DGK Deutsche Geodätische Kommission
bei der Bayerischen Akademie der Wissenschaften

Reihe C | Dissertationen | Heft Nr. 662

Ralf Laufer

Prozedurale Qualitätsmodellierung und -management für Daten – ingenieurgeodätische und verkehrstechnische Anwendungen –

München 2011

Verlag der Bayerischen Akademie der Wissenschaften
in Kommission beim Verlag C. H. Beck

ISSN 0065-5325 ISBN 978-3-7696-5074-7

DGK Deutsche Geodätische Kommission
bei der Bayerischen Akademie der Wissenschaften

Reihe C | Dissertationen | Heft Nr. 662

Prozedurale Qualitätsmodellierung und -management für Daten
– ingenieurgeodätische und verkehrstechnische Anwendungen –

Von der Fakultät für Luft- und Raumfahrttechnik und Geodäsie
der Universität Stuttgart
zur Erlangung der Würde eines
Doktor-Ingenieurs (Dr.-Ing.)
genehmigte Abhandlung

Vorgelegt von

Dipl.-Ing. Ralf Laufer

aus Villingen-Schwenningen

München 2011

Verlag der Bayerischen Akademie der Wissenschaften
in Kommission beim Verlag C. H. Beck

ISSN 0065-5325 ISBN 978-3-7696-5074-7

Adresse der Deutschen Geodätischen Kommission:

Deutsche Geodätische Kommission

Alfons-Goppel-Straße 11 ● D – 80 539 München

Telefon +49 – 89 – 23 031 1113 ● Telefax +49 – 89 – 23 031 - 1283 / - 1100

e-mail hornik@dgfi.badw.de ● http://www.dgk.badw.de

Hauptberichter: Prof. Dr.-Ing. habil. Volker Schwieger

Mitberichter: Prof. Dr.-Ing. habil. Dieter Fritsch

Tag der mündlichen Prüfung: 23.02.2011

ISSN 0065-5325 ISBN 978-3-7696-5074-7

Zusammenfassung

Daten und Informationen sind die fundamentalen Grundlagen für Entscheidungen. Jeder Mensch trifft im Laufe seines Lebens unzählige Entscheidungen, die auf mehr oder weniger vollständigen, objektiven und korrekten Daten und Informationen beruhen. Nicht nur bei persönlichen Entscheidungen spielt die Qualität der Informationen, die zur Abwägung zur Verfügung stehen, eine wesentliche Rolle. Insbesondere bei technischen Anwendungen in allen Bereichen des Lebens, ist die Qualität von Daten als Entscheidungsgrundlage von essenzieller Bedeutung.

Grundlage und Motivation dieser Arbeit war die konsequente Fortführung der Tätigkeiten im Bereich der Handhabung von Datenqualität am Institut für Anwendungen der Geodäsie im Bauwesen (IAGB) bzw. am Institut für Ingenieurgeodäsie (IIGS)[1] der Universität Stuttgart. Die erarbeiteten Grundlagen zur qualitätsgesicherten Bereitstellung von Daten, die in modernen Fahrerassistenzsystemen (FAS) von Bedeutung sind (vgl. Wiltschko [2004]), sollten weiter verallgemeinert und für andere Bereiche, in denen Datenqualität eine wesentliche Rolle spielt, zugänglich gemacht werden. Neben den FAS sind hier im Bereich der Verkehrsanwendungen insbesondere Mobilfunkortungsverfahren zu nennen. Datenqualität spielt seit jeher jedoch auch in anderen Bereichen, wie zum Beispiel in der Geodäsie, eine entscheidende Rolle. Die Qualität von Daten hat oftmals erheblichen Einfluss auf sicherheitsrelevante Entscheidungen und mangelnde Datenqualität kann unter Umständen zu Systemausfällen oder folgenschweren Fehlinterpretationen führen.

Das von Wiltschko [2004] aufgestellte Konzept zur Beschreibung und Bewertung der Datenqualität innerhalb informationsverarbeitender Systeme besteht aus einem Qualitätsmodell und einem Analyseverfahren. Das Qualitätsmodell umfasst sechs inhärente Merkmale, die durch geeignete Qualitätsparameter konkretisiert werden können. Durch die Wahl der Merkmale ist das Modell universell einsetzbar und für eine Vielzahl von unterschiedlichen Arten von Daten geeignet. Da das vorgestellte Modell auf einer umfangreichen Recherche beruht, konnte hier zumindest für raumbezogene Daten in der Geodäsie und Verkehrstelematik kein weiterer Forschungsbedarf festgestellt werden. Im Bereich der Geoinformationssysteme wurde nach jahrelanger Forschungsarbeit mit der ISO TS 19138 [2006] ein Modell zur Beschreibung der Datenqualität eingeführt. Dabei handelt es sich jedoch um ein speziell für den Datenaustausch und die Datenanalyse von Geodaten in Geoinformationssystemen entwickeltes Modell, welches für die hier untersuchten Daten nicht allgemeingültig genug ist.

Neben dem immer noch aktuellen Qualitätsmodell wurde von Wiltschko auch ein Analyseverfahren auf Basis Boolescher Algebra eingeführt, welches bei näherer Betrachtung einige Einschränkungen hinsichtlich der praktischen Umsetzbarkeit zeigt und Qualitätsparameter nur unzureichend oder gar nicht behandeln kann. Hier wurde konkreter Forschungsbedarf erkannt, um das bestehende Verfahren zu ergänzen oder zu verbessern und die in dieser Arbeit herausgearbeiteten Grenzen zu überwinden.

Die Recherche nach geeigneten Verfahren zur quantitativen Beschreibung von Datenqualität in Prozessen ist daher ein wesentlicher Bestandteil der Arbeit. Künstliche neuronale Netze (KNN) stellen ein mächtiges Werkzeug dar, welches auch für die umfassende Modellierung von Datenqualität zum Einsatz kommen kann. In dieser Arbeit wurde deren Eignung an zwei Beispielen im Detail untersucht.

[1] Während der Fertigstellung der Arbeit fand aufgrund einer Neuberufung eine Umbenennung des Instituts statt.

Wie die Untersuchungen ergaben, können KNN zur Modellierung von Qualitätsparametern in Prozessen eingesetzt werden. An einem einfachen geodätischen Beispiel wird gezeigt, dass eine Modellierung der Genauigkeit auch in Kombination mit Parametern der Verfügbarkeit und Vollständigkeit in zufriedenstellender Qualität erfolgen kann. Im Anschluss erfolgt der Test von KNN in einem komplexeren, verkehrstechnischen Beispiel. Dabei werden KNN erfolgreich bei der Mobilfunkortung mit realen Eingangsdaten zur Modellierung eines Qualitätsparameters eingesetzt, welcher der Beurteilung des Prozesses dient. Damit ist der Einsatz eines trainierten KNN auch in Echtzeit zur kontinuierlichen Überwachung eines laufenden Prozesses möglich.

Schließlich wird ein umfassendes Qualitätsmanagementkonzept für Daten vorgestellt, welches die im Detail beschriebenen und untersuchten Methoden, sowie das bestehende Qualitätsmodell in einen globalen Gesamtzusammenhang stellt. Das vorgestellte Konzept basiert auf dem PDCA-Gedanken (dem Konzept der ständigen Verbesserung) von Deming [1982] und kann in ein ISO-konformes Qualitätsmanagementsystem integriert werden. Das Konzept besteht einerseits aus bekannten Bausteinen zur Beschreibung und Analyse von Daten verarbeitenden Systemen, andererseits bietet es jedoch neue Möglichkeiten zur Behandlung und Simulation von Datenqualität in Prozessen mit Hilfe künstlicher neuronaler Netze.

Abstract

Data and information are the fundamental basis for decisions. Every human being makes countless decisions during lifetime, which base on more or less complete, objective and correct data and information. Quality of available information plays an important role not only within personal decisions. Quality of data as basis for decisions is of essential importance particularly in technical applications in all fields of live.

Background and motivation for this document was the consistent continuation of activities in the field of handling data quality at the Institute for Applications of Geodesy to Engineering (IAGB) respectively at the Institute of Engineering Geodesy (IIG)[2] at the University of Stuttgart. The fundamentals of allocating quality assured data which are of interest in advanced driver assistance systems (ADAS) and which have been worked out by Wiltschko [2004] should be further generalized and should be made accessible for other fields of application in which data quality plays a significant role. Besides ADAS, in the field of traffic applications, mobile phone positioning can be named in particular. However data quality plays a key role in other fields as well such as geodesy. Quality of data often has an important impact on security relevant decisions and insufficient data quality may cause system failure or lead to severe misinterpretations.

