TU Clausthal

Master Thesis Final Presentation

Benchmarking of Models for Unsupervised Segmentation of Multivariate Time Series With a Novel Homogeneity Metric

Clausthal University of Technology
Institute for Software and Systems Engineering

2022-04-07

[1]*Automotive Electronic Systems. Clemson Vehicular Electronics Laboratory: Automotive Electronic Systems. (n.d.). Retrieved March 27, 2022, from https://cecas.clemson.edu/cvel/auto/systems/auto-systems.html*

(Automotive) Software Systems

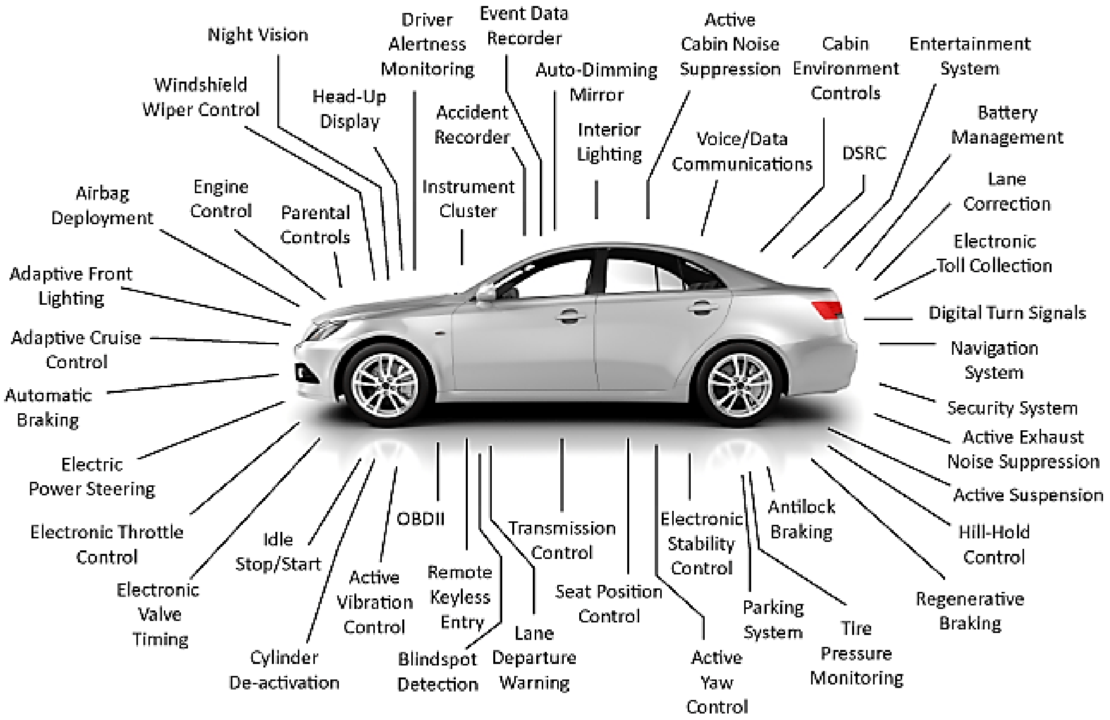
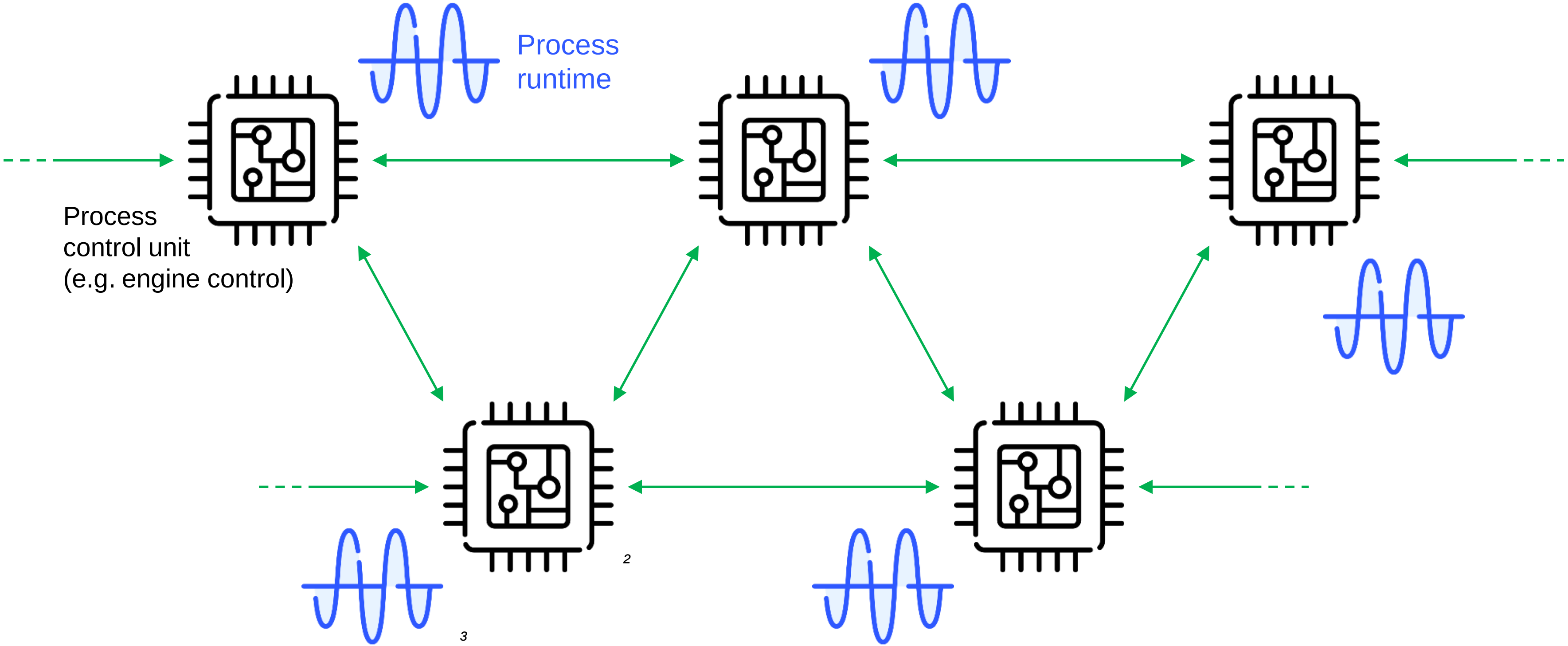


