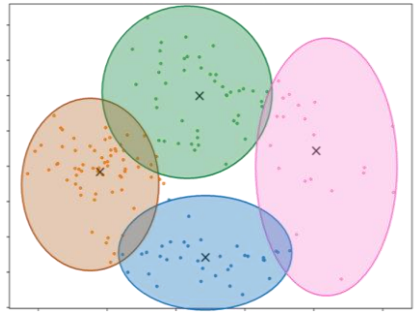
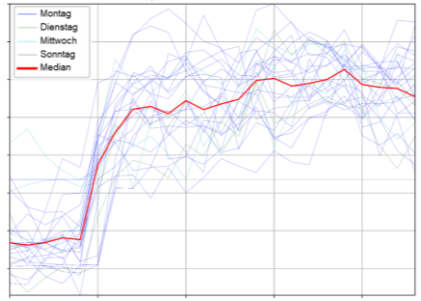
Clustering von Druckluftvolumenströmen zur Erkennung von Anomalien mit maschinellem Lernen

16. Symposium Energieinnovation, 12.-14.02.2020, Graz/Austria

Christian Dierolf, Alexander Sauer, Ivan Bogdanov

- Druckluftverwendung, -bedarf und -leckage
- Warum maschinelles Lernen?
- Untersuchte Messungen und angewendete Methoden
- Ergebnisse
- Bewertung und Ausblick





Fraunhofer IPA

Druckluftverwendung, -bedarf und -leckage

- Für die Erzeugung von Druckluft werden in Deutschland 3 % des nationalen Gesamtstrombedarfs eingesetzt.
- In Industriebetrieben ist Druckluft eine teure Energieform mit jährlich verursachenden Kosten von fast 1,5 Mrd. EUR [2] [3] [4].
- Druckluftleckagen verursachen in Deutschland bis zu einem Drittel des gesamten Druckluftverbrauchs [5]
- Ein erhöhter bzw. im Profil abweichender Druckluftverbrauch kann sich abhängig von dessen Signifikanz deutlich im Tagesprofil ausprägen.
- Lediglich auf Basis der dargestellten Zeitreihen der Verbrauchsdaten in einem Monitoringsystem fällt es schwer, eine Aussage über ein normales oder anormales Verhalten zu treffen.

Wie eignet sich unüberwachtes maschinelles Lernen im Speziellen – Clustering[1] – zur Erkennung von Anomalien und damit zur Reduzierung des Druckluftverbrauchs?

[1] Clustering beschreibt das Finden von Ähnlichkeitsstrukturen in Daten und deren anschließende Zuordnung zu Gruppen.

Warum maschinelles Lernen?

- Mit zunehmender Steigerung des Digitalisierungsgrad werden Daten immer entscheidender:
  - Unterstützung bei Vorhersagen: „Was wird passieren?“ (predictive analytics)
  - Automatisierte Entscheidungsfindung: „Wie soll reagiert werden?“ (prescriptive analytics)
- Für die Auswahl der Algorithmen stellen sich hinsichtlich dem Einsatzzweck und Modellleistung drei Kernfragen:
  - Welche Messdaten sind wertvoll, um Wissen zu generieren?
  - Welche Gesetzmäßigkeit kann aus den vorliegenden Messdaten abgeleitet werden?
  - Wie kann schnell und nachvollziehbar eine Handlungsempfehlung ermittelt werden?

Fokus im Anwendungsfall: Erkennung von charakteristischen Verläufen bzw. den Abweichungen in Form von Anomalien[1].

[1] Anomalien sind Abweichungen von einem erwarteten Verhalten, Muster oder einer Struktur.

Warum maschinelles Lernen?

Modelle des maschinellen Lernens werden aus historischen Daten trainiert

Trainingsart: Unterscheidung in überwachtes und unüberwachtes Lernen

Überwachtes Lernen: Zuordnung von Merkmalen (X) und Zielgrößen (Y bzw. Label) bekannt, da sie explizit vorgegeben wird.

- Grundverständnis der zu analysierenden Daten durch den Anwender erforderlich.
- Aufwendige Zuordnung der Daten kann manuell oder automatisiert bspw. unter Verwendung von Betriebsdaten durchgeführt werden.