The concept for describing and evaluating data quality within information processing systems proposed by Wiltschko [2004] consists of a quality model and an analyzing method. The quality model contains six inherent characteristics which can be concretized using appropriate quality parameters. The model is versatile applicable for different kinds of data with its selection of characteristics. Since the presented model is based on an extensive research, at least for space-oriented data in the field of geodesy and transport telematics, no further demand for research was identified. With the ISO TS 19138 [2006] a model for describing data quality in the field of geographic information systems was established after years of research work. Since this model is developed for data exchange and data analysis of geodata in geographic information systems in particular, it is not universally valid for the data examined in this work.

Besides the quality model which is still up to date, Wiltschko established an analyzing method based on Boolean algebra as well. After a closer examination, the analyzing method shows some constraints regarding the practical implementation and is not quite capable in handling quality parameters. For complementing or advancing the existing method and to resolve the constraints identified within this thesis, further research is required.

The research for capable methods to describe data quality quantitatively within processes is an essential part of the presented document. It turned out that artificial neural networks (ANN) are a powerful and versatile tool, which can be used as well for comprehensive modeling of data quality. Their ability has been analyzed in detail in this paper with the help of appropriate examples.

The analysis show that ANN can be used for modeling of quality parameter in processes. As shown with the help of a simple geodetic example, modeling of accuracy, also in combination with parameter of availability and completeness, can be done with sufficient quality. ANN are as well successfully applied within the second, more complex, example. Within this example, ANN are used for modeling a quality parameter from real input data, capable for evaluating the process of mobile phone positioning. Hence

[2] During finishing of the thesis a renaming took place due to a new appeal of the institute.

the application of a trained ANN for continuous quality monitoring of a running process in real time is possible.

Finally a comprehensive quality management concept will be presented which puts the extensive described and examined methods as well as the existing quality model into a global context. The presented concept bases on the PDCA theory (the concept of constant improvement) by Deming [1982] and can be integrated into an ISO conform quality management system. On the one hand the concept consists of known components for describing and analyzing data processing systems, on the other hand however it provides new possibilities for dealing with and simulation of data quality in processes using artificial neural networks.

Inhaltsverzeichnis

Zusammenfassung **3**

Abstract **5**

1 Einleitung **9**
1.1 Umfeld . . . 9
1.2 Motivation . . . 10
1.3 Aufbau der Arbeit . . . 10

2 Grundlagen **13**
2.1 Zum Begriff Datenqualität . . . 13
2.1.1 Erläuterung der Begriffe Daten und Informationen . . . 13
2.1.2 Datenarten . . . 15
2.1.3 Der Begriff Datenqualität . . . 16
2.1.4 Datenqualität aus Kundensicht . . . 18
2.2 Ein Qualitätsmodell für verkehrstechnische Anwendungen . . . 19
2.3 Charakterisierung von Prozessen in der Datenverarbeitung . . . 23
2.4 Einführung in Qualitätsmanagementsysteme . . . 25
2.4.1 Prozessorientierter Ansatz nach ISO 9000ff . . . 25
2.4.2 Total Quality Management - TQM . . . 27
2.4.3 Six-Sigma-Methode . . . 28
2.4.4 Zusammenfassung QM-Systeme . . . 30

3 Darstellungs- und Fortpflanzungsverfahren für Datenqualität **33**
3.1 Ein Verfahren zur Analyse der Informationsqualität . . . 33
3.1.1 Das Informationsflussdiagramm . . . 34
3.1.2 Rechenverfahren auf Basis Boolescher Algebra . . . 37
3.1.3 Kovarianzfortpflanzung . . . 45
3.1.4 Grenzen des bestehenden Verfahrens . . . 46
3.2 Alternative Darstellungs- und Fortpflanzungsverfahren . . . 49
3.2.1 Methoden und Verfahren der Zuverlässigkeitsanalyse . . . 49
3.2.2 Petri-Netze (PN) . . . 53
3.2.3 Monte-Carlo-Simulation . . . 55
3.2.4 Künstliche neuronale Netze - KNN . . . 57
3.2.5 Weitere Verfahren . . . 67
3.3 Gegenüberstellung der Verfahren . . . 67
3.4 Adaption der künstlichen neuronalen Netze zur Beschreibung von Qualität in Prozessen . 69

4 Anwendung von KNN zur Fortpflanzung der Datenqualität an Beispielen **75**
4.1 Beispiel aus der klassischen Geodäsie . . . 75
4.1.1 Modellierung der Genauigkeit . . . 76
4.1.2 Modellierung der Verfügbarkeit . . . 79
4.1.3 Modellierung der Vollständigkeit . . . 87
4.1.4 Modellierung der Konsistenz . . . 89

4.1.5 Modellierung der Aktualität . . . 92
4.1.6 Modellierung der Korrektheit . . . 93
4.1.7 Zusammenfassung . . . 93
4.2 Beispiel: Floating Phone Data . . . 94
4.2.1 Generierung von FPD-Trajektorien . . . 95
4.2.2 Modellierung der Querabweichung . . . 98
4.2.3 Beurteilung der Ergebnisse . . . 114
4.3 Beurteilung der Eignung von KNN . . . 116

5 Ein Qualitätsmanagementkonzept für Daten 119
5.1 Übersicht über das Konzept . . . 120
5.2 Bestandteile des Konzepts . . . 121
5.2.1 Qualitätsmodell . . . 121
5.2.2 Messmethoden . . . 123
5.2.3 Qualitative Analyseverfahren . . . 124
5.2.4 Quantitative Analyseverfahren . . . 127
5.2.5 Evaluierung . . . 129
5.2.6 Qualitätssicherung und -verbesserung . . . 129
5.3 Zusammenfassung . . . 131

6 Zusammenfassung und Ausblick 133

Literatur 136

Glossar 141

Formelzeichen 145

Anhang A: Das GUI nntool von Matlab 147

Anhang B: Das Qualitätsmodell für FPD-Trajektorien 153

Anhang C: Beispiel: Ursachen-Wirkungs-Diagramm 155

Anhang D: Beispiel: FMECA 157

Danksagung 159

Lebenslauf 161

1 Einleitung

1.1 Umfeld

In ingenieurtechnischen Anwendungen sind Daten allgegenwärtig. Sie dienen der Steuerung von Anlagen, sind Grundlage für Entscheidungen, geben Sollgrößen vor oder stellen selbst das zu erzeugende Produkt eines Prozesses dar. Daten ohne gesicherte Kenntnis ihrer Qualität sind jedoch gänzlich nutzlos, da deren Verlässlichkeit dann grundsätzlich in Frage zu stellen ist. Den Daten, die die Grundlage von Entscheidungen darstellen, wird oft implizit das Prädikat „Qualität ist ausreichend“gegeben, ohne dass diese tatsächlich hinreichend beschrieben, geschweige denn überprüfbar ist. Die Verantwortung wird so automatisch der Datenquelle bzw. dem Datenlieferanten übertragen, der für diese ungeschriebene Aussage „Qualität ist ausreichend“bürgt.

Im Bereich der verkehrstechnischen Anwendungen bestand Bedarf, Datenqualität einheitlich zu beschreiben. Sinnvoll und notwendig war daher die Schaffung eines allgemeingültigen Qualitätsrahmens, innerhalb dessen die Qualität der Daten anhand von geeigneten Parametern beschrieben wird. Dieses Qualitätsmodell dient der detaillierten Beschreibung von Datenqualität, die den Ansprüchen aller beteiligten Parteien gerecht werden und neben der Definition der Parameter auch Hinweise zu deren Ermittlung beinhalten muss.

Dieser Mangel an einem umfassenden Qualitätsmodell wurde unter anderem im Bereich der Geodaten erkannt und es gibt mittlerweile einige Qualitätsmodelle, die jeweils speziell für unterschiedliche Datenarten entwickelt wurden. Teilweise wurden diese Modelle sogar in ISO-Normen allgemeingültig ausformuliert. Das Modell von Wiltschko [2004] ist sehr allgemein gehalten und stellt hier nach wie vor eine Alternative dar, wenn keine zu enge Festlegung auf einzelne Datenarten gewünscht ist. Dieses Modell wird in der vorliegenden Arbeit zu Grunde gelegt.

Über die Verknüpfung von datenverarbeitenden Prozessen und der umfassenden Beschreibung der Datenqualität der Eingangs- und Ausgangsgrößen ist in der Literatur bislang nur wenig zu finden. Lediglich die Verknüpfung von Genauigkeitsparametern in Form von Standardabweichungen ist ein seit langem gelöstes Problem. Hier kommt unter anderem die klassische geodätische Methode der Kovarianzfortpflanzung zum Einsatz. Auch die Betrachtung der Systemzuverlässigkeit bzw. -verfügbarkeit ist bereits seit der Entwicklung der ersten großtechnischen Anlagen und der Planung von Weltraummissionen mit standardisierten Methoden möglich. Es fehlen jedoch praktikable Konzepte, wie weitere Qualitätsmerkmale wie z. B. die Vollständigkeit oder die Konsistenz dargestellt werden können und insbesondere wie eine gemeinsame Modellierung unterschiedlicher Qualitätsparameter in der Praxis erfolgen kann.

