Ausreißerresistente Messdatenfilterung für Anwendung in der Ergospirometrie

Sebastian Guttke

Leipzig, den 5. Dezember 2007

Zusammenfassung

Unter dem Motto "HTWK meets HIMALAYA" wurde in einer Kooperation zwischen der Fa. CORTEX Biophysik GmbH und der HTWK Leipzig, Fachbereich EIT, Institut AET ein finanziell von der IHK zu Leipzig gefördertes Forschungsprojekt initiiert. Die Fa. CORTEX Biophysik GmbH hat 2007 die mobile Messtechnik (Ergospirometer) für das im HIMALAYA durchgeführte Forschungsprojekt "Xtreme Everest " geliefert. In der Kooperation wurde die bis dahin noch nicht hinreichend beantwortete Fragestellung untersucht, ob und wie es möglich ist, trendaufweisende, stark verrauschte und mit großen Ausreißern behaftete Messdatensätze zu filtern, um die relevanten Informationen sicher aus den Messdaten zu extrahieren. Filter wie Median oder Mittelwert funktionieren gut, solange die Signale trendfrei und ohne Ausreißer sind. Bei Auftreten von Ausreißern oder Sprüngen im Signal werden die Ergebnisse so stark verfälscht, dass Fehlinterpretationen durch den Wissenschaftler möglich sind. Im Projekt wurde anhand realer Messdaten von Probanden ein Testsignal entwickelt. Um Worst-Case-Bedingungen zu simulieren, wurde das Testsignal in Ordinatenrichtung mit einem Rauschen von bis zu 3 Sigma der realen Messdaten und mehreren starken Ausreißern (kleiner und größer als der wahre Wert) überlagert. Auf der Abzisse wurde eine äquidistante Schrittweite der Werte festgelegt. Mit Hilfe dieses Testsignals wurde die Wirkung verschiedener Filter anhand objektiver Bewertungskriterien verglichen. Im Ergebnis konnte das Testsignal mit einem hybriden wiederholten Median Filter mit bis zu 40% kleinerem Fehler (bezogen auf Median aus 3 Werten) reproduziert werden. Das Filter entfernt bereits ab einer Filterbreite von 10 Messwerten sicher 1 bis 2 starke Ausreißer und kann auch zum Filtern großer Messreihen verwendet werden. ...

Stichworte: Signal Extraction; Robuste Regression; Ausreißer; Outliers; Artefaktbeseitigung; RM-Filter

Inhaltsverzeichnis

1 Einleitung

1.1 Vorwort zur Motivation des Filterprojektes

In der angewandten Messtechnik ist eine häufig vorkommende Aufgabe, physikalische Größen zu bestimmen, die mit heutigem Stand der Technik oder vertretbarem Aufwand nicht direkt messbar sind. Dazu werden direkt messbare physikalische Hilfsgrößen aufgezeichnet und daraus anschließend die eigentlich interessierenden Größen berechnet. Bei der Berechnung muss allerdings beachtet werden, dass sich Messfehler in den aufgezeichneten Hilfsgrößen in der Rechung zu einem resultierenden Gesamtmessfehler fortpflanzen. Da dieses Gebiet in der verfügbaren Literatur schon umfassend behandelt worden ist, soll es hier nur erwähnt werden.
Die Auswertung von Messdatensätzen kann prinzipiell entweder im Nachgang mit Hilfe diverser Software am PC (retrospektiv, hier als offline bezeichnet) oder mit bestimmter, möglichst kleiner zeitlicher Verzögerung während der laufenden Messung (hier als Online bezeichnet) erfolgen.
Die Herausforderung, Daten Online auszuwerten, besteht zum Beispiel, beim automatisierten Monitoring von Intensivpatienten, wo bestimmte physiologische Parameter überwacht, dargestellt und bei Überschreitung festgelegter Kriterien Warnmeldungen augegeben werden sollen. Die Vorraussetzungen zur Lösung dieser Aufgabe verbessern sich permanent, da die verfügbaren Mikrokontroller bei steigender Rechenleistung in immer kleinerer Bauform gefertigt werden bzw. bei gleicher Baugröße die Leistungsfähigkeit steigt. So ist es schon heute problemlos möglich, große Datensätze schnell Online zu bearbeiten, in die gewünschten Werte umzurechnen und diese auszugeben.

In diesem Beitrag soll der Fokus auf die Auswertung und Darstellung von Messdatensätzen gerichtet werden, die Trends ausweisen, mit starkem Rauschen und unbekannten zufälligen Störungen stark verfälscht sind. Zufällig meist stark verfälschte Messwerte werden oft auch als Ausreißer oder Artefakte bezeichnet, da sie sich nicht in einen Trend einpassen, den man oftmals schon in den unbearbeiteten Messdaten feststellen kann. Es sollen Möglichkeiten gefunden und diskutiert werden, wie die eigentlich interessanten Informationen solcher Messreihen unverfälscht dargestellt werden könnén. Die größte Her-

ausforderung besteht in der Automatisierung der Ausreißererkennung, damit schon während der Datenerfassung erkannt werden kann, ob ein neuer Messwert plausibel ist und dargestellt werden kann oder als unplausibel eingestuft und entfernt werden muss.
Im Allgemeinen werden in der Praxis meist aus Zeitgründen solche, dem Wissenschaftler unplausibel erscheinenden Messwerte vor der eigentlichen Auswertung einfach aus der Messreihe entfernt und erst dann z.B. über eine Regressionsanalyse versucht, die mathematischen Zusammenhänge der Messdaten heraus zu finden. Diese Vorgehensweise birgt allerdings die Gefahr, dass mit den vermeintlichen Ausreißern wertvolle, relevante Informationen aus der Messreihe entfernt werden. Bei programmierten, automatisch arbeitenden Ausreißer-Entdeckungsregeln (Filter) gibt es die Möglichkeit von zu großer oder zu geringer Empfindlichkeit bei der Identifikation von Ausreißern. Im ersten Fall würden Messwerte als Ausreißer erkannt, die keine sind und somit wertvolle Information gelöscht oder im zweiten Fall grobe Ausreißer nicht als solche identifiziert und die Messergebnisse dadurch stark verfälscht, so dass Fehlinterpretationen der Messergebnisse wahrscheinlich sind. Die richtige Balance für eine Ausreißer-Endeckungsregel zu finden, ist keine einfache Aufgabe und kann nur für ein spezielles Problem zugeschnitten gelingen.
Grundsätzlich sollte vor der Verwendung eines Datenfilters genügend Arbeit in die Analyse der Rohdaten und die Ursachenforschung für Ausreißer investiert werden, denn jeder physikalische oder technische Effekt, der während der Produkt-Entwicklungsphase nicht verstanden und aus Zeitgründen vernachlässigt worden ist, kann bei der Serienproduktion mit höheren Stückzahlen in Erscheinung treten und dann zu hohen Reklamationszahlen und damit verbundenen hohen Folgekosten führen.

1.2 Definition des Begriffes "Ausreißer"

[1] Zur Definition von Ausreißern (engl.: outlier) gibt es grundsätzlich ähnliche Aussagen in der Literatur, die jedoch im Detail etwas voneinander abweichen. Intuitiv verstehen die meisten Wissenschaftler Ausreißer als Messwerte die nicht erwartungsgemäß sind und sich nicht in die Messreihe einpassen. In Barnett and Lewis [1], Seite 4 wurde sinngemäß formuliert, dass ein Ausreißer eine Beobachtung (oder eine Teilmenge von Beobachtungen) innerhalb eines Datensatzes ist, die außerhalb des zentralen Bereichs einer Verteilung

[1] Der Inhalt dieses Kapitels ist im Wesentlichen ein übersetzter Auszug aus Heft [4]

liegt (Extremwert) und daher mit dem Rest der Daten schwer vereinbar zu sein scheint. In Hawkins [3] wird sinngemäß ausgeführt, dass Ausreißer Beobachtungen sind, die so stark vom Rest der Messwerte abweichen, dass sie den Verdacht erregen, von einem anderen Mechanismus erzeugt worden zu sein. Es gibt in Beckmann and Cook [2], Seite 121 eine weitere etwas abweichende Definition, dass Ausreißer Werte sind, die von der erwarteten Verteilungsform abweichen. Die konsequente Anwendung dieser Definition bewirkt, dass empfindlicher bewertet wird und auch weniger eindeutig abweichende Werte als Ausreißer deklariert werden.
Im Internet-Lexikon "ILMES"[5] wird folgende Definition gegeben: "Unter Ausreißern versteht man einen Datenpunkt, der relativ weit weg von den übrigen Fällen eines (eindimensionalen) Datenbündels bzw. einer (zwei- oder mehrdimensionalen) Datenwolke liegt." Ausreißer werden oft wie folgt identifiziert:

- inferenzstatistischer Kriterien (z. B. Fälle, die mehr als 2 oder 3 Standardabweichungen vom Mittelwert entfernt liegen),
- anhand deutlich größerer Werte in einschlägigen Maßzahlen (siehe Residuen), und zum Teil aufgrund
- visueller Inspektion

1.3 Modelle zur Ausreißererzeugung

[2] Es gibt verschiedene mathematische Modelle, die beschreiben sollen, wie sich Systeme verhalten, bei denen sich Ausreißer unter den Daten befinden. Zwei bekannte Modelle sind das

- SLIPPAGE - Modell und das
- MIXTURE - Modell.

Diese zwei Modelle bilden zwar nicht die realen stochastischen Ausreißer erzeugenden Effekte ab, helfen aber ein Verständnis zu erlangen, wie Ausreißer in der Praxis Einfluss nehmen.

[2] wie 1

Das **SLIPPAGE - Modell** ist das am weitesten verbreitete mathematische Modell. Hierbei wird angenommen, dass es eine Anzahl von n Beobachtungen gibt, aus denen $n - r$ Werte der Normalverteilung zugeordnet werden können und r Werte aus einer anderen Verteilungsform stammende Ausreißer sind. Bei der Normalverteilung $N(\mu, \sigma^2)$ ist μ der Mittelwert und σ die Standardabweichung (um den Mittelwert) bzw. σ^2 die Streuung (Abweichung vom Mittelwert).

Allgemein lautet die Funktion der Normalverteilung:

$$f(x) = n_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

Normiert auf $\mu = 0$ und $\sigma = 1$ ergibt sich die Funktion der Normalverteilung zu:

$$f_{\mu=0;\sigma=1}(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} \tag{2}$$

Mit diesem Modell könnten bis zu r Ausreißer erzeugt werden. Zwei einfache abgewandelte für die Betrachtung der Herkunft von Ausreißern verwendbare Verteilungsformen sind:

1. die in der Lage verschobene Normalverteilung (engl. location-shift-model) mit $N(\mu + a, \sigma^2)$ bei der der Mittelwert um a von μ abweicht oder

2. die in der der Streuung abweichende Normalverteilung (engl. scale-shift-model) mit $N(\mu, b \cdot \sigma^2)$ bei der die Streuung um den Faktor b größer oder kleiner als die der Normalverteilung ist.

Beim **Mixture - Modell** wird davon ausgegangen, dass unverfälschte Daten mit der Wahrscheinlichkeit von $1 - p$ von einer Verteilung $G1$ und die Ausreißer mit der Wahrscheinlichkeit p von einer anderen Verteilung $G2$ stammen. Die n Beobachtungen stammen damit von

einer resultierenden Mischverteilung

$$(1 - p) \cdot G1 + (p) \cdot G2 \tag{3}$$

mit der konstanten Wahrscheinlichkeit $0 \leq p \leq 1$. Die Anzahl der verfälschten Werte ist eine Zufallszahl. Es gibt keine Ausreißer, wenn $p = 0$ und ausschließlich verfälschte Werte, wenn $p = 1$ wird.

1.4 Motivation für die Untersuchung von Ausreißern

[3] Mathematische Modelle und die praktische Erfahrung zeigen die Notwendigkeit die vorliegenden Daten immer auf mögliche Ausreißer zu überprüfen. Manchmal befinden sich eigentliche Ausreißer derart in den Daten verborgen, dass sie nicht als solche in Erscheinung treten. Da Ausreißer aber oft eine bestimmte Ursache haben, könnten wenige erkennbare Ausreißer die "Spitze eines Eisberges" sein. Das Herausfinden der Ursache kann viele zusätzlich in der Datenmenge verborgene Ausreißer identifizieren. Eine genaue Kenntnis der Ursache führt zu besserem Systemverständnis und kann das Entwickeln von Prozessoptimierungen fördern. Ein Zitat aus [4]: "**Ausreißer sind manchmal Edelsteine, da sie die nützlichste Information der Untersuchung enthalten**." (die Weiterentwicklung der Technik betreffend)

1.5 Mögliche Ursachen für die Entstehung von Ausreißern

- Ausreißer können z.B. aus einer Messung mit zu geringer Auflösung oder Messfehlern stammen. So kann eine Messung lediglich ein oder zwei isolierte zufällige Ausreißer aufweisen oder aber auch viele Ausreißer, die nicht verdächtig bzw. auffällig sind.

