IDENTIFIKATION UND BEHANDLUNG VON AUSREIßERN IN FLUGBETRIEBSDATEN FÜR MACHINE LEARNING MODELLE

S. Baumann, M. Gnisia, P. Feifel, U. Klingauf
Technische Universität Darmstadt, Institut für Flugsysteme und Regelungstechnik
Otto-Berndt-Straße 2, 64289 Darmstadt, Deutschland

Kurzfassung

Operationelle Flugbetriebsdaten können teilweise auffällige Datenpunkte enthalten, die sich nicht den charakteristischen Verlauf des aufgezeichneten Parameters zuordnen lassen und man als Ausreißer bezeichnet. Zur Bereinigung von Ausreißern und zur Steigerung der Datenqualität sind geeignete Methoden zur Identifizierung, Kategorisierung und Behandlung (bspw. durch Korrektur) notwendig, die in diesem Beitrag zusammengetragen, diskutiert und anhand exemplarischer Untersuchung an realen Flugbetriebsdaten analysiert werden. Der vorliegende Beitrag widmet sich der Identifizierung und Behandlung von Ausreißern für Sensorparameter in Flugbetriebsdaten (Full Flight Data). Dabei erfolgt eine Einteilung unterschiedlicher Strategien, um Ausreißer zu identifizieren, wobei auch Vor- und Nachteile dieser Methoden diskutiert werden. Bei der Ausreißerbehandlung stehen zwei Aspekte im Fokus: einerseits wird das Ausmaß der Verzerrung der Datenreihe durch die Ausreißer analysiert und andererseits erfolgt eine Einteilung in unterschiedliche Ausreißerarten zur geeigneten Korrektur der Ausreißer. Erläutert werden geeignete Filter und Glättungsansätze sowie Algorithmen-basierte Verfahren aus dem Bereich Machine Learning. Für deren Bewertung werden statistische Metriken wie Lagemaße als Gütekritierien anhand von Testdaten bewertet. Durch die angebrachten Vorgehensweisen können Ausreißer entsprechend ihrer Art passend identifiziert und behandelt werden, sodass die Qualität der Datenreihen gesteigert werden kann. Damit kann ein Beitrag zur geeigneten Auswahl von Modellierungsparameter geleistet und die Güte der iterativen Lernmodelle gesteigert werden.

Keywords

Ausreißeridentifikation; Ausreißerbehandlung; Vorverarbeitung; Datenanalyse; Flugbetriebsdaten; Full Flight Data; Machine Learning; k-means; MAD; Hampel; DFFITS

1. EINFÜHRUNG UND RELEVANZ

Moderne Luftfahrzeuge zeichnen während des Fluges aufgrund einer Vielzahl applizierter Sensoren einen sehr umfangreichen und detaillierten Datenrahmen auf. Diese Daten lassen sich bereits in-situ während des Fluges zur Flugzustandsüberwachung sowie für nachgelagerte Leistungsanalysen im Postprocessing nutzen.

Eine Vielzahl aktuell in Anwendung befindlicher bspw. physikalischer Modelle oder statischer Buchwertmethoden setzen vergleichbare Einflussfaktoren sowie Randbedingungen als Qualitätsmerkmale an die Datenbasis voraus. Moderne, datenbasierte Analytikmethoden mit bspw. Machine Learning Algorithmen aus dem Bereich der künstlichen Intelligenz benötigen eine ausreichend hohe Datenqualität, um bspw. valide Aussagen erzeugen und Muster in den Daten finden zu können. Diese Anforderung geht mit einem praktischen Ziel einher, genauere Berechnungsmodelle mit datenbasierten Methoden der künstlichen Intelligenz zu entwickeln. Diese können gegenüber klassischen physikalischen Modellen Genauigkeitsvorteile bieten und Einflussparameter durch iterative Lernverfahren weniger aufwendig abbilden.

Einer ausreichend hohen Datenqualität stehen jedoch nicht-systemdynamische Parameterschwankungen wie Ausreißern entgegen. Diese bergen für die angesprochenen Anwendungen erhebliche Nachteile, da verzerrte Datenbasen Modellergebnisse signifikant negativ beeinflussen können (siehe garbage in, garbage out - Idiom). Der vorliegende Beitrag widmet sich der Identifizierung und Behandlung von Ausreißern in aufgezeichneten Sensordaten von Flugzeugen, analysiert Ausmaße möglicher Verzerrungen und charakterisiert Ausreißerarten, sodass die Qualität der Datenreihen gesteigert werden kann.

BASU und MECKESHEIMER [1] haben sich mit parametrischen Formen der Ausreißerdetektion und der Ausreißerbehandlung von Flugbetriebsdaten bereits beschäftigt. Hierbei wurden anhand ausgewählter Sensorparameter wie der Flughöhe sowie der Rollwinkel des Flugzeuges aus Flugdatenaufzeichnungen Auswirkungen unterschiedlicher Schwellwerte und Schrittweiten in experimentellen Studien untersucht. Es hat sich gezeigt, dass diese Methoden in vielen Anwendungen erfolgsversprechend sind, aber auch Verbesserungspotentiale aufgrund von Einschränkungen bei der Signalverarbeitung bieten. Diesem Umstand soll mit dem vorliegenden Beitrag Rechnung getragen werden, indem parametrische und auch nicht-parametrische Methoden in der Anwendung auf Flugbetriebsdaten untersucht werden. Weiterhin werden die Auswirkungen der unterschiedlichen Vorbehandlungen auf die datenbasierte Modellbildung zur Schätzung des Treibstoffflusses eines Flugzeuges untersucht.

Im Folgenden wird in Kapitel 2 die verwendete Datenbasis für die Untersuchungen zu Beginn vorgestellt. Anschließend erfolgt in Kapitel 3 eine Definition und Kategorisierung von Ausreißern. Auf dieser Grundlage werden in Kapitel 4 unterschiedliche Methoden zur Ausreißeridentifikation und Behandlung vorgestellt, diskutiert und für eine weitere Untersuchung ausgewählt. In

Kapitel 5 erfolgt eine Untersuchung der ausgewählten Methoden und Bewertung der Performanz einer Vorverarbeitung für datenbasierte Modellierungen mit den zur Verfügung stehenden Testdaten. Der Beitrag schließt mit einer Zusammenfassung und einem Ausblick zukünftiger Arbeiten in Kapitel 6.

2. DATENBASIS

Das amerikanische Projekt Discovery in Aeronautics Systems Health (DASHlink) der National Aeronautics and Space Administration (NASA) ermöglicht Wissenschaftlern einen gemeinsam Austausch von Forschungsergebnissen im Bereich Luft- und Raumfahrt. Der Fokus liegt hierbei auf Methoden in Bereich Data Science und Data Mining sowie der damit verbundenen methoden- und anwendungsbasierten Forschung (siehe [2]).

Bei den verwendeten Daten handelt es sich um sogenannte Full Flight Daten, welche die Aufzeichnungen einer Vielzahl von Sensoren am Flugzeug darstellen. Diese Aufzeichnungsparameter beinhalten Informationen zum Flugzustand und Triebwerksverhalten, Luftdaten mit Einflüssen der Umgebung, Navigations- und Positionsangaben sowie Steuereingaben und Systemparameter aus dem Flight Management System in Form von Zeitreihen. Überwiegend handelt es sich hierbei um Zustandsinformationen der technischen Systeme sowie auch um Zustandsüberwachungsdaten. Teilweise liegen hoch-dynamische Sensorparameter mit Abtastraten von bis zu 16 Hertz vor. Die Daten sind jedoch anonymisiert, um Rückschlüsse auf die Operation einer Fluggesellschaft oder dedizierte Flugzeuge zu vermeiden. Die Datenbasis beinhaltet insgesamt über 180.000 Aufzeichnungen aus den Jahren 2001 bis 2003 in Nordamerika.

Die Deskription der Daten, als wichtiges Element in Standardprozessen wie dem Knowledge Discovery Process (KDP) und im Speziellen dem Cross Industry Standard Process for Data Mining (CRISP-DM), legt nahe, dass Auffälligkeiten in den Zeitreihen der Parameter vorliegen. Zunächst ist jedoch eine Abgrenzung und Definition von Ausreißern notwendig. Der vorliegende Beitrag beschäftigt sich im Weiteren mit Ausreißern, die aufgrund der Verwendung mit Flugbetriebsdaten in entsprechenden Zeitreihen vorkommen.