In der Literatur sind Qualitätsmanagementkonzepte in vielen Variationen zu finden. Allerdings mangelt es an Beschreibungen, wie das Management von Datenqualität konkret umgesetzt werden kann. Der Darstellung eines formalen Zusammenhangs zwischen den originären Daten, der Ermittlung und Beschreibung deren Qualität sowie die Darstellung der Qualität in Prozessen, die es zunächst zu analysieren gilt, bis hin zur Qualitätssicherung und der Evaluierung, wird bislang kaum Bedeutung geschenkt. Es fehlt ein umfassendes Konzept, welches die einzelnen Bestandteile des Managements von Datenqualität in einen Zusammenhang stellt.

1.2 Motivation

Ziel der vorliegenden Arbeit war die Suche nach geeigneten Verfahren zur Modellierung von Datenqualität in Prozessen und deren Einbindung in ein umfassenden Qualitätsmanagementkonzept für Daten. Das Verfahren sollte dabei das bereits bestehende Verfahren auf Basis Boolescher Algebra ergänzen oder ersetzen und dessen Einsatzgrenzen überwinden. Dazu sollten diese zunächst festgestellt werden, um die Anforderungen an ein alternatives Verfahren konkretisieren zu können. Es war außerdem anzustreben, unter der Vielzahl bestehender Verfahren, die in der Literatur bereits zu finden sind, ein geeignetes zu finden und dieses an die Erfordernisse entsprechend anzupassen.

Die Zusammenstellung eines umfassenden Konzeptes zum Management von Datenqualität, welches sich in den globalen Zusammenhang eines unternehmerischen Qualitätsmanagements eingliedern lässt, war ebenfalls erklärtes Ziel der Arbeit. Dabei sollten bereits am Institut entwickelte Bausteine ebenso Verwendung finden wie das neue Verfahren zur Modellierung von Datenqualität, welches im ersten Teil der Arbeit zu finden und zu testen war.

1.3 Aufbau der Arbeit

Zunächst werden im Kapitel 2 wichtige Begriffe im Umfeld der Datenqualität für die Verwendung in der Arbeit erläutert und die Bedeutung der Datenqualität herausgearbeitet. Dabei wird auch das bewährte Qualitätsmodell von Wiltschko [2004] vorgestellt, welches auch weiterhin die Grundlage der Beschreibung von Datenqualität bildet. Schließlich erfolgt im Abschnitt 2.4 eine Einführung in die Qualitätsmanagementsysteme und es werden -ohne Anspruch auf Vollständigkeit- einige der bekanntesten Systeme kurz mit ihren wesentlichen Eigenschaften vorgestellt.

In Kapitel 3 wird das von Wiltschko [2004] entwickelte Verfahren zur Modellierung von Qualität in Prozessen, bestehend aus einem grafischen Darstellungs- und einem Fortpflanzungsverfahren (Wiltschko spricht hier allgemeiner von „Rechenverfahren“), vorgestellt und an einem einfachen Beispiel näher erläutert. Schließlich werden die Grenzen des Verfahrens mit Hinblick auf die weiteren Untersuchungen herausgearbeitet. In der Konsequenz schließt sich die Recherche nach alternativen Darstellungs- und Fortpflanzungsverfahren im Abschnitt 3.2 an, die mit einer Gegenüberstellung der in Betracht gezogenen Verfahren abgeschlossen wird. Die künstlichen neuronalen Netze (KNN), das Verfahren mit dem offensichtlich größten Potenzial, werden am Ende des Kapitels im Detail beschrieben und deren Adaption an die hier vorliegende Problematik wird untersucht. Die praktische Umsetzung der KNN mit Hilfe der Software Matlab und der darin enthaltenen KNN-Toolbox wird im Anhang A ausführlich erläutert. Damit dient dieser Abschnitt als wichtige theoretische Vorarbeit und als Anleitung für das anschließende Kapitel, in dem die Eignung der KNN an Beispielen evaluiert wird.

Die praktische Anwendbarkeit der KNN zur Modellierung von Qualität in Prozessen wird im Kapitel 4 an zwei Beispielen aus dem geodätischen und dem verkehrstechnischen Umfeld untersucht. Zunächst dient ein simuliertes Beispiel dazu, die Möglichkeiten zur Darstellung und Abbildung von einzelnen verschiedenen Typen von Qualitätsparametern zu testen. Insbesondere konnten hier erste praktische Erfahrungen in der Netzdimensionierung und der erforderlichen Anzahl an Trainingsbeispielen gewonnen werden. Diese Erfahrungen konnten im zweiten praktischen Beispiel anhand realer Daten umgesetzt werden, welches im Abschnitt 4.2 ausführlich dargestellt ist. Bei dem Datenverarbeitungsprozess, der als zweites Beispiel diente, handelte es sich um die Generierung von Fahrzeugtrajektorien im Straßennetz aus Mobilfunkdaten. Der Qualitätsparameter *mittlere Querabweichung*, mit dem die Qualität der erzeugten Trajektorien beurteilt werden kann, wurde mit Hilfe von KNN modelliert.

Nach den detaillierten praktischen Untersuchungen zum Nachweis der Eignung der KNN, wird die neue Methode im Kapitel 5 schließlich in ein Qualitätsmanagementkonzept integriert. Das vorgestellte Konzept beinhaltet das Qualitätsmodell, mit dem die Qualität unterschiedlicher Datenarten umfassend beschrieben werden kann und erläutert die verschiedenen Messmethoden, die grundsätzlich zur Quantifizierung von Datenqualität zur Auswahl stehen. Als weiterer Kernbestandteil des Konzeptes werden Verfahren zur Analyse von Datenverarbeitungsprozessen vorgeschlagen und teilweise an einzelnen Beispielen erläutert. Dabei stehen sowohl bekannte Verfahren zur qualitativen Analyse als auch quantitative Verfahren zur Modellierung von Datenqualität zur Verfügung. Neben dem bereits bestehenden Verfahren auf Basis Boolescher Algebra wird hier nach den erfolgreichen Tests auch der Einsatz von KNN empfohlen. Die Verknüpfungen zu den beiden weiteren Bestandteilen des Konzeptes, der Qualitätssicherung bzw. -verbesserung und der Evaluierung von Datenqualität werden ebenfalls in diesem Kapitel erläutert.

2 Grundlagen

In diesem Kapitel werden alle zum Verständnis des weiteren Vorgehens erforderlichen Grundlagen erörtert. Es werden wichtige Begriffe erläutert und die Datenqualität aus den verschiedenen Sichtweisen betrachtet. Schließlich wird ein bestehendes Qualitätsmodell vorgestellt und die Datenqualität bzw. das Management von Datenqualität in den globalen Zusammenhang eines modernen Qualitätsmanagements gestellt.

2.1 Zum Begriff Datenqualität

Trotz der vermeintlichen Trivialität des Begriffs *Daten* wird hier eine kurze Betrachtung für das Verständnis aller weiteren Ausführungen als sinnvoll erachtet. Die Erläuterung und Abgrenzung der Begriffe *Daten* und *Informationen* wird in diesem Abschnitt ebenso dargestellt, wie der Zusammenhang der im allgemeinen Sprachgebrauch verwendeten Begriffe *Datum*, *Daten*, *Datensatz*, *Datenmenge* und *Datenart*. Nachfolgend wird als Grundlage für die Modellierung von Qualität in Prozessen der Datenverarbeitung (DV) ein allgemeiner Überblick über Datenarten und Arten der Datenverarbeitung im ingenieurwissenschaftlichen Umfeld gegeben. Die Überlegungen zur Datenqualität aus Kundensicht und der allgemeinen Frage, welche Kategorien von Kunden es für Daten geben kann, tragen wesentlich zum Gesamtverständnis und zur Definition des im Kapitel 5 vorgestellten Qualitätsmanagementkonzeptes für Daten bei.

2.1.1 Erläuterung der Begriffe Daten und Informationen

Die Begriffe *Daten* und *Informationen* sind nicht immer einfach voneinander zu unterscheiden, stellen jedoch zwei sehr unterschiedliche Blickwinkel dar. Als *Daten* bezeichnet man die symbolische Repräsentation eines Sachverhalts bzw. einer physikalischen Erscheinung (Capurro [1978]; Fank [2001]). Der Begriff der *Information* im ingenieurwissenschaftlichen Kontext hingegen impliziert bereits eine Bedeutung über deren reine physikalische Existenz hinaus und deutet damit auf einen mehr oder weniger definierten Verwendungszweck der vorliegenden Daten hin. Informationen sind somit kurz und prägnant formuliert *„[...] Daten mit Bedeutung“* (DIN EN ISO 9000 [2005]). Ebenfalls sehr treffend und kurz werden nach Fank [2001] „*Informationen durch den Erkenntniswert von Daten repräsentiert*“.