- In der Medizintechnik können Ausreißer auch Artefakte, scheinbare Befunde sein, die, durch technische Störungen des Messaufbaus oder durch den Probanden selbst verursacht, das eigentliche Messsignal verfälschen.

- Die Annahme einer falschen Verteilung für die Messdatenanalyse kann scheinbare Ausreißer erzeugen, die bei Verwendung einer passenden Verteilung keine sind. Ein häufiger Fehler in der Datenanalyse ist die Annahme, dass Messdaten normal verteilt sind, obwohl sie von einer wesentlich anderen Verteilung stammen. Eine gründliche Untersuchung der Ausreißer kann auch hier zu neuen statischen Erkenntnissen führen.

- Gelegentliche Ausreißer können ein Hinweis auf eine gewisse Struktur in den Daten sein. So kann es z.B. sein, dass es eine Wiederholung von Ausreißern gibt, die erst er-

[3] Der Inhalt dieses Kapitels ist im Wesentlichen ein übersetzter Auszug aus Heft [4]

sichtlich wird, wenn man die Daten über einen längeren Zeitraum betrachtet (Effekte in speziellen Schichten bei der Arbeit, Tageszeiten, Tagen oder auch Monaten).

- Gelegentlich zeigt ein ungewöhnlicher Messwert, dass solche Werte möglich sind. Unter Umständen führt das Erforschen der genauen Umstände der Entstehung dieses Wertes zu bedeutenden Verbesserungen in Prozessen oder zur Entwicklung alternativer Produkte.

1.6 Schlussfolgerungen zum Umgang mit Ausreißern

Fußnote [4]

- Ausreißer sollten als Teil der gesammelten Daten sorgfältig mit gespeichert und auch die Messbedingungen exakt mit dokumentiert werden. Gründliche Datenerfassung mit Aufzeichnung der Ausreißer fungiert gewissermaßen als Erinnerung, dass man die Ursache für Ausreißer noch nicht behoben hat. Unterdrückung der Ausreißer bei der Messdatenerfassung kann die Aufmerksamkeit für ein potentielles ernsthaftes Problem oder auch Weiterentwicklungspotential reduzieren.

- Ausreißer deren Ursachen nicht eindeutig erklärt werden können, sollten in der Datenanalyse verwendet werden, gerade wenn sie Fehler zu sein scheinen.

- Gleichzeitig sollten Ausreißer deren Ursache eindeutig geklärt ist nicht aufgezeichnet oder nicht in der Datenanalyse verwendet werden, da sie zu falschen Schlüssen aus den Messdaten führen können. Die Ursache sollte, wenn technisch möglich und finanzierbar, behoben werden.

1.7 Behandlung von Ausreißern

[5] Je nach der Zielstellung gibt es verschiedene Möglichkeiten Ausreißer mit Hilfe statistischer Methoden zu behandeln. Der Übersicht wegen sollen hier einige Möglichkeiten erwähnt werden.

1. Die Kennzeichnung von Ausreißern kann mit Z-Scores, modifizierten Z-Scores und dem relativ bekannten Boxplot erfolgen.

2. Unterbringung (engl. accomodation) von Ausreißern bedeutet, dass die Berechnung von Kenngrößen wie Mittelwert und Streuung so zugeschnitten (engl. trimmed) wird, dass auftretende Ausreißer den tatsächlichen Wert nicht mehr verfälschen. Der Mittelwert und andere klassische Schätzer (engl. estimators) haben keinen Schutz bzw. keine Widerstandfähigkeit (engl. resistance) gegen Ausreißer, so dass die Schätzwerte bei Auftreten von Ausreißern sofort stark verfälscht werden. Beim Zuschnei-

[4] Der Inhalt dieses Kapitels ist im Wesentlichen ein übersetzter Auszug aus B.Iglewicz [4]
[5] wie 4

den oder auch Trimmen werden die Werte zuerst der Größe nach sortiert und dann die größten und die kleinsten Werte abgeschnitten. Danach werden der Mittelwert und die Streuung (zugeschnitten) berechnet. (siehe B.Iglewicz [4], S.20-21)

3. Identifikation von Ausreißern mit Hilfe statistischer Tests: Es gibt für verschiedene Annahmen, Hypothesen und Verteilungsformen jeweils zugeschnittene Tests, die in der Literatur näher beschrieben sind. Zu nennen wären hier der einfach zu programmierende und auch bei Normalverteilung mit mehreren Ausreißern gut geeignete Test **"Generalized EDS Procedure "** und alternativ auch der **"Dixon-Type-Test"**, die beide bei kleinen Stichproben anwendbar sind. (siehe auch Wikipedia [6])

Bei anderen von der Normalverteilung abweichenden Verteilungsformen wie z.B.

- Lebenszeitverteilungen (Lognormal-, Weibull-, Exponentialverteilung (dem Spezialfall mit konstanter Fehlerrate) oder
- censored data,

gibt es jeweils angepasste Methoden die zur Ausreißererkennung dienen. (siehe auch: B.Iglewicz [4], S.43, ff. oder Weibull-Homepage [6])

1.8 Filtermöglichkeiten für ausreißerbehaftete Datensätze

- Robuste Lineare Regression

  Ausreißer beeinflussen den Verlauf einer Regressionsgeraden stärker als die restlichen Werte. Deshalb ist es sinnvoll, eine robuste Regression anzuwenden, die den Einfluss der Ausreißer minimiert oder sogar ausschließt und den Verlauf der Regressionsgeraden mit minimalem Fehler ermittelt. Die robuste lineare Regression wird vom Autor als eine effiziente Methode angesehen, um aus den Messdaten offline und mit bestimmter Zeitverzögerung auch Online evt. vorhandene Ausreißer sicher zu eliminieren und trendbehaftete Signale aus Messdaten zu extrahieren. Um solche Messdaten filtern zu können, muss die lineare Regression mit einer optimal gewählten Fensterbreite lokal angewendet und dann als Zeitfenster[6] gleitend über die ganze Messreihe verschoben werden. Je Position können dann ein oder mehrere Schätzwerte berechnet werden. Das Zeitfenster kann so auch in Sprüngen über den Datensatz verschoben und je Position die benötigte Anzahl von Schätzwerten berechnet werden.

  Mögliche, für die lineare Regression anwendbare Methoden sind die "Hebelwirkung mit der Schätzwert-Matrix" (engl. leverage & hat-matrix), die Schätzung von Werten mit Auslassung einzelner Werte (engl. deletion), studentisierte Residuen (engl. studentized residuals), mehrmaliger oder wiederholter Median (engl. repeated median).

[6] z.B. Fensterbreite über 10 Werte, Datensatz mit 1000 Werten

Nach einer Spektralanalyse der Messdaten lassen sich auch Grenzwerte für maximal zulässige Anstiege zwischen den einzelnen Messwerten bzw. die höchsten im Nutzsignal enthaltenen Frequenzen festlegen. Die Messpunkte mit unzulässig hohen Anstiegen können als unplausible Werte aus der Datenauswertung ausgeklammert und durch plausible Werte ersetzt werden. Weitere Möglichkeiten zum Filtern von Daten sind

- Die digitale Tiefpassfilterung mit FIR oder IIR-Filter
  Mit dieser Methode werden die Messergebnisse mit einem digitalen Tiefpassfilter geglättet und somit starke Sprünge im Signal (Ausreißer) gedämpft. Mit dieser Methode werden alle Frequenzanteile oberhalb der Grenzfrequenz mit einer bestimmten Filtercharakteristik eliminiert.

- Hybrides Filter als Kombination aus robuster lin. Regression und FIR oder IIR-Tiefpass
  Mit dieser Methode werden die Tiefpassfilterung und die robuste linéare Regression kombiniert verwendet, um ein optimale Filtereigenschaften und minimale Filterbreiten zu erreichen.

- Robuster Filter (Glätter) basierend auf der gewichteten L1-Regression

- FMH - FIR Median Hybrid Filter
  Bei diesem Filter wird zur Bestimmung des Schätzwertes $FMH(x(t))$ der Median aus z.B. 3 Werten bestimmt, die von 3 nacheinander laufenden Filtern errechnet wurden. Die Reihenfolge kann z.B. ein Mittelwert mit der Filterbreite k vor dem aktuellen Messwert $x(t)$, der Messwert $x(t)$ selbst und ein weiterer Mittelwertfilter der Breite k nach dem Messwert $x(t)$ sein. Dieses Filter ist schon sehr resistent gegenüber Ausreißern und relativ leicht zu programmieren.

- RMH - Repeated-Median-Hybrid-Filter
  Ähnlich wie beim FMH Filter wird hier der Median aus z.B. 3 Werten ermittelt, die ein Median aus k- Werten vor dem aktuellen Messwert $x(t)$ (jeweils ermittelt über lokale robuste lineare Regression), der Messwert $x(t)$ selbst und ein Median aus k-Werten nach dem Messwert $x(t)$ (jeweils ermittelt über lokale robuste lineare Regression) sein können. Es können auch mehr als 3 Filter ausgewertet werden. Es ist vorteilhaft, den aktuellen Messwert mit zu verwenden, da hierduch die Flan-

ken besser erhalten bleiben. Je nach Anforderung und den vorliegenden Daten kann der aktuelle Messwert k auch durch den Mittelwert des Zeitfensters ersetzt werden (RMMH-**R**epeated **M**edian **M**ean **H**ybrid-Filter). (siehe hierzu Gather, U., Bernholt, Th., Fried, R.,(2006)[8] und Davis, P.L., Gather, U., Fried, R.(2004)[9])

- Adaptive Wiener Filterung
  Mit diesem Filter werden die Daten mit Hilfe eines korrelierten Referenzsignals gefiltert. Diese Methode ist adaptiv und die Filterkoeffizienten werden kontinuierlich über eine Rückkopplung angepasst. Das Wienerfilter bietet den großen Vorteil, dass die Flanken von tatsächlich im Nutzsignal vorkommenden Sprüngen nicht mehr als nötig verschliffen werden.

2 Robuste einfache lineare Regression

2.1 Grundlegendes zur Regression

Mit einer Regressionsanalyse soll der Zusammenhang zwischen einer abhängigen Variablen y und einer unabhängigen Variablen x mit Hilfe statistischer Methoden untersucht und mathematisch beschrieben werden. Der Zusammenhang kann durch verschiedene mehr oder weniger komplexe mathematische Funktionen beschrieben werden. Die einfachste Möglichkeit ist die Verwendung einer linearen Funktion $y = b \cdot x + a$. Eine bekannte, oft angewandte Methode zur Bestimmung und Optimierung der Parameter a und b ist die mathematische Methode der Minimierung der Summe der Fehlerquadrate (engl. least sqares). Hierbei wird die Regressionsgeraden so in die Datenpunktwolke platziert, dass die Summe der quadratischen Fehler[7] minimal wird. Diese Methode ist aber so anfällig gegen Ausreißer, dass sich der Anstieg der Geraden bereits bei Auftreten nur eines größeren Ausreißers negieren kann (siehe Abbildung 1).

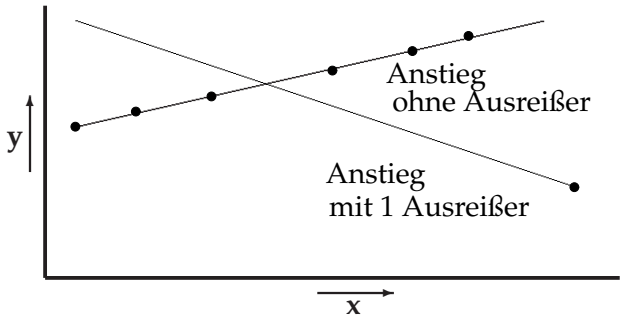

Abbildung 1: Least-Square-Regressionsgeraden mit einem Ausreißer

[7] Quadratischer Abstand eines jeden Datenpunktes zur Regressionsgeraden geht in die Summe ein

Als eine gegen Ausreißer robuste Regression soll hier ein Verfahren beschrieben werden, welches die Regressionsgeraden mit Hilfe der Repeated-Median-Methode in eine Wolke von Datenpunkten legt. Die allgemeine Geradengleichung lautet in diesem Fall:

$$\widehat{y}_{RM} = b_{RM} \cdot x + a_{RM} \tag{4}$$

RM – **R**epeated **M**edian-Index (wiederholter Median)
$\widehat{y}_{RM}$ – Schätzwertvariable als Funktion von x
b_{RM} – Anstieg der Regressionsgeraden (wiederholter Median)
a_{RM} – Offset der Regressionsgeraden (wiederholter Median)

Der Zirkumflex (auch Dach oder Hut genannt) über dem $\widehat{y}_{RM}$ beschreibt, dass der Wert $\widehat{y}_{RM}$ ein Schätzwert ist. Er berechnet sich aus der robusten Regressionsgeraden, d.h. dem unabhängigen Wert x multipliziert mit dem Anstieg b_{RM} und Addition des Produktes mit dem Offset a_{RM}. Die einzelnen zeitdiskreten Werte werden mit y_i und x_i [8] bezeichnet $(1 \leq i \leq k)$. k ist die Anzahl der betrachteten Messwerte und kann auch als Filterbreite bezeichnet werden. Damit lautet die Gleichung zur Berechnung von Schätzwerten mit dem Repeated Median:

$$\widehat{y}_{RM}(i) = a_{RM} + b_{RM} \cdot x_i \tag{5}$$

Zur Bestimmung des Anstieges b_{RM} werden aus einem Fenster, welches sich über k x-Werte erstreckt, alle möglichen Anstiege (siehe Abb. 2 und Abb. 3) zwischen allen k Werten berechnet.

