Selecting Time Series Clustering Methods based on Run-Time Costs

Andreas Schörgenhumer
Paul Grünbacher
Hanspeter Mössenböck

12.11.2020
Motivation
Motivation

Time Series

Clustering
Motivation
Motivation

**Benefits:**
- General data insights
- Cluster-specific tools
- Better prediction and forecasting models
Clustering

- **Raw**-based clustering
Clustering

- **Raw-based clustering**

```
<table>
<thead>
<tr>
<th>t_0</th>
<th>t_1</th>
<th>t_2</th>
<th>t_3</th>
<th>t_4</th>
<th>t_5</th>
<th>t_6</th>
<th>t_7</th>
<th>t_8</th>
<th>t_9</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>13</td>
<td>9</td>
<td>1</td>
<td>2</td>
<td>10</td>
<td>12</td>
<td>13</td>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>
```

Clustering
Clustering

• **Raw**-based clustering

• **Feature**-based clustering
Clustering

• **Raw**-based clustering

\[
\begin{array}{cccccccccc}
  t_0 & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 \\
  12  & 13  &  9  &  1  &  2  & 10  & 12  & 13  &  7  &  4  \\
\end{array}
\]

Clustering

• **Feature**-based clustering

Feature Extraction

Clustering

3
Clustering

- **Raw**-based clustering

<table>
<thead>
<tr>
<th>t₀</th>
<th>t₁</th>
<th>t₂</th>
<th>t₃</th>
<th>t₄</th>
<th>t₅</th>
<th>t₆</th>
<th>t₇</th>
<th>t₈</th>
<th>t₉</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>13</td>
<td>9</td>
<td>1</td>
<td>2</td>
<td>10</td>
<td>12</td>
<td>13</td>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>

- **Feature**-based clustering

  **Feature Extraction**

<table>
<thead>
<tr>
<th>Kurtosis</th>
<th>Autocorrelation</th>
<th>Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.1</td>
<td>0.95</td>
<td>2.5</td>
</tr>
</tbody>
</table>
Clustering

- **Raw-based clustering**

**Benefits:**
- Inspect time series properties
- Dimensionality reduction
- Handle unequal time series lengths

- **Feature-based clustering**

<table>
<thead>
<tr>
<th>Kurtosis</th>
<th>Autocorrelation</th>
<th>Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.1</td>
<td>0.95</td>
<td>2.5</td>
</tr>
</tbody>
</table>

Clustering
Clustering Methods

• **Method** = triplet of (model, features, variant)
Clustering Methods

- **Method** = triplet of (model, features, variant)

<table>
<thead>
<tr>
<th>Model</th>
<th>Features</th>
<th>Variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>k-means</td>
<td>raw</td>
<td>no post-processing</td>
</tr>
<tr>
<td>BIRCH</td>
<td>feature set A</td>
<td>clip [0, 1]</td>
</tr>
<tr>
<td>Agglomerative</td>
<td>feature set B</td>
<td>drop correlated</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Clustering Methods

- **Method** = triplet of (model, features, variant)

<table>
<thead>
<tr>
<th>Model</th>
<th>Features</th>
<th>Variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>k-means</td>
<td>raw</td>
<td>no post-processing</td>
</tr>
<tr>
<td>BIRCH</td>
<td>feature set A</td>
<td>clip [0, 1]</td>
</tr>
<tr>
<td>Agglomerative</td>
<td>feature set B</td>
<td>drop correlated</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- Which one has the best clustering **quality**?
### Clustering Methods

**Method** = triplet of (model, features, variant)

<table>
<thead>
<tr>
<th>Model</th>
<th>Features</th>
<th>Variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>k-means</td>
<td>raw</td>
<td>no post-processing</td>
</tr>
<tr>
<td>BIRCH</td>
<td>feature set A</td>
<td>clip [0, 1]</td>
</tr>
<tr>
<td>Agglomerative</td>
<td>feature set B</td>
<td>drop correlated</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- Which one has the best clustering **quality**?
- What are the run-time **costs**?
Assessing Quality

• Any external evaluation metric
• Requirement: Labeled data
Assessing Quality

- Any **external evaluation metric**
- Requirement: **Labeled data**
Assessing Costs

• Idea: Use run-time complexities
Assessing Costs

- **Idea:** Use run-time complexities
- **Problem:** Identical estimates

```python
def func1(n=1000):
    x = []
    for i in range(n):
        x.append(i)
    return x
```

```python
from numba import jit

@jit
def func2(n=1000):
    x = []
    for i in range(n):
        x.append(i)
    return x
```

```python
def func3(n=1000):
    return [i for i in range(n)]
```

```python
import numpy as np

def func4(n=1000):
    return np.arange(n).tolist()
```
Assessing Costs