3. DEFINITION UND ABGRENZUNG VON AUSREIßERN

Die Herausforderung Ausreißer zu identifizieren ergibt sich bereits aus dem Fehlen einer exakten mathematischen Definition. Im Bereich der Zeitreihenanalyse kann eine bspw. annähernd stationäre Zeitreihe nach HEIJ, C. ET AL. [3] durch konstante Metriken zur Beschreibung von Lage und Streuung der Stichprobe wie Mittelwert und Varianz charakterisiert werden. Weiterhin ist die Kovarianz zwischen vorangegangenen Werten und dem aktuellen Wert des Ensembles lediglich von der zeitlichen Differenz bzw. dem Abstand zweier Beobachtungen abhängig. Ausreißer heben sich von dieser linearen Beschreibungsweise des einfachen Zeitreihenmodells maßgeblich ab.

Unter dem Begriff Ausreißer können Datenpunkte einer Datenreihe oder einer Stichprobe subsummiert werden, die im Gegensatz zu den übrigen Werten auffällig sind und / oder von diesen maßgeblich abweichen (vgl. [4]). BRILLINGER [5] beschreibt Ausreißer als Beobachtungen in einer Datenreihe, die auffallend weit von einem zentralen Wert bzw. Lageparameter entfernt liegen und ungewöhnliche Werte im Verhältnis zum Großteil der Daten aufweisen. Häufig berechnete Größen wie Mittelwerte und kleinste Fehlerquadrate können durch solche Werte drastisch beeinflusst werden.

Für grafische Darstellungen von Häufigkeitsverteilungen sowie zum Vergleich von Datensätzen werden in der deskriptiven Statistik beispielsweise Boxplots nach TUKEY [6] verwendet, die aus Fünf-Punkte-Beschreibungen resultieren. Darin werden Streu- und Lagemaße, zu denen Median, Quartile und Extremwerte bspw. als Maximum und Minimum zählen, dargestellt. Eine exemplarische Darstellung eines Boxplots findet sich in BILD 1.

Durch eine Box wird die Lage von insgesamt 50 % der Datenpunkte repräsentiert, wobei die Intervallbreite als sogenannter Interquartilsabstand bezeichnet wird. Dieser ist resistenter ggü. Ausreißern als die Standardabweichung bei den Streuungsmaßen. Innerhalb der Box wird der Median gekennzeichnet, der selbst ebenfalls robuster ggü. Ausreißern ist als bspw. der Mittelwert der Stichprobe bei den Lageparametern (vgl. [7]). Die übrigen 50 % der Datenpunkte der Stichprobe können im Boxplot unterschiedlich dargestellt werden. Eine in der Wissenschaft weit verbreitete Darstellung beschränkt den Abstand von unterem bzw. oberem Quartil zu den Whiskern auf maximal das Eineinhalbfache des Interquartilsabstandes. Die Lage der Whisker ist dabei abhängig von den Daten und auf Datenpunkte der Stichprobe referenziert. Beinhaltet der Datensatz keine Datenpunkte außerhalb des zuvor genannten Abstandes von den Quartilen, werden die Whisker als Maximum und Minimum der Stichprobe festgelegt. Hierbei können keine Ausreißer identifiziert werden, da diese Darstellung die Gesamtheit der Stichprobe einschließt.

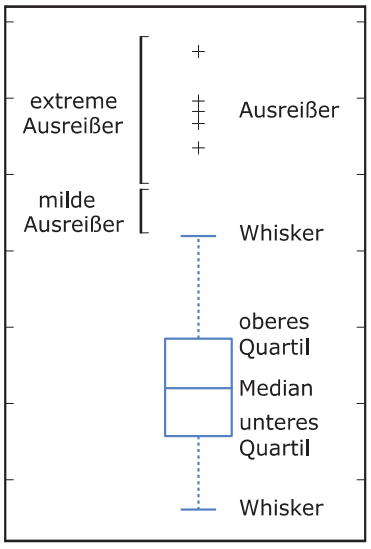

BILD 1. Qualitative Darstellung eines Boxplots mit Lage- und Streumaßen sowie klassifizierten Ausreißern

Datenpunkte außerhalb der Whisker werden in der Interquartilsabstands-Methode nach TUKEY [6] als auffällige Datenpunkte klassifiziert, bei denen es sich voraussichtlich um Ausreißer handelt. Datenpunkte, die dabei jeweils zwischen dem Eineinhalbfachen des Interquartilsabstandes von der Box (also dem ersten und dritten Quartil) entfernt liegen, werden als milde Ausreißer, Datenpunkte, die um mehr als Dreifache des Interquartilsabstandes von der Box entfernt liegen, als extreme Ausreißer bezeichnet. Dies stellt eine einfache Methode zur Klassifizierung von Ausreißern in Datenreihen dar. Mit dieser Methode lässt sich jedoch nicht ausschließen, dass Datenpunkte, die nicht als Ausreißer in einem Boxplot gekennzeichnet werden, ebenfalls fehlerbehaftet sind. Ebenfalls muss angemerkt werden, dass die Anzahl der identifizierten Ausreißer von der Größe der Stichprobe abhängt. Kleinere Stichprobenumfänge zeigen zudem selbst extreme Ausreißer weniger an. Unabhängig von der Größe der Stichprobe können so Fehler erster Art, sogenannte α (Alpha)-Fehler, auftreten, sodass auch bei perfekt normalverteilten Stichproben Ausreißer identifiziert werden (siehe [6]). Die Interquartilsabstands-Methode nach TUKEY ist daher vielmehr als informeller sowie visueller Test bzw. Hinweis anstelle einer Problemdeutung der zu Grunde liegenden Stichprobe zu verstehen (siehe [8]).

Nach AGUINIS ET AL. [9] lassen sich insgesamt vier unterschiedliche Arten von Ausreißern in Zeitreihen abgrenzen: additive outlier, innovation outlier, level shift outlier und temporary changes outlier. Der erste Ausreißertyp kennzeichnet sich durch eine deutliche Abweichung zu den anderen Datenpunkten in der Zeitreihe. Bei einer Verbindung der Datenreihe würde der Ausreißer so zu einem spitzen Verlauf führen (sog. influential time series additive outlier). Führt dieser Ausreißer allerdings dazu, dass auch nachfolgende Datenpunkte durch ihn beeinflusst werden, so handelt es sich um einen influential time series innovation outlier. Verursacht ein Datenpunkt einen abrupten und anhaltenden Stufensprung und verschiebt die Zeitreihe somit auf ein anderes Niveau, wird dies als influential level shift outlier bezeichnet.

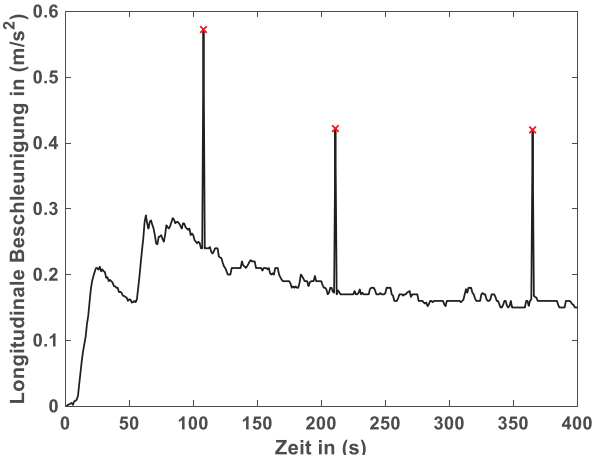

BILD 2. Exemplarische Zeitreihe der longitudinalen Beschleunigung eines Flugzeuges beim Start mit Ausreißern

In Darstellung BILD 2 ist ein Ausreißer-belastetes Merkmal beispielhaft aus einem Flugbetriebsdatensatz dargestellt. Die mit einem roten Kreuz markierten Werte können als auffällige Beobachtungen gedeutet werden und für eine Ausreißeridentifikation und Ausreißerbehandlung in Betracht gezogen werden. Der spitze Verlauf im Bereich der Ausreißer unterstreicht die Klassifikation nach AGUINIS ET AL. [9].

Im Folgenden werden Methoden zur Detektion und Behandlung von Ausreißern anhand von statistisch parametrischen Methoden und auch nicht-parametrischen Verfahren mit Hilfe von Algorithmen aus dem Bereich Machine Learning vorgestellt.

4. METHODEN ZUR AUSREIßERDETEKTION UND BEHANDLUNG

Um die Auswirkungen von Ausreißern auf die Deskription von Datensätzen, nachgelagerten Analysen oder Modellbildungen zu reduzieren, werden Methoden benötigt, um Ausreißer zu erkennen und diese geeignet zu behandeln. Nachfolgend werden unterschiedliche Strategien vorgestellt und nachfolgend deren Einsatz auf Flugbetriebsdaten untersucht.