Die Ergebnisse werden in eine Matrix eingetragen. Die Zellen in der Hauptdiagonalen sind nicht definiert oder Null, da bei gleichem Zeilen- und Spaltenindex wegen der Differenz Null im Nenner kein Anstieg berechnet werden kann. Im nächsten Schritt wird je Zeile der Median der $k-1$ Anstiege bestimmt und anschließend aus den k Zeilenmedianen der Spaltenmedian b_{RM}. Dieser Repeated Median b_{RM} wird als Anstieg der Regressionsgeraden verwendet.

[8] x_i-unabhängige und y_i-abhängige Messwerte

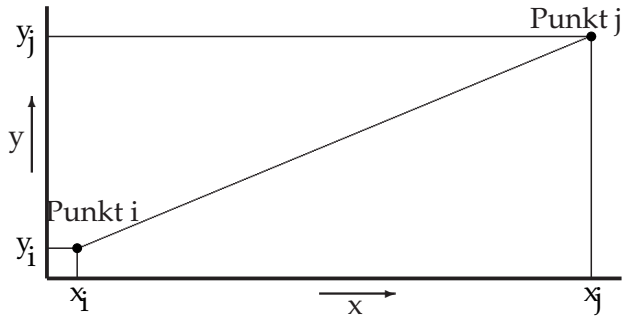

Abbildung 2: Geraden zwischen 2 Punkten

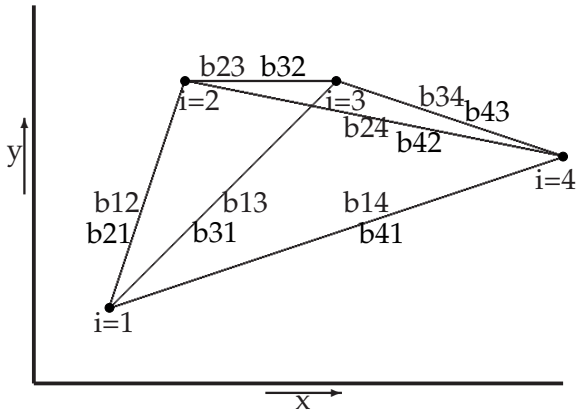

Abbildung 3: Anstiegsberechnung

Anstiegsmatrix						1. Median (Zeile)
X	b_{12}	b_{13}	b_{14}	b_{15}	->	Zeilenmedian 1
b_{21}	X	b_{23}	b_{24}	b_{25}	->	Zeilenmedian 2
b_{31}	b_{32}	X	b_{34}	b_{35}	->	Zeilenmedian 3
b_{41}	b_{42}	b_{43}	X	b_{45}	->	Zeilenmedian 4
b_{51}	b_{52}	b_{53}	b_{54}	X	->	Zeilenmedian 5
						V
2. Median (Spalte):						b_{RM}

Tabelle 1: Ablaufschema zur Bestimmung des Anstieges

Anschließend wird die Geradengleichung nach a_i umgestellt und für jeden der k Messwerte ein Offset $a_i = y_i - b_{RM} \cdot x_i$ ermittelt. Auch aus den k Offsetwerten a_i wird der Median bestimmt und als Offset a_{RM} für die Regressionsgeraden verwendet. Mit der resultierenden robusten Geradengleichung können je nach den Erfordernissen $1 \ldots k$ Schätzwerte für die k Messwerte gleichzeitig ermittelt werden.

Bei bekannter oder abzuschätzender Schrittweite Δx lassen sich anhand der Regressionsgeraden auch weitere, in der Zukunft liegende Werte schätzen. Es ist dann möglich, anhand bestimmter Kriterien zu entscheiden, ob ein neuer Messwert plausibel und zulässig oder als unplausibel einzustufen ist.

2.2 Beispielrechnung mit 10 Werten

Die einzelnen Rechenschritte sollen jetzt anhand eines praxisnahen Beispiels, welches in MATLAB programmiert wurde, anschaulich erläutert werden. Es wird die Zeitfunktion eines Gasvolumenstromes $\frac{\Delta V_{O2}}{\Delta t}$ betrachtet, die zunächst noch **keine großen Ausreißer** enthält. Der Beispieldatensatz besteht aus $k = 10$ Messwerten [9]. Die Filterlänge k schließt hier alle 10 Messwerte ein. Der Anschaulichkeit wegen, werden die Beispielwerte des Gasvolumenstromes verallgemeinert mit y_i und die Zeitwerte mit x_i bezeichnet $1 \leq i \leq 10$. Die Einheit $\frac{Liter}{min}$ wird im folgenden Abschnitt für eine bessere Übersichtlichkeit nicht mitgeführt.

i	x_i	y_i
1	4	0,968
2	6	1,202
3	8	0,805
4	10	0,743
5	14	0,387
6	17	0,475
7	20	0,421
8	24	0,463
9	27	0,52
10	31	0,109

Abbildung 4: Beispieldaten

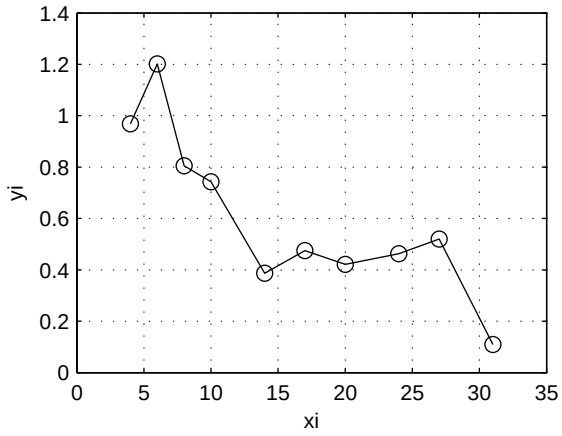

Abbildung 5: Beispielgraph

Die Berechnung der einzelnen Anstiege erfolgt nach folgender Formel (vgl. Abb. 2 und Abb. 3).

$$b_{ij} = \frac{y_j - y_i}{x_j - x_i} \tag{6}$$

Zuerst wird der Index j jeweils über dem Index i erhöht und die einzelnen Anstiege berechnet. Bei $i = j$ ist kein Anstieg berechenbar, wegen der resultierenden Division durch Null. Bei der Berechnung ergibt sich beispielsweise der Anstieg b_{ij} mit $i = 1$ und $j = 2$ zu

[9] In der Praxis sind die Daten digitalisierte Werte, die von einem A/D-Wandler zeit- und amplitudendiskret (wertdiskret) erfasst und in einem Mikroprozessor verarbeitet werden.

$b_{12} = \frac{y_j - y_i}{x_j - x_i} = \frac{1{,}202 - 0{,}968}{6-4} = \frac{0{,}234}{2} = 0,117\,.$

Einige weitere Beispiele für den Anstieg b_{ij} sollen hier noch aufgezählt werden.

$b_{11} = 0; b_{12} = 0.1170; b_{13} = -0.0407; \ldots b_{110} = -0.0318$

$b_{21} = 0; b_{22} = 0; b_{23} = -0.1985; \ldots b_{210} = -0.0437$ usw. bis $b_{910} = -0.1028$

Damit ist die obere Dreiecksmatrix zeilenweise mit den einzelnen Anstiegen gefüllt. Anschließend wird der Index i jeweils über dem Index j erhöht und die untere Dreiecksmatrix spaltenweise gefüllt. Die mit Nullen gefüllte Hauptdiagonale wird bei der folgenden

b_{ij}	$j =$	2	3	4	5	6	7	8	9	10
$i = 1$	0	0.1170	-0.0407	-0.0375	-0.0581	-0.0379	-0.0342	-0.0252	-0.0195	-0.0318
2	0	0	-0.1985	-0.1147	-0.1019	-0.0661	-0.0558	-0.0411	-0.0325	-0.0437
3	0	0	0	-0.0310	-0.0697	-0.0367	-0.0320	-0.0214	-0.0150	-0.0303
4	0	0	0	0	-0.0890	-0.0383	-0.0322	-0.0200	-0.0131	-0.0302
5	0	0	0	0	0	0.0293	0.0057	0.0076	0.0102	-0.0164
6	0	0	0	0	0	0	-0.0180	-0.0017	0.0045	-0.0261
7	0	0	0	0	0	0	0	0.0105	0.0141	-0.0284
8	0	0	0	0	0	0	0	0	0.0190	-0.0506
9	0	0	0	0	0	0	0	0	0	-0.1028
10	0	0	0	0	0	0	0	0	0	0

Tabelle 2: Obere Dreiecksmatrix mit Anstiegswerten

b_{ij}	$j = 1$	2	3	4	5	6	7	8	9	10
$i = 1$	0	0.1170	-0.0407	-0.0375	-0.0581	-0.0379	-0.0342	-0.0252	-0.0195	-0.0318
2	0.1170	0	-0.1985	-0.1147	-0.1019	-0.0661	-0.0558	-0.0411	-0.0325	-0.0437
3	-0.0407	-0.1985	0	-0.0310	-0.0697	-0.0367	-0.0320	-0.0214	-0.0150	-0.0303
4	-0.0375	-0.1147	-0.0310	0	-0.0890	-0.0383	-0.0322	-0.0200	-0.0131	-0.0302
5	-0.0581	-0.1019	-0.0697	-0.0890	0	0.0293	0.0057	0.0076	0.0102	-0.0164
6	-0.0379	-0.0661	-0.0367	-0.0383	0.0293	0	-0.0180	-0.0017	0.0045	-0.0261
7	-0.0342	-0.0558	-0.0320	-0.0322	0.0057	-0.0180	0	0.0105	0.0141	-0.0284
8	-0.0252	-0.0411	-0.0214	-0.0200	0.0076	-0.0017	0.0105	0	0.0190	-0.0506
9	-0.0195	-0.0325	-0.0150	-0.0131	0.0102	0.0045	0.0141	0.0190	0	-0.1028
10	-0.0318	-0.0437	-0.0303	-0.0302	-0.0164	-0.0261	-0.0284	-0.0506	-0.1028	0

Tabelle 3: Gesamte Matrix gefüllt mit Anstiegswerten

Auswertung zur Bestimmung des Zeilenmedians $med(b_i)$ ignoriert oder die gesamte obere Dreiecksmatrix zuvor um eine Stelle nach links verschoben, damit die Nullen in der

Matrix entfernt sind. Die k – te (rechte) Spalte muss danach noch entfernt werden, da deren Inhalte sonst doppelt vorkommen. Der Median würde damit eventuell nicht richtig bestimmt werden.

b_{ij}	j = 1	2	3	4	5	6	7	8	9
i =1	0.1170	-0.0407	-0.0375	-0.0581	-0.0379	-0.0342	-0.0252	-0.0195	-0.0318
2	0.1170	-0.1985	-0.1147	-0.1019	-0.0661	-0.0558	-0.0411	-0.0325	-0.0437
3	-0.0407	-0.1985	-0.0310	-0.0697	-0.0367	-0.0320	-0.0214	-0.0150	-0.0303
4	-0.0375	-0.1147	-0.0310	-0.0890	-0.0383	-0.0322	-0.0200	-0.0131	-0.0302
5	-0.0581	-0.1019	-0.0697	-0.0890	0.0293	0.0057	0.0076	0.0102	-0.0164
6	-0.0379	-0.0661	-0.0367	-0.0383	0.0293	-0.0180	-0.0017	0.0045	-0.0261
7	-0.0342	-0.0558	-0.0320	-0.0322	0.0057	-0.0180	0.0105	0.0141	-0.0284
8	-0.0252	-0.0411	-0.0214	-0.0200	0.0076	-0.0017	0.0105	0.0190	-0.0506
9	-0.0195	-0.0325	-0.0150	-0.0131	0.0102	0.0045	0.0141	0.0190	-0.1028
10	-0.0318	-0.0437	-0.0303	-0.0302	-0.0164	-0.0261	-0.0284	-0.0506	-0.1028

Tabelle 4: Nachbearbeitete Matrix gefüllt mit Anstiegswerten

Nun wird für jeder Zeile $i = 1 \ldots 10$ der Zeilenmedian $med(b_i)$ ermittelt. Die einzelnen Werte für $med(b_i)$ mit $i = 1 \ldots 10$ sind:

$-0.0342; -0.0558; -0.0320; -0.0322; -0.0164; -0.0261; -0.0284; -0.0200; -0.0131; -0.0303$

Aus diesen 10 Medianen $med(b_i)$ wird der Abstieg b_{RM} mit wiederholtem Median ermittelt.

$$b_{RM} = med(med(b_j));\ j = 1 \ldots 9 \tag{7}$$

Dieser Anstieg wird für die Regressionsgeraden verwendet. Im Beispiel ergibt sich $b_{RM} = -0.0284$. Anschließend wird ein Offset a_{RM_i} für jeden der k Messwerte mit folgender Formel berechnet:

$$a_{RM_i} = y_i - b_{RM} \cdot x_i \tag{8}$$

Aus diesen k Offsetwerten a_{RM_i} wird der Median $a_{RM} = med(a_{RM_i})$ ermittelt. Der Offset ergibt sich zu $a_{RM} = 1.0378$. Damit lautet die Gleichung der robusten Regressionsgeraden für die $k = 10$ Beispielwerte

$$\mathbf{\hat{y}_{RM} = a_{RM} + b_{RM} \cdot x = 1{,}0378 - 0{,}0284 \cdot x}$$

Die Widerstandsfähigkeit (engl. resistance) der Regression gegenüber Ausreißern nimmt mit steigender Filterbreite k zu (vergleiche Kap. 2.6). Nachteilig wirkt sich aber aus, dass mit steigendem k schnelle Änderungen im Signal zunehmend verschliffen werden. Weitere Eigenschaften werden am Ende des Abschnittes 2.6 diskutiert.