Figure 1 – Automotive control units[1]

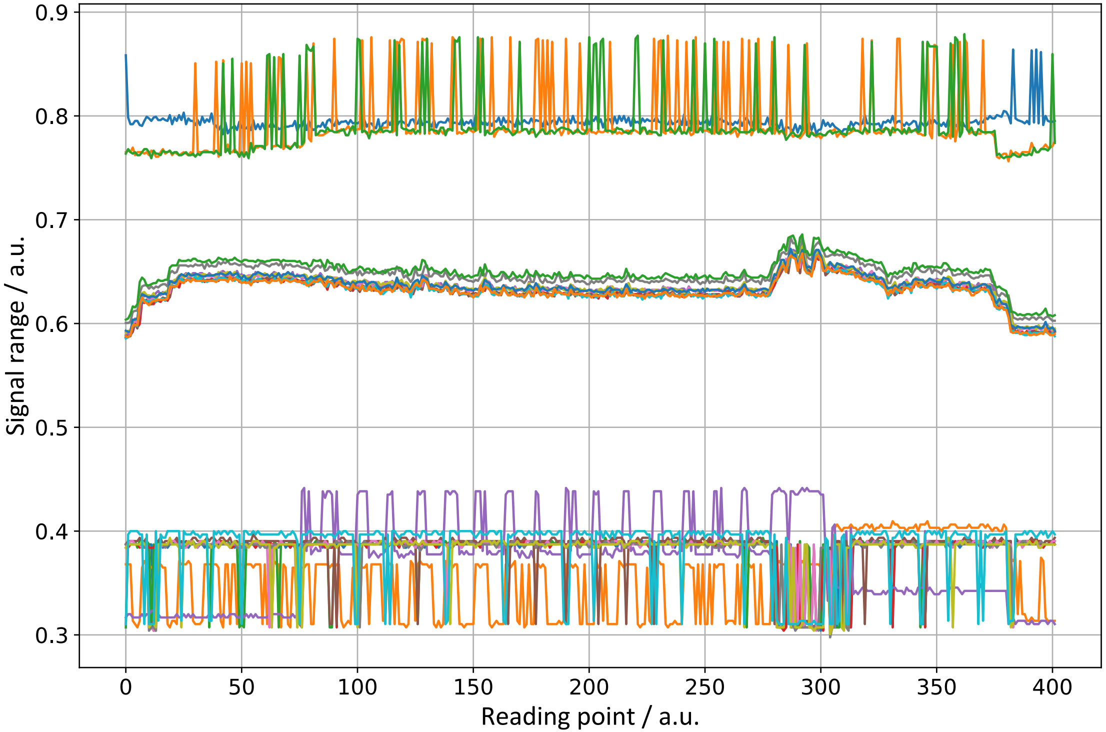
[2]Icon made by Vitaly Gorbachev from Flaticon

[3]Icon made by Freepik from Flaticon

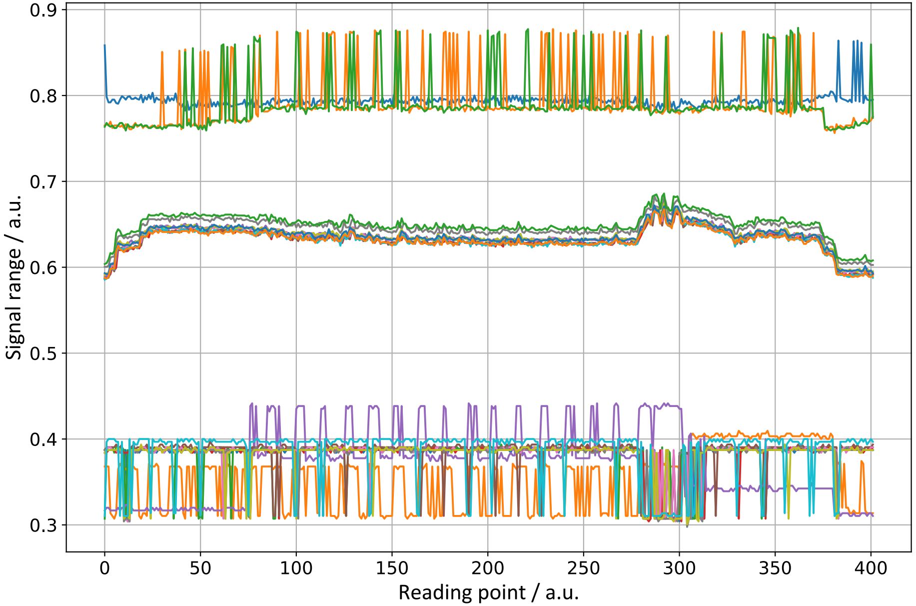
Control Units in (Automotive) Software Systems