4.1. Ausreißerdetektion

Ausreißer lassen sich in weniger komplexen Fällen in der Phase des Data Understanding anhand einer Datendeskription bspw. durch graphische Analyse von Auffälligkeiten ermitteln. Nachteile liegen hierbei jedoch in der a posteriori sowie subjektiven Bewertung.

Nach AGUINIS ET AL. [9] können drei weitere Möglichkeiten der Ausreißerdetektion festgehalten werden: single-construct techniques, multiple-construct techniques und influence techniques. Erste Methode beinhaltet bspw. die bereits vorgestellte Methode nach TUKEY mit Hilfe des Interquartilsabstandes oder auch Verfahren, die auf dem Abstand von Lageparameter, die mit Hilfe von Metriken über die Standardabweichung gebildet werden, basieren. Unter multiple-construct techniques finden sich Distanz-basierte Verfahren wie bspw. Streudiagramme. Weiterhin zählt unter anderem auch das sogenannte k-Clustering hierzu, was bspw. mit dem k-means Algorithmus den Datensatz über eine Kennzahl (Gütefunktional) in *k* Klassen einteilt, die bspw. anhand der euklidischen Distanz zwischen den Datenpunkten und sogenannten Cluster-Zentren ermittelt werden. Unter influence techniques fallen bspw. Verfahren wie die Cook Distanz, die zur Bestimmung von Einflüssen einzelner Datenpunkte auf ein Regressionsmodell, bspw. ein ordinary least squares Verfahren (kurz OLS), verwendet wird.

Problematisch bei einer Ausreißerdetektion mit Hilfe konventioneller statistischer Methoden wie beispielsweise dem Interquartilsabstand oder weiteren Distanz-basierten Verfahren ist deren Modellabhängigkeit. Bei zeitlich veränderlichen Prozesssignalen und großen Ensemblereihen kann es zu einem Modellversagen kommen (vgl. [1]).

4.2. Ausreißerbehandlung

Für die Behandlung von Ausreißern, können in Anlehnung an RUNKLER [4] unterschiedliche Möglichkeiten in Abhängigkeit des Datensatzes und des Anwendungsfalls hilfreich sein. Hierbei ist auch der Datenumfang ein entscheidendes Bewertungskriterium. Für die Verfahren werden in diesem Beitrag drei Kategorien zusammengefasst:

- Erstellung von Fehlerlisten
- Korrektur oder Schätzung von Merkmalswerten
- Entfernen von Datenpunkten (ggfs. auch Fehlerwertsetzung durch bspw. not a number, NaN) oder ganzer Merkmale aus dem Datensatz.

Eine Behandlung der Daten durch Transformation bzw. Schätzen der Merkmale kann negative Implikationen von fehlerbehafteten Daten mindern. Jedoch ist zu beachten, dass zwar eine statistische Verteilungsform stärker angenähert wird, jedoch multivariate Interdependenzen zwischen den Merkmalen oder auch das zugrunde liegende Gesamtverhalten der Daten dadurch manipuliert werden. Diesem Aspekt ist bei der Bewertung Rechnung zu tragen. Das Entfernen von Datenpunkten oder ganzer Merkmale kann bei großen Datensätzen und hoher Korrelationen zwischen ähnlichen Parametern ein effizientes Instrument sein, um hohe Aufwände der Datenvorverarbeitung bereits im Vorfeld vermeiden zu können. Im Gegensatz zur Transformation kann zwar das Risiko vermieden werden, Scheinwerte bzw. künstliche Parameterinterdependenzen zu erzeugen, bei geringen Datenumfängen können jedoch maßgebliche Systeminformationen verloren gehen oder das Globalverhalten ungünstig verzerrt werden. Eine parallele Untersuchungsmethodik, welche die Merkmalswerte mit und ohne Ausreißerbehandlung, bspw. durch Schätzung von Merkmalswerten, untersucht, ist ebenfalls denkbar. Wie bei der Ausreißerdetektion können auch nicht-parametrische statistische Verfahren zum Einsatz kommen, da diese robuster gegenüber Ausreißern ausfallen können (vgl. [10]).

Es kann festgehalten werden, dass es für die Ausreißerbehandlung informierte Entscheidungen unter Verwendung von Expertenwissen bedarf und aufgrund der Individualität der Problemstellungen keine generalisierten Wege oder Ansätze „off the shelf“ zur Verfügung stehen.

4.3. Auswahl und Vorstellung ausgewählter Methoden zur Ausreißerdetektion

Im Folgenden werden die in diesem Beitrag untersuchten Methoden zur Ausreißeridentifikation vorgestellt.

MAD / Hampel Filter

Ein Medianfilter, im Deutschen auch als Hampel Filter oder im Englischen als median absolute deviation filter, kurz MAD, bezeichnet, kann Ausreißer aus einem Datensatz effizient entfernen. Er basiert auf einem gleitenden Fenster. Es handelt sich hierbei um einen Entscheidungsfilter, der bei Identifikation eines Ausreißers anhand von Schwellwerten den Datenpunkt durch den Median des Fensterensembles ersetzt. Die Entscheidung, ob ein Messwert als Ausreißer klassifiziert wird, wird anhand des Abstandes zwischen Datenpunkt zum gleitenden Median getroffen. Dieser Filter kann sowohl durch die Fensterbreite als auch durch den Schwellwert, also dem n_σ-fachen der Standardabweichung σ_i zwischen Datenpunkt und gleitenden Median, eingestellt werden. Der lokale Median $\tilde{x}_i$ eines gleitenden Fensters mit der Breite k wird mit

$$\tilde{x}_i = \text{median}(x_{i-k}, x_{i-k+1}, \dots, x_i, \dots, x_{i+k-1}, x_{i+k}) \quad (1)$$

berechnet und die median absolute deviation (MAD) mit

$$\text{MAD} = |x_i - \tilde{x}_i| \quad (2)$$

bestimmt. Der Hampel Filter identifiziert einen Ausreißer für einen gegebenen Schwellwert, ausgedrückt in der Entscheidungsregel

$$|x_i - \tilde{x}_i| > n_\sigma \sigma_i \quad (3)$$

über die Anzahl von Standardabweichungen und ersetzt diesen durch den Median des Fensters. Häufig wird n_σ mit dem Wert drei belegt, sodass zur Bewertung das 3σ(Sigma)-Gesetz herangezogen und durch die Bewertung anhand des Medians variiert wird. Die Fensterbreite ist das Doppelte des Einstellparameters k, welcher die Anzahl an benachbarten Instanzen auf jeder Seite des Datenpunktes x_i angibt. [11]

DFFITS

Auffallende Beobachtungen in einem Datenset (Ausreißer) können feststellbare Auswirkungen auf ein Regressionsmodell und damit auf eine Parameterschätzung besitzen. Bei DFFITS, abgekürzt aus dem Englischen differences in fitted values, handelt es sich um ein Ausreißermaß, welches den Einfluss einzelner Datenpunkte auf ein linear aufgestelltes Regressionsmodell in der Form

$$y = X \cdot \beta + b \quad (4)$$

bewertet. Hierbei stellt y den Regressand, X die Designmatrix der erklärenden Variablen, β den Vektor der Gewichtungsfaktoren und b einen Vektor von Konstanten dar. Das Differenzmaß DFFITS ist ähnlich zur zuvor erwähnten Cook Distanz und gehört zur sog. Delete-1 Statistik. Das Ausreißermaß bei DFFITS wird für jede Beobachtung i im Datensatz durch die Formel

$$\text{DFFITS}_\text{i} = \frac{y - y_{(i)}}{\sigma_{(i)}\sqrt{h_{ii}}} \quad (5)$$

berechnet. Der Zähler auf der rechten Seite beschreibt die Differenz aus dem Schätzwert der Regression mit und ohne (durch den Index (i) gekennzeichnet) den i-ten Datenpunkt im Verhältnis zum Produkt aus der Standardabweichung ohne den i-ten Datenpunkt und dem Diagonalelement h_{ii} (Englisch leverage) der sogenannten Hat-Matrix (siehe [12]). Es handelt sich dabei um eine Residuen-erzeugende Matrix, die sich durch Multiplikation mit dem Schätzvektor

zum Residualvektor ergibt. Das Differenzmaß wird nun als Anzahl der Standardabweichungen, welche ein Schätzwert des Regressionsmodells ohne den i-ten Datenpunkt verändert, quantifiziert. Ein Datenpunkt gilt als einflussreich und damit als Ausreißer, wenn der absolute Wert die Ungleichung

$$|\mathrm{DFFITS_i}| > 2 \cdot \sqrt{\frac{p}{n}} \tag{6}$$

erfüllt, wobei p die Anzahl der Parameter und n die Länge der Zeitreihe darstellt. [12]

k-means

Ein Verfahren zur Ausreißerdetektion sollte auch ohne vorherige Festlegung von Referenzfällen oder Schwellwerten in der Lage sein, zuverlässig, adaptiv und robust Ausreißer zu identifizieren.