In der Regel handelt es sich daher bei den vom Anbieter gelieferten Daten aus Sicht der Nutzer um Informationen, da die gewünschten Daten meist einem bestimmten Zweck dienen. Aus Sicht der Anbieter bzw. Lieferanten hingegen spielt die Verwendung der Daten zunächst keine Rolle. Dies ändert sich, sobald der Lieferant spezielle Bedürfnisse der Nutzer an die Datenqualität erfüllen muss. Hier ist der Übergang zwischen den beiden Begriffen fließend. Im Rahmen dieser Arbeit erscheint es daher nicht sinnvoll, die beiden Begriffe Daten und Informationen immer streng nach ihrer, vom Kontext abhängigen Bedeutung, zu unterscheiden. Die Ausführungen und Untersuchungen zur Darstellung von Daten- bzw. Informationsqualität in Prozessen und deren Management gelten in gleicher Weise für beide Begriffe. Insbesondere impliziert die Aufstellung eines Qualitätsmodells, als Grundlage für die Modellierung der Datenqualität, bereits die weitere Verwendung der Daten zu einem bestimmten Zweck. Daher werden die Begriffe in der gesamten Arbeit synonym verwendet.

Produkte werden in der DIN EN ISO 9000 [2005] als Ergebnis von Prozessen definiert, wobei vier übergeordnete Produktkategorien (Dienstleistungen, Hardware, Software, verfahrenstechnische Produkte) unterschieden werden. In der ISO wird weiter erläutert: *„Software besteht aus Informationen [und] ist üblicherweise immateriell [...]“*. Damit können Daten ganz allgemein als Ergebnisse von Prozessen und damit als Produkte angesehen werden, die im ISO-Verständnis der Produktkategorie „Software“zuzuordnen sind.

Des Weiteren ist die korrekte Trennung und Verwendung der Begriffe für Datenbestandteile, einzelne Daten und Mengen von gleichartigen Daten wichtig zum Verständnis. Daher erfolgt hier eine kurze Definition der für die weiteren Ausführung wichtigen Begriffe:

- Datum
- Daten
- Datensatz
- Datenmenge
- Datenart

Unter einem *Datum* wird im allgemeinen Sprachgebrauch häufig ein Kalenderdatum verstanden und nicht der Singular von *Daten*, wie der lateinische Wortstamm vorgibt. Daher werden für den Singular von Daten häufig Synonyme wie Datenelement oder Datenbestandteil zur Vermeidung des Wortes Datum in diesem Zusammenhang verwendet. Der Begriff des Datenbestandteils ist jedoch nicht klar definiert, da darunter sowohl ein einzelnes Datum als kleinste Einheit einer Datenmenge verstanden werden kann, als auch ein Bestandteil eines einzelnen Datums. Der Begriff des Datenelements spielt bei der Definition von Metadatenmodellen eine wichtige Rolle. Daher wird in dieser Arbeit bewusst der Begriff *Datum* als einzelner atomarer Bestandteil von Daten verwendet. Ein Datum wird in diesem Zusammenhang oft durch eine Maßzahl mit einer Einheit dargestellt. Neben SI-Einheiten wie z. B. [m], [kg], [s], kann es sich dabei auch um Prozentangaben oder Anzahlen handeln. Ein Datum kann jedoch auch ein qualitatives Attribut wie z. B. die Farbe, der Wohnort oder die Postleitzahl sein, derartige Daten werden in dieser Arbeit jedoch nicht weiter betrachtet.

Ein *Datensatz* ist als Zusammenfassung von verschiedenartigen Daten zu einer zusammengehörigen Einheit definiert. Dies ist gleichbedeutend mit einer Karteikarte oder einer Zeile in einer Datenbank. Es gibt immer eine klare Definition des Formats und damit der notwendigen Datenfelder in dem jeweiligen Datensatz. In einer Adressdatenbank bilden z. B. jeweils Name, Vorname, Straße, Hausnummer, Postleitzahl und Ort einen Datensatz. Beim polaren Anhängen von Neupunkten an einen Tachymeterstandpunkt könnte ein Datensatz beispielsweise aus Richtungswinkel und Strecke bestehen. Konsequenterweise kann hier, im Unterschied zum Datum als atomarer Datenbestandteil, von einer molekularen Struktur gesprochen werden.

Unter *Datenmenge* wird hier eine Menge von gleichartigen Elementen verstanden. Eine Datenmenge kann sowohl eine Menge aus gleichartigen Daten (im Sinne der Mehrzahl von Datum), als auch aus Datensätzen darstellen.

Zur Unterscheidung verschiedener Arten von Daten wird konsequent der Begriff *Datenart* verwendet. Der Begriff Datentyp bleibt dem Umfeld der Informatik vorbehalten und bezeichnet eine zulässige Menge von Werten zusammen mit den darauf definierten mathematischen Operationen wie z. B. integer oder float (nach Fischer und Hofer [2008]).

Insbesondere im Hinblick auf die Beschreibung der Qualität von Daten spielt die strenge Unterscheidung und bewusste Verwendung dieser Begriffe eine wesentliche Rolle. Naturgemäß können einige Aspekte der Datenqualität tatsächlich nur für den Plural des Begriffs Datum sinnvoll sein (z. B. die Vollständigkeit) und sich andere wiederum auch oder nur zur Beschreibung eines einzelnen Datums eignen.

Beispielsweise kann sich die Korrektheit sowohl auf ein Datum als auch auf einen Datensatz beziehen. Grundsätzlich ist jedoch -sofern es sinnvoll erscheint- die Zusammenfassung derartiger, auf einzelne Daten bezogener Qualitätsparameter auf größere Datenmengen möglich. So kann die Korrektheit einzelner, singulär beurteilter Messungen über den Ort oder die Zeit, zu einem Parameter zusammengefasst werden (z. B.: *95 % der Messungen in den letzten 24 h waren korrekt*).

2.1.2 Datenarten

Der Begriff *Daten* ist nur sehr schwer einzugrenzen. Alle Attribute, die Eigenschaften von materiellen oder immateriellen Dingen oder Personen beschreiben und in einer Tabelle digital oder analog gespeichert werden können, stellen Daten dar. Im Lexikon der Informatik werden Daten als

> *„[...] alles, was sich in einer für die Datenverarbeitungsanlage, den Computer, erkennbaren Weise codieren, speichern und verarbeiten lässt [...]“*

bezeichnet (Fischer und Hofer [2008]). Damit wird der Begriff der Daten extrem weit gefasst und konkrete Einschränkungen können nur anwendungsbezogen erfolgen. Nach Winkler [2006] umfassen Daten ganz allgemein unter anderem sogar Programme, Dateien, Ziffern, Zahlen, Zeichen und Parameter. Zur Festlegung der Datenarten, für die das in dieser Arbeit vorgestellte Qualitätsmanagementkonzept Anwendung finden kann, werden daher zunächst die wesentlichen unterschiedlichen Sichtweisen erläutert, die in der Literatur zu finden sind, nach denen eine Einteilung von Daten erfolgen kann. Schließlich erfolgt ein grober Umriss der Datenarten, deren Qualität mit den im weiteren Verlauf der Arbeit vorgestellten Methoden beschreibbar sind, soweit dies möglich und sinnvoll ist.

Umfeld GIS und Kartographie: Im Umfeld der Geoinformationssysteme (GIS) werden Vektor-, Raster- sowie Geometrie- und Sachdaten unterschieden. Sachdaten sind dabei alle nicht-geometrischen Elemente, wie z. B. Texte, Zahlen, Messwerte, Nummern, Namen, Eigenschaften. Nach der Erfassung und damit auch der Art der Repräsentation von lagebezogenen Daten werden Vektor- und Rasterdaten unterschieden. Die relative Form und Lage der Bestandteile eines Objekts wird mit Hilfe der Geometriedaten beschrieben. Wesentliches Merkmal der Datenhaltung in GIS ist die Zusammenfassung zusammengehöriger Daten zu Objekten (z. B. Flurstück bestehend aus Geometriedaten, Sachdaten und Vektordaten) (Bill [1999]). In der Kartographie werden Geodaten oft in Raumbezugsdaten (beinhalten Geometrie, Raster- oder Vektordaten und den Raumbezug), Semantische Daten (Informationen über Art und Menge der Objekte) sowie zeitbezogene Daten wie z. B. der Zeitpunkt der Erfassung eingeteilt (Hake u. a. [2002]).