2.3 Beispiel mit 1 Ausreißer

Im folgenden **neuen** Beispiel soll die Widerstandsfähigkeit der robusten Regression anhand eines Datensatzes mit $k = 10$ Messwerten getestet werden, der einen extremen Ausreißer enthält. Die robuste Regressions- und die Geraden mit der minimalen Summe der Fehlerquadrate werden verglichen. Es ist gut zu erkennen, dass trotz des extremen Ausreißers der Trend der Messreihe mit der robusten Regressionsgeraden sicher erkannt wird. Mit der Methode der kleinsten Fehlerquadrate wird der Trend fehlerhaft, stark vom Ausreißer beeinflusst ermittelt.

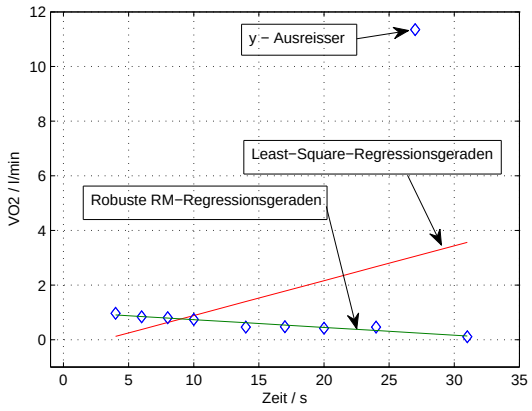

Abbildung 6: Repeated-Median-Regression

$\frac{Zeit}{s}$	Messwert	Schätzwert
x_i	y_i	$\widehat{y}_{RMi}$
4	0.9680	0.9067
6	0.8400	0.8491
8	0.8050	0.7915
10	0.7430	0.7339
14	0.4600	0.6187
17	0.4750	0.5323
20	0.4210	0.4459
24	0.4630	0.3307
27	**11.350**	**0.2443**
31	0.1090	0.1291

Abbildung 7: Mess- und Schätzwerte

2.4 Erzeugung eines Testsignals

Der Vorteil eines synthetischen Testsignals gegenüber realen Messdaten ist, dass es aus verschiedenen bekannten Komponenten zusammensetzt ist. So sind die überlagerten Störungen (Rauschen und Ausreißer) und der Verlauf des eigentlichen Nutzsignals genau quantifiziert. Damit lassen sich Filtereigenschaften anhand von Kennwerten objektiv bewerten und verschiedene Filter können miteinander vergleichen werden. Das Testsignal soll einem realen Sauerstoffverbrauchssignal in $\frac{Liter}{min}$ entsprechen, welches mit einem Ergospirometer an einem Probanden gemessen wurde.
In einer zwischen 6 und 15 Minuten dauernden Messung wird die Belastung eines Probanden (die zu erbringende Leistung) sprungförmig erhöht und jeweils für eine gewisse Zeit konstant gehalten. Während der Messung werden kardiovaskuläre Parameter wie EKG, Blutdruck und Puls (ergometrische) sowie pulmonale Parameter wie Vitalkapazität, Atemvolumenstrom, Sauerstoff- und Kohlendioxidkonzentration (spirometrische) ermittelt. Zum Ende der Messung wird die Belastung abgestellt und der Patient kann sich erholen, während er sich noch ein bis zwei Minuten unter minimaler Belastung bewegt. Danach ist die Messung beendet. Es werden die Zeit und die anderen Größen simultan gemessen. Die Zeitabstände sind bei den Messwerten nicht äquidistant. Mit jedem Atemzug, d.h. etwa alle 2 bis 3 Sekunden wird ein Messwert aufgenommen und ausgewertet (Breath-by-Breath-Methode). Für das nachzubildende physiologische Testsignal wurde mit guter Näherung ermittelt, dass es der sprunghaften Erhöhung der Belastung mit einen Tiefpass 1. Ordnung gedämpft folgt. Die Zeitkonstante τ ist jedoch nicht konstant, sondern sie vergrößert sich abhängig von der Leistungsfähigkeit des Probanden mit jedem Leistungssprung um ca. 10 Sekunden. ($\frac{\tau}{s} = 4, 14, 24, 34, 44, 54, 64, 74$)
Das simulierte Nutzsignal besteht aus einer Funktion mit mehreren gedämpften Sprüngen und idealisiert aus zeitlich äquidistanten Messwerten. Als Störungen werden dem Signal ein normal verteiltes Rauschen (d.h. $Mittelwert = 0\ \frac{Liter}{min}$) von $\sigma = 0,3\ \frac{Liter}{min}$ in der Zeit von $t = 0s \ldots 570s$, dann ein kleineres Rauschen von $\sigma = 0,05\ \frac{Liter}{min}$ bei $t = 570s \ldots 1200s$ und schließlich wieder das Rauschen mit $\sigma = 0,3\ \frac{Liter}{min}$ bei $t = 1200s \ldots 2048s$. Die Ausreißer in positiver und negativer Y-Richtung werden dem physiologischen Testsignal im Bereich von $t = 570s \ldots 1200s$ überlagert.

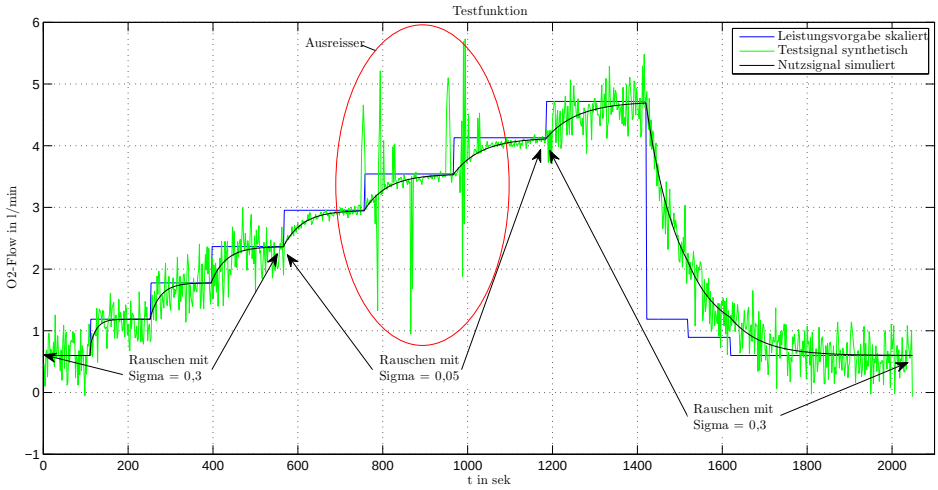

Abbildung 8: Testsignal

2.5 RM-Filterung eines Datensatzes

Wenn das Repeated Median-Filter für Messreihen [10] mit $i = 1 \ldots N$ und $N > k$ Messwerten [11] angewendet werden soll, muß das Filter (als Zeitfenster) mit der Breite k in bestimmter Schrittweite über die Messreihe verschoben und für jede Position eine lokale lineare robuste Regressionsgeraden berechnet werden (vergleiche Bild 9). Abhängig von der gewählten Schrittweite besteht die Möglichkeit, je Position nur einen oder gleichzeitig mehrere Schätzwerte zu berechnen.

Es wird in dieser Anwendung die Schrittweite 1 betrachtet und entsprechend nur je ein Schätzwert pro Zeitfensterposition berechnet. Welcher der k möglichen Schätzwerte berechnet werden soll, kann im Prinzip frei gewählt werden. Die Wahl der Position des berechneten Schätzwertes hat aber einen Einfluss auf die Verzögerungszeit zwischen Mess- und Berechnungszeitpunkt des Wertes sowie auf die Eigenschaften des Filters (Filtercharakteristik). Hier werden der aktuellste Schätzwert[12] $(i+k-1)$ und der mittlere Schätzwert[13] $(i+\frac{k}{2})$ verglichen. Je Filterposition wird der gewählte zu berechnende Schätzwert in einer Matrix in eine Spalte neben der Messwertspalte eingetragen. Für die ersten Schätzwerte müssen solange die Rohmessdaten verwendet werden, bis die erste Regressions-

[10] hier z.B. das verrauschte und ausreißerbehaftete Testsignal
[11] k - Breite des Repeated-Median-Filters
[12] der jüngste, aktuellste der $k = 10$ möglichen Schätzwerte
[13] Der mittlere der k=10 möglichen Schätzwerte; wenn k ungerade gewählt ist, wird $\frac{k}{2}$ aufgerundet

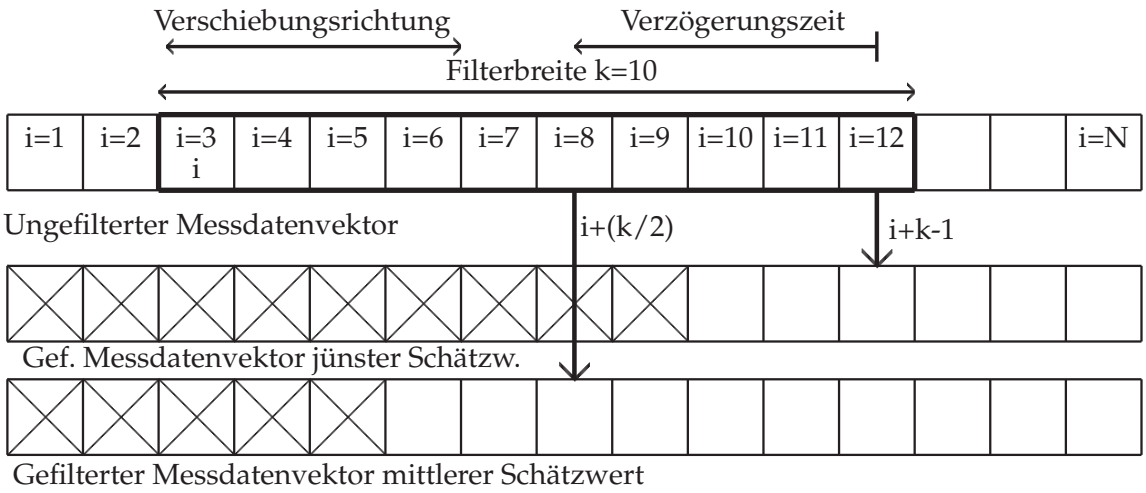

Abbildung 9: Verschiebung des Filters über einen Datensatz

geraden verfügbar ist. Danach können rückwirkend die Schätzwerte robust berechnet werden. Nach der ersten Regression wird i um 1 auf $i = 2$ erhöht und im nächsten Schritt die Messwerte $2 \ldots 11$ betrachtet. Die Werte $i + k - 1 = 11$ und $i + \frac{k}{2} = 7$ werden berechnet. Im Ergebnis (Diagramm 10) mit dem berechneten letzten Schätzwert $(i + k - 1)$ lässt sich die Robustheit des Filters schon gut erkennen, obwohl dieser Fall noch relativ stark von mehreren aufeinander folgenden großen Ausreißern beeinflusst wird.

In einem weiteren Versuch wurde anstelle des letzten jeweils der mittlere Schätzwert des Zeitfensters $(i + \frac{k}{2})$ errechnet und in die Matrix eingetragen. Der Signalverlauf ist wesentlich stärker gedämpft. Das Filter erzeugt eine besser geglättete Kurve und wird noch weniger von Ausreißern beeinflusst. Allerdings ist in diesem Fall auch die Empfindlichkeit (Dynamik) bei Sprüngen im Signal stärker gedämpft.

In Abbildung 13 soll der Verlauf der zwei RM-Filterkurven für Filterlängen $k = 10$ und $k = 25$ mit mittig berechnetem Schätzwert verglichen werden. Als Referenz für den Signalverlauf dient das physiologische Testsignal.