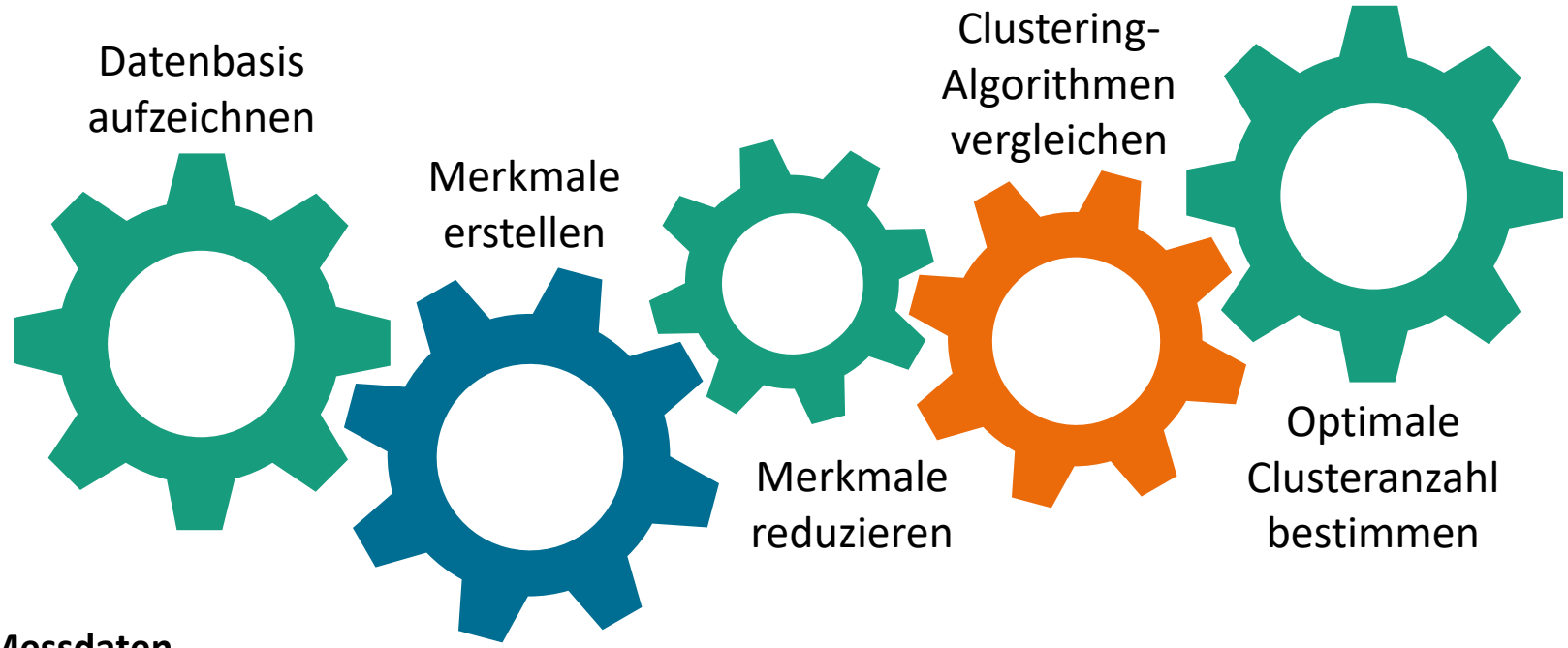
Unüberwachtes Lernen: Charakterisiert durch die Verwendung einer unbekannten nicht festgelegten Zuordnung der Daten. [6]

- Kein zusätzlicher Aufwand für den Anwender, da die Zuordnung nicht vorab festgelegt wird.
- Bei dieser Art des Lernens wird versucht, typische Daten bzw. Datengruppen wie bspw. „Maschine in Leerlauf“ im Signalrauschen zu erkennen.

Unüberwachtes maschinellem Lernen geeignet, da keine zusätzlichen Informationen (wie Produktionsauslastung oder Betriebszustände der Kompressoren) verwendet werden.

Untersuchte Messungen und angewendete Methoden

Vorgehensweise und Beschreibung der Messdaten



Beschreibung der Messdaten

- Gesamtvolumenstrom einer Druckluftanlage, zentral gemessen nach der Druckluftaufbereitung.
- Sechs Kompressoren mit insgesamt 130 kW Nennleistung und einem Druckluftspeicher.
- Zwei-Schicht-Betrieb, Messzeitraum von fast sechs Monaten Messdaten, Messintervall von 15 Minuten.