Machine Learning Algorithmen aus dem Bereich des unüberwachten Lernens zeichnen sich dadurch aus, dass aus dem Datensatz Regelmäßigkeiten (Muster) extrahiert werden, in dem die Datenpunkte unterschiedlichen Klassen zugewiesen werden. Im Gegensatz dazu werden beim überwachten Lernen in der Trainingsphase des Algorithmus Informationen über die Klassenzugehörigkeit, bspw. ob es sich um einen Ausreißer handelt oder nicht, mitgegeben. Die Zuweisung beim unüberwachten Lernen erfolgt bspw. durch eine Bewertung anhand einer Distanzmetrik. [13]

In diesem Beitrag wird als Vertreter dieses Bereichs der k-means Algorithmus zum Clustering der Datenvektoren näher untersucht. Die Funktionsweise des k-means Algorithmus kann anhand des Pseudo-Codes in BILD 3 nachvollzogen werden.

1 **Input**: Anzahl der Clusterzentren k; Datenvektor
2 **Output**: Clusterzuweisung mit Zentroiden
3 Initialisiere k Clusterzentren
4 **repeat**
5 ***for*** alle Beobachtungen n
6 Berechne die Distanzen der Datenpunkte zu allen Clusterzentren
7 Zuordnung der Beobachtungen zu einem nächstgelegenen Schwerpunkt / Zentroid anhand der geringsten Distanz
8 Neuzuordnung der Beobachtungen zu einem neuen Clusterzentrum, sofern die Summe der innerhalb des Clusters liegenden Distanzen verringert werden kann
9 Berechnen des Durchschnitts der Beobachtungen in jedem Cluster, um k neue Schwerpunkte zu erhalten.
10 **until** Clusterzuordnung ändert sich nicht mehr oder maximale Anzahl der Iterationen ist erreicht

BILD 3. Pseudo-Code des k-means Algorithmus (in Anlehnung an [14] und [15])

Es handelt sich bei k-means um einen iterativen, partitionierenden Algorithmus, der jeder Beobachtung eines von k Clustern zuordnet. Die Custer sind über Zentroide definiert. Zu Beginn ist die Anzahl der Clusterzentren dem Algorithmus zu übergeben und es werden k Referenzvektoren zu den Clusterzentren initialisiert. Anschließend wird die Distanz der einzelnen Instanzen (Merkmale) zu den Referenzvektoren gebildet, so lange, bis das Distanzmaß ausreichend minimiert werden konnte und die Ergebnisse konvergieren. Die Merkmale werden dann einem Cluster zugewiesen, dessen neues Clusterzentrum sich aus dem Durchschnitt aller Merkmale dieses Clusters berechnet. Damit kann eine von Merkmalen unabhängige Anwendung erreicht werden. [14]

Hybridmethode

Aufgrund eines stark nicht normalverteilten Charakters der Merkmale in dem verwendeten Datensatz bzw. im Allgemeinen der aufgezeichneten Sensorparameter über einen Flug oder auch in dedizierten Flugphasen eines Flugzeuges ist die Anwendung statistischer Verfahren nur bedingt bzw. mit Einschränkungen möglich. Bei einer Bewertung von Streuintervallen anhand der Standardabweichung σ (Sigma) der Normalverteilung, bspw. von ± 3σ, erlauben die verschiedensten statistischen Verteilungen der Merkmale keine theoretische Begründung des Vorgehens.

In einer für diesen Beitrag entwickelten hybriden Methode erfolgt eine Kombination des 3σ-Gesetzes sowie dem zuvor vorgestellten k-means Algorithmus zum Clustering. In BILD 4 ist der Pseudo-Code des Verfahrens aufgeführt.

1 **Input**: Anzahl der Clusterzentren k; Datenvektor
2 **Output**: Identifizierte Ausreißer; bereinigtes Signal; Clusterzuweisung
3 ***for*** Beobachtungen $t = \mathrm{n} - 1$
4 Bildung der absoluten Differenzen zwischen allen Datenpunkte der Zeitreihe in der Form

$$\mathrm{Diff}_t = x_{t+1} - x_t \tag{7}$$

5 Bildung der empirischen Verteilungsfunktion (Engl. cumulative density function, CDF)
6 Identifikation von x_t als Ausreißer anhand des 3-σ-Gesetzes, wenn

$$(|\mathrm{Diff}_t| \wedge |\mathrm{Diff}_{t+1}|) > 3\sigma \wedge \mathrm{Diff}_t \cdot \mathrm{Diff}_{t+1} < 0 \tag{8}$$

erfüllt
angewendet auf die berechneten Differenzen
7 ***if*** Ausreißer identifiziert ***then***
Clustering des Originalsignals k-means, $k = 3$
8 ***if*** Übereinstimmung von Datenpunkten in Clustern mit größten/minimalsten Abstand des Zentroids zu den anderen Clusterzentren mit Werten aus Schritt 6 ***then***
Markierung des Datenpunkts als Ausreißer

BILD 4. Pseudo-Code der entwickelten Hybridmethode

Beide Verfahren werden auf die Differenzen zwischen den Datenpunkten untereinander einer Zeitreihe eines Sensorsignals angewendet. Aufgrund der Differenzierung ersten Grades der Zeitreihe ist die Anwendung zur Detektion von sogenannten additiven Ausreißern gut

durchführbar. Angewendet auf das Originalsignal liefert der k-means Algorithmus dabei auch unabhängig vom Merkmal eine belastbare Unterscheidung zwischen systemdynamischen Verhalten und Ausreißer. Dennoch besteht mit der Verwendung eines k-means Algorithmus weiterhin die Notwendigkeit der Vorgabe der Cluster-Anzahl. Für die mit dem Hybridverfahren gebildeten Differenzen ersten Grades kann in wesentlich besserer Näherung im Vergleich zum Ausgangssignal eine asymptotische Normalverteilung angenommen werden. Dies kann anhand der empirischen Verteilungsfunktion, im Englischen als cumulative density function (CDF) bezeichnet, verdeutlicht werden (siehe Kapitel 5). Ein potentieller Ausreißer wird bei diesem Verfahren erkannt, wenn sich seine Differenzen zu benachbarten Datenpunkten in der Zeitreihe außerhalb eines 3-Sigma-Bandes befinden. So kann verhindert werden, dass große, jedoch real auftretende Differenzen und damit Anstiege oder Abfälle zwischen benachbarten Messwerten fälschlicherweise als Ausreißer charakterisiert werden. Dabei müssen die Bedingungen

$$|\mathrm{Diff}_t| = |x_t - x_{t-1}| > 3 \cdot \sigma \tag{9}$$

$$|\mathrm{Diff}_{t+1}| = |x_{t+1} - x_t| > 3 \cdot \sigma \tag{10}$$

$$\mathrm{Diff}_t \cdot \mathrm{Diff}_{t+1} < 0 \tag{11}$$

erfüllt sein. Die geforderte Negativität der Verknüpfungen der benachbarten Differenzen soll sicherstellen, dass ein extremer level shift nicht fälschlicherweise als Ausreißer identifiziert wird. Vielmehr soll ein kurzzeitiges Springen des Parameterwertes als Ausreißer identifiziert werden.

Bei der Detektion von mehreren additiven Ausreißern zwischen mehreren Datenpunkten einer Zeitreihe, kommt der k-means Algorithmus zur Anwendung. Dieser clustert das Originalsignal in mehrere Gruppen und ordnet die Abstände zwischen Datenpunkten einem Mittelwert eines Clusters zu. Wenn hierbei bestätigte Ausreißer in dem Cluster mit dem größten Abstand zu den anderen Clusterzentren fallen und gleiche Werte wie potentiell hintereinander gelegene Ausreißer aufweisen, können auch diese Datenpunkte zusätzlich als Ausreißer identifiziert werden. Abschließend erfolgt die Behandlung der Ausreißer bspw. durch eine lineare Approximation der Nachbarwerte oder das Ersetzen durch einen (gleitenden) Median über eine definierte Fensterbreite bzw. eine Erweiterung der Werte auf den folgenden oder den darauffolgenden Nachbarn, sofern mehrere additive Ausreißer zusammenliegen.