• Idea: Use run-time complexities
• Problem: Identical estimates may yield different run times

```python
import numpy as np

def func1(n=1000):
    x = []
    for i in range(n):
        x.append(i)
    return x

from numba import jit

@jit
def func2(n=1000):
    x = []
    for i in range(n):
        x.append(i)
    return x

def func3(n=1000):
    return [i for i in range(n)]

def func4(n=1000):
    return np.arange(n).tolist()  
```

Measure Run Time (10000 runs)

- O(n) ~675ms
- O(n) ~380ms -44%
- O(n) ~285ms -25%
- O(n) ~145ms -49%
Assessing Costs

- Idea: Use run-time complexities
- Problem: Identical estimates may yield different run times

```python
def func1(n=1000):
    x = []
    for i in range(n):
        x.append(i)
    return x
```

```python
def func2(n=1000):
    x = []
    for i in range(n):
        x.append(i)
    return x
```

```python
def func3(n=1000):
    return [i for i in range(n)]
```

```python
import numpy as np

def func4(n=1000):
    return np.arange(n).tolist()
```

**General issues:**
- Compiler optimizations
- Language mixtures (Java Native Interface)
- Language intrinsics (Python loop iteration vs. list comprehension)
Assessing Costs

Measure **actual run time** \( r \) on a concrete machine
Assessing Costs

Measure **actual run time** $r$ on a concrete machine

- Given: Set of $n$ time series of length $t$, sets of $p$ features

$r_{\text{Method}} =$
Assessing Costs

Measure **actual run time** $r$ on a concrete machine

- Given: Set of $n$ time series of length $t$, sets of $p$ features

$$r_{\text{Method}} = r_{\text{Model}}$$

Measure Model
Fitting: $n, p$
Assessing Costs

Measure **actual run time** $r$ on a concrete machine

- Given: Set of $n$ time series of length $t$, sets of $p$ features

\[ r_{\text{Method}} = r_{\text{Model}} + r_{\text{Features}} \]

- Measure Model Fitting: $n, p$
- Measure Feature Calculation: $n, t$
Assessing Costs

Measure **actual run time** \( r \) on a concrete machine

- Given: Set of \( n \) time series of length \( t \), sets of \( p \) features

\[
\begin{align*}
    r_{\text{Method}} &= r_{\text{Model}} + r_{\text{Features}} + r_{\text{Variant}} \\
    \text{Measure Model} &\quad \text{Fitting: } n, p \\
    \text{Measure Feature} &\quad \text{Calculation: } n, t \\
    \text{Measure Variant} &\quad \text{Calculation: } n, p
\end{align*}
\]
Assessing Costs

Measure **actual run time** $r$ on a concrete machine

- Given: Set of $n$ time series of length $t$, sets of $p$ features

\[ r_{\text{Method}} = r_{\text{Model}} + r_{\text{Features}} + r_{\text{Variant}} \]

- Robust: Measure multiple times $\rightarrow$ set of measurements $R$
Assessing Costs

Measure **actual run time** $r$ on a concrete machine

- Given: Set of $n$ time series of length $t$, sets of $p$ features

\[ r_{\text{Method}} = r_{\text{Model}} + r_{\text{Features}} + r_{\text{Variant}} \]

- Measure Model Fitting: $n, p$
- Measure Feature Calculation: $n, t$
- Measure Variant Calculation: $n, p$

- Robust: Measure multiple times $\rightarrow$ set of measurements $R$

\[ r = \frac{1}{|Q|} \sum_{r' \in Q} r' \quad Q = \{ r' \in R | r' \geq q_l(R) \land r' \leq q_u(R) \} \]
Assessing Costs

Measure **actual run time** $r$ on a concrete machine

- Given: Set of $n$ time series of length $t$, sets of $p$ features

$$ r_{\text{Method}} = r_{\text{Model}} + r_{\text{Features}} + r_{\text{Variant}} $$

- Robust: Measure multiple times $\rightarrow$ set of measurements $R$

$$ r = \frac{1}{|Q|} \sum_{r' \in Q} r' \quad Q = \{ r' \in R | r' \geq q_l(R) \land r' \leq q_u(R) \} $$

lower quantile

upper quantile
## Time Series Characteristics (TSC)