Untersuchte Messungen und angewendete Methoden

Merkmalserstellung und Merkmalsreduktion

Merkmalserstellung

- Unterteilung der Messdaten in 163 Tagesmessungen.
- Für jede Tagesmessung wird ein Merkmalssatz von 78 unterschiedlichen Merkmalen (d) erstellt.
  - Zeitbereich (72): Minima, Maxima und Mittelwerte der Stundenwerte.
  - Frequenzbereich (6): Minima, Maxima und Mittelwerte der der Amplituden und Frequenzen aus dem Tagesspektrum.

Merkmalsreduktion

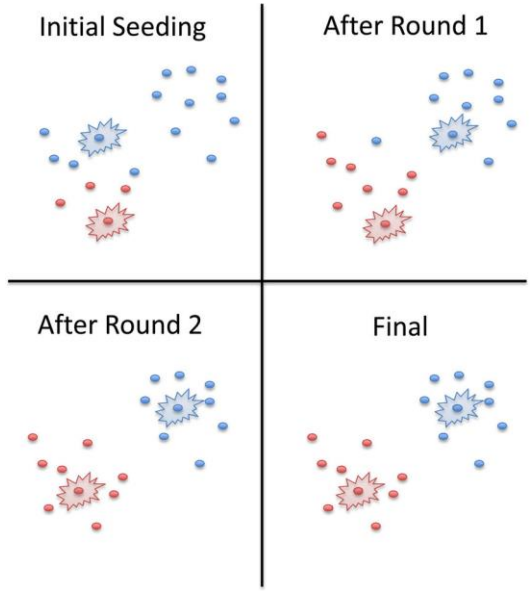
- Reduktion der Merkmale, um den aufgespannten Merkmalsraum ausreichend zu befüllen. Für jedes unterschiedliche Merkmal ist die Potenzierung der Daten empfohlen (10^d)[1]. [8]
- Im Anwendungsfall sind die Merkmale nicht manuell, sondern über die Hauptkomponentenanalyse (PCA) reduziert [9]. Die Signifikanz jedes Merkmals nicht ausweisbar.[2]

[1] Problematik: „Curse of Dimensionality“
[2] Die Merkmale sind auf die definierte Anzahl an Hauptkomponenten informationsverdichtet.

Untersuchte Messungen und angewendete Methoden

Clustering-Algorithmen am Beispiel von kMeans

- Nach der Reduzierung der 78 Merkmale auf deren zwei bzw. drei Hauptkomponenten
- Die Hauptkomponenten der Merkmalssätze werden im kMeans-Algorithmus wie folgt geclustert [10], exemplarisch in Abbildung rechts:
  - I. Zufällige Mittelpunkte mit der Anzahl der definierte Clustern werden gewählt.
  - II. Jeder Datenpunkt wird dem am nächsten liegenden Cluster-Mittelpunkt zugeordnet.
  - III. Das geometrische Zentrum der zugeordneten Punkte wird bestimmt und der Mittelpunkt dorthin verschoben.
  - IV. Nach anfänglich zufälliger Wahl der Mittelpunkte (I) werden die Schritte II bis IV so lange wiederholt, bis sich die Mittelpunkte unwesentlich verändern.



K-means clustering am Beispiel von zwei Clustern [15]

Untersuchte Messungen und angewendete Methoden

Bestimmung der optimalen Clusteranzahl

- KMeans liefert im Vergleich zu den anderen untersuchten Algorithmen (Gaussian mixture models [11] und Ward hierarchical clustering [12]) die aussagekräftigsten Clusterergebnisse. Das zeigt sich im Clusterplot und in ähnlichen Verläufen des Volumenstroms von Tagesmessungen desselben Clusters.

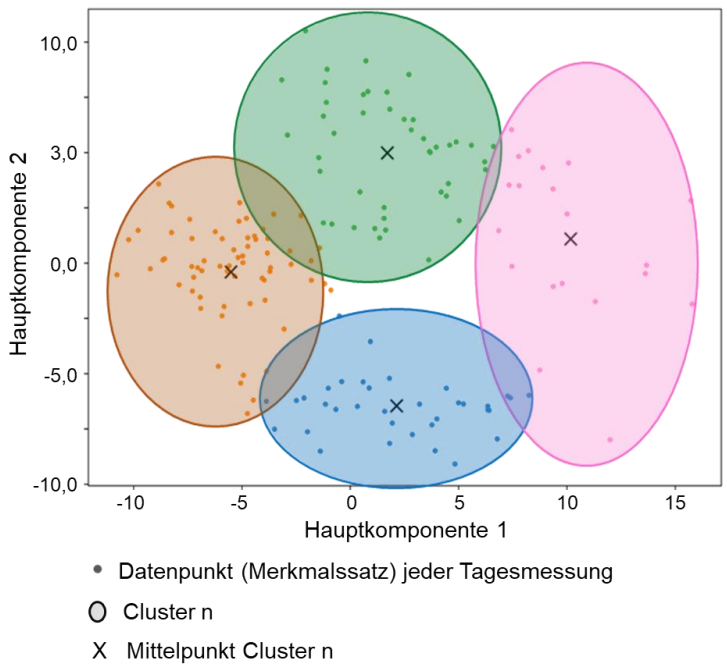
- Die optimale zu findende Clusteranzahl[1] wird durch die Analyse des Silhouettenkoeffizients ermittelt.
  - Maß der Ähnlichkeit eins Clusterpunkts zu seinem zugeordneten Cluster im Vergleich zu anderen Clustern
  - Korrekte Zuordnung: Wert nahe +1.
  - Falsche Zuordnung: Wert nahe -1.