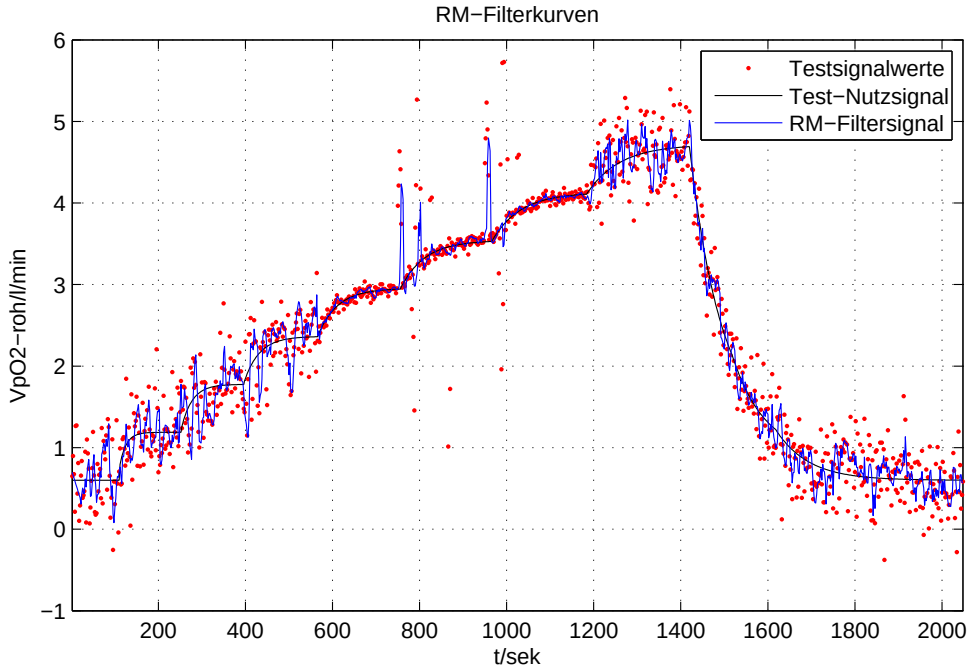

Abbildung 10: Anwendung der RM-Filterung mit k=10 letzter W.

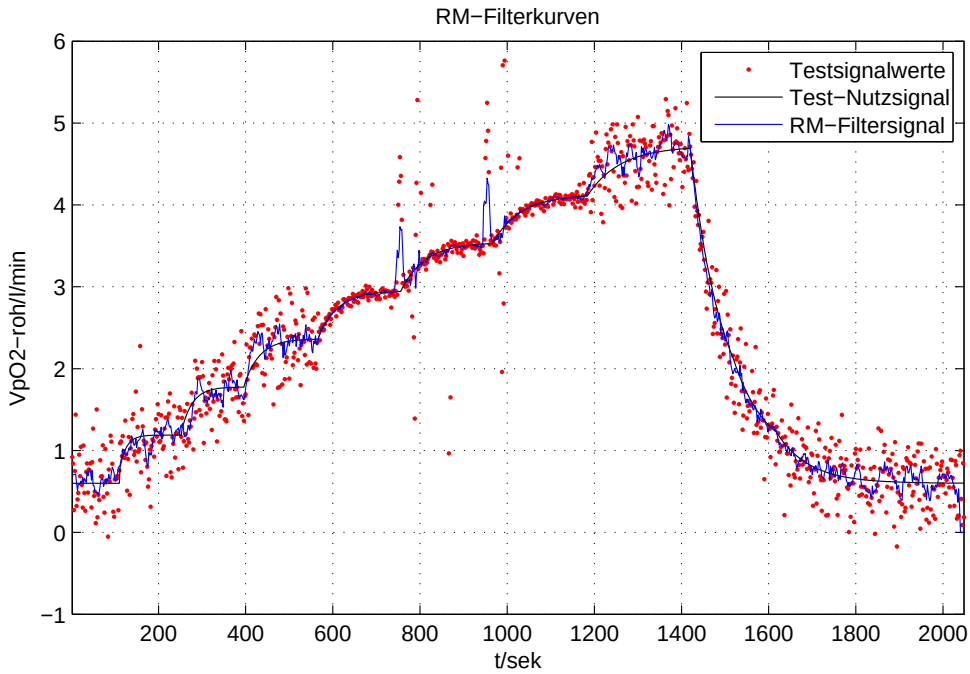

Abbildung 11: Anwendung der RM-Filterung mit k=10 mittl. W.

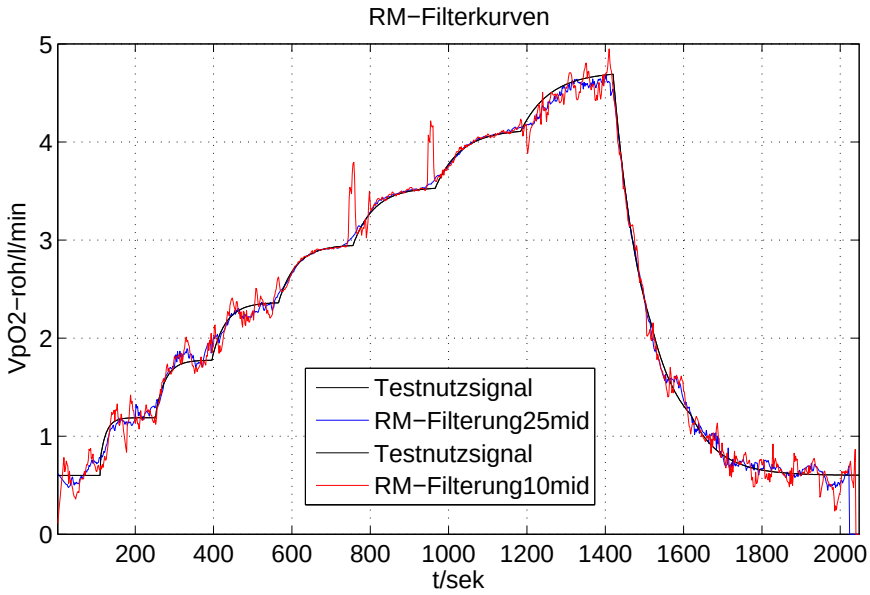

Abbildung 12: Anwendung der RM-Filterung mit k=10 mittl. W.

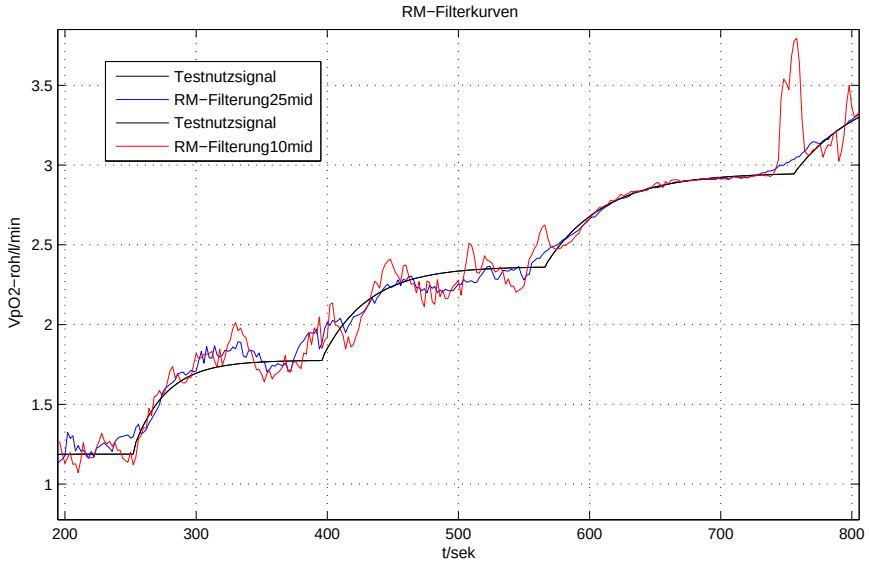

Abbildung 13: Vergrößerung Vergleich RM-Filterung mit k=10, 25 mittl. W.

2.6 Objektive Filterbewertung

Als Kenngröße zur objektiven Bewertung der Filtereffizienz und zum Vergleich verschiedener Filter soll die Fehlerleistung P_F eingeführt werden. Als Fehlerleistung wird der Effektivwert des Fehlersignals bezeichnet, das als Differenz aus dem gefilterten Signal und dem physiologischen Testsignal entsteht. Der quadratische Fehler ist in jedem Messzeitpunkt die Differenz von Schätzwert und originalem Messwert. Die Formel zur Berechnung der Fehlerleistung wird abgeleitet von der allgemeinen Formel 9 zur Bestimmung des Effektivwertes eines Signals.

$$s_{eff} = \sqrt{\frac{1}{T} \cdot \int_0^T s^2(t) \cdot dt} \tag{9}$$

Berechnung des Effektivwertes eines Signals

s – Periodisches zeitkontinuierliches Signal
T – Periodendauer des Signals
s_{eff} – Effektivwert des Signals

Die aus Formel 9 abgeleitete Formel 10 zur Fehlerleistungsberechnung ist für wert- und zeitdiskrete Messwerte geeignet, bei denen die Zeitabstände nicht äquidistant sind.

$$P_F = \sqrt{\frac{1}{t_{N-1} - t_1} \cdot \sum_{i=1}^{N-1} (\widehat{y}_i - y_i)^2 \cdot (t_i - t_{i-1})} \tag{10}$$

Berechnung Effektivwert des Fehlersignals

P_F – Fehlerleistung
$(\widehat{y}_i - y_i)^2$ – Quadratischer Fehler aus Schätzwert−Originalwert
$(t_i - t_{i-1})$ – Zeitdifferenz zwischen zwei Werten
N – Anzahl der Messwerte
$t_N - t_1$ – Betrachteter Zeitraum, vergleichbar der Periodendauer

Wenn die Zeitabstände äquidistant sind, vereinfacht sich die Berechnung wie in Formel 11 dargestellt.

$$P_F = \sqrt{\frac{1}{N-1} \cdot \sum_{i=1}^{N-1} (\widehat{y}_i - y_i)^2} \tag{11}$$

Berechnung Effektivwert des Fehlersignals (äquidistante Messwerte)

Im Ergebnis ist erkennbar, dass sich ein Fehlerminimum[14] bei einer Filterlänge von 25 Messwerten einstellt. Es ist auch deutlich ersichtlich, dass es vorteilhaft ist, den Schätzwert jeweils in der Mitte des Filters zu berechnen. Die zeitliche Verzögerung von einer halben Filterlänge muss dabei in Kauf genommen werden. Bei einer Filterlänge oberhalb 25 steigt der Fehler wieder an, durch das zunehmend integrierende Verhalten des RM-Filters.

Filterlänge	Fehlerleistung	
	$P_{F\ last}$	$P_{F\ mid}$
5	0.318	0.2316
10	0.2160	0.1318
15	0.1772	0.1102
20	0.1434	0.0970
25	0.1377	0.1005
30	0.1060	0.0943
35	0.1134	0.1060
40	0.1070	0.1111
45	0.1083	0.1179
50	0.1151	0.1225
70	0.1619	0.1474

Abbildung 14: Fehlerleistung

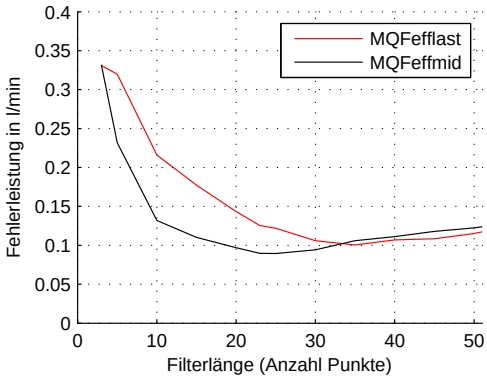

Abbildung 15: Verlauf Fehlerleistung

2.7 Spektralanalysen mit DFT

Das Testsignal (siehe Kap. 2.4) soll nun mit Hilfe der **D**iskreten-**F**ourier-**T**ransformation (DFT) auf die im Signal enthaltenen Frequenzanteile untersucht werden. Dazu wird im ersten Schritt das Spektrum des physiologischen Testsignals ohne Rauschen und ohne Ausreißer betrachtet. Im zweiten Schritt wird das Testsignal mit Ausreißern überlagert und dessen Spektrum mit dem des physiologischen Testsignals verglichen. Im dritten Schritt soll zusätzlich das Rauschen überlagert und wiederum das Spektrum betrachtet werden. Damit das durch DFT berechnete Spektrum keinen Gleichanteil enthält, wird zu-

[14] gemessen mit dem oben erleuterten physiologischen Testsignal

vor vom physiologischen Testsignal der Mittelwert subtrahiert. Die Periodendauer soll auf die Zeit zwischen dem ersten und letzten Wert festgelegt werden, damit das Signal als scheinbar periodisches Signal ausgewertet werden kann. Wenn es zwischen dem letzten und dem nächsten ersten Wert einen großen Sprung gibt, hat das ausgewertete Signal eine andere Form und es wird ein fehlerhaftes Spektrum berechnet. Beim physiologischen Testsignal beträgt diese Differenz nur 2 $\frac{ml}{min}$ was bei einer Amplitude von $\frac{4,5}{2}$ $\frac{l}{min}$ ein minimaler Fehler ist.
Im Ergebnis (Abb. 17) ist deutlich zu erkennen, dass das Nutzsignal eine Grundfrequenz von $f_0 = 0,488\ mHz$ hat, was einer Periodendauer von $T_0 = 2048\ s$ entspricht und als höchste noch relevante im physiologischen Nutzsignal auftretende Frequenz kann $f_{max} = 0,01\ Hz$ definiert werden. Danach kann die Grenzfrequenz eines Tiefpassfilters eingestellt werden. Die Grundfrequenz f_0 ist in dieser Anwendung nicht von großem Interesse, da sie nur von Länge und Verlauf der Messung bestimmt wird. Die physiologisch interessanteren Informationen enthalten die Frequenzanteile, die durch die Leistungssprünge verursacht werden. Diese Leistungssprünge müssen möglichst originalgetreu reproduziert werden.

