Kapitel 6

Datenprüfung

Die vorliegende Arbeit ist eine statistische Arbeit. Das bedeutet, dass alle zu nutzenden Daten einer intensiven Prüfung unterzogen werden mussten. Nur so ist abzusichern, dass das entwickelte Modell später zuverlässig arbeitet. Diese Prüfung erfolgte in zwei Schritten. Im ersten Schritt wird auf unplausible Einzelwerte, im zweiten auf systematische Fehler geprüft. Dafür konnten zwei schon vorliegende Fortran-Module genutzt werden. Bei der Suche nach unplausiblen Einzelwerten ist dieses ein vom Autoren [Schneider 2001] für das Umweltbundesamt im Herbst 2001 entwickeltes Modul (Abschnitt 6.1). Es prüft dort im realtime Modus den ständigen Dateneingang der DAL-Lieferungen. Das Programm zur Prüfung der systematischen Fehler wurde von [Enke 2002] ebenfalls für das Umweltbundesamt entwickelt (Abschnitt 6.2). Mit den vorliegenden Modulen ist es möglich, in operationeller Arbeitsweise:

- fehlerhafte Werte zu markieren,
- als fehlerhaft erkannte Werte durch Ausfallkennungen zu ersetzen und
- als fehlerhaft erkannte Werte und Ausfallwerte durch Schätzwerte zu ersetzen.

Dies erfolgt nach einheitlich objektiven Kriterien. Der Einsatz der Module zur Prüfung der Datenarchive ist ebenfalls möglich. Hiermit kann an beiden Seiten eines statistisch arbeitenden Systems Sicherheit erzeugt werden. Die Entwicklung und die operative Vorhersage erfolgen mit geprüften Daten.

6.1 Methodik zur Suche nach unplausiblen Einzelwerten

Für die Erstellung des Validierungsmodules wurden folgende Schritte unternommen. Zunächst wurden einige statistische Parameter des Datenkollektives ermittelt. Diese werden während der Validierung als Prüfgrößen herangezogen. Im Prozess der Validierung werden die folgenden fünf Prüfschritte durchgeführt:

- Prüfung auf Werte kleiner Null bzw. größer 360 $\mu g/m^3$,
- Ausreißerprüfung,
- Prüfung des Anstiegs bzw. des Rückgangs zwischen aufeinander folgenden Halbstundenwerten,
- Prüfung in der Fläche,
- und die Prüfung mit einer approximierten Zeitreihe.

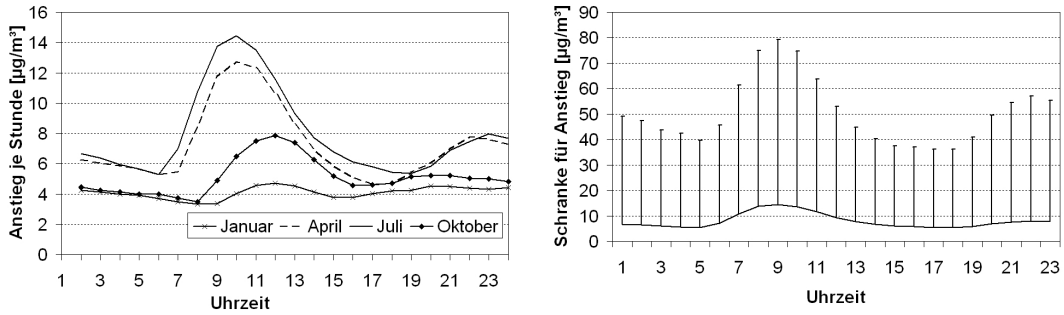
6.1.1 Statistische Parameter

Mit den vom UBA bereitgestellten Datenarchiven wurden die statistischen Parameter berechnet. Das Datenarchiv, das für die Ableitung zur Verfügung stand, umfasst 10 Jahre mit insgesamt 480 Stationen. Getrennt voneinander wurden jeweils der Mittelwert und die Standardabweichung für:

- die Ozonkonzentration,
- den Anstieg bzw. die Reduktion innerhalb einer Stunde
- und die Differenz zwischen Teststation und den Nachbarstationen gleichen Typs ermittelt.