Recorded Process Runtimes from Automotive Software System

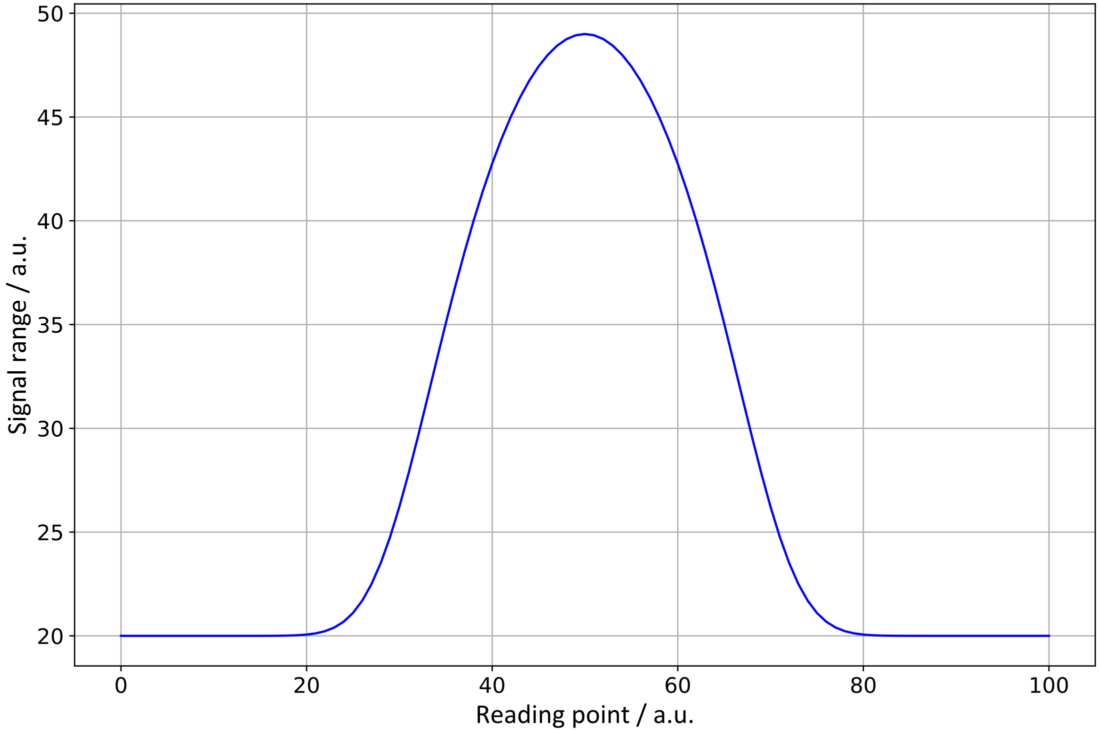


Recorded Process Runtimes from Automotive Software System

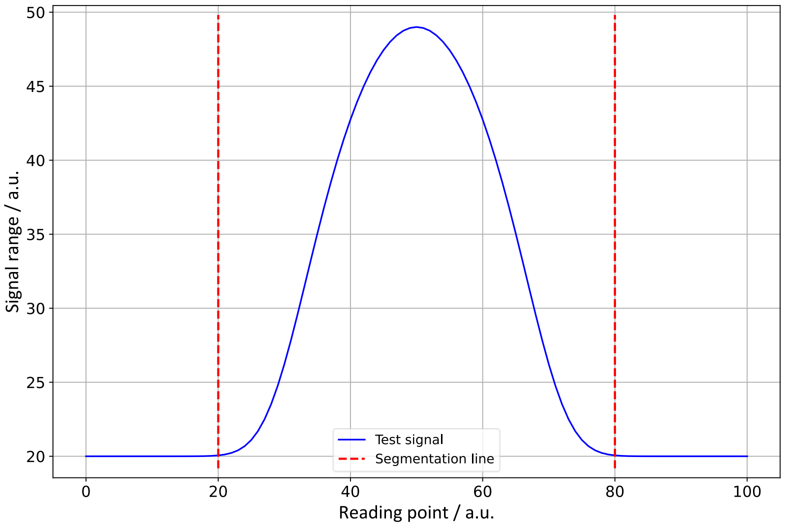


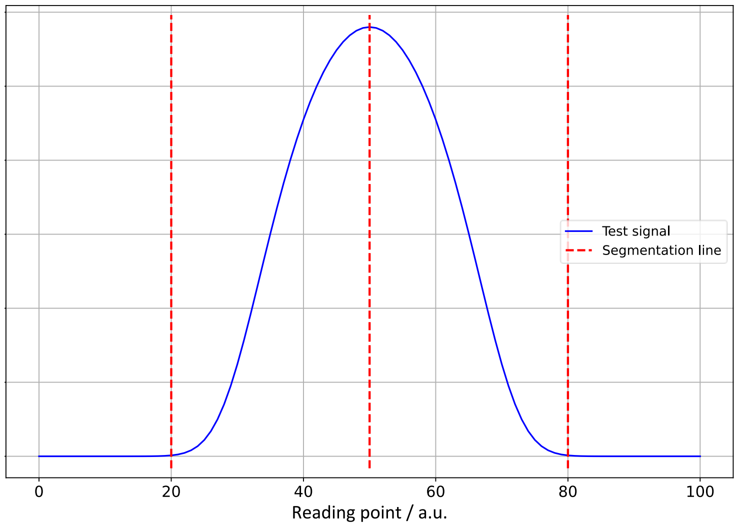
1. How can homogeneity in segments of a multivariate time series be defined and measured?
2. How can the ideal number of segments in a multivariate time series be determined without supervision?
3. How can precise segmentation indices for maximizing homogeneity in segments be determined?