5. BEWERTUNG DER EFFIZIENZ DER DATENVORVERARBEITUNG

Im Folgenden werden die Ergebnisse der ausgewählten Methoden zur Ausreißerdetektion und Behandlung vorgestellt und diese anhand der Datenbasis diskutiert. Anhand von unbehandelten und vorverarbeiteten Datensätzen wird auch ein Vergleich in Bezug auf eine Diagnose-Aufgabe anhand von Modellierungen mit iterativen Lernverfahren mit vorverarbeiteten Daten sowie mit Originärdaten gezogen.

5.1. Darstellung der Effektivität der Methoden auf unterschiedliche Parameterverläufe

Es kann gezeigt werden, dass sich anhand der verwendeten Testdaten additive Ausreißer häufig durch eine systematische Wiederholung mit identischen Datenwerten auszeichnen. Diese Systematik erlaubt auch eine Ausweitung der Detektion auf aneinander folgende additive Ausreißer, die jedoch noch keinen Niveausprung (Engl. level shift) darstellen. Die Elimination eines Niveausprungs aufgrund einer fälschlichen Identifikation als Ausreißer (sog. falsch – positiv Kategorisierung) soll ausdrücklich vermieden werden. Im Folgenden werden ausgewählte Ergebnisse anhand unterschiedlicher Sensorparameter eines Flugzeuges mit den angebrachten Verfahren dargestellt und diskutiert.

MAD und DFFITS

In BILD 5 ist eine Ausreißeridentifikation mit einem Hampel Filter mit den Einstellparametern $k = 10$ und $n_\sigma = 5$ anhand der kalibrierten Fluggeschwindigkeit über eine gesamte Flugmission dargestellt. Es ist zu erkennen, dass sich Änderungen der Fluggeschwindigkeit zwischen zwei Datenpunkten in den unterschiedlichen Flugphasen wie Start oder Landung sowie im Reiseflug (Intervall zwischen etwa 3.000 und 9.000 Sekunden) deutlich unterscheiden können. Hierfür empfiehlt es sich, die Schwellwerte entsprechend der Parameterdynamik anzupassen (vgl. [1]).

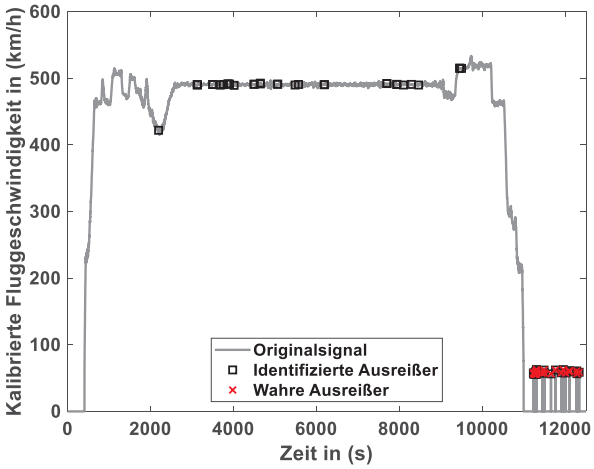

BILD 5. Identifikation von Ausreißern mit einem Hampel Filter anhand der Fluggeschwindigkeit über eine gesamte Flugmission

Anhand von Expertenwissen kann festgestellt werden, dass das Sensorsignal bis auf den zeitlich fortgeschrittenen Bereich, ab ca. 11.000 Sekunden, keine maßgeblichen Anomalien aufweist. Der Hampel Filter identifiziert jedoch fälschlicherweise Ausreißer im Verlauf der Fluggeschwindigkeit während des Reisefluges. Verdeutlicht wird dies in BILD 6, in dem ein Ausschnitt aus dem Reiseflug vergrößert dargestellt ist. In blau sind zudem die oberen und unteren Grenzbereiche aufgrund der errechneten Schwellwerte des Fensters in der Farbe blau gekennzeichnet. Der Bereich zwischen diesen Grenzen

kann aufgrund der Zeitreihe des Originalsignals sehr schmal ausfallen, sodass die Methode hierbei sehr sensitiv ggü. Parameterveränderungen ausfallen kann und Ausreißer fälschlicherweise identifiziert werden können.

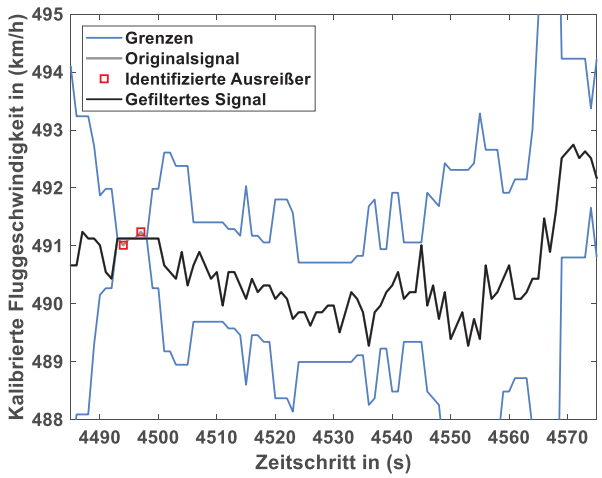

BILD 6. Grenzbereiche eines Hampel Filters im Ausschnitt der Fluggeschwindigkeit während des Reisefluges mit nach Experteneinschätzung fälschlicherweise identifizierten Ausreißern

Die Grundlage für die Anwendung der DFFITS-Methode ist das Regressionsmodell

$$\mathrm{DFFITS}_i = \beta \tag{12}$$

Dahinter steht die Annahme, dass die Differenzen zwischen benachbarten Datenpunkte innerhalb der Parameter über die Zeit betrachtet konstant bleiben. Die Differenzen zwischen den Parameterwerten gehen letztendlich in das Regressionsmodell ein. Damit stellen erst größere Unterschiede Anomalien bzw. Ausreißer dar, sodass diese Methode den klassischen Medianfilter übertreffen kann. In BILD 7 ist die Ausreißeridentifikation mit dieser Methode dargestellt. Es lassen sich hierbei auch verschiedene Niveausprünge (Engl. level shifts) erkennen, bspw. um die Zeitschritte 125 und 950 Sekunden, die mit dieser Methode nicht fälschlicherweise als Ausreißer identifiziert werden.

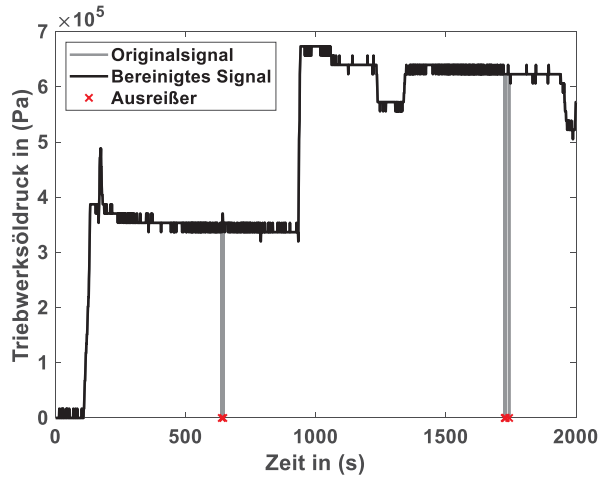

BILD 7. Identifikation von Ausreißern mit DFFITS anhand des Triebwerköldruckes von Start bis Steigflug

k-means und Hybridmethode

BILD 8 zeigt die Ausreißeridentifikation mit dem vorgestellten k-means Algorithmus. Auch hier werden für die Bewertung die Differenzen zwischen den Datenpunkten herangezogen. Mit dieser Methode lassen sich Stufensprünge und auch sog. influential level shift Ausreißer effektiv identifizieren.

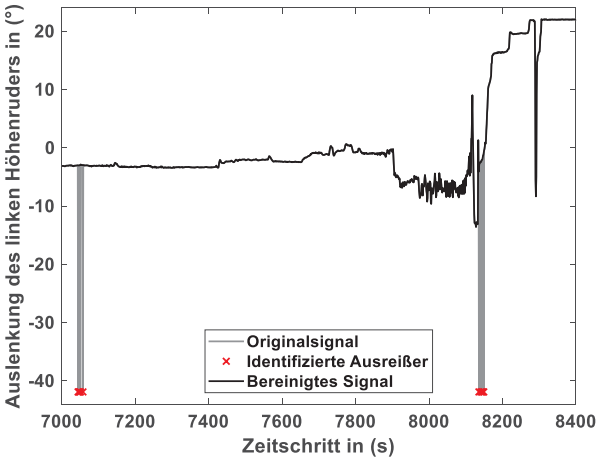

BILD 8. Ausreißeridentifikation mit k-means am Beispiel des Höhenruders gegen Ende einer Flugdatenaufzeichnung mit Erreichen der Abstellposition

Stets problematisch bei einer Ausreißeridentifikation sind jedoch auch mehrere, aufeinanderfolgende additive Ausreißer. Bspw. werden die Differenzen zwischen diesen Datenpunkten aufgrund ihrer Nähe zueinander nicht mehr als „kritisch“ durch eine 3σ-Gesetzmäßigkeit in der Bewertung der Differenzen ersten Grades eingestuft und können daher nicht mehr identifiziert werden. Abhilfe kann hierbei das Clustering des Originalsignals mit k-means schaffen, was bei einer Kombination beider Methoden zur in diesem Beitrag bezeichneten Hybridmethode führt.