Optimale Anzahl an Clustern ist erreicht, wenn der Mittelwert der Silhoutten-koeffizienten über alle geclusterten Trainingsbeispiele den Maximalwert erreicht.

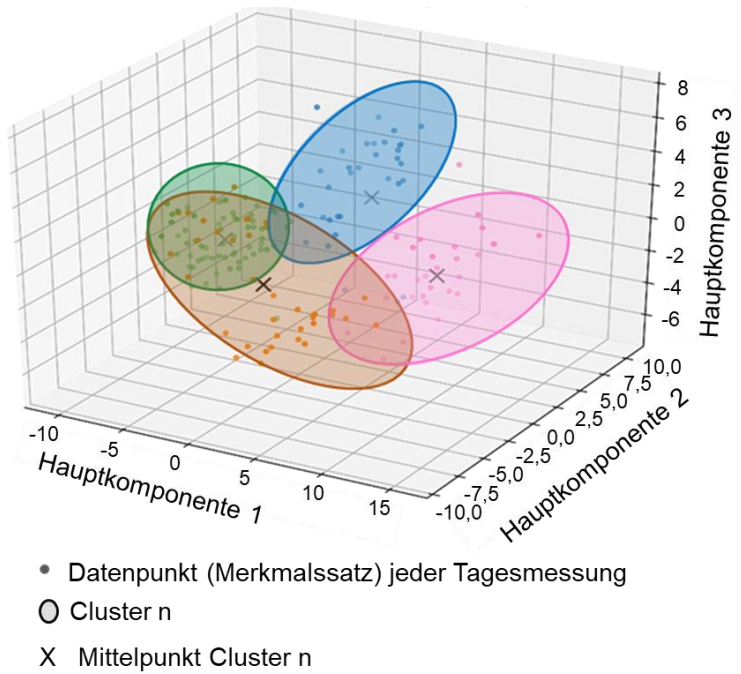
[1] Eine weitere untersuchte Methode ist die Ellenbogen-Methode. Sie vergleicht die prozentuale Varianz gegenüber der Anzahl an Clustern. In der grafischen Analyse ist die optimale Anzahl diese, bei der die Kurve einen ellenbogenähnlichen Knick aufweist [14]. Sie lieferte in der Untersuchung vergleichbare Ergebnisse basierend auf den verwendeten Daten.

Ergebnisse

KMeans Clustering der zwei bzw. drei Hauptkomponenten der Merkmalsätze für vier Cluster (optimale Anzahl)