Es erfolgt eine getrennte Berechnung über alle Stationstypen[1], Monate und Stunden. Der Einsatz des Mittelwertes trotz bekannter Schiefe der Verteilung ist unproblematisch, da im Folgenden mit Mittelwert plus X-mal die Standardabweichung gerechnet wird. Dabei fallen die Unterschiede zwischen Mittelwert und Median nicht ins Gewicht.
Abbildung 6.1 zeigt die deutliche Abhängigkeit der Prüfgrößen von Monat und Uhrzeit. So sind ein typischer Tagesgang beim möglichen Anstieg der Ozonkonzentrationen innerhalb einer Stunde und die unterschiedliche Höhe der Steigerung bei den dargestellten Monaten zu erkennen (Abbildung 6.1a).

[1]Die Einteilung erfolgte entsprechend [Enke u. a. 1998] in Stadt, Land, Verkehr und Berg.



(a) Anstieg (b) Prüfwerte Juli

Abbildung 6.1. (a) Beispielhafte Darstellung der mittleren Anstiege von Stunde zu Stunde (Stationstyp: Stadt) gemittelt über alle Stationen. Jede Kurve stellt einen Monat dar. (b) Beispielhafte Darstellung von Prüfwerten bei Ozon. Dargestellt wurde der Monat Juli mit dem Mittelwert plus 6-mal die Standardabweichung für den Anstieg von Stunde zu Stunde (Stationstyp: Stadt).

6.1.2 Die Prüfschranken

Für Temperaturen unter einer festgelegten Schwellentemperatur (25 °C) wurden feste Prüfschranken für jede Prüfung festgelegt. Das heißt zum Beispiel, dass bei der Ausreißerprüfung jeder Messwert mit einer Summe aus Mittelwert und der 6-fachen Standardabweichung verglichen wird.

$$Pruefwert = Mittelwert + 6 \times Standardabweichung \tag{6.1}$$

Der Mittelwert und die Standardabweichung entsprechen den oben berechneten Parametern für jeden Monat in Abhängigkeit von der Uhrzeit und dem Stationstyp. In Abbildung 6.1b ist das 6-fache der Standardabweichung als ein Fehlerbalken zum Mittelwert hinzugefügt worden. Dieser zeigt, dass nicht nur der Mittelwert einen deutlichen Tagesgang hat, sondern auch die Standardabweichung.

Für Temperaturen über 25 °C werden die Prüfgrenzen an die für den aktuellen Tag prognostizierte Maximum Temperatur angepasst. Mittels einer empirisch gefundenen Polynomfunktion 2. Grades erhöhen sich die Grenzen entsprechend Gleichung 6.2. Die prognostische bzw. diagnostische Maximum Temperatur wird vom Deutschen Wetterdienst (DWD) für die Ozonprognose bereitgestellt. Jede zu prüfende Station wird dazu einem der Prognosegitterpunkte zugeordnet. Überschreitet die prognostizierte bzw. diagnostische Maximum Temperatur 25° C so wird der vorgeschriebene Faktor unter Nutzung der Gleichung 6.2 erhöht.

$$F_{Stdabw,var} = 6 + 2 \times (0,004 \times T^2 - 0,124 \times T + 0,58) \tag{6.2}$$

Hierbei ist T die prognostizierte bzw. diagnostische Maximum Temperatur für einen in der Nähe liegenden Gitterpunkt des DWD-Modells.

6.1.3 Der Ablauf der Prüfung

Unsinnige Werte

Als erster Schritt erfolgt die Ersetzung aller negativen Werte durch die Ausfall-Kennung (-999.00). Weiterhin wird geprüft, ob der Wert die Konzentration von 360 $\mu g/m^3$ überschreitet. Wenn ja, so erfolgt eine Ersetzung durch die Ausfall-Kennung. Ozonkonzentrationen über 360 $\mu g/m^3$ (Stundenmittel) sind im Raum Deutschland laut dem UBA in den letzten zehn Jahren nicht mehr aufgetreten.

Ausreißerprüfung

Monats- und Uhrzeit-spezifisch erfolgt eine Prüfung aller Einzelwerte. Als nächstes wird auf Ausreißer bzw. Extremwerte geprüft. Die Prüfung erfolgt nur nach oben hin. Übersteigt ein Wert die Schranke 'Mittelwert plus x-mal Standardabweichung', so wird dieser speziell gekennzeichnet.

An einigen Stationen werden die folgenden Prüfroutinen nicht genutzt, da diese aufgrund ihrer Nähe zu Industrieanlagen sprunghafte Konzentrationsverläufe aufweisen können. Diese Stationen werden von der Routine gesondert behandelt.