In Abbildung 19 ist das Spektrum des mit Ausreißern überlagerten Testsignals dargestellt. Es ist ersichtlich, dass es im gesamten betrachteten Frequenzbereich eine Überlagerung des physilogischen Testsignalanteils durch die Ausreißer gibt. Die Anteile oberhalb f_{max} sind unproblematisch, da man diese mit einem Tiefpassfilter wirkungsvoll eliminieren kann. Problematisch sind alle Frequenzanteile, die sich mit dem Nutzsignalanteilen überlappen. Dieses Problem könnte vor einer Tiefpassfilterung durch eine robuste Regression minimiert werden.

Im dritten Schritt, den Abbildungen 20 und 21 ist der Signalverlauf und das Spektrum des vollständigen Referenztestsignals dargestellt. Das Nutzsignal ist vollkommen vom Rauschen überlagert und díe ursprüngliche Nutzinformation im Spektrum nicht mehr erkennbar.

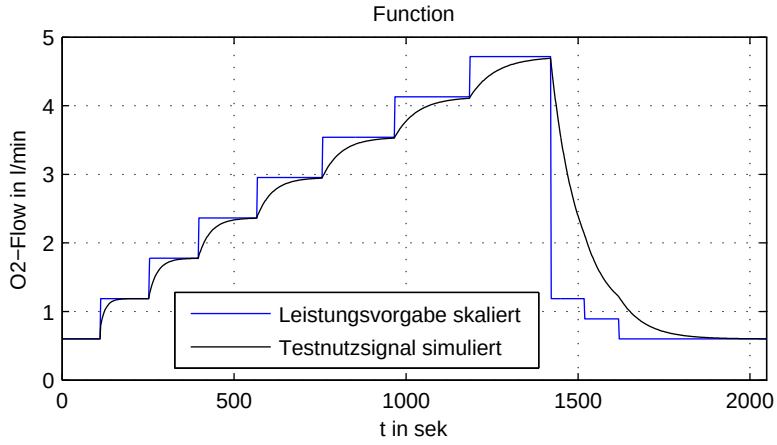

Abbildung 16: Signalverlauf des Testnutzsignals

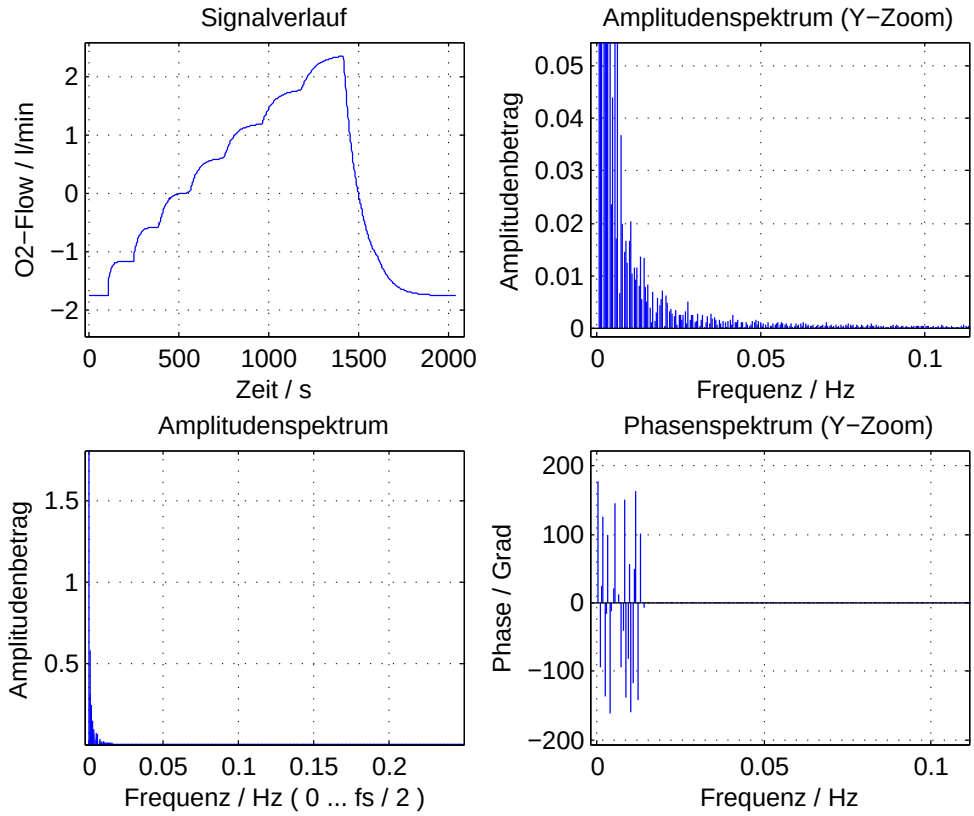

Abbildung 17: Amplitudenspektrum des Testnutzsignals

Wie bereits in Abbildung 10 erkennbar war, bereitet starkes Rauschen im Signal dem RM–Filter die größten Schwierigkeiten. Im Bereich des schwachen Rauschens mit den überlagerten Ausreißern kann das Nutzsignal trotz der großen Ausreißer sehr gut repro-

duziert werden.

Fazit: Es müssen neben der nachträglichen Signalfilterung mit einem Softwarefilter möglichst technische Maßnahmen entwickelt und umgesetzt werden, damit das Signal–Rausch–Verhältnis schon bei der Messung verbessert wird. Bei sehr schwachem Rauschen und einigen Ausreißern kann die Nutzinformation noch sicher aus den Messdaten extrahiert werden.

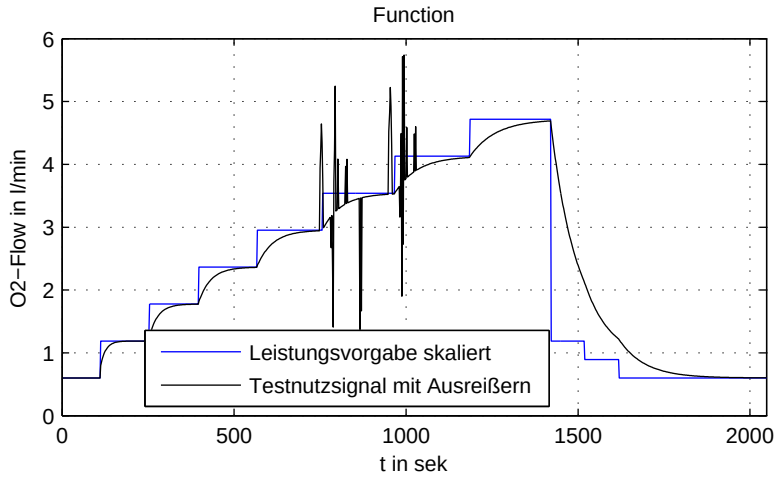

Abbildung 18: Signalverlauf des Testnutzsignals mit Ausreißern

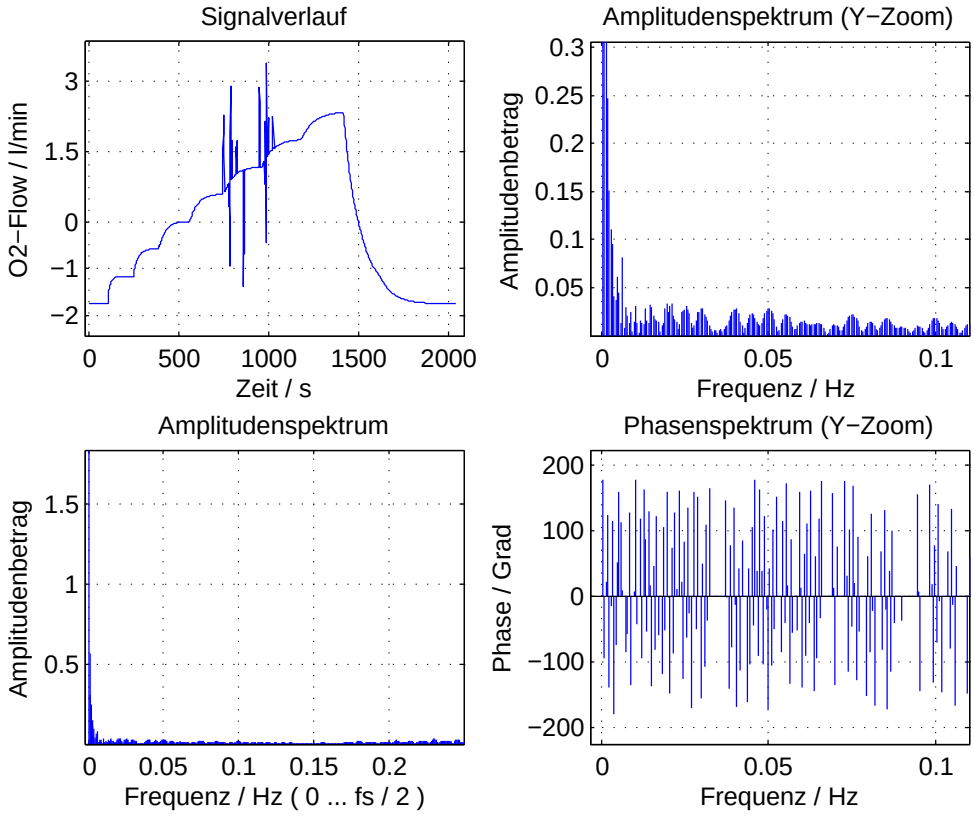

Abbildung 19: Amplitudenspektrum des gestörten Testnutzsignals

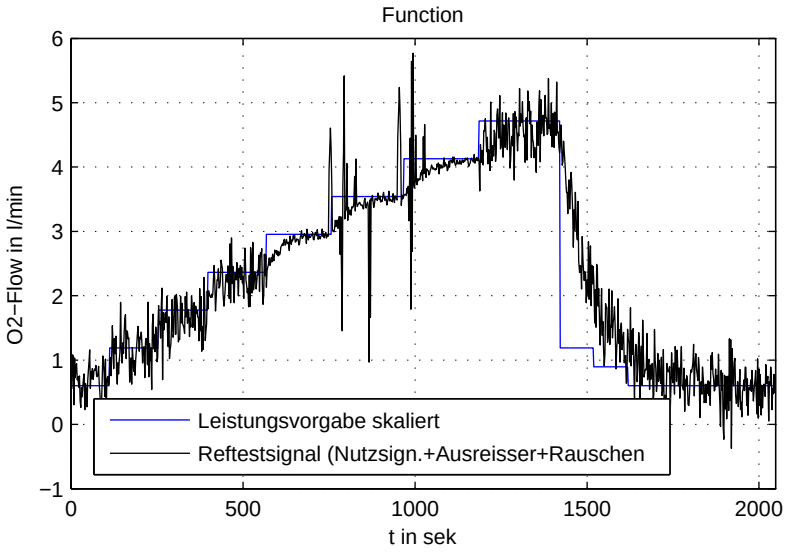

Abbildung 20: Signalverlauf des Reftestsignals

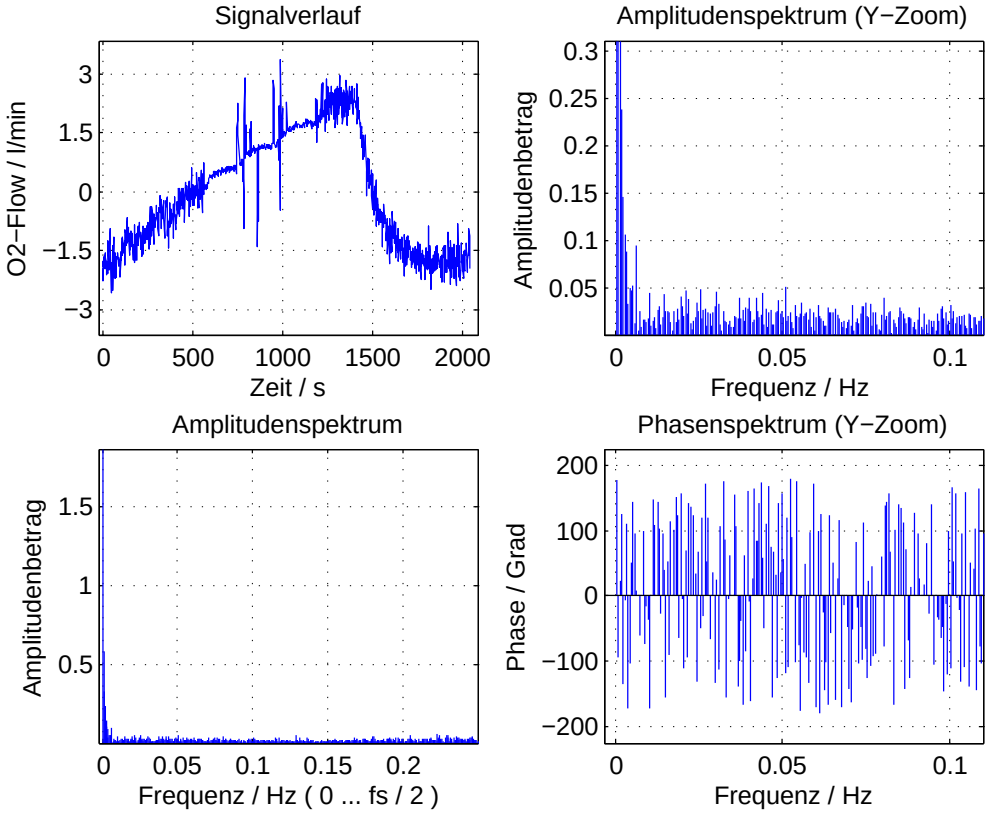

Abbildung 21: Amplitudenspektrum des Reftestsignals

2.8 Hybrides RM-Filter

Nachdem das Spektrum des verrauschten und ausreißerbehafteten Signales bekannt ist, kann das RM-Filter zu einem hybriden RM-Filter weiterentwickelt werden. Für das hybride RM-Filter hat sich die **Filterlänge 16** als optimal erwiesen, d.h. durch einen Tiefpass kann eine kleinere Filterlänge für das RM-Filter gewählt werden, um gleiche bzw. sogar bessere Ergebnisse wie mit dem einfachen RM-Filter zu erreichen. Dem RM-Filter ist ein digitales IIR-Tiefpassfilter mit folgenden Eigenschaften nachgeschaltet worden.