KMeans Clustering der zwei Hauptkomponenten der Merkmalsätze für vier Cluster



KMeans Clustering der drei Hauptkomponenten der Merkmalsätze für vier Cluster

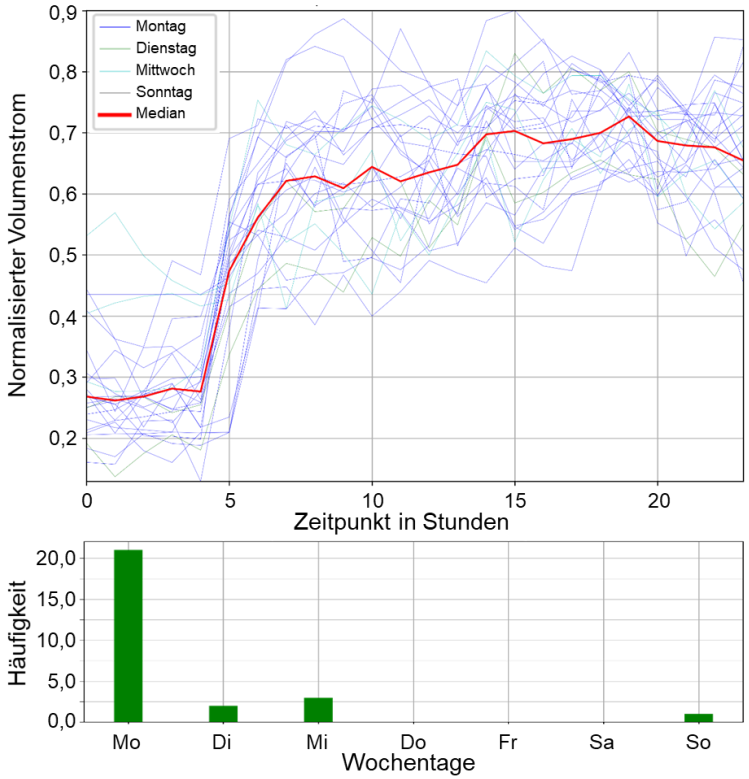
Validierung der Ergebnisse auf nachfolgender Folie:

- Tagesmessungen des jeweiligen Clusters als Zeitreihen.
- Normalisierung der Signalverläufe auf den Maximalwert aller Messpunkte.
- Median aus allen Signalverläufen eines zugeordneten Clusters ist in Rot dargestellt (charakteristischer Volumenstromverlauf).
- Die Häufigkeit der auftretenden Wochentage als Histogramm.

Ergebnisse

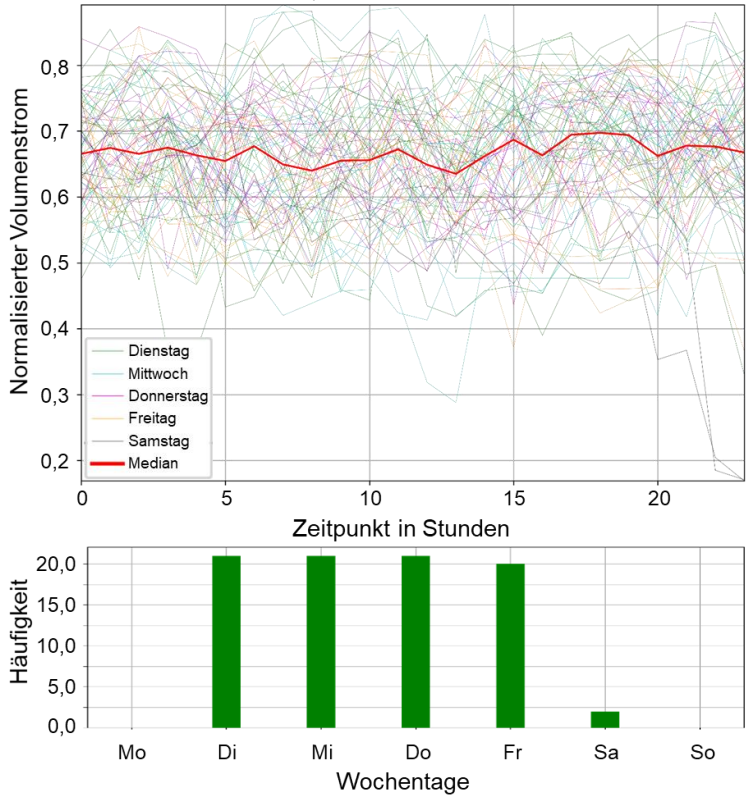
Drei Hauptkomponenten

Cluster 1: Montag



Charakteristisches Profil des Volumenstrom für Montag mit Median (rot) mit signifikantem Anstieg ab etwa 4 Uhr.