Analysis of Signals

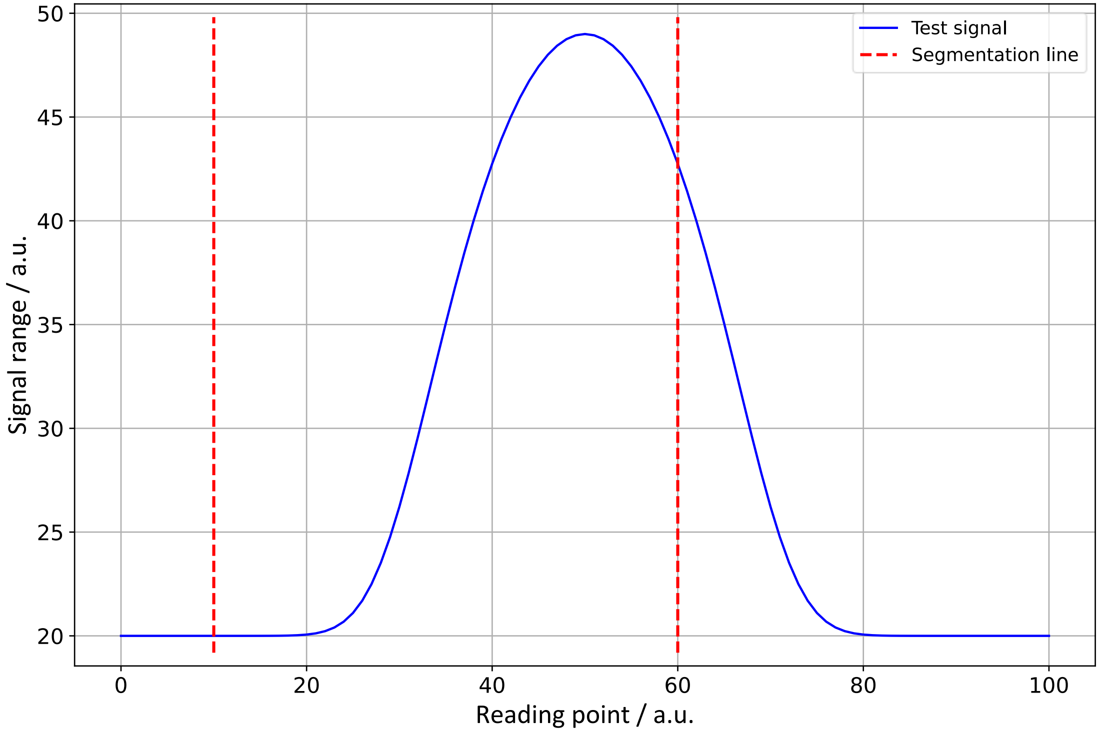


Analysis of Signals





Analysis of Signals



TU Clausthal

Analysis of Multivariate Signal

$$\textit{Homogeneity} \coloneqq \frac{\left(\frac{\sum_{p=1}^{l}\sum_{s=1}^{S}\left(\max\left(s_p\right) - \min\left(s_p\right)\right) * n_s}{\sum_{p=1}^{l}\left(\max(p) - \min(p)\right) * n} + \frac{\sqrt{n} * S}{1.7n}\right)}{1 + \frac{\sqrt{n}}{1.7}}$$

s – segment
p – process
n – total number of data points (i.e. columns) per process
l – total number of processes (i.e. signals) in multivariate data set
S – total number of segments in data set
s_p – segment *s* in process *p*
n_s – number of data points in segment *s*

Analysis of Multivariate Signal

$$Homogeneity \coloneqq \frac{\left(\frac{\sum_{p=1}^{l}\sum_{s=1}^{S}\left(\max(s_p)-\min(s_p)\right)*n_s}{\sum_{p=1}^{l}(\max(p)-\min(p))*n} + \frac{\sqrt{n}*S}{1.7n}\right)}{1+\frac{\sqrt{n}}{1.7}}$$

s – segment
p – process
n – total number of data points (i.e. columns) per process
l – total number of processes (i.e. signals) in multivariate data set
S – total number of segments in data set
s_p – segment *s* in process *p*
n_s – number of data points in segment *s*

Analysis of Multivariate Signal

$$Homogeneity := \frac{\left(\frac{\sum_{p=1}^{l}\sum_{s=1}^{S}\left(\max\left(s_p\right)-\min\left(s_p\right)\right)*n_s}{\sum_{p=1}^{l}\left(\max(p)-\min(p)\right)*n}+\frac{\sqrt{n}*S}{1.7n}\right)}{1+\frac{\sqrt{n}}{1.7}}$$

s – segment
p – process
n – total number of data points (i.e. columns) per process
l – total number of processes (i.e. signals) in multivariate data set
S – total number of segments in data set
s_p – segment *s* in process *p*
n_s – number of data points in segment *s*

TU Clausthal

Analysis of Multivariate Signal

$$Homogeneity := \frac{\left(\frac{\sum_{p=1}^{l}\sum_{s=1}^{S}\left(\max\left(s_p\right)-\min\left(s_p\right)\right)*n_s}{\sum_{p=1}^{l}\left(\max(p)-\min(p)\right)*n}+\frac{\sqrt{n}*S}{1.7n}\right)}{1+\frac{\sqrt{n}}{1.7}}$$