<table>
<thead>
<tr>
<th>Group</th>
<th>Subgroup</th>
<th>#Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distributional</td>
<td>Dispersion</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Dispersion (blockwise)</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Duplicates</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Distribution</td>
<td>16</td>
</tr>
<tr>
<td>Temporal</td>
<td>Dispersion</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Dispersion (blockwise)</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Similarity</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Frequency</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Linearity</td>
<td>44</td>
</tr>
<tr>
<td>Complexity</td>
<td>Entropy</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Complexity (miscellaneous)</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Flatness</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Peaks</td>
<td>8</td>
</tr>
<tr>
<td>Statistical Tests</td>
<td>-</td>
<td>2</td>
</tr>
</tbody>
</table>
## Time Series Characteristics (TSC)

<table>
<thead>
<tr>
<th>Group</th>
<th>Subgroup</th>
<th>#Features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Distribution</strong></td>
<td>Dispersion</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Dispersion (blockwise)</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Duplicates</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Distribution</td>
<td>16</td>
</tr>
<tr>
<td><strong>Temporal</strong></td>
<td>Dispersion</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Dispersion (blockwise)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Similarity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Frequency</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Linearity</td>
<td>44</td>
</tr>
<tr>
<td><strong>Complexity</strong></td>
<td>Entropy</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Complexity (miscellaneous)</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Flatness</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Peaks</td>
<td>8</td>
</tr>
<tr>
<td><strong>Statistical Tests</strong></td>
<td>-</td>
<td>2</td>
</tr>
</tbody>
</table>

4 main groups
13 subgroups
## Time Series Characteristics (TSC)

<table>
<thead>
<tr>
<th>Group</th>
<th>Subgroup</th>
<th>#Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distributional</td>
<td>Dispersion</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Dispersion (blockwise)</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Duplicates</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Distribution</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Dispersion</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Linearity</td>
<td>44</td>
</tr>
<tr>
<td>Complexity</td>
<td>Entropy</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Complexity (miscellaneous)</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Flatness</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Peaks</td>
<td>8</td>
</tr>
<tr>
<td>Statistical Tests</td>
<td>-</td>
<td>2</td>
</tr>
</tbody>
</table>

43 characteristics
167 features with parameterization
Evaluation

• Data: **UCR** time series classification archive
  o 128 datasets
  o Various domains (synthetic, sensors, motion, image, ECG, etc.)
Evaluation

- **Data:** UCR time series classification archive
  - 128 datasets
  - Various domains (synthetic, sensors, motion, image, ECG, etc.)
- **External evaluation metric:** ARI (adjusted Rand index)
Evaluation

- **Data:** UCR time series classification archive
  - 128 datasets
    - Various domains (synthetic, sensors, motion, image, ECG, etc.)
- External evaluation metric: **ARI** (adjusted Rand index)
- Run-time measurement: **30** runs, quantile range \([0.1, 0.9]\)
Evaluation: Methods

• Models:
Evaluation: Methods

• Models:
  o k-means
  o BIRCH
  o Agglomerative clustering (Ward’s linkage + Euclidean distance)
  o Agglomerative clustering (weighted average linkage + Euclidean dist.)
  o Agglomerative clustering (weighted average linkage + cosine distance)
Evaluation: Methods

• **Models:**
  - k-means
  - BIRCH
  - Agglomerative clustering (Ward’s linkage + Euclidean distance)
  - Agglomerative clustering (weighted average linkage + Euclidean dist.)
  - Agglomerative clustering (weighted average linkage + cosine distance)

• **Features:**
  - 4 main groups + 13 subgroups + all TSC
  - Raw time series data
Evaluation: Methods

• **Models:**
  - k-means
  - BIRCH
  - Agglomerative clustering (Ward’s linkage + Euclidean distance)
  - Agglomerative clustering (weighted average linkage + Euclidean dist.)
  - Agglomerative clustering (weighted average linkage + cosine distance)

• **Features:**
  - 4 main groups + 13 subgroups + all TSC
  - Raw time series data

• **Variants** (for features):
  - Dropping correlated features
  - Clipping to $[0, 1]$ + logarithm-based clipping + tangent-based clipping
  - All combinations of dropping + clipping variants
  - No post-processing
Evaluation: Methods

- **Models:**
  - k-means

- Clipping to [0, 1] + logarithm-based clipping + tangent-based clipping
- All combinations of dropping + clipping variants
- No post-processing
Evaluation: Methods