Cluster 2: Dienstag bis Freitag

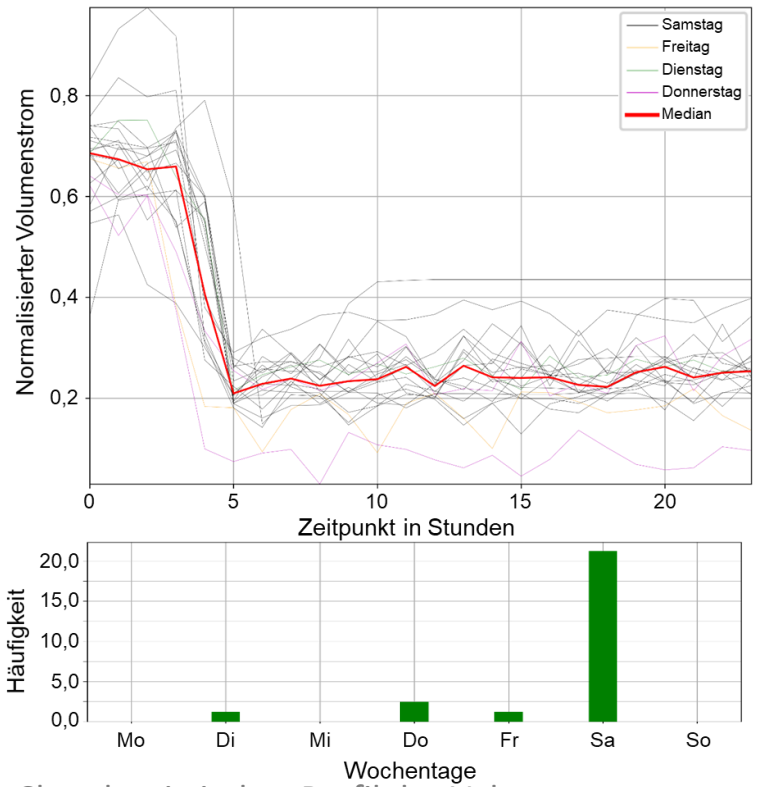


Charakteristisches Profil des Volumenstrom für Dienstag bis Freitag mit Median (rot).

Ergebnisse

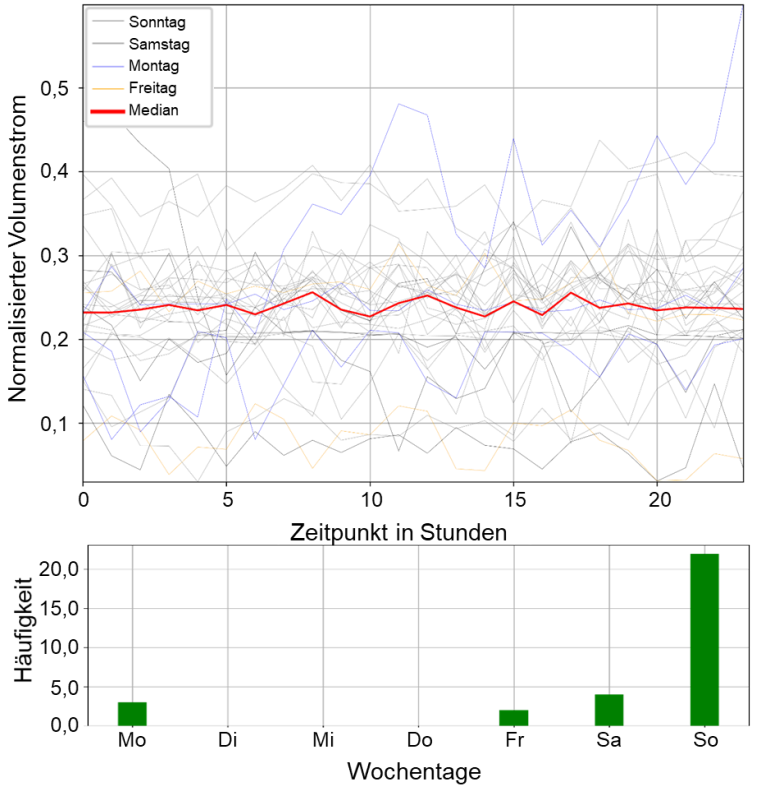
Drei Hauptkomponenten

Cluster 3: Samstag



Charakteristisches Profil des Volumenstrom für Samstag mit Median (rot) mit signifikantem Abfall ab etwa 4 Uhr.

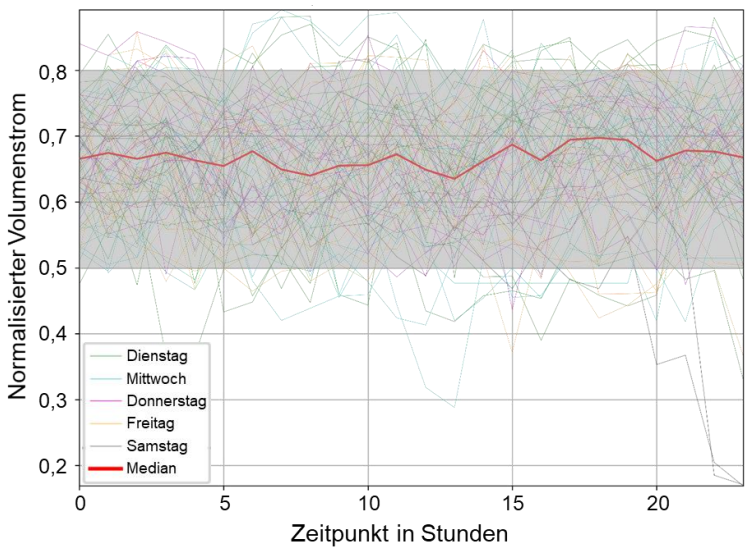
Cluster 4: Sonntag



Charakteristisches Profil des Volumenstrom für Sonntag mit Median (rot).