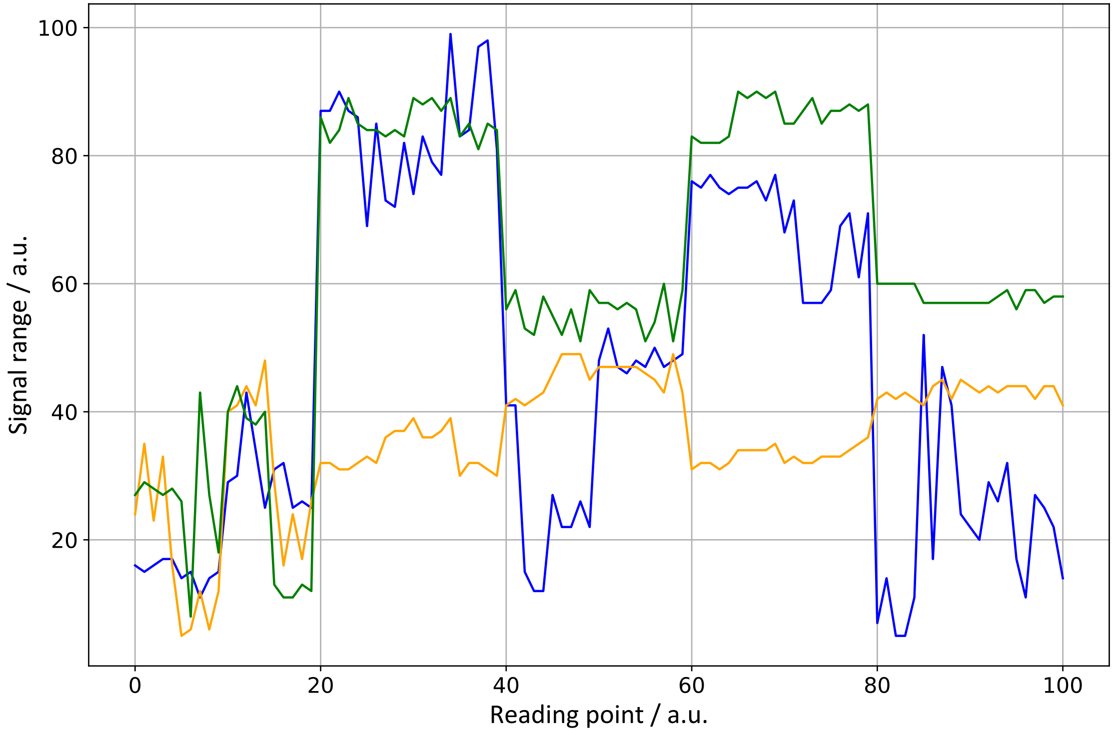
s – segment
p – process
n – total number of data points (i.e. columns) per process
l – total number of processes (i.e. signals) in multivariate data set
S – total number of segments in data set
s_p – segment *s* in process *p*
n_s – number of data points in segment *s*

Analysis of Multivariate Signal

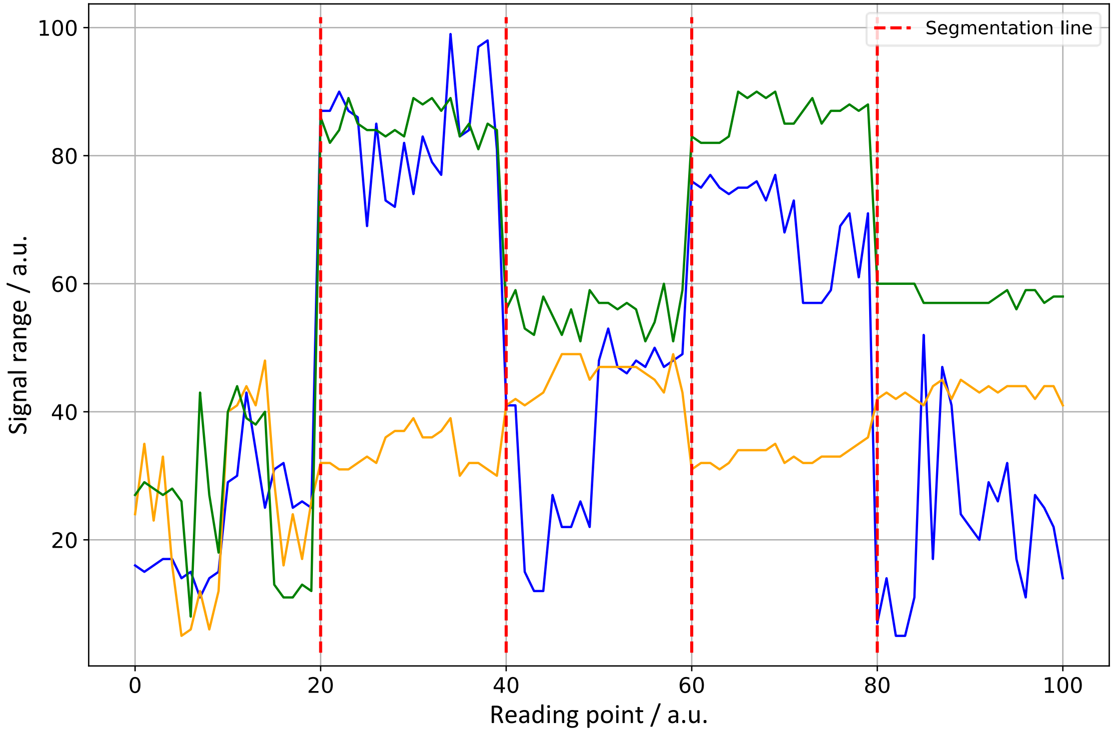
$$Homogeneity \coloneqq \frac{\left(\frac{\sum_{p=1}^{l}\sum_{s=1}^{S}\left(\max\left(s_p\right)-\min\left(s_p\right)\right)*n_s}{\sum_{p=1}^{l}\left(\max(p)-\min(p)\right)*n}+\frac{\sqrt{n}*S}{1.7n}\right)}{1+\frac{\sqrt{n}}{1.7}}$$