- $IIR - Filter$; (**I**nfinite **I**mpulse **R**esponse, Vorteil: weniger Filterkoeffizienten)
- $butter(TPn, TPfg/fs,'low')$; (Filter mit Butterworth-Charakteristik)
- $TPn = 1$; (Ordnung n des Tiefpassfilters)
- $TPfg = 0.01$; (Grenzfrequenz fg des TP-Filters)
- $fs = 0.25$; (Samplefrequenz der Messdaten)

Es genügt bereits ein stabil abgestimmtes IIR-Tiefpassfilter erster Ordnung, um das Ergebnis Abbildung **??** zu erreichen. Zur allgemeinen Information sollen die Eigenschaften eines Butterworth-Filters kurz aufgezählt werden:

- Linearer Frequenzgang unterhalb der Grenzfrequenz, d.h. maximal flachen Frequenzgang im Durchlassbereich
- Schnelles Abknicken bei Grenzfrequenz, steigend mit der Filterordnung n
- Starkes Überschwingen bei der Sprungantwort, steigert sich mit höherer Ordnung n
- Dämpfung von $n \cdot (-6dB)/Oktave$ (Frequenzverdopplung) bzw. $n \cdot (-20dB)/Dekade$ (Zehnerpotenz)

In Abbildung 22 ist das Ergebnis der Filterung mit hybridem RM-Filter dargestellt. Die Fehlerleistung des besten Ergebnisses beträgt $Fehlerleistung = 0,0702 \cdot \frac{l}{min}$. Verglichen mit der Fehlerleistung der ungefilterten Messwerte von $Fehlerleistung = 0.3403 \cdot \frac{l}{min}$ ergibt sich eine Verbesserung um einen Faktor von $\frac{0.3403 \cdot l/min}{0.0702 \cdot l/min} = \mathbf{4,85}$ oder anders betrachtet

eine $1 - \frac{0.0702 \cdot l/min}{0.3403 \cdot l/min} \cdot 100\% = \mathbf{79{,}4\%}$ genauere Extraktion der Nutzinformation aus den Messwerten verglichen mit den ungefilterten Messwerten.

Der Vergleich mit den Standardfiltern ergibt folgende Resultate:

- **Mittelwert:** $Fehlerleistung = 0.1197 \cdot \frac{l}{min}$ und den Faktor $\frac{0.1197 \cdot l/min}{0.0702 \cdot l/min} = 1,7$ bzw. die Verbesserung der Extraktion von $1 - \frac{0.1197 \cdot l/min}{0.3403 \cdot l/min} \cdot 100\% = 64,8\%$
- **5aus7 Mittelwert:** $Fehlerleistung = 0.1595 \cdot \frac{l}{min}$ und den Faktor $\frac{0.1595 \cdot l/min}{0.0702 \cdot l/min} = 2,3$ bzw. die Verbesserung der Extraktion von $1 - \frac{0.1595 \cdot l/min}{0.3403 \cdot l/min} \cdot 100\% = 53,1\%$
- **Median aus 3 Werten:** $Fehlerleistung = 0.2907 \cdot \frac{l}{min}$ und den Faktor $\frac{0.2907 \cdot l/min}{0.0702 \cdot l/min} = 4,14$ bzw. die Verbesserung der Extraktion von $1 - \frac{0.2907 \cdot l/min}{0.3403 \cdot l/min} \cdot 100\% = 14,5\%$

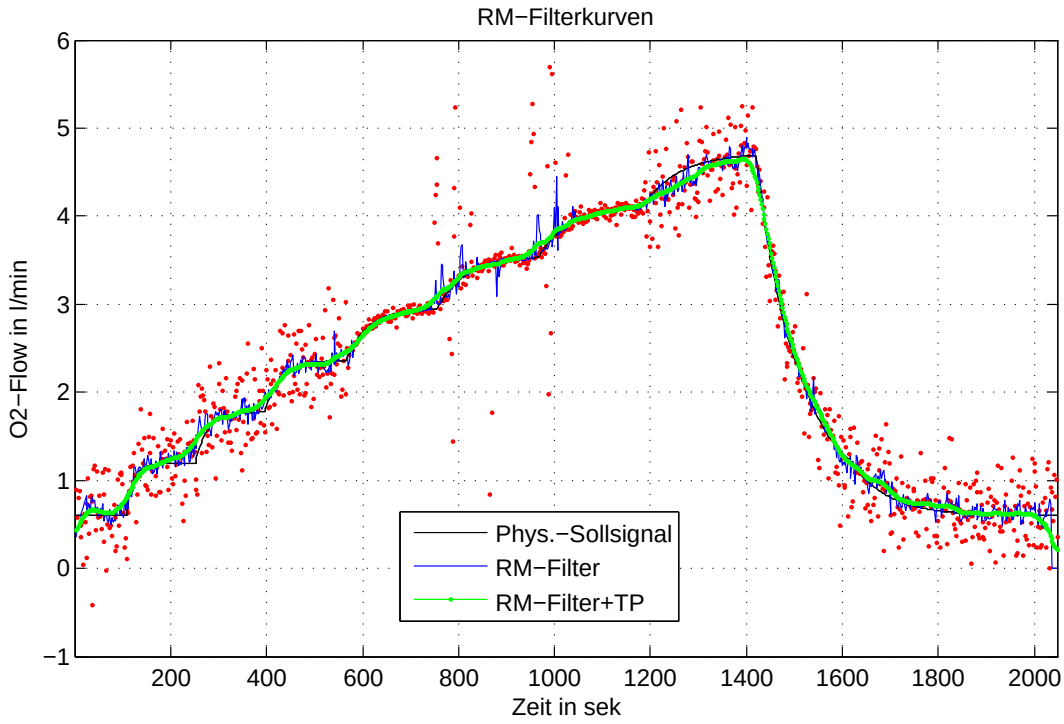

Abbildung 22: Ergebnis Hybride-RM-Filterung k=16

2.9 Online-RM-Filter

Bei der robusten Online-Filterung sind besondere Randbedingungen zu beachten. Die gemessenen Daten müssen direkt nach der Datenerfassung verarbeitet und möglichgst verzögerungsfrei auf einer Anzeige dargestellt werden. Das Filter muss diskrete, zeitlich nicht äquidistante Messwerte verarbeiten können. Basierend auf zurückliegenden Messwerten ist zu entscheiden, ob der neueste Messwert plausibel ist und weiterverarbeitet werden kann oder ob der Messwert ein Ausreißer ist, der durch einen geeigneten Schätzwert ersetzt werden soll. Für den Online-Modus wurde ein Repeated Median Filter entwickelt, bei dem der Anstieg der Fehlersumme[15] als Kriterium dient, um direkt nach dem Einlesen des Messwertes zu entscheiden, ob die aktuelle Regressionsgeraden weiterverwendet werden kann oder ob wegen einer Trendänderung in den Messwerten eine neue berechnet werden muss.

Die Idee ist, dass bei optimal berechneter Regressionsgeraden und gleichverteiltem überlagertem Rauschen der kummulierte Fehler (im Mittel) zu Null wird. Hier kann der Trend beibehalten werden. Wenn die letzte Regressionsgeraden durch Trendänderung nicht mehr passend ist, wächst der kummulierte Fehler in positiver oder negativer Richtung. Der Anstieg der Betragsfunktion des kummulierten Fehlers kann ausgewertet werden. Bei (mehreren) Ausreißern in einer Richtung ist der Fehlerzuwachs wesentlich größer und man kann dies gut im Verlauf der Betragsfunktion des Summenfehlers erkennen und auswerten. In so einem Fall soll auch keine neue Regressionsgeraden berechnet werden.

Es muss ein Kompromiss gefunden werden, bei dem mehrere aufeinander folgende Austreißer beseitigt werden können und trotzdem ein neuer zugrundeliegender Tend erkannt wird.

Das Filter optimal abzustimmen ist eine anspruchsvolle Aufgabe. Zum Beispiel musste bei ungünstiger Wahl der Filterkennwerte und bei bestimmten Sprüngen im Messsignal festgestellt werden, dass das Filter nur noch Schätzwerte eines falschen Trends ausgab und dem eigentlichen Trend der Messreihe nicht mehr folgte. Das Ausschließen dieses Fehlverhaltens ist möglich, aber bei der Rekonstruktion des Nutzsignals muss ein größerer Fehler in Kauf genommen werden. Als Maßnahme wurde hier eine zusätzliche Überwachung

[15] Summe der Fehler aus neustem Messwert minus zugehörigem Schätzwert

implementiert, die vermeiden soll, dass sich das Filter "festfährt". Im Folgenden Abschnitt soll die Realisierung des Online-RM-Filters näher erläutert werden. (siehe hierzu Blockschaltbild 23)

2.10 Beschreibung des Online-RM-Filters

Im ersten Schritt müssen alle notwendigen Parameter, Grenzwerte und Variablen definiert und mit einem Startwert gesetzt werden. Die Startwerte sollen den Variablen nur einen definierten Wert geben. Der erste gemessene Wert wird direkt abgespeichert. In der Ausgangssituation, vor Beginn der Messung, hat das Filter eine festgelegte Startfilterlänge von z.B. 3...5 Werten. Die ersten Messwerte müssen direkt angezeigt werden, da keine Regressionsgeraden berechnet werden kann, solange die Anzahl der Messwerte kleiner als die Startfilterlänge ist. Sobald genügend Messwerte [16] erfasst worden sind, wird die erste robuste Regressionsgeraden berechnet. Danach wird der mittlere Zeitabstand zwischen den letzten [17] Messwerten ermittelt. Es wird ein Mittelwert errechnet, da die Messpunkte zeitlich nicht äquidistant verteilt sein können. Mit der Regressionsgeraden und dem mittleren Zeitabstand wird im nächsten Schritt der zukünftige Schätzwert $y(n+1)$ berechnet. Danach wird der Fehler F, also die Differenz von Messwert und Schätzwert, errechnet. Dieser Fehler wird zur bereits gebildeten Summe der zurückliegenden Fehler (Fehlersumme FS) addiert und gespeichert. Dann wird über z.B. 6 zurückliegende Werte[18] der Anstieg der Fehlersumme berechnet.

Jetzt wird geprüft, ob der Fehlersummenanstieg innerhalb eines vorher definierten Wertebereiches zwischen dem oberen und dem unteren Schwellwert liegt. Wenn der Anstieg zwischen den Schwellwerten liegt, soll im nächsten Programmzyklus eine neue Regressionsgeraden berechnet werden, da sich wahrscheinlich ein neuer Trend in den Messwerten ergeben hat. Dazu wird das Neuberechnungsbit gleich Eins gesetzt. Wenn der Fehlersummenanstieg kleiner als der untere Schwellwert ist, kann die aktuelle Regressionsgeraden weiter verwendet werden, da der Fehlerzuwachs klein und der aktuelle Trend noch richtig berechnet ist. Das Neuberechnungsbit wird Null gesetzt. Wenn der Fehlersummenanstieg größer als der obere Schwellwert ist, liegt mit hoher Wahrscheinlichkeit ein Ausreißer vor,

[16] Anzahl gleich der Startfilterlänge
[17] Es wird jeweils über der Filterlänge der Mittelwert bestimmt.
[18] Die Anzahl der Werte kann variiert und optimiert werden.

der nicht angezeigt werden soll. Das *Neuberechnungsbit* wird Null gesetzt und keine neue Regressionsgeraden berechnet.