Mit der zuvor vorgestellten 3σ-Gesetzmäßigkeit wird eine Differenzbetrachtung mit benachbarten Datenpunkten und dem k-means kombiniert. Die Clusteranzahl wird mit drei Zentroiden initialisiert, um den Verlauf geeignet zu klassifizieren. Ausreißer, die bereits mit der 3σ-Gesetzmäßigkeit markiert wurden und sich nun in einem Cluster mit übereinstimmenden Werten wiederfinden, können damit als aufeinanderfolgende additive Ausreißer identifiziert und behoben werden.

In BILD 9 ist die Identifikation von Ausreißern anhand der Position des linken Querruders des Flugzeuges aus den Testdaten dargestellt.

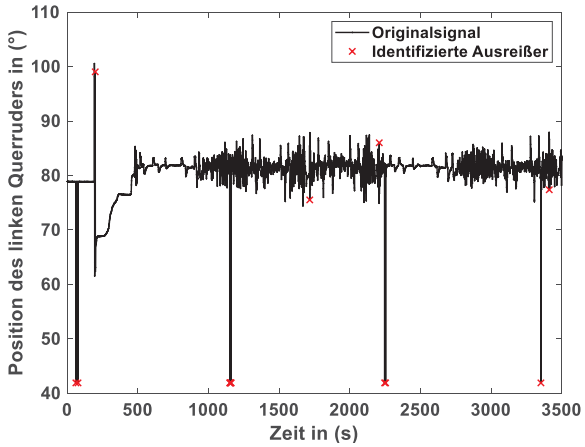

BILD 9. Identifikation von Ausreißern mit der entwickelten Hybridmethode anhand der Position des linken Querruders eines Flugzeuges

In BILD 10 sind die gebildeten Differenzen des Signals z-skaliert auf die Standardabweichung bezogen. Zudem sind die Schwellwerte bei $\pm 3\sigma$ eingezeichnet. Zur Identifikation wahrer Ausreißer wird nun noch das Clusterergebnis des k-means Algorithmus herangezogen und nur die markierten Ausreißer in BILD 9 behandelt. Dies erfolgt mit einem Medianfilter.

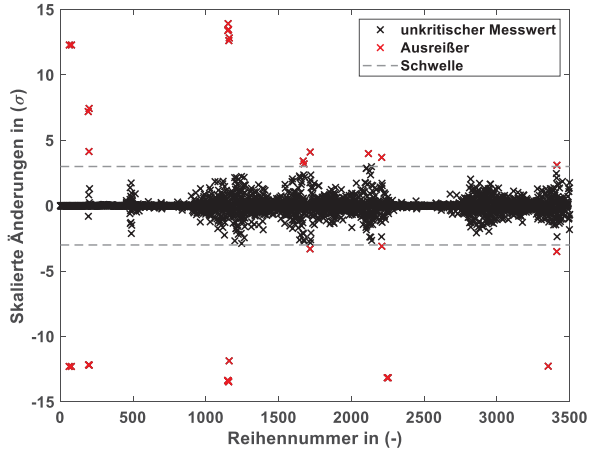

BILD 10. Ausreißeridentifikation über angepasste Sigma Methode

Die Effektivität der Identifikation mit dieser Methode und unter Bildung der Differenzen ersten Grades kann mit Hilfe einer kumulierten Verteilungsfunktion (CDF), wie in BILD 11 abgebildet, bewertet werden. Das Originalsignal hat schiefsymmetrische Gestalt und es bilden sich teilweise Ansätze von Plateaus in der Verteilungsfunktion aus. Aufgrund der Betrachtung der Differenzen ersten Grades ergibt sich eine der Normalverteilung im Mittelwert, in Symmetrie (Engl. skewness) sowie Stetigkeit ähnlichere Verteilungsfunktion, was in der Behandlung und Bewertung der Systemdynamik des aufgezeichneten Parameters und damit auch bei der Identifikation von Ausreißern maßgebliche Vorteile bieten kann (vgl. [6]).

Die Kurvensteilheit und Kurtosis nimmt zwar im Vergleich mit dem Originalsignal zu, jedoch zeugt dies von einer geringeren Standardabweichung und einer bereits günstigeren Tendenz zur Anwendung von Mittelwertschätzern (vgl. [16]).

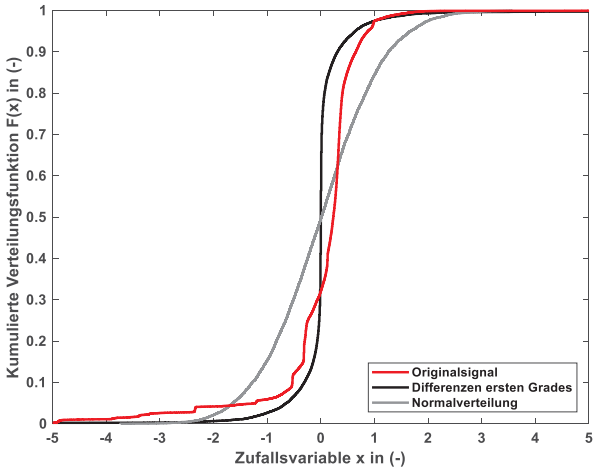

BILD 11. Vergleich der kumulierten Verteilungsfunktionen aus dem Originalsignal der Position des linken Querruders, der aus diesem Signal gebildeten Differenzen ersten Grades sowie einer Normalverteilung

Als Vorteile der entwickelten Hybridmethode bieten sich Zeitreihen mit nahezu stationären Differenzen ersten Grades an. Die Differenzen ersten Grades bieten eine geeignete Grundlage für die Ausreißeridentifikation als das Originalsignal, da damit eine geeignete Stationarität bewertet werden kann. Statistische Belegbarkeit kann bspw. über einen sog. unit-root-test, bspw. augmented Dickey-Fuller Test, hergestellt werden. Dieser prüft, ob ein Parameter in einer Zeitreihe nicht stationär ist und eine Einheitswurzel (1 ist Nullstelle des charakteristischen Polynoms) besitzt. Die Nullhypothese bezeichnet im Allgemeinen das Vorhandensein einer Einheitswurzel und die alternative Hypothese ist entweder durch Stationarität, Trendstationarität oder explosive Wurzel gekennzeichnet. Das vorgestellte Verfahren kann so eine Unabhängigkeit von der Normalverteilung eines Parameters erzeugen und ist zudem frei von individuell auszuwählenden (Filter-) Einstellparametern.

Nachteile des Verfahrens ergeben sich jedoch aus der Abhängigkeit durch Annahme der Normalverteilung der Differenzen als Grundlage des 3σ-Gesetzes. Die Erkennung mehrerer additive Ausreißer beläuft sich hierbei grundlegend auf die Gleichheit der Zeitreihenwerte innerhalb der identifizierten Cluster mit minimalen oder maximalen Zentroiden. Durch eine weitere Parametrisierung könnte hierbei noch eine Optimierung vorgenommen werden. Auch Differenzen nahe des 3σ-Schwellwertes können fälschlicherweise als Ausreißer identifiziert werden. Hierbei kann es dann zu einer Behandlung der Ausreißer kommen, die einer Glättung des Signals ähnlich kommt.

5.2. Qualitative Bewertung der Effektivität der angebrachten Maßnahmen

Um abschließend die Wirksamkeit der Methoden beurteilen zu können, kann der sogenannte Jaccard Koeffizient J herangezogen werden (vgl. [1]). Dieser bewertet die Ausreißeridentifikation der Methoden anhand einer Wahrheitsmatrix und den dazugehörigen Kategorien. Zu diesen gehören falsch negativ (FN) und falsch positiv (FP) klassifizierte Ergebnisse. Diese werden in Relation zu den richtig detektierten Ausreißern, mit wahr negativ bezeichnet (kurz WN) in der Form

$$J = \frac{1}{1 + \frac{FP + FN}{WN}} \tag{13}$$

gesetzt. Zur ersten Kategorie gehören Datenwerte, welche Ausreißer sind, jedoch nicht durch die Methode identifiziert werden, zur zweiten Kategorie zählen Datenwerte, die keine Ausreißer sind, aber durch die Methode fälschlicherweise als Ausreißer markiert werden.