s – segment
p – process
n – total number of data points (i.e. columns) per process
l – total number of processes (i.e. signals) in multivariate data set
S – total number of segments in data set
s_p – segment *s* in process *p*
n_s – number of data points in segment *s*

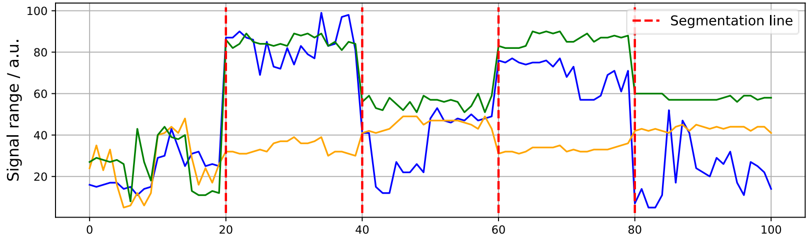
Analysis of Multivariate Signal



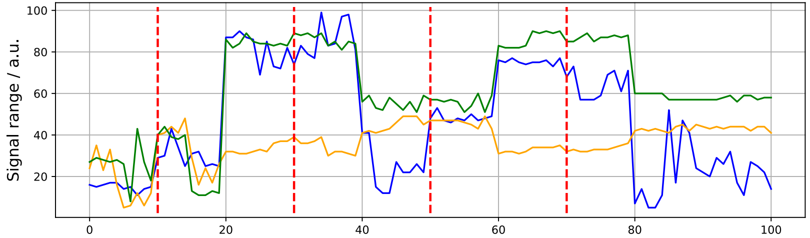
Analysis of Multivariate Signal



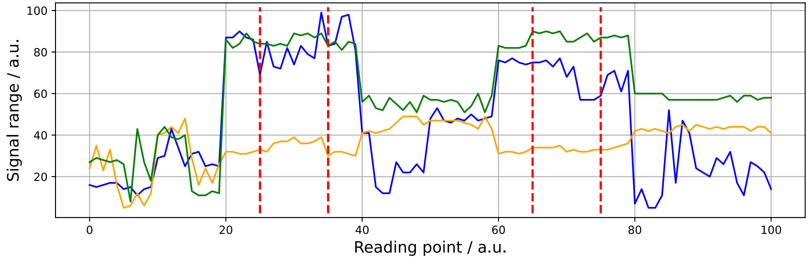
Analysis of Multivariate Signal



Homogeneity: 0.08



Homogeneity: 0.12

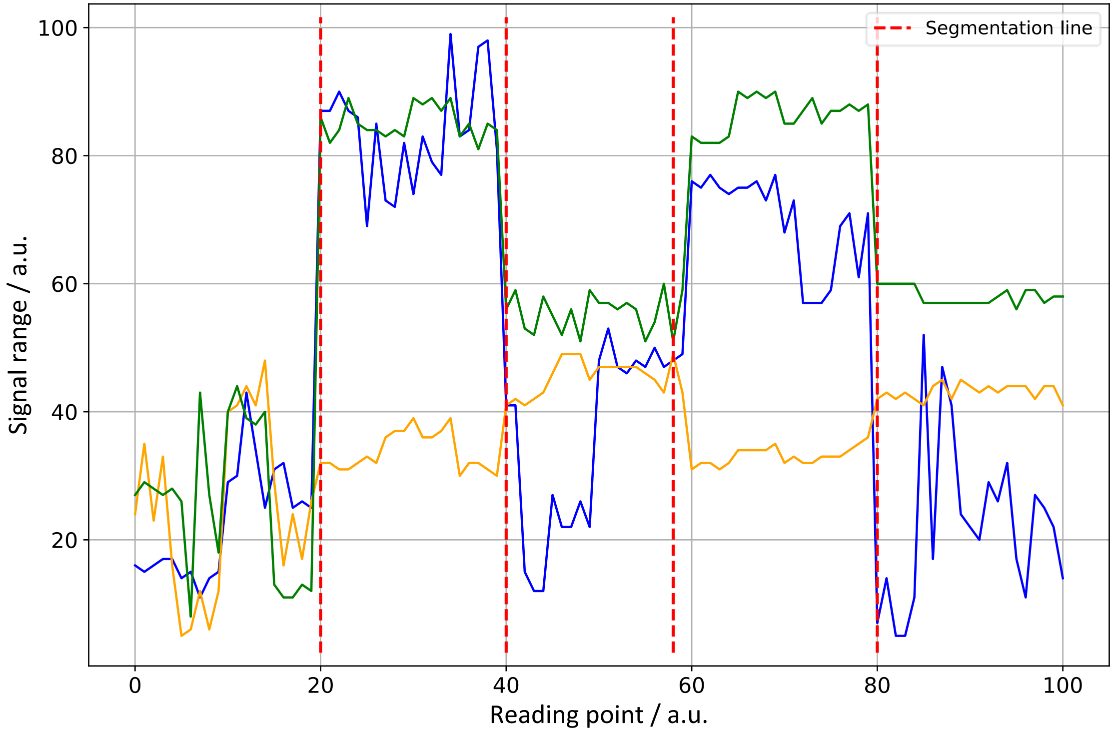