Umfeld Organisationsmanagement: Aus unternehmerischer Sicht werden drei Gruppen von Daten unterschieden: Sachdaten, Personaldaten und Auftragsdaten. Die Sachdaten oder auch Erzeugnisdaten umfassen die Eigenschaften der Produkte sowie die Betriebsmitteldaten wie z. B. das Alter der Maschinen, Instandhaltungsdaten usw. Die Personaldaten enthalten alle relevanten Daten des Personals wie Name, Anschrift oder Gehaltsgruppe. Alle wichtigen Informationen zur Bearbeitung eines Auftrags sind den Auftragsdaten zugeordnet. Darunter fallen z. B. die Auftragsnummer, Angaben über die Auftragsmengen und vereinbarte Lieferfristen (Lexis [2010]).

Umfeld Informatik, Datenbanken: Im Umfeld der Informatik werden häufig Stamm- oder Bestandsdaten und Bewegungs- bzw. Änderungsdaten unterschieden. Stammdaten ändern sich nur selten, Bewegungsdaten hingegen haben eine kurze Lebensdauer (Winkler [2006]). Beispielsweise sind Kundendaten wie Name, Geburtsdatum und Adresse als Stammdaten und eine laufende Bestellung des Kunden als Bewegungsdaten anzusehen. Weitere, sehr einfache Unterscheidungen nach

der Aussagekraft der Daten können nach Claus und Schwill [2006] durch die Einteilung in qualitative/quantitative, aktive/passive oder numerische/alphanumerische Daten erfolgen. Auch die Einteilung nach unterschiedlichem Aggregationsniveau ist üblich. So werden Zeitreihen als über die Zeit wiederholt erhobene Daten (meist mit regelmäßiger Frequenz), Querschnittsdaten als zeitlich ungeordnete Daten und Paneldaten als eine Kombination aus den ersten beiden Datenarten definiert (Gabler Verlag [2010]). Die Befragung der stets gleichen Personengruppe zu verschiedenen Zeitpunkten liefert beispielsweise Paneldaten.

Wie diese Aufstellung zeigt, unterscheiden sich die Auffassungen und Einteilungen des Begriffs *Daten* zum Teil erheblich. Die Unterteilung ist dabei sehr anwendungsgebunden und unterscheidet sich stark mit dem Umfeld, in dem die Daten betrachtet werden. Eine abschließende Definition der Daten, die mit Hilfe des in der Arbeit vorgestellten Qualitätsmanagementkonzeptes behandelt werden können, kann hier daher nicht erfolgen. Die Einschränkung auf Daten aus dem Umfeld GIS und Kartographie ist nicht sinnvoll. Die Einteilungen für Datenarten aus den Bereichen Organisation und Informatik sind sehr anwendungsbezogen und tragen daher wenig zur Eingrenzung der Datenarten bei, die mit dem im Kapitel 5 vorgestellten Konzept behandelt werden können.

Allgemein kann formuliert werden, dass das im Kapitel 5 vorgestellte Konzept zum Management von Daten, die in Prozessen der Datenverarbeitung als Eingangs- und Ausgangsdaten vorkommen können, geeignet ist. Dies sind ausschließlich numerisch beschreibbare Daten wie Messwerte oder Berechnungsergebnisse, denen oft auch eine physikalische Einheit zugeordnet ist. Datenarten mit lediglich qualitativem Charakter, das heißt reine Sachdaten, wie beispielsweise die Farbe, werden ausgeschlossen. Damit beschränkt sich die Anwendung des Konzeptes weitgehend auf geometrische Daten, wie sie im GIS- und Kartographieumfeld, der Geodäsie und in anderen ingenieurtechnischen Disziplinen auftreten. Für die im Umfeld des Organisationsmanagements und der Informatik auftretenden Datenarten ist das Konzept nur in Einzelfällen geeignet, jedoch nicht für komplette, dort beschriebene Datenarten.

2.1.3 Der Begriff Datenqualität

Die Qualität ist eines der entscheidenden Kaufkriterien für jegliche Art von Produkt. Bereits lange vor der bewussten Verwendung des Qualitätsbegriffs war die Einhaltung der spezifischen, dem Produkt zugesicherten Eigenschaften entscheidend bei einem Kaufabschluss. Im Laufe der Industrialisierung und der Automatisierung der Fertigung wurde die Qualität der Zwischen- und Endprodukte immer mehr der entscheidende Faktor für den Erfolg eines Unternehmens. Lange Zeit wurde der Begriff *Qualität* jedoch nur auf die spezifischen Eigenschaften von Produkten angewandt, die Verallgemeinerung auf Daten oder gar auf Prozesse im Unternehmen blieb aus. Hier werden Daten entsprechend der DIN EN ISO 9000 [2005] als Produkte betrachtet, daher lassen sich für Daten in ähnlicher Weise Kundenanforderungen bzw. Qualitätsanforderungen formulieren wie für Produkte im ursprünglichen Sinne.

Qualität ist im allgemeinen Sprachgebrauch ein positiv belegter Begriff und beschreibt die Eigenschaften einer Sache oder Dienstleistung. Der weltweit bekannte Werbeslogan *„Quality made in Germany“* hat den deutschen Unternehmen über Jahrzehnte hohe Gewinne garantiert, obwohl er im Grunde keinerlei Bewertung der gepriesenen Qualität enthält. Lediglich der Zusatz *„made in Germany“* impliziert eine Wertung, die nur auf der über lange Jahre entwickelten Assoziation der restlichen Welt: *made in Germany = hohe Qualität* beruht. Dabei ist nicht alles, was in Deutschland produziert und in aller Welt verkauft wird, auch von hoher Qualität. Der Begriff *Qualität* muss daher immer mit einem bewertenden Adjektiv wie gut, schlecht, hoch, niedrig usw. verwendet werden, um vergleichende Aussagen machen zu können.

Neben diesen relativen Aussagen ist deren Quantifizierung für eine universelle Beschreibung und Bewertung der Datenqualität erforderlich. Die Grundlage dazu legt die DIN EN ISO 9000 [2005] in der die Qualität kurz und prägnant definiert wird als

> *„Grad, in dem ein Satz inhärenter Merkmale Anforderungen erfüllt."*

Diese Definition enthält bereits die durch die Organisation selbst oder durch den Kunden gestellten Anforderungen, die überprüft werden müssen. Damit ist die Beschreibbarkeit von Merkmalen eine wesentliche Voraussetzung zur Beurteilung von Qualität und damit zur Beurteilung des Grades der Erfüllung aller Anforderungen. Diese Merkmale können sowohl qualitativer (z. B. Farbe oder Form) als auch quantitativer (z. B. Gewicht oder Größe) Natur sein.

Im Umfeld der Geoinfomationen wird die Produktqualität und damit im speziellen die Datenqualität in der Norm DIN EN ISO 19113 [2005] definiert als

> *„Qualität ist die Gesamtheit von Merkmalen eines Produkts bezüglich ihrer Eignung, vorgegebene und unausgesprochen enthaltene Erfordernisse zu erfüllen."*

Diese auf Produkte angepasste Definition ist immer noch sehr allgemein gehalten und gilt daher neben den Geodaten auch für jegliche andere Art von Daten. Diese Auffassung von Datenqualität wird in den weiteren Überlegungen für die hier betrachteten, numerisch beschreibbaren Arten von Daten (vgl. Abschnitt 2.1.2) zu Grunde gelegt und dient als Basis für das im Kapitel 5 vorgestellte Qualitätsmanagementkonzept für Daten.

Damit steht fest, dass auch bei Daten die Einhaltung der zugesicherten Eigenschaften mit dem Begriff *Qualität* assoziiert werden kann. Die Datenqualität kann anhand einer Reihe von Qualitätsmerkmalen umfassend beschrieben werden, wie im Abschnitt 2.2 gezeigt wird. Einige sehr anschauliche Beispiele sind im Folgenden aufgeführt:

- Die Vollständigkeit von Wanderkarten ist entscheidend bei der Orientierung.
- Die Korrektheit von Fahrplandaten des öffentlichen Verkehrs ist entscheidendes Kriterium für dessen erfolgreiche Nutzung.
- Die Verfügbarkeit von Niederschlagsmessungen ist erforderlich für die Hochwasservorhersage.
- Die Genauigkeit geodätischer Messungen ist Voraussetzung für die erfolgreiche Überwachung von Hangrutschungen.
- Die Aktualität der digitalen Kartengrundlage kann für eine korrekte Fahrzeugnavigation entscheidend sein.

Daneben spielt die Datenqualität auch beim Vergleich und dem Austausch von Daten eine entscheidende Rolle. Insbesondere für Datenlieferanten wie z. B. Vermessungsbüros oder Anbieter digitaler Karten ist eine einheitliche Beschreibung der Datenqualität innerhalb eines spezifizierten Datenmodells dringend erforderlich. Internationale Zusammenarbeiten, wie beispielsweise beim Bau eines Airbus oder bei der Installation und dem Betrieb des Satellitennavigationssystems Galileo, sind unmöglich ohne einheitliche Beschreibungen der Qualität unterschiedlichster Arten von Daten.