• **Models:**
  - k-means
  - BIRCH
  - Agglomerative clustering (Ward’s linkage + Euclidean distance)
  - Agglomerative clustering (weighted average linkage + Euclidean dist.)
  - Agglomerative clustering (weighted average linkage + cosine distance)

• **Features:**
  - 18 feature sets + raw
    - 4 main groups + 13 subgroups + all TSC
    - Raw time series data

• **Variants** (for features):
  - 8 variants
    - Dropping correlated features
    - Clipping to [0, 1] + logarithm-based clipping + tangent-based clipping
    - All combinations of dropping + clipping variants
    - No post-processing
Evaluation: Methods

• **Models:**
  - k-means
  - BIRCH
  - Agglomerative clustering (Ward's linkage + Euclidean distance)
  - Agglomerative clustering (weighted average linkage + Euclidean dist.)
  - Agglomerative clustering (weighted average linkage + cosine distance)

• **Features:**
  - 4 main groups + 13 subgroups + all TSC
  - Raw time series data

• **Variants (for features):**
  - Dropping correlated features
  - Clipping to [0, 1] + logarithm-based clipping + tangent-based clipping
  - All combinations of dropping
  - No post-processing

\[ 5 \times 18 \times 8 + 5 = 725 \text{ methods} \]
Results: Quality-Cost-Trade-off

Dataset: ElectricDevices
Type: Device
Samples: 16637
Time series length: 96
Results: Quality-Cost-Trade-off

Dataset: ElectricDevices
Type: Device
Samples: 16637
Time series length: 96
Results: Quality-Cost-Trade-off

Dataset: ElectricDevices
Type: Device
Samples: 16637
Time series length: 96
Results: Quality-Cost-Trade-off

Dataset: ElectricDevices
Type: Device
Samples: 16637
Time series length: 96
## Results: Quality-Cost-Trade-off

<table>
<thead>
<tr>
<th>Model</th>
<th>Features</th>
<th>Variant</th>
<th>ARI</th>
<th>Run time [s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>l</td>
<td>complexity</td>
<td></td>
<td>0.35</td>
<td>98.25</td>
</tr>
<tr>
<td>l</td>
<td>c_entropy</td>
<td>01_d</td>
<td>0.32</td>
<td>44.10</td>
</tr>
<tr>
<td>k</td>
<td>c_entropy</td>
<td>log_d</td>
<td>0.32</td>
<td>35.51</td>
</tr>
<tr>
<td>l</td>
<td>d Dispersion_b</td>
<td>log</td>
<td>0.32</td>
<td>15.30</td>
</tr>
<tr>
<td>l</td>
<td>d Dispersion_b</td>
<td>01</td>
<td>0.32</td>
<td>15.27</td>
</tr>
<tr>
<td>l</td>
<td>t Dispersion_b</td>
<td>01</td>
<td>0.31</td>
<td>14.49</td>
</tr>
<tr>
<td>k</td>
<td>t Dispersion_b</td>
<td>tan_d</td>
<td>0.31</td>
<td>6.55</td>
</tr>
<tr>
<td>k</td>
<td>t Dispersion_b</td>
<td>tan</td>
<td>0.31</td>
<td>6.53</td>
</tr>
<tr>
<td>k</td>
<td>d Dispersion_b</td>
<td>log</td>
<td>0.31</td>
<td>6.38</td>
</tr>
<tr>
<td>k</td>
<td>d Dispersion_b</td>
<td>01_d</td>
<td>0.31</td>
<td>6.38</td>
</tr>
<tr>
<td>k</td>
<td>d Dispersion_b</td>
<td>01</td>
<td>0.31</td>
<td>6.36</td>
</tr>
</tbody>
</table>

### Dataset:
- **Type:** ElectricDevices
- **Samples:** 16637
- **Time series length:** 96

Models: l = linkage, k = m-means. Features: group_subgroup (group abbreviated to first letter), _b = blockwise. Variants: empty = no post-processing, 01 = clip [0, 1], log = logarithm-based clipping, tan = tangent-based clipping, _d = variant + dropping correlated features.
Results: Quality-Cost-Trade-off

<table>
<thead>
<tr>
<th>Model</th>
<th>Features</th>
<th>Variant</th>
<th>ARI</th>
<th>Run time [s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>temporal</td>
<td>d</td>
<td>0.49</td>
<td>1.38</td>
</tr>
<tr>
<td>l</td>
<td>t Dispersion b</td>
<td>log</td>
<td>0.46</td>
<td>0.14</td>
</tr>
<tr>
<td>l</td>
<td>c flatness</td>
<td>01</td>
<td>0.44</td>
<td>0.13</td>
</tr>
<tr>
<td>l</td>
<td>c flatness</td>
<td></td>
<td>0.44</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Dataset: FaceFour
Type: Image
Samples: 112
Time series length: 350
Results: Quality-Cost-Trade-off