Die Untersuchung zur Wirksamkeit der angebrachten Methoden wurde heuristisch und qualitativ auf den Parametern für Triebwerksöldruck, Höhen- und Querruderausschlag, kalibrierte Fluggeschwindigkeit und Longitudinalbeschleunigung für fünf ausgewählte Flüge durchgeführt. Es lässt sich damit feststellen, dass Jaccard Koeffizienten für die Hybridmethode durchweg im Bereich $J > 0{,}9$ erreicht werden. Vereinzelt können Medianfilter und die Identifikation mittels DFFITS ähnlich gute Ergebnisse erzielen. Dennoch sind diese Ergebnisse von den einzelnen Einstellparametern wie Fensterbreite und Clusteranzahl sehr stark abhängig. Bei Medianfiltern und Clusteralgorithmen sollten diesbezüglich für jedes Signal die Einstellparameter optimiert werden. Überwiegend liegen jedoch die Vergleichsmetriken der Methoden mit MAD/Hampel oder auch k-means bei Jaccard Koeffizienten von etwa $J \approx 0{,}5$ oder schlechter. Auf den verwendeten Testdaten können damit signifikante Unterschiede für die Wirksamkeit der unterschiedlichen Identifikationsmethoden für Ausreißer nachgewiesen werden und es lässt sich die Hybridmethode mit guten bis sehr guten Ergebnissen herausstellen.

5.3. Bewertung der Performanz für datenbasierte Modellierungen

Für die abschließende Bewertung der Performanz zur Ausreißeridentifikation und Behandlung werden Ergebnisse einer datenbasierten Modellierung mit iterativen Lernmodellen vorgestellt. Für die Erläuterung der Modelle aus dem Bereich des maschinellen Lernens sowie zur Vorgehensweise des Trainings der Lernnetzwerke und den Metriken zur Validierung wird auf [17] verwiesen.

Als Modelleingänge für die nachfolgenden Analysen dienen insgesamt 24 Merkmale (Prädiktoren). Zu diesen gehören Steuerflächen der primären Flugsteuerung wie Quer-, Seiten- und Höhenruderausschläge, sekundäre Steuerflächen wie die Klappenpositionen, Höhe, Steig-/Sinkrate, Fluglage und Anstellwinkel, Beschleunigungen, Geschwindigkeit, Treibstoffgewicht, Druck und Temperatur sowie Windgeschwindigkeit und Windrichtung. Das Modellierungsergebnis (Engl. response) stellt den kumulierten Treibstoffffluss aller Triebwerke dar.

Als Lernmodelle werden Entscheidungsbäume zur Klassifikation und Regression, ebenfalls eine Ensemble-Methode (Engl. abgekürzt bagged trees für bootstrap aggregating, siehe [18]) mit einem sogenannten random forest („Wald“ aus unkorrelierten Entscheidungsbäumen) sowie ein neuronales Netzwerk untersucht. Diese wurden bereits in einem vorangegangenen Beitrag mit relevanten Metriken für die Performanzbewertung vorgestellt und diskutiert (siehe [17]).

Es werden unterschiedliche Modelle für die Reiseflugphase sowie eine Flugdurchführung mit den Phasen Steigen, Reisen und Sinken antrainiert. Bei den Entscheidungsbäumen beträgt die minimale Anzahl von Blättern 36 und es wird eine 5-fache Kreuzvalidierung verwendet. Die vorwärts gerichteten neuronalen Netze weisen eine versteckte Schicht mit fünf Neuronen auf. Als Trainingsfunktion wird die Bayes Regularisierung gewählt und die Anpassung der Gewichte erfolgt über Rückpropagierung. Vom Trainingsdatensatz dienen randomisiert ausgewählte Anteile von 10 % und 20 % zum Testen während des Trainings und zum Validieren des trainierten Netzes. Zudem wird ein weiterer, im Englischen sogenannter hold out, Datensatz mit 250.000 Instanzen zur abschließenden Validierung herangezogen. Die Daten werden für das Training des neuronalen Netzes zudem normalisiert.

Als Bewertungsmetrik wird nachfolgend die Wurzel der mittleren Fehlerquadratsumme (Engl. root mean square error, kurz RMSE) verwendet. Dieser bildet sich aus den einzelnen Fehlerinkrementen $\mathrm{error_i}$ zwischen Modellergebnis $x_{\mathrm{Response,i}}$ und realer Zielgröße $x_{\mathrm{Target,i}}$ über

$$\mathrm{error}_i = x_{\mathrm{Target},i} - x_{\mathrm{Response},i} \tag{14}$$

zu

$$\mathrm{RMSE} = \sqrt[2]{\frac{1}{n}\sum_{i=1}^{N}(\mathrm{error}_i)^2} \tag{15}$$

wobei n die Anzahl an Beobachtungen darstellt. Exemplarisch werden nachfolgend die Ergebnisse mit neuronalem Netz für eine Flugmission, bestehend aus den Flugphasen Steigen, Reisen und Sinken, vorgestellt und weitere Ergebnisse aus der Analyse mit anderen Lernmodellen beschrieben (dann ohne explizite tabellarische oder bildliche Darstellung der Ergebnisse).

In TAB 1 ist die Validierungsmetrik für die Diagnoseergebnisse der Treibstoffflussprofile mit den unterschiedlichen Modellbildungen (ohne Vorverarbeitung (Originärdaten) sowie mit der Hybridmethode) mit Hilfe des neuronalen Netzes aufgeführt.

VALIDIERUNGSPHASE	MODELL	RMSE [kg/h]	p-Wert | h r
TRAINING	Originärdaten	146	<0.01 | 1 0,02
	Hybridmethode	141	
HOLD OUT	Originärdaten	163	<0.01 | 1 0,05
	Hybridmethode	158	
KREUZVERGLEICH HOLD OUT	Originärdaten in Hybridmodell	588	<0.01 | 1 0,08

TAB 1. Vergleich der Validierungsmetriken für die Modellierung mit einem neuronalen Netz für gesamte Flugmissionen, bestehend aus den Flugphasen Steigen, Reisen und Sinken

Im Vergleich der unterschiedlichen Modellergebnisse, welche auf der Verwendung von Originärdaten und den vorverarbeiteten Datensätzen beruhen, zeigen sich Unterschiede in den Fehlermetriken (Verlauf des Absolutfehlers error_i sowie bei der Wurzel der mittleren Fehlerquadratsumme RMSE). Diese weichen jedoch nicht sehr merklich voneinander ab. Allgemein kann mit dieser Untersuchung festgestellt werden, dass bei den vorverarbeiteten Datensätzen mit der Hybridmethode durchweg bessere Modellergebnisse im Training und Test mit iterativen Lernmodellen erreicht werden.

Anhand eines für die Testaufgabe geeigneten Wilcoxon-Vorzeichen-Rangtests kann überprüft werden, ob es sich um signifikant unterschiedliche Ergebnisse handelt. Bei dieser Teststatistik handelt es sich um eine nichtparametrische, verteilungsfreie Überprüfung zentraler Tendenzen im Median für zwei gekoppelte Stichproben (vgl. [19]). Trotz geringer absoluter Unterschiede bei den Werten für den RMSE lassen sich damit dennoch signifikante Unterschiede im Verlauf der Modellierungsfehler $\mathrm{error_i}$ nachweisen. Der Test zeigt Signifikanzniveaus unterhalb von einem Prozent an. Der entsprechende p-Wert wird bei diesem Vorgehen anhand einer z-Statistik in Abhängigkeit der Anzahl an Instanzen gebildet und ist in TAB 1 angegeben (vgl. [19]). Damit kann die Nullhypothese, dass die zu untersuchenden Stichproben aus einer Verteilung mit gleichen Medianen stammen, verworfen werden. Der ebenfalls in TAB 1 angegebene h-Wert von eins zeigt an, dass der Test die Nullhypothese zurückweist und es damit signifikante Unterschiede zwischen den Medianen auf dem Signifikanzniveau kleiner 1 % gibt. Es ist demnach ein ausreichend statistischer Beleg gegeben, dass der Medianwert der mittleren Fehlerquadratsumme eines Modells ohne Ausreißeridentifikation und Behandlung größer ist als der Medianwert mit einer entsprechenden Datenvorverarbeitung. Die Signifikanz lässt jedoch noch keine Aussagen über das Ausmaß des Effektes zu. Dies kann anhand der standardisierten Ermittlung einer Effektstärke r bewertet werden (vgl. [19]). Für die Validierung mit einem Hold Out Set ergeben sich hierfür nach Cohen (nach [19]) jedoch geringe $(r < 0{,}2)$ Effektstärken.