Der Schritt hin zu einer ganzheitlichen Betrachtung von Qualität unter Einbeziehung sämtlicher Bereiche eines Unternehmens und weg von der reinen Produktqualität, wurde erst in den letzten Jahrzehnten des 20. Jahrhunderts als notwendig erachtet. Er ist für eine reibungsfreie und effektive Zusammenarbeit zwischen Unternehmen, zwischen Unternehmen und Kunden, insbesondere jedoch innerhalb einzelner Unternehmen zwingend erforderlich.

2.1.4 Datenqualität aus Kundensicht

Die Kundenorientierung ist der erste Grundsatz eines Qualitätsmanagements nach DIN EN ISO 9000 [2005]. Dort heißt es „*Organisationen hängen von ihren Kunden ab und sollten daher [...] deren Anforderungen erfüllen und danach streben, deren Erwartungen zu übertreffen*“. Die wesentliche Rolle der Kundenzufriedenheit beim Erfolg einer Organisation lässt sich an diesem ersten Grundsatz klar erkennen. Sie ist daher auch der Antrieb und Motivation für ein umfassendes Qualitätsmanagement von Daten als Produkt.

Aus diesem Grund ist es sinnvoll, die Arten von möglichen Abnehmern von erfassten und/oder weiterverarbeiteten Daten etwas näher zu betrachten. Deren Auffassung von Qualität ist Grundlage für eine solide und dauerhafte Geschäftsbeziehung zwischen Lieferant und Abnehmer von Daten.

In der öffentlich verfügbaren Spezifikation PAS 1071 [2007] (PAS: Publicly Available Specification) wird eine Vielzahl von Nutzergruppen für Geodaten genannt. Unter anderem werden Energieversorger, Landesverteidigung, Wasserwirtschaft, Statistik, Telekommunikation und nicht zuletzt der Bürger als Konsument als Nutzergruppen aufgeführt. Die Liste ließe sich noch lange fortsetzen, da in nahezu jeder Branche Geodaten als Planungsgrundlage dienen. Durch die Verallgemeinerung der Datenarten, die mit den in dieser Arbeit vorgestellten und getesteten Methoden behandelt werden können, wird der potenzielle Nutzerkreis nochmals erweitert.

Mit Hinblick auf die umfangreiche Beschreibung von Datenqualität mit Hilfe des umfassenden Modells, wie es in Abschnitt 2.2 vorgestellt wird, ist eine sinnvolle Auswahl nach dem Umfang der für den Kunden erforderlichen Qualitätsbeschreibung zweckmäßig. Viele Qualitätsparameter, die während der Erfassung, Aufbereitung oder Verarbeitung einen hohen Informationsgehalt haben, sind für den Endnutzer bedeutungslos und/oder nicht zu verstehen. Andere Detailinformationen zur Datenqualität sind möglicherweise nur für interne Nutzer bestimmt und sollen aus Wettbewerbs- oder Datenschutzgründen den Kunden nicht übermittelt werden. Wie die Erfahrung zeigt, wollen viele Nutzer auch nur eine einfache und klare Information über die Verwendbarkeit der Daten, die beispielsweise ampelartig erfolgen kann.

Im Rahmen des Projektes Do-iT[1], welches im Abschnitt 4.2 noch näher erläutert wird, wurden im Jahr 2006 Umfragen zu dieser Problematik durchgeführt. Ziel der Erhebung war es, die Meinung der motorisierten Verkehrsteilnehmer zu Art und Umfang der TMC[2]-basierten Übermittlung von Verkehrsstörungen zu eruieren. In den als persönliche Interviews durchgeführten Befragungen an Autobahnraststätten konnte ein großes Interesse an zusätzlichen Informationen zu Staumeldungen festgestellt werden. Über 80 % der Befragten wünschten sich neben der Lage der Verkehrsstörung zusätzlich eine Information über deren Länge und Tendenz. Jedoch interessiert sich nur etwa ein Drittel für die Genauigkeit der Angaben oder deren Herkunft. Parallel zu der Befragung der Endnutzer wurde auch eine kleine Anzahl Dienstanbieter per Email befragt, deren Produkte als Eingangsdaten Verkehrsinformationen benötigen. Hier zeigte sich erwartungsgemäß ein anderes Bild. Bei den Anbietern besteht ein großes Interesse an sehr detaillierten Qualitätsinformationen (Do-iT [2006]). Die beiden Befragungen, an der etwa 150 Verkehrsteilnehmer als Endnutzer und sieben Dienstanbieter teilgenommen haben, hat deutlich gezeigt, dass der Inhalt und Umfang der mit den eigentlichen Daten mitgelieferten Qualitätsinformation stark von dem Nutzerprofil abhängt. Eine Einteilung in die folgenden Nutzer- oder Kundengruppen ist daher sinnvoll:

[1] Do-iT steht für ***D**ata **o**ptimization for **i**ntegrated **T**elematics*. Der Schwerpunkt des vom BMWi geförderten Projektes war die Entwicklung und Evaluierung von Methoden zur Generierung von Verkehrsinformationen aus Mobilfunkdaten (Ramm und Schwieger [2008]).

[2] TMC steht für **T**raffic **M**essage **C**hannel, dabei handelt es sich um kostenlos über UKW empfangbare kodierte Verkehrsinformationen, die von Autoradios ausgegeben und von onboard-Navigationsgeräten interpretiert werden können

Interne Kunden sind Kunden in der eigenen Organisation, die direkt oder indirekt an der Produktion oder Aufbereitung der Daten beteiligt sind. Diese verfügen in der Regel über die erforderliche Fachkenntnis, um alle im Qualitätsmodell aufgeführten Parameter verstehen zu können. Hier ist auch die Weitergabe interner Informationen im Regelfall nicht problematisch. Im Einzelfall ist bereits eine Selektion einzelner Parameter vorzunehmen, wenn beispielsweise einzelne Aspekte der Datenqualität dem Management vorgestellt werden sollen.

Externe Kunden sind Kunden, die die Daten weiter aufbereiten oder als Eingangswerte für die Generierung von Folgeprodukten benötigen. Diese Kunden sind meist an einer detaillierteren Qualitätsinformation interessiert, um die Fortpflanzung in die Qualität ihres Produktes besser beurteilen zu können.

Zwischenhändler und Endkunden sind die Abnehmer der fertig aufbereiteten Daten. Sie verkaufen die Daten an die Endkunden bzw. die Endkunden nutzen die Informationen für ihre Zwecke und/oder in ihren Anwendungen. Eine Weiterverarbeitung der Daten ist in der Regel nicht vorgesehen. Probleme bei der Erfassung, Herstellung, Bereitstellung der Daten interessieren diese Kunden ebenfalls nicht, daher müssen viele Qualitätsparameter, die zur Beurteilung der genannten Arbeitsschritte erforderlich sind, nicht an diese Kunden weitergegeben werden.

In der Regel kann der Grad der Detaillierung der Qualitätsinformation durch die Auswahl der jeweils geeigneten Parameter aus dem für das jeweilige Produkt entwickelten Qualitätsmodell geregelt werden. Es ist jedoch auch denkbar, die für den jeweiligen Kunden relevanten Aspekte der Datenqualität ganz oder teilweise in eine Art von Qualitätskennzahl (z. B. Tauglichkeit oder Verwendbarkeit) einfließen zu lassen. Diese Konzentration auf eine oder wenige Kenngrößen vereinfacht die Beurteilung der Daten. Diese Kenngröße kann beispielsweise als eine Art gewichtetes Mittel der relevanten Parameterwerte ermittelt und dann in Form einer Ampel-Information dem Kunden mitgeteilt werden.

Grundsätzlich kann der Detaillierungsgrad der Qualitätsbeschreibung auch vom Geschäftsmodell des Anbieters abhängen. Gewisse Qualitätsdetails sind möglicherweise für einzelne Nutzer von besonderem Interesse. Wenn deren Bereitstellung zusätzlichen Aufwand verursacht, wird dies in der Regel an den Kunden weitergegeben.

Im Folgenden wird zunächst ein bestehendes Modell zur Beschreibung von Datenqualität vorgestellt, bevor kurz auf die Entwicklungsgeschichte des Begriffs „Qualitätsmanagement“bis hin zur Entstehung moderner Qualitätsmanagementkonzepte eingegangen wird.

