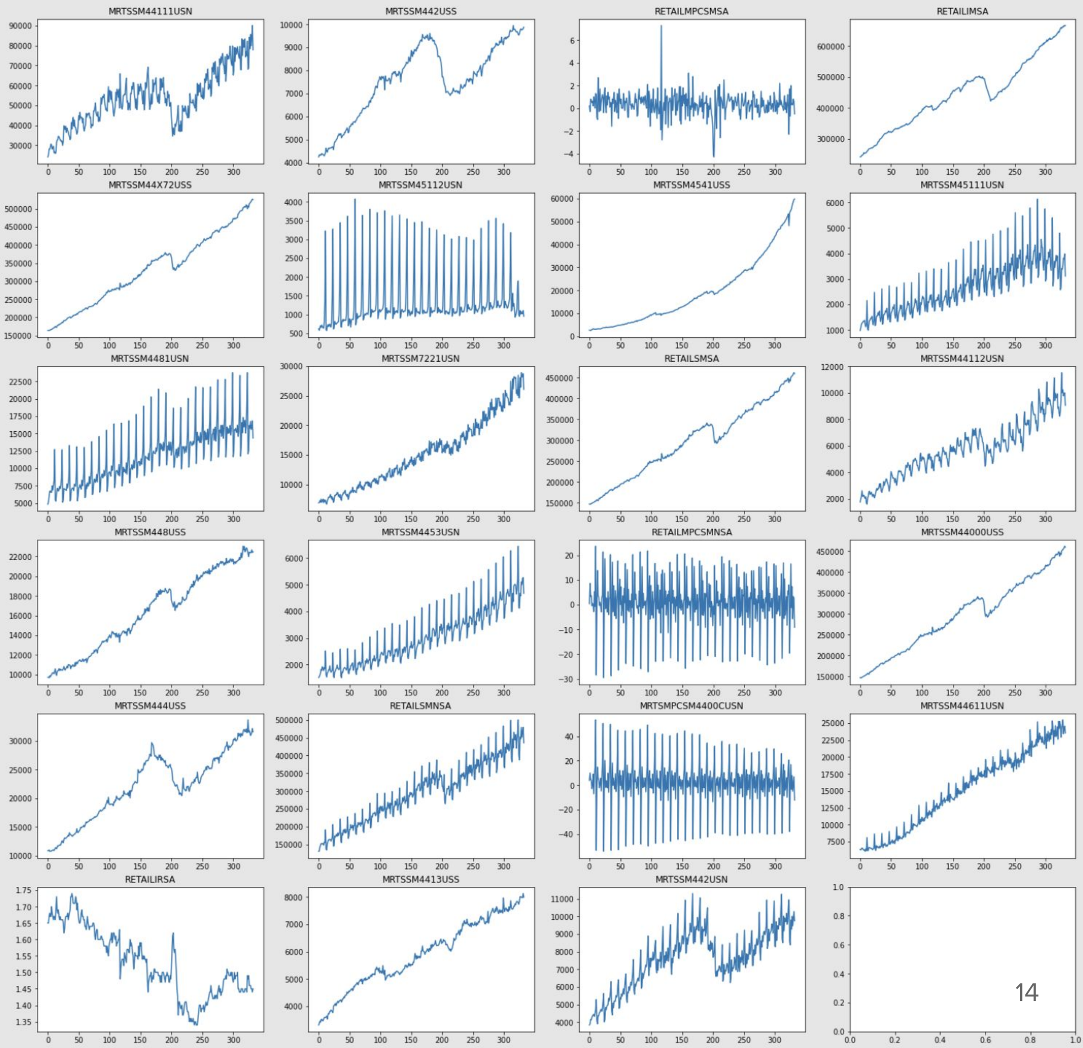


Clustering

❖ Cluster using K Means - simple yet effective

❖ Use ***tsfresh*** to feature engineering, converting time series to high dimensional feature matrices