<table>
<thead>
<tr>
<th>Model</th>
<th>Features</th>
<th>Variant</th>
<th>ARI</th>
<th>Run time [s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>t_similarity</td>
<td>01_d</td>
<td>0.52</td>
<td>4.02</td>
</tr>
<tr>
<td>b</td>
<td>t_linearity</td>
<td></td>
<td>0.49</td>
<td>0.61</td>
</tr>
<tr>
<td>lw</td>
<td>t_linearity</td>
<td>01</td>
<td>0.36</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Dataset: ItalyPowerDemand
Type: Sensor
Samples: 1096
Time series length: 24
Results: Quality-Cost-Trade-off

<table>
<thead>
<tr>
<th>Model</th>
<th>Features</th>
<th>Variant</th>
<th>ARI</th>
<th>Run time [s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>temporal</td>
<td>01_d</td>
<td>0.39</td>
<td>10.87</td>
</tr>
<tr>
<td>b</td>
<td>t_linearity</td>
<td>log</td>
<td>0.39</td>
<td>3.35</td>
</tr>
<tr>
<td>b</td>
<td>t_linearity</td>
<td>01</td>
<td>0.37</td>
<td>3.29</td>
</tr>
</tbody>
</table>

**Dataset:** FiftyWords
**Type:** Image
**Samples:** 905
**Time series length:** 270
Conclusion

- Clustering **method selection** based on actual run-time costs:
  - Models
  - Features
  - Variants
Conclusion

• Clustering **method selection** based on actual run-time costs:
  o Models
  o Features
  o Variants

• User can selected method via **quality-cost trade-off**
Conclusion

• Clustering **method selection** based on actual run-time costs:
  o Models
  o Features
  o Variants

• User can selected method via **quality-cost trade-off**

• Future work: Apply to **other areas**:
  • Classification
  • Forecasting
  • …
Selecting Time Series Clustering Methods based on Run-Time Costs

Andreas Schörgenhumer
Paul Grünbacher
Hanspeter Mössenböck

12.11.2020
# TSC: Distributional

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Characteristic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dispersion</td>
<td>kurtosis</td>
<td>measure of tailedness</td>
</tr>
<tr>
<td></td>
<td>skewness</td>
<td>measure of asymmetry</td>
</tr>
<tr>
<td></td>
<td>shift</td>
<td>mean minus the median of those values that are smaller than the mean</td>
</tr>
<tr>
<td>Dispersion (blockwise)</td>
<td>lumpiness</td>
<td>variance of the variances of blocks</td>
</tr>
<tr>
<td></td>
<td>stability</td>
<td>variance of the mean of blocks</td>
</tr>
<tr>
<td>Duplicates</td>
<td>normalized_duplicates_max</td>
<td>number of duplicates that have the maximum value of the data</td>
</tr>
<tr>
<td></td>
<td>normalized_duplicates_min</td>
<td>number of duplicates that have the minimum value of the data</td>
</tr>
<tr>
<td></td>
<td>percentage_of_reoccurring_datapoints</td>
<td>number of unique duplicates compared to the number of unique values</td>
</tr>
<tr>
<td></td>
<td>percentage_of_reoccurring_values</td>
<td>number of duplicates compared to the length of the data</td>
</tr>
<tr>
<td></td>
<td>percentage_of_unique_values</td>
<td>number of unique values compared to the length of the data</td>
</tr>
<tr>
<td>Distributio n</td>
<td>quantile</td>
<td>threshold below which $x%$ of the ordered values of the data are, giving a hint on the distribution</td>
</tr>
<tr>
<td></td>
<td>ratio_beyond_r_sigma</td>
<td>ratio of values that are more than a factor $r \cdot \sigma$ away from the mean</td>
</tr>
<tr>
<td></td>
<td>ratio_large_standard_deviation</td>
<td>ratio between the standard deviation and the (max – min) range of the data (based on the &quot;range rule of thumb&quot;)</td>
</tr>
</tbody>
</table>
## TSC: Temporal