2.2 Ein Qualitätsmodell für verkehrstechnische Anwendungen

Die Bemühungen nach einer Vereinheitlichung der Qualitätsbeschreibung von Daten führten unter anderem zu dem von Wiltschko [2004] vorgestellten und später von Wiltschko und Kaufmann [2005] weiterentwickelten Qualitätsmodell, welches ursprünglich aus sieben inhärenten Merkmalen besteht, die alle wesentlichen Aspekte der Datenqualität abdecken. Eine weitere Konkretisierung durch den Merkmalen zugeordnete Parameter ist erst nach Kenntnis der zu beschreibenden Datenart mit ihren spezifischen Eigenschaften sinnvoll. Das Modell ist grundsätzlich geeignet, alle wesentlichen Aspekte der Datenqualität umfassend zu beschreiben. Die Verallgemeinerung von den an Fahrerassistenzsystemen beteiligten Datenarten hin zu diversen andersartigen Daten, wurde dabei von Wiltschko zum Teil bereits berücksichtigt. Die Tabelle 2.1 zeigt die Merkmale, welche das Grundgerüst des Qualitätsmodells darstellen. Die detaillierten Vorüberlegungen zur Herleitung des Modells sind in Wiltschko [2004] beschrieben. Hier wird jedoch zu Gunsten der Allgemeingültigkeit bereits auf die Unterscheidung von metrischer und semantischer Genauigkeit verzichtet, die insbesondere im GIS-Umfeld besondere Bedeutung hat. Die Einteilung

in die verbleibenden sechs Merkmale, die in der linken Spalte dargestellt sind, hat sich in mehrjähriger Praxis als geeignet erwiesen, um die Qualitätsaspekte von Daten umfassend darzustellen. In der rechten Spalte sind die Merkmale in Kurzform definiert. Die Merkmale sind in die Zuverlässigkeit, die Integrität und die Genauigkeit beschreibende Merkmale gruppiert.

Tabelle 2.1: Definition der Qualitätsmerkmale für Informationen (nach Wiltschko [2004])

Die Zuverlässigkeit beschreibende Merkmale		
Verfügbarkeit **Availability**	VE	Gibt das Ausmaß des Vorhandenseins der Information zu einem definierten Zeitpunkt an einem bestimmten Ort an
Aktualität **Timeliness**	AK	Gibt das Ausmaß der Übereinstimmung der Information mit der sich zeitlich ändernden konzeptionellen Realität an
Die Integrität beschreibende Merkmale		
Vollständigkeit **Completeness**	VO	Gibt das Ausmaß des Vorhandenseins sämtlicher zur Beschreibung der konzeptionellen Realität erforderlichen Informationen an
Konsistenz **Consistency**	KO	Gibt das Ausmaß der Übereinstimmung der Information mit dem Informationsmodell an
Korrektheit **Correctness**	KR	Gibt das Ausmaß der Übereinstimmung der Information mit der konzeptionellen Realität bei vorausgesetzter Aktualität an
Die Genauigkeit beschreibendes Merkmal		
Genauigkeit **Accuracy**	GE	Gibt den Zusammenhang zwischen dem ermittelten (meist gemessenen) und dem wahren bzw. plausibelsten Wert an

Es wurde hier, trotz einer teilweise abweichenden Sichtweise des Autors im Bezug auf die Definitionen der Merkmale, an den ursprünglichen Definitionen festgehalten, da auch das bestehende Analyseverfahren, welches im Kapitel 3.1 vorgestellt wird, auf dieser Sichtweise basiert. Insbesondere die Definition der Aktualität und die Voraussetzung dieser bei der Definition der Korrektheit sind aus Sicht des Autors diskussionswürdig. Am Ende des Kapitels erfolgt eine kurze kritische Beurteilung des vorgestellten Modells.

Um Einschränkungen zu vermeiden, wurden die Definitionen der Qualitätsmerkmale bewusst sehr allgemein gehalten. Damit wird die Allgemeingültigkeit des Modells zur Beschreibung verschiedenartiger Daten erhalten. In Anlehnung an die ausführlichen und umfassenden Beschreibungen der Merkmale in der Arbeit von Wiltschko werden die Merkmale im Folgenden noch genauer definiert. Dies ist eine notwendige Voraussetzung zum Verständnis der weiteren Ausführungen und zum Nachvollziehen der Schlussfolgerungen im Verlauf der gesamten Arbeit. Das Qualitätsmodell, bestehend aus den sechs Merkmalen und den datenindividuell definierten Parametern, bildet damit die Basis für alle weiteren Argumentationsketten und für die Einbettung und Behandlung der Datenqualität in einem umfassenden, unternehmerischen Qualitätsmanagementkonzept.

Nachfolgend werden die Eigenschaften der einzelnen Qualitätsmerkmale in Anlehnung an die Erläuterungen in der Arbeit von Wiltschko noch weiter ausgeführt und durch die Hinzunahme von Qualitätsparametern an Beispielen veranschaulicht.

Verfügbarkeit (VE): Die **Verfügbarkeit** beschreibt die Existenz der Daten *an einem bestimmten Ort zu einem bestimmten Zeitpunkt*. Hier wird die Verfügbarkeit mit der **technischen Verfügbarkeit** eines DV-Systems wie z. B. eines Sensors oder eines Datenservers gleichgesetzt. Ist das betrachtete System zur Zeit nicht verfügbar, so gilt das auch für die zugehörigen Ausgangsdaten in Form von DV-Ergebnissen oder gespeicherten Daten. Die Beurteilung der Verfügbarkeit unter Berücksichtigung des Zeitaspekts ist ein Unterscheidungsmerkmal zur Vollständigkeit von Daten, die ebenfalls die Existenz von Daten oder einzelnen Datensätzen beschreibt. In der Regel dient die Verfügbarkeit jedoch nur der Beschreibung der Qualität der übergeordneten Datenebene und wird nicht für einzelne Datensätze angegeben. Sie bezieht sich vielmehr auf den eigentlichen Datenstrom von der Quelle bzw. dem Anbieter bis hin zum Nutzer der Daten.

Aktualität (AK): Das zweite, zeitabhängige Qualitätsmerkmal neben der Verfügbarkeit stellt die **Aktualität** dar. Die Aktualität dient der Beschreibung der *anwendungsbezogenen Bedeutung der Information in der Gegenwart.* Zur Beurteilung der Aktualität ist immer eine Anwendung erforderlich, hinsichtlich derer Informationen als aktuell oder veraltet angesehen werden können. Abstufungen sind auch möglich und unter Umständen sinnvoll.
Wiltschko hat die Aktualität bei neu generierten Ausgangsdaten per Definition immer als erfüllt angesehen und betont die Bedeutung der Aktualität bei der Speicherung von Daten und Informationen in Datenbanken. Die Merkmale Verfügbarkeit und Aktualität werden bei Wiltschko als Zuverlässigkeitsmerkmale zusammengefasst, da sie den zeitbezogenen Aspekt der Qualität betrachten.

Vollständigkeit (VO): Zusammen mit der Konsistenz und der Korrektheit stellt die **Vollständigkeit** eines der drei Integritätsmerkmale für Daten dar. Sie ist ein *Maß für das Vorhandensein aller zur Beschreibung der konzeptionellen Realität erforderlichen Daten* (Beispiel: Zur Prognose der Verkehrslage sind die Daten aller Montage des Jahres erforderlich, fehlen einzelne Tage, so ist die Vollständigkeit verletzt). Die Vollständigkeit kann nur zur Beschreibung von Daten oder Datensätzen angewandt werden und nicht auf ein einzelnes Datum. Ist ein einzelnes Datum nicht vollständig, so liegt laut Definition ein Mangel in der Konsistenz vor. Zusätzlich muss zwischen der Vollständigkeit und der Konsistenz eines Datensatzes unterschieden werden. Sofern das Datenmodell die einzelnen Bestandteile eines Datensatzes explizit vorgibt, wird das Fehlen eines oder mehrerer Bestandteile als Mangel in der Konsistenz betrachtet, andernfalls ist der Datensatz nicht vollständig.

Konsistenz (KO): Als Bestandteil der Integritätsmerkmale wird die **Konsistenz** als *Grad der Übereinstimmung der Daten mit dem Datenmodell* interpretiert. Dabei kann das Datenmodell sowohl Wertebereich oder Format eines einzelnen Datums definieren als auch die Bestandteile eines Datensatzes vorgeben (zum Beispiel muss eine Position aus X- und Y-Koordinate bestehen und darf sich nur in einem gewissen Intervall bewegen). Die Festlegung eines Datenmodells ist somit die Grundlage zur Beurteilung der Datenkonsistenz. Die Wahrung der Konsistenz von Daten wird in der Praxis vielfach durch die Erarbeitung von Normen erleichtert (Beispiel: GDF-Standard zur Beschreibung und zum Austausch von digitalen Straßendaten).

Korrektheit (KR): Die **Korrektheit** ist ein *Maß für die Übereinstimmung der Daten mit der konzeptionellen Realität zum Zeitpunkt der Beurteilung*. Damit kann die Korrektheit klar von einer Verletzung der Aktualität getrennt werden. Der Zeitbezug spielt hier keine Rolle, es muss zur Beurteilung lediglich die konzeptionelle Realität zum Beurteilungszeitpunkt bekannt sein. Witschko hat hier die Aktualität stets vorausgesetzt.
Auf der Bewertungsebene, die der reinen Beschreibung folgt, kann in einigen Fällen durch Beurteilung anhand eines Schwellwertes die Genauigkeit in eine Beschreibung der Korrektheit umgewandelt werden (Beispiel: Liegt die Standardabweichung einer ermittelten Position innerhalb

eines Grenzwertes, den die Anwendung vorgibt und kann eine systematische Abweichung ausgeschlossen werden, so ist die Position korrekt).