Fraunhofer IPA

Vier Cluster – Optimale Clusteranzahl

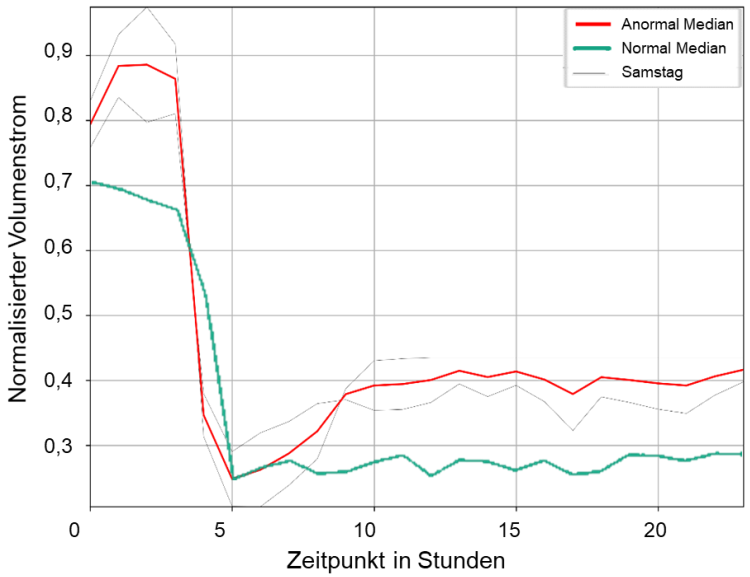


Charakteristisches Cluster für Dienstag bis Freitag mit exemplarischen Toleranzband von 0,5 bis 0,8 des normalisierten Volumenstroms (grau).

Mit den beschriebenen Methoden ist es möglich:

- Anomalien mit unüberwachtem maschinellem Lernen zu clustern und anschließend im Signalverlauf des Volumenstroms tageweise zu interpretieren.
- Mit den Medianverläufe eines jeweiligen Clusters in Abbildung links, kann ein Toleranzband durch den Anwender entsprechend definiert werden.
- Weicht ein Tagesprofil einem vorgegebenen Toleranzband ab, ist dies im Signalverlauf als Anomalie erkennbar.

Neun Cluster – Erhöhung der Clusteranzahl



Signalverlauf Cluster für Samstag normal (grün) und Signalverlauf Cluster Samstag anormal (rot) als Clustermedian.

Die Abweichung von der optimalen Clusteranzahl ist nur dann sinnvoll, wenn die geclusterten Verläufe plausible Ergebnisse liefern.
Bspw. zwei unterschiedliche Cluster für Samstag in Abbildung links.

- Die Erhöhung der Clusteranzahl führt dazu, dass Anomalien durch neun unterschiedliche Cluster erkannt werden können.
- Zwei der Cluster gruppieren bspw. Signalverläufe von Samstagen so, dass:
  - häufig auftretende ähnlich normale und
  - weniger häufig auftretende anormale Signalverläufe

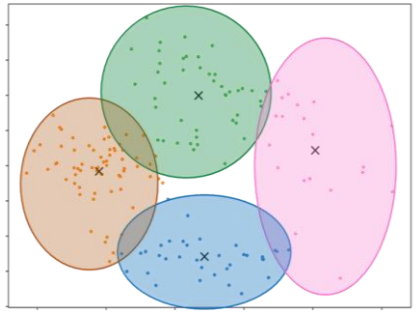
  jeweils darin enthalten sind.

[1] Direkt: Ohne den Vergleich des Verlaufs der Zeitreihe mit definiertem Toleranzband mit vier Clustern.