Anhand eines Vergleiches der Fehlerverläufe und der mittleren Fehlerquadratsumme zwischen Training und Test lässt sich eine gute Generalisierbarkeit der gelernten Modelle für die unterschiedlichen Flugphasen ausmachen. Damit zeigen sich keine signifikanten Einbußen der Modellgüte. Ein Vergleich der Determinationskoeffizienten sowie des mittleren absoluten Fehlers, hier nicht explizit dargestellt, lässt ebenfalls keine signifikanten Unterscheidungen zu. Ein deutlicher Anstieg der Fehlermetrik ist jedoch zu verzeichnen, wenn das Modell, welches mit vorverarbeiteten Daten trainiert wurde, auf Originärdaten getestet wird (siehe TAB 1). Im Vergleich mit den Modellierungen mittels Entscheidungsbäumen zeigen sich auch hier bereits signifikante Unterschiede der Fehlerwerte beim Training, da hier mit der Teststatistik bereits die Nullhypothese verworfen werden kann.

Anhand der Ergebnisse lässt sich festhalten, dass es sich bei der für diese Untersuchung verwendeten Datenbasis um nur moderat mit Ausreißern behaftete Parameterwerte handelt. Der Fokus dieses Beitrages liegt im Gegensatz zu systematischen Ausreißern jedoch auf der Identifikation von zufälligen, nicht systemdynamischen Ausreißern und Anomalien. Aufgrund des Umfanges und der benötigten Automatisierung der Verfahren und Analysen ist eine Abschätzung des Effizienzgewinns der Vorverarbeitungsmethoden daher a priori nur sehr eingeschränkt bis gar nicht möglich. Dennoch überwiegen die Vorteile solcher Verfahren, was die qualitative Bewertung in diesem Beitrag untermauert.

6. FAZIT UND AUSBLICK

Die Intention dieses Beitrages ist die Vorstellung und Untersuchung von Methoden zur effizienten Identifikation und Behandlung von Ausreißern in Flugbetriebsdaten. Dadurch soll die Datenqualität erhöht sowie die Genauigkeit bei anschließenden Untersuchungen und bei der Verwendung der Daten zum Aufbau iterativ lernender Modelle im maschinellen Lernen gesteigert werden.

Anhand von statistischen Metriken, Verteilungsanalysen, einer qualitativen Bewertung anhand des Jaccard Koeffizienten sowie der Bewertung von Trainings- und Testperformanzen in iterativen Lernmodellen zeigt der vorliegende Beitrag, dass die Ausreißeridentifikation und Behandlung mit der entwickelten Hybridmethode am effektivsten im Vergleich mit der angebrachten Methodenauswahl ausfällt. Dahingehend zeigen die Ergebnisse signifikante Unterschiede bei der Verwendung von Daten, die mit der entwickelten Hybridmethode vorverarbeitet werden. Diese kann vielversprechend zur Identifikation und Behandlung von Ausreißern in Flugbetriebsdaten angewendet werden, bspw. auch für mehrere aufeinander folgende additive und innovative Ausreißer, ohne Niveauverschiebungen fälschlicherweise als Ausreißer zu klassifizieren. Demgegenüber eignen sich bei Signalen mit wesentlich höheren Rauschanteilen eher Filteralgorithmen zur Glättung.

Es hat sich bei dieser Untersuchung als unabdingbar herausgestellt, geeignete Kenntnisse über den zu behandelnden Parameterverlauf zu besitzen, um wahre Ausreißer von einer übrigen und realen Systemdynamik abgrenzen zu können.

KONTAKTADRESSEN

Sebastian Baumann, M. Sc.
baumann@fsr.tu-darmstadt.de

Prof. Dr.-Ing. Uwe Klingauf
klingauf@fsr.tu-darmstadt.de

DANKSAGUNG

Dieser Beitrag ist in Zusammenarbeit der Autoren im Rahmen eines Promotionsvorhabens sowie durch studentische (Abschluss-)Arbeiten am Institut für Flugsysteme und Regelungstechnik an der TU Darmstadt entstanden. Der vorliegende Beitrag trägt auch zur Bearbeitung von ähnlich gelagerten Herausforderungen bei. Die Autoren bedanken sich daher an dieser Stelle beim Bundesministerium für Wirtschaft und Energie (BMWi) für die Unterstützung im Teilvorhaben RetroEff (Retrofit Technologiebewertung für effiziente und sparsame Flugzeugflotten). Dieses beinhaltet die Entwicklung und Erprobung einer datenbasierten Bewertungsmethodik zur Nachweisführung der Effektivität von Maßnahmen zur Senkung des Kerosinverbrauches im Rahmen des Luftfahrtforschungsprogramms (LuFo V-2) durch die TU Darmstadt. Neben dem Institut für Flugsysteme und Regelungstechnik finden sich im Projekt RetroEff die Projektpartner Lufthansa Technik AG und DLR Lufttransportsysteme. Das Projekt endet am 31.07.2019.

LITERATURANGABEN

[1] Basu, S.; Meckesheimer, M.: Automatic outlier detection for time series: an application to sensor data. In: Knowledge and Information Systems, Vol. 11, Iss. 2, S.137-154; (2007)

[2] National Aeronautics and Space Administration (Hrsg.): NASA DASHlink. Abrufbar unter: https://c3.nasa.gov/dashlink/ (abgerufen am 31.07.2016)

[3] Heij, C.; de Boer, P.; Franses, P. H.; Kloek, T.; van Dijk, H. K.: Econometric methods with applications in business and economics. Oxford University Press, Oxford (2004)

[4] Runkler, T. A.: Data Mining. Springer Vieweg, Wiesbaden (2015)

[5] Brillinger, D. R.: Data Analysis, Exploratory. In: International Encyclopedia of Political Science. Badie, B.; Berg-Schlosser, D.; Morlino, L. [Hrsg.]. SAGE Publications, Thousand Oaks (2011)

[6] Tukey, J. W.: Exploratory Data Analysis. Pearson Publishing, Cambridge (1977)

[7] Kosfeld, R.; Eckey, H. F.; Türck, M.: Deskriptive Statistik. Springer Verlag, Wiesbaden (2016)

[8] Dawson, R.: How Significant Is A Boxplot Outlier? In: Journal of Statistics Education, Vol. 19 (2) (2011)

[9] Aguinis, H.; Gottfredson, R. K.; Joo, H.: Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. In: Organizational Research Methods, Vol. 16, Iss. 2, S. 270–301 (2013)

[10] Hodge, V.; Austin, J.: A survey of outlier detection methodologies. In: Artificial intelligence review, Vol. 22, Iss. 2, S. 85–126 (2004)

[11] Pearson, R. K.; Neuvo, Y.; Astola, J.; Gabbouj, M.: Generalized Hampel Filters. In: EURASIP Journal on Advances in Signal Processing. 2016:87. Springer International Publishing (2016)

[12] Belsley, D. A.; Kuh, E.; Welsch, R. E.: Regression Diagnostics. Wiley Interscience, Hoboken (2004)

[13] Salgado, C. M.; Azevedo, C.; Proenca, H.; Vieira, S. M.: Noise versus Outliers. In: Secondary Analysis of Electronic Health Records. MIT Critical Data [Hrsg.]. Springer (2016)

[14] Alpaydin, E.: Maschinelles Lernen. Oldenbourg Wissenschaftsverlag, München (2008).

[15] Lloyd, S. P.: Least Squares Quantization in PCM. In: IEEE Transactions on Information Theory Vol. 28, S. 129–137 (1982).

[16] Komorowski, M.; Marshall, D. C.; Salciccioli, J. D.; Crutain, Y.: Exploratory Data Analysis. In: Secondary Analysis of Electronic Health Records. MIT Critical Data [Hrsg.]. Springer (2016)

[17] Baumann, S.; Klingauf, U.: Prognose des Treibstoffverbrauches eines Flugzeuges mit Hilfe von machinellen Lernalgorithmen. In: 66. Deutscher Luft- und Raumfahrtkongress. Deutsche Gesellschaft für Luft- und Raumfahrt, München (2017)

[18] Witten, I. H.; Frank, E.; Hall, M. A.: Data Mining. Morgan Kaufmann [Hrsg.]. Elsevier Inc., Burlington (2011)

[19] Field, A.: Discovering Statistics Using IBM SPSS Statistics. SAGE Publications Inc., London (2015)