Prüfung des Anstiegs bzw. des Rückgangs zwischen aufeinander folgenden Halbstundenwerten

Jeder Wert wird mit seinem zeitlich vorhergehenden Wert verglichen. Entsprechend der Änderungsrichtung wird der Betrag der Änderung geprüft. Übersteigt die Änderung die Schranke 'Mittelwert plus x-mal Standardabweichung' (siehe Abbildung 6.1 (b)), erfolgt eine Markierung des Wertes. Fehlt der vorhergehende Wert, so werden die Werte bis zu zwei Stunden zurück genutzt. Bei der Verwendung von Werten mit einem zeitlichen Abstand von mehr als einer Stunde wird die Schranke entsprechend modifiziert (es erfolgt eine Interpolation über den entsprechenden Zeitraum).

Prüfung in der Fläche

Mit den aus der Stationsliste gesuchten nächstliegenden Stationen kann eine Prüfung in der Fläche erfolgen. Hierbei wird der zu validierende Wert mit den aktuellen Werten der Nachbarstationen[2] verglichen und die minimale Differenz gesucht. Diese wird dann gegen eine für den Stationstyp spezifische Schranke 'Mittelwert plus x-mal Standardabweichung' geprüft. Wird die Schranke überschritten, so wird der Wert durch die Ausfall-Kennung ersetzt. Da nicht immer

[2]Gewählt werden in den vier Himmelsrichtungsquadranten (NO, SO, SW und NW) je die zwei naheliegendsten Stationen gleichen Stationstyps, mit maximal 150 km Abstand.

alle acht Stationen zu allen Zeitpunkten mit Werten besetzt sind, wurde die Bedingung eingebaut, dass bei mindestens drei Nachbarstationen Werte vorhanden sein müssen.

Prüfung mit einer approximierten Zeitreihe

Hier wird der zu prüfende Wert mit einem approximierten Wert verglichen. Übersteigt die absolute Differenz zwischen beiden den Wert 30 $\mu g/m^3$, so wird der Wert gekennzeichnet. Ist die Differenz größer als 50 $\mu g/m^3$, so ist dieser Wert nicht mehr als sinnvoll anzusehen und er wird durch die Ausfall-Kennung ersetzt. Diese Grenzwerte gelten für Temperaturen unter 25 °C. Darüber werden diese Grenzwerte entsprechend der Gleichung 6.3 erhöht.

$$GW_{var} = 30(50) + 30(50) \times (0,004 \times T^2 - 0,124 \times T + 0,58) \qquad (6.3)$$

Hierbei ist T die prognostizierte bzw. diagnostische Maximum Temperatur für einen in der Nähe liegenden Gitterpunkt des DWD-Modells. Zusätzlich zu der Veränderung der Grenzwerte wurde noch eine weitere Bedingung eingebaut. Ab Temperaturen von ≥25 °C erfolgt eine Markierung und Ausfall-Setzung nur, wenn auch die Prüfung des Anstieges bzw. der Reduktion einen möglichen Fehler erkannt hat.

Basis für die Approximierung ist eine Screening-Regressions-Analyse. Hierbei wird für jede Station, auf Basis des Datenarchivs, die beste Kombination an Stationen gesucht, mit denen der Wert an der Station approximiert werden kann. Vor der Analyse sind alle Messwerte vom Jahres- und Tagesgang bereinigt worden. Fehlt ein Messwert bei einer der für die Regression genutzten Stationen, so wird deren Regressionsfunktion genutzt, um wiederum deren Wert zu approximieren. Fehlt auch dort ein Messwert, so wird das Mittel der Abweichungen der insgesamt bis zu 8 Nachtbarstationen angesetzt. Mit diesen approximierten Vergleichswerten wird der Messwert geprüft.

Bei drei der fünf Prüfungen werden bemängelte Werte markiert (Ausreißerprüfung, Prüfung des Anstiegs bzw. des Rückgangs zwischen aufeinander folgenden Halbstundenwerten und Prüfung mit einer approximierten Zeitreihe). Als letzter Schritt erfolgt nun eine Abwägung der markierten Fehlertypen. Tritt bei zwei Tests die Kennzeichnung *fragwürdig* auf, so wird dieser Wert durch die Ausfall-Kennung ersetzt.