Bewertung und Ausblick

Charakteristische Tagesprofile können zur **Kosteneinsparung verwendet** werden:

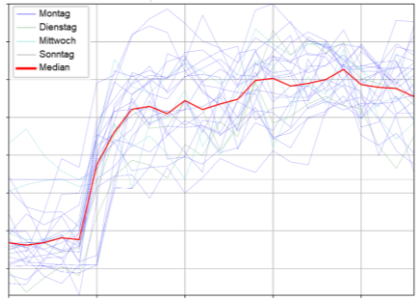
- Bedarfsprognosen für Folgetage,
- Monitoring von Betriebszeiten (wertschöpfend und nicht wertschöpfend).



Das Clustering kann durch **zusätzliche Kontextinformation verbessert** werden:

- Untermessungen wie Einzelmessung der Kompressoren,
- charakteristische Stundenprofile pro Tag,
- Verknüpfung der Messdaten mit der Produktionsplanung oder Produktionsauslastung.

Dies gilt es in nachfolgenden Forschungsvorhaben weiter zu untersuchen.



Danke für Ihre Aufmerksamkeit

M. Sc. Christian Dierolf
Wissenschaftlicher Mitarbeiter
Industrielle Energiesysteme

Telefon +49 711 970-3650
Christian.Dierolf@ipa.fraunhofer.de

Dipl.-Ing. Ivan Bogdanov
Wissenschaftlicher Mitarbeiter
Industrielle Energiesysteme

Telefon +49 711 970-1338
Ivan.Bogdanov@ipa.fraunhofer.de

Wir produzieren Zukunft

Nachhaltig. Personalisiert. Smart.

Sie bleiben wettbewerbsfähig

Nachhaltig. Flexibel. Wirtschaftlich.

Referenzen

[1] R. Gloor, „Energieeinsparungen bei Druckluftanlagen in der Schweiz - Programm Elektrizität Forschungsprojekt,“ Gloor Engineering, Sufers, 2000.

[2] Bayerisches Landesamt für Umweltschutz, „Effiziente Druckluftsysteme,“ 2004. [Online]. Verfügbar unter: https://www.bestellen.bayern.de/application/eshop_app000003?SID=580600777&ACTIONxSESSxSHOWPIC(BILDxKEY:%27lfu_klima_00028%27,BILDxCLASS:%27Artikel%27 ,BILDxTYPE:%27PDF%27). [Zugriff am 28 Januar 2020].

[3] Umweltbundesamt, „Entwicklung des Stromverbrauchs nach Sektoren,“ 2018. [Online]. Verfügbar unter: https://www.umweltbundesamt.de/sites/default/files/medien/384/bilder/dateien/2_und_4_datentabelle_eev_2018-02-14.pdf. [Zugriff am 22 Januar 2020].

[4] Statista GmbH, „Nettostromverbrauch in Deutschland in den Jahren 1991 bis 2018 (in Terawattstunden). Statista,“ Februar 2019. [Online]. Verfügbar unter: https://de.statista.com/statistik/daten/studie/164149/umfrage/netto-stromverbrauch-in-deutschland-seit-1999/. [Zugriff am 11 Oktober 2019].

[5] Zentrale Koordinierungsstelle KEFF, „KEFFIZIENZ-LEITFADEN DRUCKLUFT,“ Zentrale Koordinierungsstelle KEFF, Stuttgart, 2017.

[6] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, Sebastopol: O'Reilly Media, Inc., 2019.

[7] M.-A. Richard, H. Fortin und M. Fournier, „Daily load profiles clustering: a powerful tool for medium-sized industries demand side management,“ in ACEEE summer study on Energy Efficiency in Industry, Denver, 2017.

[8] M. Verleysen und D. François, „The Curse of Dimensionality in Data Mining and Time,“ Computational Intelligence, pp. 758-770, 2005.

[9] L. Smith, „A tutorial on Principal Components Analysis,“ Februar 2002. [Online]. Verfügbar unter: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf. [Zugriff am 31 Oktober 2019].

[10] D. Arthur und S. Vassilvitski, „k-means++: The advantages of careful seeding,“ in Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics , 2007.

[11] J. Karter, Statistical Methods With Matlab, CreateSpace Independent Publishing Platform, 2016. ISBN: 978-1-5395-4238-4

[12] F. Murtagh und P. Legendre, „Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm,“ Journal of Classification, pp. 274-295, 18 Oktober 2014.

[13] L. Kaufman und P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, 2009.

[14] R. L. Thorndike, „Who belongs in the family?,“ in Psychometrika, Springer-Verlag, 1953, pp. 267-276.

[15] Justin T. Page, Zach Liechty, Mark D Huynh und Joshua A Udall, „BamBam: Genome sequence analysis tools for biologists”, BMC Research Notes 7(1):829, 2014