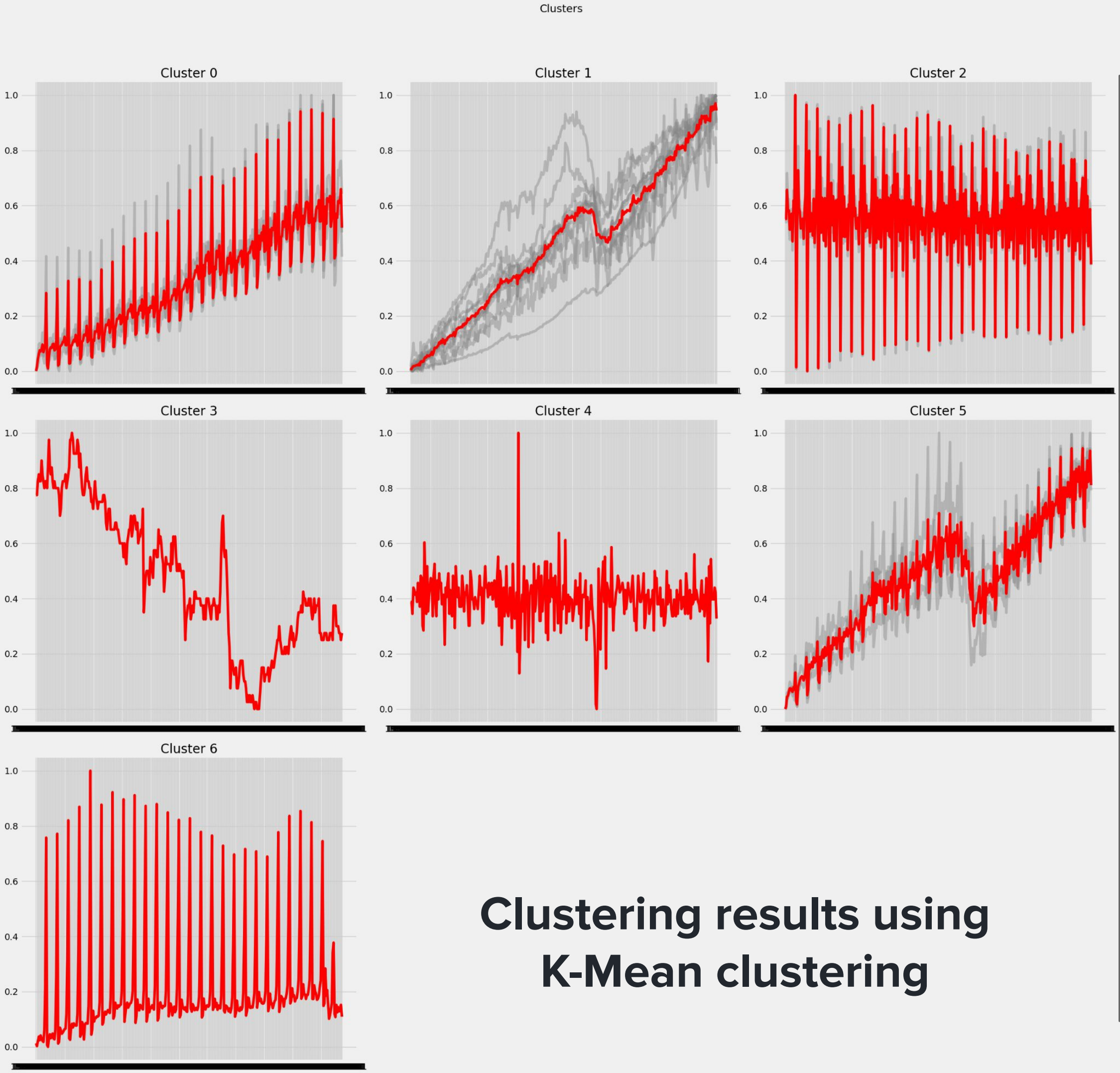
❖ Find optimal number of clusters using Elbow method



Clustering

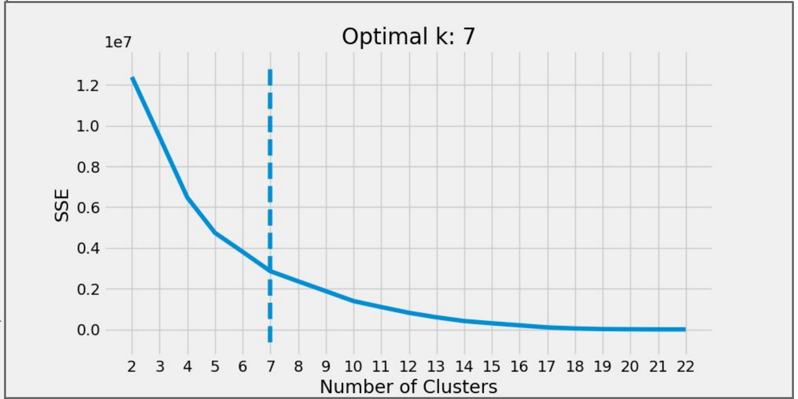
- ❖ From a time series has length of 333 timesteps, tsfresh converts it into a feature vector with 778 dimensions, each dimension is a statistic of that time series ➔ suitable for representation learning
- ❖ Some of the 778 features are:
  - ➢ Min, max, mean, variance, median, quantile,... of the series
  - ➢ Entropy, AR coefficients, autocorrelation,.... of the series
  - ➢ Wavelet transform, Fourier transform,..
  - ➢ .....
- ❖ K Means Clustering are then trained on these extracted features

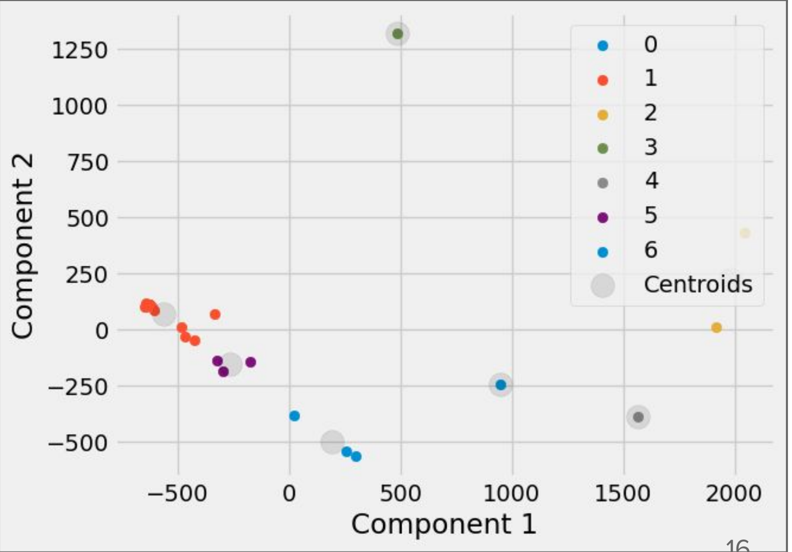
https://tsfresh.readthedocs.io/en/latest/text/list_of_features.html



Clustering results using K-Mean clustering

Finding optimal number of clusters using Elbow method





2D features projection using PCA

Terms need look at

- ACF, PACF
- ARIMA, SARIMA
- KMeans
- STL decomposition, rolling mean, std
- Data interpolation
- Elbow method
- Tsfresh features
- PCA ?
- Cross validation on time series
- Stationary
- scalability, adaptation, model selection, generalisation