Abbildung 6.2 zeigt beispielhaft den Verlauf der gemeldeten 30 Minuten Mittelwerte der Station Rinteln (DENI041) am 3. Juni 2003. Die weißgepunkteten Säulen sind von der Routine als fehlerhaft markiert worden. Als Ursache für die Markierung gibt das log-File bei den ersten beiden Säulen eine Überschreitung der Schwellenwerte bei der approximierten Reihe und bei den letzten drei Säulen eine Überschreitung der Abstandsschwelle im Vergleich mit den Nachbarstationen (Abstand zwischen 95 und 132 $\mu g/m^3$) an.

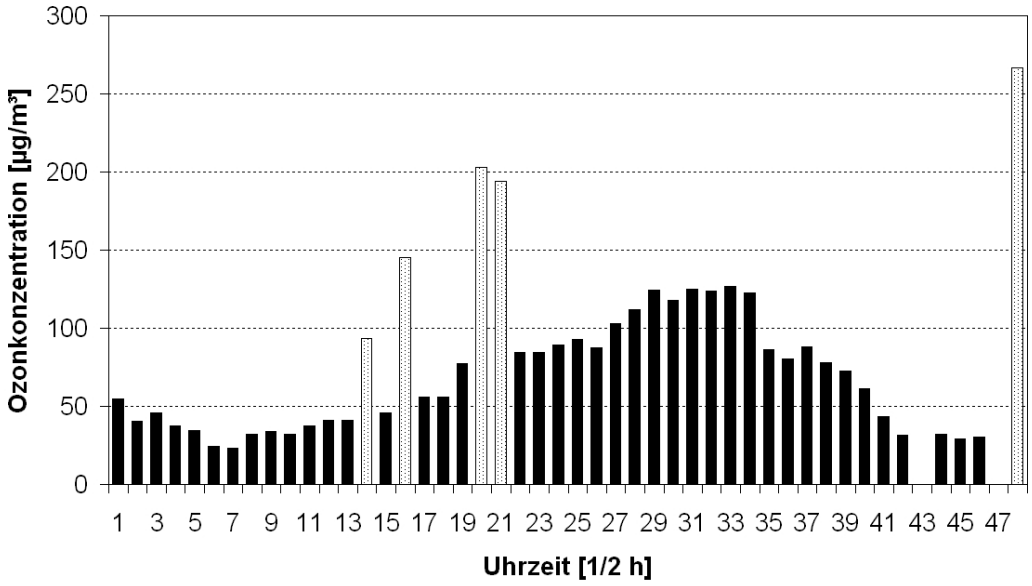


Abbildung 6.2. Gemeldete 30 Minuten Mittelwerte der Station Rinteln (DENI041) vom 3. Juni 2003. Die weißgepunkteten Säulen sind von der Prüfroutine als fehlerhaft markiert und vor der Webpräsentation aussortiert worden.

Die Entwicklung dieser Routine erfolgte im Herbst 2001. Im Rahmen der Entwicklung wurden umfangreiche Tests mit den vorliegenden Datensätzen des Zeitraumes 1993 bis 2000 durchgeführt. Seit der Fertigstellung im November 2001 ist das Programm im operationellen Einsatz am Umweltbundesamt. Dort werden die stündlich einlaufenden Datenlieferungen vor der grafischen Aufbereitung für die Internetdarstellung geprüft.

Im Laufe des Einsatzes sind einige Änderungen notwendig geworden. So war anfänglich ein Prüfwert aus dem Mittelwert plus 4-mal Standardabweichung als Prüfgrenze festgelegt worden. Im Sommer 2002 zeigte sich jedoch, dass dieser Faktor erhöht werden muss. Dies galt insbesondere für die Flächenprüfung. Im selben Schritt erfolgte die Einführung der exponentiell steigenden Faktoren bei Temperaturen ab 25 °C. Es wird vermutet, das an Tagen mit sehr hohen Ozonwerten ($> 180\ \mu g/m^3$) die räumlichen Korrelationen nachgelassen haben. Die Spitzenkonzentrationen scheinen somit zu einem großen Teil lokal begründet zu sein. Einige weitere Änderungen, wie zum Beispiel der Einbau eines Modules zum Abfangen von sich über längere Zeit ($\Delta t \geq 4$ Stunden) nicht ändernden Werten, sind geplant. Eine Aktualisierung der statistischen Beziehungen ist für Herbst 2003 vorgesehen.