Homogeneity: 0.13

Time Series Segmentation Models

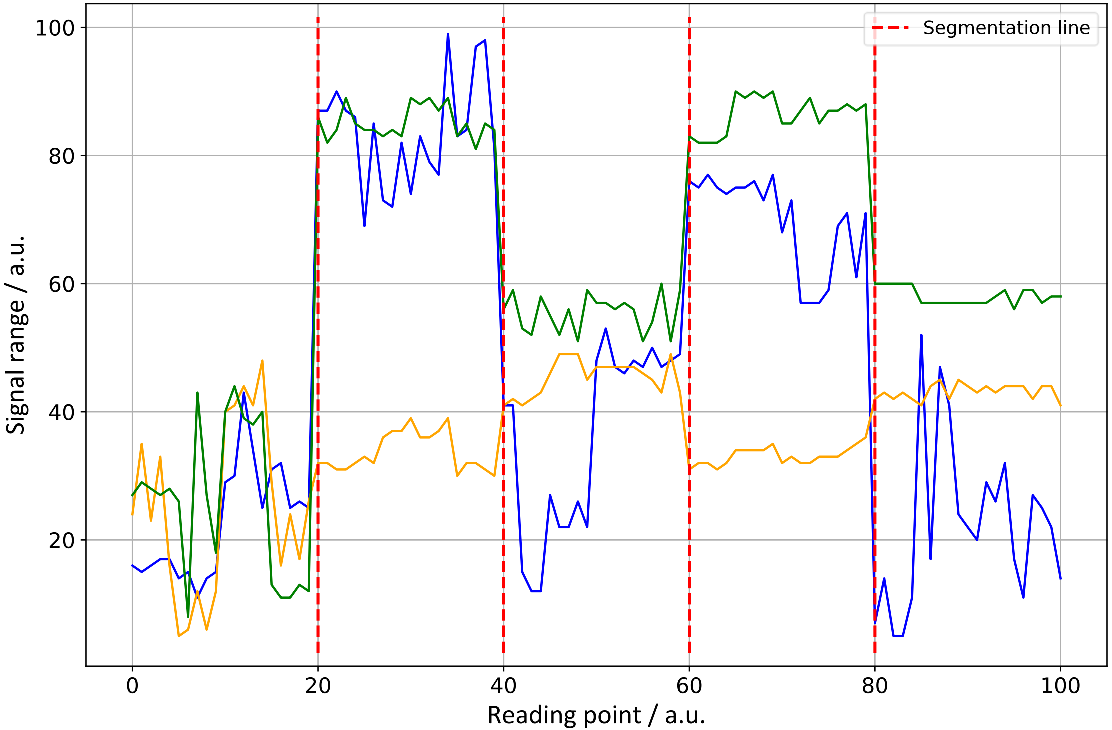
- The time series segmentation models used in this work are state-of-the-art statistical models for analyzing and clustering (i.e. segmenting) multivariate time series
- Two models are used to segment multivariate time series:
  - Principal component analysis based fuzzy clustering (PCAFC)
  - Multiple hidden Markov model for regression (MHMMR)
- Both models are first validated for their capability of segmenting multivariate time series before applying them to multivariate time series