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Characteristic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dispersion</strong></td>
<td>mean_abs_change</td>
<td>average absolute difference of two consecutive values</td>
</tr>
<tr>
<td></td>
<td>mean_second_derivative_central</td>
<td>measure of the rate of the rate of change</td>
</tr>
<tr>
<td><strong>Dispersion (blockwise)</strong></td>
<td>level_shift</td>
<td>maximum difference in mean between consecutive blocks</td>
</tr>
<tr>
<td></td>
<td>variance_change</td>
<td>maximum difference in variance between consecutive blocks</td>
</tr>
<tr>
<td><strong>Similarity</strong></td>
<td>hurst</td>
<td>measure of long-term memory of a time series, related to auto-correlation</td>
</tr>
<tr>
<td></td>
<td>autocorrelation</td>
<td>correlation of a signal with a lagged version of itself</td>
</tr>
<tr>
<td><strong>Frequency</strong></td>
<td>periodicity</td>
<td>power (intensity) of specified frequencies in the signal (based on the periodogram)</td>
</tr>
<tr>
<td></td>
<td>agg_periodogram</td>
<td>results of user-defined aggregation functions (e.g., fivenum) calculated on the periodogram</td>
</tr>
<tr>
<td><strong>Linearity</strong></td>
<td>linear_trend_slope</td>
<td>measure of linearity: slope</td>
</tr>
<tr>
<td></td>
<td>linear_trend_rvalue2</td>
<td>measure of linearity: $r^2$ (coefficient of determination)</td>
</tr>
<tr>
<td></td>
<td>agg_linear_trend_slope</td>
<td>variance-aggregated slopes of blocks</td>
</tr>
<tr>
<td></td>
<td>agg_linear_trend_rvalue2</td>
<td>mean-aggregated $r^2$ of blocks</td>
</tr>
<tr>
<td></td>
<td>c3</td>
<td>measure of non-linearity (originally from the physics domain)</td>
</tr>
<tr>
<td></td>
<td>time_reversal_asymmetry_statistic</td>
<td>asymmetry of the time series if reversed, which can be a measure of non-linearity</td>
</tr>
</tbody>
</table>
# TSC: Complexity

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Characteristic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entropy</td>
<td>binned_entropy</td>
<td>fast entropy estimation based on equidistant bins</td>
</tr>
<tr>
<td></td>
<td>kullback_leibler_score (KL score)</td>
<td>maximum difference of KL divergences between consecutive blocks, where the KL divergence is a measure of how two probability distributions differ</td>
</tr>
<tr>
<td></td>
<td>index_of_kullback_leibler_score</td>
<td>relative location where the maximum KL score was found</td>
</tr>
<tr>
<td>Complexity (misc.)</td>
<td>cid_ce</td>
<td>measure of complexity invariance</td>
</tr>
<tr>
<td></td>
<td>permutation_analysis</td>
<td>measure of complexity through permutation</td>
</tr>
<tr>
<td></td>
<td>swinging_door_compression_rate</td>
<td>compression ratio of the signal under a given error tolerance $\epsilon$</td>
</tr>
<tr>
<td>Flatness</td>
<td>normalized_crossing_points</td>
<td>number of times a time series crosses the mean line (based on fickleness)</td>
</tr>
<tr>
<td></td>
<td>normalized_above_mean</td>
<td>number of values that are higher than the mean</td>
</tr>
<tr>
<td></td>
<td>normalized_below_mean</td>
<td>number of values that are lower than the mean</td>
</tr>
<tr>
<td></td>
<td>normalized_longest_strike_above_mean</td>
<td>relative length of the longest series of consecutive values above the mean</td>
</tr>
<tr>
<td></td>
<td>normalized_longest_strike_below_mean</td>
<td>relative length of the longest series of consecutive values below the mean</td>
</tr>
<tr>
<td></td>
<td>flat_spots</td>
<td>maximum run-length of values when divided into quantile-based bins</td>
</tr>
<tr>
<td>Peaks</td>
<td>normalized_number_peaks</td>
<td>number of peaks, where a peak of support $n$ is defined as a value which is bigger than its $n$ left and $n$ right neighbors</td>
</tr>
<tr>
<td></td>
<td>step_changes</td>
<td>number of times the time series significantly shifts its value range</td>
</tr>
</tbody>
</table>
# TSC: Statistical Tests

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Characteristic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>adf</td>
<td>augmented Dickey-Fuller (ADF) test for unit root presence</td>
</tr>
<tr>
<td></td>
<td>kpss</td>
<td>Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test for stationarity</td>
</tr>
</tbody>
</table>