Ein nachgeschaltetes Modul kann bei Bedarf dann alle Lücken im Datenarchiv füllen. Die Beschreibung des dafür genutzten Moduls erfolgt in Kapitel 6.3. Hierdurch wird für alle Stationen eine Vorhersage trotz einzelner Ausfälle ermöglicht.

6.2 Methodik zur Suche nach systematischen Fehlern

Die Methodik zur Suche nach systematischen Fehlern ist in [Enke 2002] detailliert beschrieben worden. Der folgende Bereich (2 Abschnitte) wurde mit kleineren Anpassungen direkt aus [Enke 2002] übernommen.

Auch wenn ein Datenkollektiv der Ozonkonzentration von nicht plausiblen Werten und Ausreißern nach der im Abschnitt 6.1 dargestellten Validierungsstufe bereinigt ist, können im zeitlichen Verlauf einer Reihe trotzdem Trends und Strukturbrüche auftreten, die sowohl reale Ursachen (meteorologische Bedingungen, Änderungen der Emission der Vorläuferstoffe u.a.m.) haben als auch artifiziell (z.B. Änderungen der Probenahme- und Messtechnik) bedingt sein können. Deshalb ist es notwendig, die meteorologisch oder emissionsseitig bedingten Einflüsse jahreszeitspezifisch und stündlich aus der Messreihe zu eliminieren. Eine Möglichkeit hierzu bietet die Ersetzung der zu untersuchenden Zeitreihe mit Erwartungswerten, die aus anderen Ozonmessreihen berechnet werden (simulierte Zeitreihe). Für die operationelle Datenprüfung setzt man dabei voraus, dass die meteorologischen Bedingungen sich in den Ozonkonzentrationen umliegender Stationen genügend abbilden. Um nicht nur die Situation eines konkreten Tages für die Interpolation zu verwenden, wie dies bei rein räumlichen Interpolationsverfahren (Shepard-Verfahren, Thin Plate Spline, Triangulation oder IDW-Verfahren) der Fall ist, sondern auch die statistischen Beziehungen zwischen verschiedenen Beobachtungsreihen und der zu ersetzenden Reihe, wird ein stufenweises Screening-Regressionsverfahren eingesetzt. In Anlehnung an den Alexandersson-Test [Alexandersson 1986], wird die Differenz zwischen der insgesamt simulierten Messreihe (Erwartungswerte) und den Originalwerten einer Station verwendet. Die Zeitreihe der Erwartungswerte selbst, ist dagegen nicht dargestellt. In dieser Differenzenreihe treten durch die weitgehende Eliminierung meteorologischer und emissionsseitig bedingter Einflüsse mögliche Strukturbrüche, Trends oder Messfehler deutlicher hervor, als dies bei den Originalreihen der Fall ist.

Eine korrekte Eliminierung meteorologischer Einflüsse, wie sie im Rahmen des UBA F+E Vorhabens: 297 42 848 -Analyse historischer Datenreihen und Entwicklung einer Methode zur quasi-wetterbereinigten Trendanalyse von bodennahem Ozon- Teilbericht 3 -Wetterbereinigung der Jahre 1980 bis 1997- [Enke 2001a] unter Einbeziehung meteorologischer Informationen durchgeführt wurde, ist hier aufgrund der Komplexität und des großen Aufwandes eines solchen Vorgehens nicht möglich. Man kann jedoch davon ausgehen, dass sich in den täglichen Schwankungen der Ozonmessreihen ein Großteil der Variabilität des Wettergeschehens widerspiegelt. Die Synchronität der Tagesgänge zwischen nahe beieinander liegenden Stationen drückt also die Ähnlichkeit des Wettergeschehens auf regionaler Ebene aus. Andersherum betrachtet, ist in der Differenzenreihe zwischen zwei oder mehreren nahe beieinander liegenden Ozonreihen der Wettereinfluss größtenteils eliminiert.

Hier werden auch Untersuchungen mit simulierten Fehlern vorgestellt. Aufgrund des rechentechnischen Anspruches der Routine läuft diese nur in der Nacht am UBA. Somit kann keine Prüfung im Rahmen der Prognose erfolgen. Alle Daten die zur Entwicklung der Ozonprognose genutzt wurden, sind von der Routine geprüft worden.