Im folgenden Schritt ist eine Sicherung implementiert, um zu vermeiden, dass sich das Filter "festfährt". Bei Wahl von Filterlängen von 25 bis 30 Werten besteht ein Risiko, dass bei großen Signalsprüngen (in Größenordnung von Ausreißern) der Grenzwert OGW dauerhaft überschritten wird und das Filter keine neue Regressionsgeraden mehr berechnet, sondern permanent ungeeignete Schätzwerte ausgibt. Um dies zu vermeiden, wird geprüft, ob sich der Betrag der Fehlersumme und der Betrag des Produktes aus $Maxfehlerfaktor$ und aktuellem Schätzwert gleichen. Dieser Fehlerfall kann lediglich bei sehr kleinen Messwerten eintreten oder wenn das Filter vollkommen falsche Schätzwerte berechnet, die weit von den Messwerten entfernt liegen. Der $Maxfehlerfaktor$ sollte in der hier untersuchten Anwendung zwischen 0,1 und 1,0 gewählt werden, damit ein evt. aufgetretener Fehlerfall schnell abgestellt wird. Der Grenzwert könnte aber auch größer gewählt werden, wenn das Filter gut abgestimmt ist. Zum Abstellen des Problems wird die Fehlersumme auf Null gesetzt und das *Neuberechnungsbit* auf Eins, damit im nächsten Zyklus einen neue Regressionsgeraden berechnet wird.
Im vorletzten Schritt wird die *Filterlaenge* je Schleifendurchlauf um Eins erhöht, bis die $Endfilterlaenge$ erreicht ist. Im letzten Schritt wird n um Eins erhöht und zu Beginn des nächsten Durchlaufes der neue Messwert $(n+1)$ eingelesen.
Nachdem die $Endfilterlnge$ erreicht ist, wird nicht automatisch je Schleifenzyklus eine neue Regressionsgeraden berechnet, sondern anhand des Fehlersummenanstieges entschieden, ob es nötig ist oder nicht.

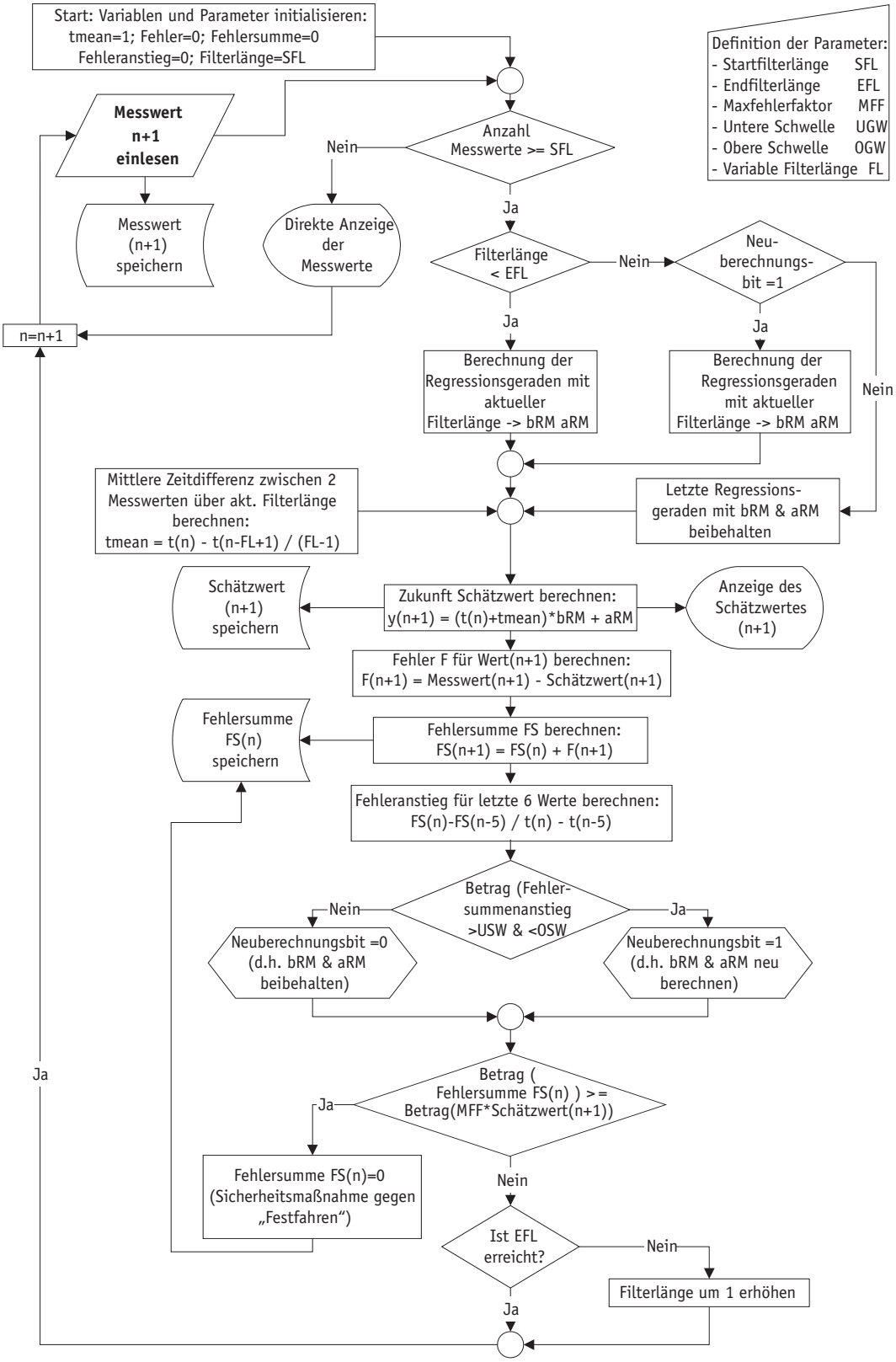

Abbildung 23: Blockschaltbild Online-Filter

2.11 Ergebnisse mit Online-RM-Filter

Das objektiv beste Ergebnis mit der $Fehlerleistung = 0.1008 l/min$ wird erreicht, wenn die Filterparameter folgendermaßen gewählt werden (siehe Blockschaltbild 24):

- Startfilterlaenge = 3; (minimale Filterlänge bei Start der Online-Filterung)
- Maxfilterlaenge=32; (Endfilterlänge, Filter wird robuster, wenn länger)
- Maxfehlerfaktor=1; (notwendig, um Summenfehler zurückzusetzen im Fehlerfall)
- USW=0.007; (unterer Schwellwert, aktuelle Regressionsgeraden passt gut)
- OSW=0.3; (Oberer Schwellwert, ein Ausreißer liegt wahrscheinl. vor)

Die ungefilterten Messwerte haben eine $Fehlerleistung = 0.3403 l/min$. Damit ergibt sich eine Optimierung um einen Faktor $\frac{0.3403 l/min}{0.1008 l/min} = 3,4$ gegenüber den ungefilterten Messwerten, die bisher bei Echtzeitanzeige im Online-Modus ausgegeben werden mussten.

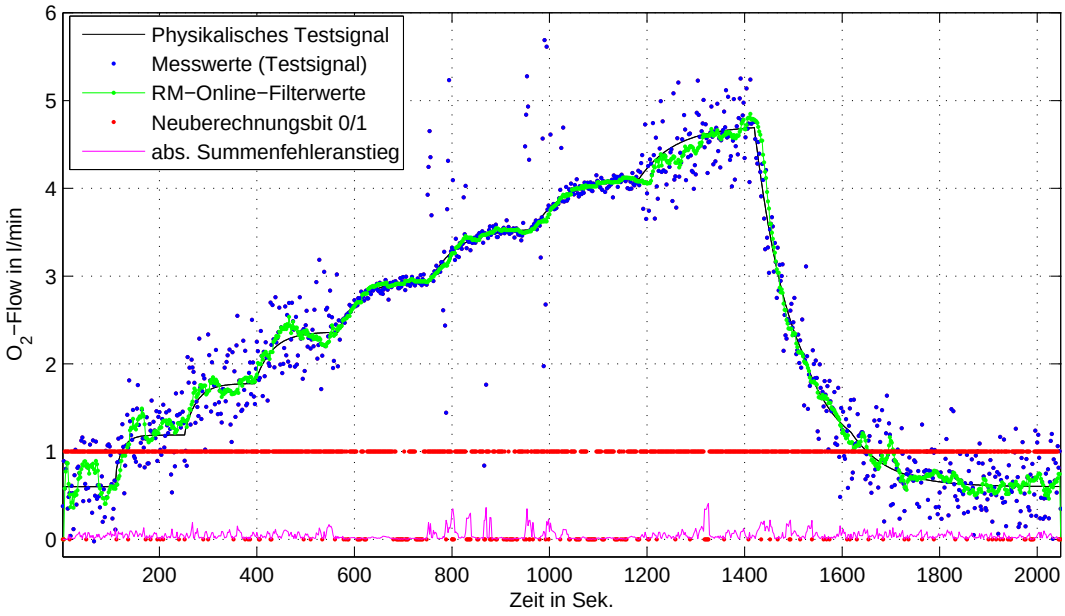

Abbildung 24: Ergebnis Online-Filterung k=32

Literatur

[1] Barnett,V., Lewis, T.: *Outliers in statistical data*. 3rd ed. p. cm. (1994)

[2] Beckman, R.J., Cook, R.D.: *Outliers*. Technometrics, (1983)

[3] Hawkins, D.M.: *Identification of Outliers*. Chapman & Hall London 1. Auflage (1980)

[4] Iglewicz, B., Hoaglin, D.C.: *How to detect and handle outliers*. VOLUME 16, 2.Aufl. (1993)

[5] Ludwig-Mayerhofer, W.: *ILMES-Internet Lexikon der Methoden der empirischen Sozialforschung*. Universität Siegen, http://www.lrz-muenchen.de/~wlm/ein_voll.htm

[6] Internetenzyklopädie: *http://de.wikipedia.org/wiki/Statistischer_Test, Stichwort: Statistische Tests*

[7] www.weibull.com

[8] Fried, R., Bernholt, T., and Gather, U. (2006): *Repeated Median and Hybrid Filters*, Computational Statistics and Data Analysis.

[9] Davies, P. L., Fried, R., and Gather, U. (2004): *Robust Signal Extraction for On-line Monitoring Data*, Journal of Statistical Planning and Inference, 122, 6578.

Weiterführende Literatur

[10] Moschytz, G., Hofbauer, M. (2000): *Adaptive Filter*, Springer, Berlin

[11] Donoho, D. L., and Huber, P. J. (1983): *The Notation of Breakdown Point, in A Festschrift for Erich Lehmann, eds. P. J. Bickel, K. Doksum, and J. L. Hodges Jr., Belmont, CA: Wadsworth*, pp. 157,184.

[12] Edgeworth, F. Y. (1887): *A New Method of Reducing Observations Relating to Sev- eral Questions*, Phil. Mag., 24, 184191.

[13] Einbeck, J., and Kauermann, G. (2003): *Online Monitoring with Local Smoothing Methods and Adaptive Ridging*, Journal of Statistical Computation and Simulation, 73, 913929.

[14] Ellis, S. P., and Morgenthaler, S. (1992): *Leverage and Breakdown in L1-Regression*, Journal of the American Statistical Association, 87, 143148.

[15] Fan, J., and Hall, P. (1994): *On Curve Estimation by Minimizing Mean Absolute Deviation and its Implications*, Annals of Statistics, 22, 867885.

[16] Fan, J., Hu, T.-C., and Truong, Y. K. (1994): *Robust Nonparametric Function Estimation*,Scandinavian Journal of Statistics, 21, 433446.

[17] Gather, U., Schettlinger, K., and Fried, R. (2006): *Online Signal Extraction by Robust Linear Regression, Computational Statistics*, to appear.

[18] Giloni, A., and Padberg, M. (2004): *The Finite Sample Breakdown Point of L1-regression*, SIAM Journal of Optimization, 14, 10281042.

[19] HÄardle, W., and Gasser, T. (1984): *Robust Non-parametric Function Fitting*, Journal of the Royal Statistical Society, Ser. B, 46, 4251.

[20] Hastie, T. and Loader, C. (1993): *Local Regression: Automatic Kernel Carpentry*, Statistical Science, 8, 120129.

[21] He, X., Jureckova, J., Koenker, R. and Portnoy, S. (1990): *Tail Behavior of Regression Estimators and Their Breakdown Points*, Econometrica, 58, 11951214.

[22] Imho®, M., Bauer, M., Gather, U., and Fried, R. (2002): *Pattern Detection in Intensive Care Monitoring Time Series with Autoregressive Models*, Influence of the Model

[23] Fan, J. (1992): *Design-adaptive Nonparametric Regression*, Journal of the American Statistical Association, 87, 9981004.

[24] Davies, P. L., and Gather, U. (2005): *Breakdown and Groups,Annals of Statistics*, 33, 9771035.