Time Series Segmentation Models - PCAFC

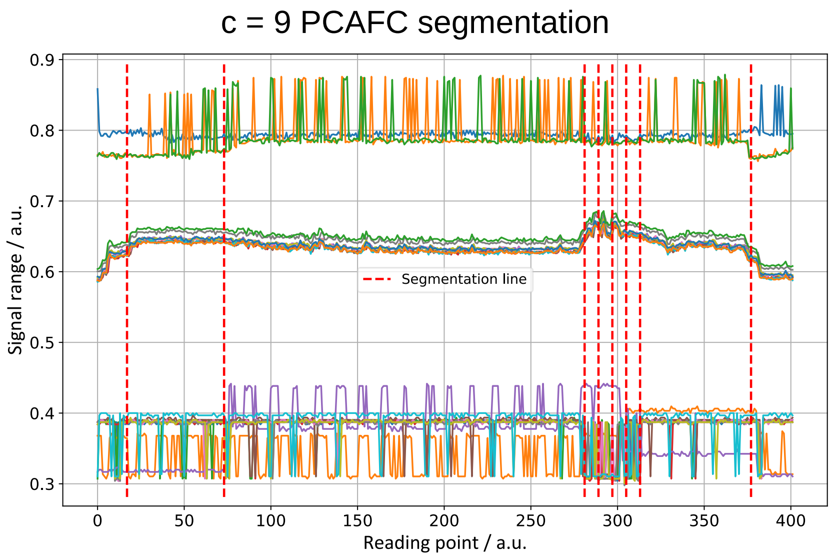


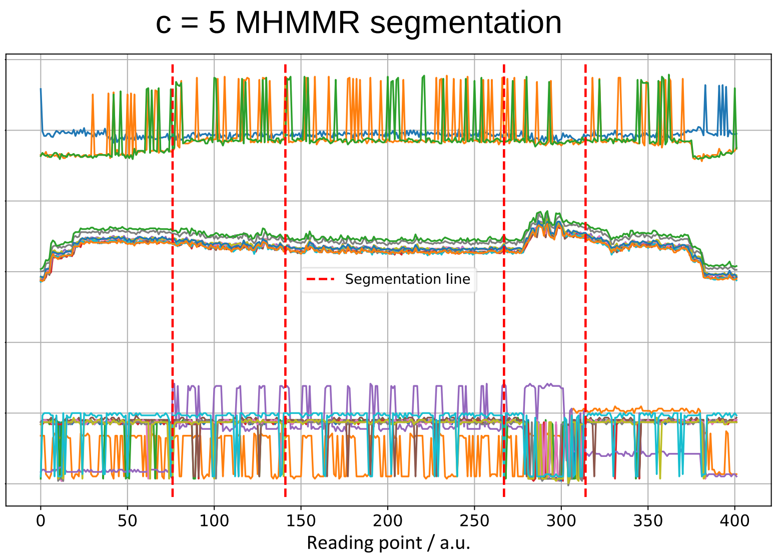
TU Clausthal

Time Series Segmentation Models - MHMMR



Exemplary PCAFC and MHMMR segmentations

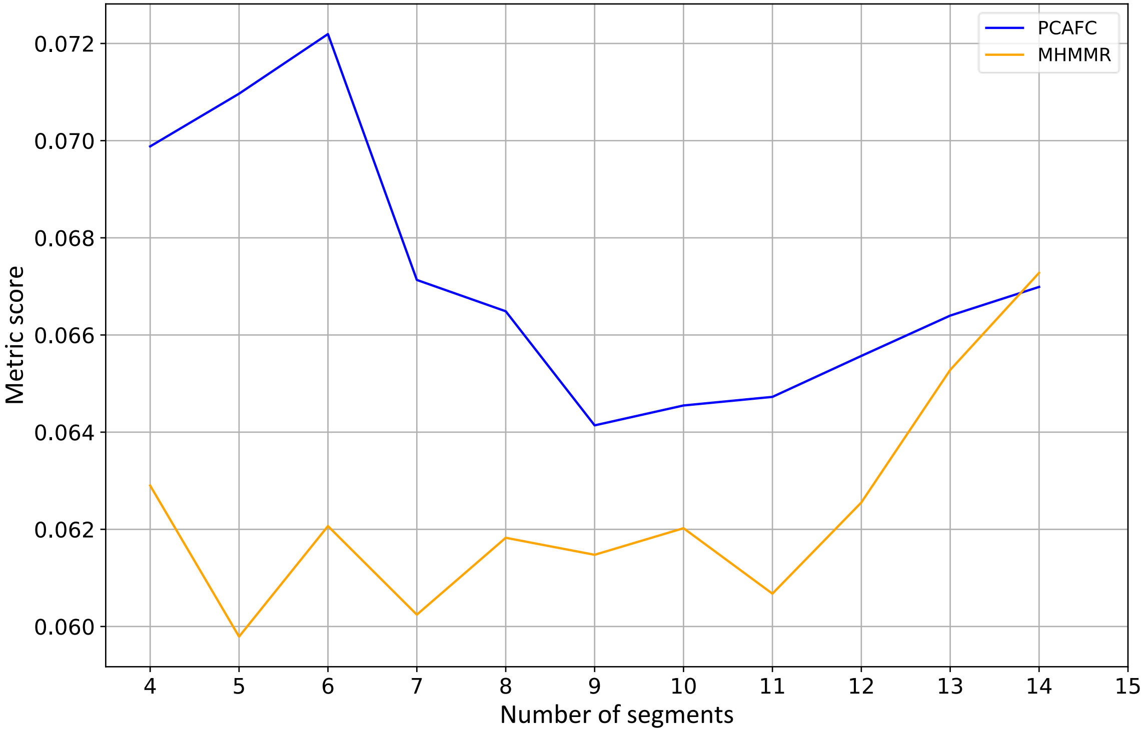




Time Series Segmentation Models - Evaluation

# of segments	Segmentation algorithm	
	PCAFC	MHMMR
Four	0.070	0.063
Five	0.071	**0.060**
Six	0.072	0.062
Seven	0.067	0.060
Eight	0.066	0.062
Nine	**0.064**	0.061
Ten	0.065	0.062
Eleven	0.065	0.061
Twelve	0.066	0.063
Thirteen	0.066	0.065
Fourteen	0.067	0.067

Time Series Segmentation Models - Evaluation



Conclusion and Future Outlook

- This work was motivated by the problem of state extraction from (i.e. segmentation of) multivariate time series.
- A novel metric for defining and measuring homogeneity for segmentations was defined and two segmentation models, namely PCAFC and MHMMR, were presented.
- The segmentation models were first applied to a multivariate test signal and then to a multivariate time series taken out of industry data.
- Findings: Both segmentation models show relevant results, which could be successfully measured with the novel homogeneity metric
- Future Outlook: The presented tools, models, and metric should be incorporated into future works in the field of multivariate time series analysis

Recommendable References

- “Segmentation of Multivariate Time-series”. In: *Cluster Analysis for Data Mining and System Identification*. Basel: Birkhäuser Basel, 2007, pp. 253–273. ISBN: 978-3-7643-7988-9. doi: https://doi.org/10.1007/978-3-7643-7988-9_6.
- Faicel Chamroukhi et al. “Model-based clustering with Hidden Markov Model regression for time series with regime changes”. In: *The 2011 International Joint Conference on Neural Networks (2011)*, pp. 2814–2821.

Thank you for your attention!