6.3 Auffüllen von Datenlücken

Vor dem Auffüllen von Datenlücken muss erst die Frage beantwortet werden, warum es nötig ist, Lücken zu füllen. Für die Entwicklung einer statistischen Routine ist es notwendig, einen ausreichend großen Datensatz mit Datenmaterial zu haben. Nur so ist möglich, eine Aufteilung in Entwicklungs- und Testkollektiv durchzuführen. Wenn zu viele Datensätze unvollständig wären, wäre keine Entwicklung möglich.

Zum Auffüllen der Datenlücken werden zwei Methoden herangezogen. Ist die Datenlücke kleiner oder gleich drei Stunden, so wird die Lücke durch lineare Interpolation aufgefüllt. Ist die Zeitspanne größer als drei Stunden, so werden approximierte Werte zum Auffüllen genommen.

Die Grundlage der Approximation von Werten ist eine Screening-Regressions-Analyse. Dabei wird für jede Station die Kombination an Stationen gesucht, mit der via Regression am besten der Messwert approximiert werden kann. Grundlage dafür ist die Idee, dass in der Hauptsache großräumige Ursachen für die aktuelle lokale Ozonsituation verantwortlich sind. Durch die Bereinigung aller dazu genutzten Werte vom Tages- und Jahresgang wird mit den Abweichungen von der *normalen* mittleren Stationssituation gerechnet.

Zur Berechnung des approximierten Wertes werden am jeweiligen Tag zur entsprechenden Uhrzeit die Messwerte der vorher in der Analyse ermittelten Stationen genutzt. Fehlt einer dieser Werte, so wird dieser mittels der für diese Station vorliegenden Regressionsfunktion ermittelt. Fehlt auch dort ein Messwert, so wird das Mittel der Abweichungen der insgesamt bis zu 8 Nachtbarstationen angesetzt.

6.4 Pro und Contra zur Datenveränderung

Grundsätzlich ist jede nachträgliche Veränderung von gemessenen Werten eine Datenverfälschung. Dies beginnt schon mit der Datenprüfung. Jeder dort vorliegende Wert muss gegebenenfalls bis zu seiner Falsifikation grundsätzlich als richtig angesehen werden.

Zitat 1 Ich glaube nur an Statistiken, die ich selbst gefälscht habe.
Winston Churchill zugesprochen

6.4 Pro und Contra zur Datenveränderung

Zitat 2 Wenn man Zahlen richtig foltert, gestehen sie, was man will.
Peter E. Schuhmacher

Daher muss jede Änderung der vorliegenden Daten begründbar sein. Und die Änderung selbst darf keine Änderung der *statistischen Eigenschaften* des Datenkollektives zur Folge haben.

Nun ist die Arbeit mit statistischen Routinen auf der Basis von inkonsistenten und lückenhaften Stichproben aber sehr problematisch. Somit ist die Wahl des *kleinsten Übels* nötig. Für die Bearbeitung dieses Projektes bedeutet dies, dass:

a) Messwerte kleiner Null bei Konzentrationsmessungen als nicht real angesehen und damit entfernt werden;

b) Messwerte die mit naturwissenschaftlichen Ansätzen nicht erklärbar sind (bzw. die Aufgrund der Rahmenbedingungen in Deutschland untypisch sind), als *fragwürdig* markiert werden und, falls sie bei mehr als einem Verfahren *fragwürdig* sind, entfernt (auf Ausfall gesetzt) werden;

c) falls Zeitreihen von Messwerten zeitweise einen systematischen Versatz aufweisen, diese in begründeten Fällen in diesem Zeitraum systematisch verschoben werden;

d) angenommen wird, dass Datenlückenfüllung via Approximation mit nahezu statistischen Mittelwerten keinen großen Einfluss auf das Datenkollektiv hat.

Bei allen diesen Veränderungen wurde darauf Wert gelegt, dass die Anzahl der Veränderungen in Bezug auf die Gesamtzahl der Daten minimal sind, damit die statistischen Eigenschaften der Zeitreihe erhalten bleiben. Es ist jedoch klar, dass damit diese Arbeit streng genommen angreifbar ist. Allerdings erfolgte eine Kennzeichnung der approximierten Werte. So ist jederzeit feststellbar, woher der jeweilige Wert kommt.

Der Einsatz der Lückenauffüllung erfolgte nur während der Entwicklung. Auf eine Nutzung während des operationellen Einsatzes wurde verzichtet.

Teil II

Angewandte Methoden und Ergebnisse