Genauigkeit (GE): Die **Genauigkeit** gibt den quantitativen Zusammenhang einer als Zahlenwert darstellbaren Größe mit deren Erwartungswert an. Die Genauigkeit kann im Allgemeinen nur von gemessenen oder berechneten Größen angegeben werden. Grundsätzlich ist der angebbare Zahlenwert immer mit einer Einheit verknüpft, die ebenfalls angegeben werden muss.

Dieses Qualitätsmodell wurde allgemeiner formuliert, als es später in der ISO TS 19138 [2006] bzw. in der PAS 1071 [2007] erfolgt ist. Jedoch ähneln sich die Modelle in weiten Teilen. Die Merkmale werden in den Spezifikationen mit Elementen und die Parameter mit Subelementen bezeichnet. Insgesamt werden in dem ISO/PAS-Modell nur fünf Merkmale definiert, die Verfügbarkeit wird dort bereits vorausgesetzt. Alle weiteren Merkmale finden sich in den Elementen und Subelementen des Modells wieder. Bei beiden Modellen handelt es sich um offene Modelle, die um weitere geeignete Parameter bzw. Subelemente erweitert werden können, lediglich die Merkmale, respektive die Datenelemente stehen bei beiden fest. Die Qualitätselemente des ISO/PAS-Modells sind jedoch bereits auf Geodaten im engen Kontext der Geoinformationssysteme zugeschnitten, daher wird in dieser Arbeit an dem von Wiltschko und Kaufmann entwickelten, allgemeineren Modell festgehalten.

Die Anpassung des universellen Qualitätsmodells an einzelne Datenarten erfolgt mit Hilfe geeigneter Qualitätsparameter, die das Modell konkretisieren. Grundsätzlich kann jedes Qualitätsmerkmal dabei durch beliebig viele Parameter beschrieben werden. Die Definition der geeigneten Parameter erfolgt dabei in der Regel nach Bedarf und auf Anregung aller beteiligten Parteien. Damit entsteht für jede vorkommende Datenart ein eigenes Qualitätsmodell, welches bei Bedarf jederzeit um weitere Qualitätsparameter erweitert werden kann. Die Parameter stellen immer quantitative Beschreibungen dar, daher ist zu jedem Parameter eine geeignete Messmethode erforderlich. Diese erläutert, wie der entsprechende Parameter bestimmt werden kann. Eine allgemeine Zusammenstellung der verschiedenen Messmethoden ist im Kapitel 5.2 dargestellt, in dem das Qualitätsmodell als Bestandteil eines umfassenden Qualitätsmanagementkonzepts für Daten nochmals kurz Erwähnung findet.

Zur Anwendung des Qualitätsmodells auf konkrete Daten erfolgte bei Wiltschko zusätzlich eine Einteilung der Merkmale gemäß der Reihenfolge, in der eine Beurteilung stattfinden sollte, in primäre (VE, VO, KO), sekundäre (KR und AK) und tertiäre Qualitätsmerkmale (GE). Da die Verfügbarkeit und die Vollständigkeit maßgeblich die Existenz der Information betreffen, genießen Sie die höchste Priorität ebenso wie die Konsistenz, da bei einer Konsistenzverletzung ein Informationsverlust droht (Wiltschko [2004]). Die Korrektheit der Daten, und dieser nachgeordnet die Aktualität, bilden nach Wiltschko die Sekundärmerkmale. Erst wenn alle anderen Merkmale (positiv) beurteilt wurden, spielt die Genauigkeit der Informationen eine Rolle. Diese logische Reihenfolge ist insbesondere bei der Anwendung des Analyseverfahrens, welches im Abschnitt 3.1 im Detail vorgestellt wird, von Bedeutung.

Für die weiteren Ausführungen sind die exakten Definitionen der Qualitätsmerkmale nicht zwingend erforderlich. Daher wird an den Definitionen des vorgestellten und am Institut für Anwendungen der Geodäsie im Bauwesen entwickelten Modells festgehalten. Dennoch muss hier erwähnt werden, dass einige Definitionen aus heutiger Sicht überarbeitet werden müssten. Die Definition der Aktualität, wie sie in Tabelle 2.1 dargestellt ist, lässt den aktuellen Zeitbezug vermissen. Passender wäre hier beispielsweise das Alter einer Information als Zeitraum von der Entstehung bis zum Beurteilungszeitpunkt zu definieren. Dies würde der allgemeinen Auffassung des Begriffes „Aktualität“entsprechen und ließe eine Interpretation in Abhängigkeit von Art und Anwendung der Daten zu. Damit ist auch die Definition, die Aktualität für Ausgangsdaten immer auf $100\,\%$ zu setzen, nur beim Prozess der Datenerfassung sinnvoll.

Des Weiteren ist die Voraussetzung von Aktualität zur Definition der Korrektheit nicht erforderlich. Die Korrektheit wird immer durch den Vergleich der zu beurteilenden Ist-Daten mit geeigneten Soll-Daten erzielt. Damit ist die Korrektheit nach Auffassung des Autors getrennt von dem Merkmal Aktualität zu betrachten. An dem Ergebnis des Soll-Ist-Vergleichs ändert die Zeitspanne seit dessen Durchführung nichts.

Die Überarbeitung der Merkmalsdefinitionen des Qualitätsmodells bleibt jedoch weiteren Arbeiten vorbehalten und hat für diese Arbeit keine Relevanz.

2.3 Charakterisierung von Prozessen in der Datenverarbeitung

Mit Hinblick auf die Modellierung von Datenqualität in Prozessen ist die Identifizierung möglicher Arten der Datenverarbeitung (DV) sinnvoll, wie sie in technischen Systemen auftreten können. Datenverarbeitung wird in Wahrig [2002] allgemein als Begriff aus der EDV aufgefasst und mit

> „*Sammeln, Sichten, Speichern, Bearbeiten u. Auswerten von Informationen, die als Größen u. Werte miteinander in Beziehung gesetzt werden können*“

erläutert. Damit werden bereits schon einige aus der Geodäsie bekannte, wichtige Arten der Datenverarbeitung, die Datenerfassung (*Sammeln*) und Speicherung, die Prüfung/Kontrolle (*Sichten*) sowie Evaluierung und Interpretation der Daten (*Auswerten*) konkret genannt. Alle weiteren Verarbeitungsarten werden unter dem Begriff *Bearbeiten* zusammengefasst. In Wiltschko [2004] werden die Arten der DV als *Applikationsformen* bezeichnet und in die folgenden sechs Gruppen eingeteilt, in denen sich sinngemäß ebenfalls die Begriffe der oben dargestellten Definition wiederfinden:

- Erfassung
- Verarbeitung
- Verwendung
- Übertragung
- Verzweigung
- Kontrolle

Diese Aufteilung wurde im Hinblick auf das von Wiltschko entwickelte Verfahren zur Analyse von Informationsqualität in Prozessen vorgenommen, welches im Kapitel 3.1 im Detail vorgestellt wird. Mit Ausnahme der Verwendung ziehen alle DV-Arten auch Veränderungen der Daten oder zumindest der Datenqualität nach sich. Welche Auswirkungen die Applikationsformen auf die Datenqualität im Einzelnen haben, und wie die Darstellung im sogenannten Informationsflussdiagramm erfolgt, wird ebenfalls im Kapitel 3.1 erläutert.

Die folgende Abbildung 2.1 gibt einen plakativen Überblick über die Vielzahl der möglichen DV-Arten. Teilweise ergeben sich Überschneidungen der Begriffe. Diese sind jedoch im Rahmen dieser Übersicht nicht von Bedeutung. Die diversen Verarbeitungsarten wurden bereits sinnvoll gruppiert. Die Grafik erhebt keinen Anspruch auf Vollständigkeit, je nach Umfeld und Datenarten sind eine Vielzahl weiterer Verarbeitungsarten denkbar.

Die primäre Datenerfassung beinhaltet neben der primären digitalen oder analogen Erfassung von Messwerten oder Phänomenen auch die Speicherung und Sicherung der Daten durch Kopieren. Die Sichtung der Daten beinhaltet sowohl deren Ordnung (Gruppierung, Katalogisierung, Klassifikation) als auch deren Abfrage oder Kontrolle (Evaluierung, Prüfung). Die Bearbeitung von Daten kann auf viele Arten erfolgen und beispielsweise die Datenveredelung, die Verkleinerung der Datenmenge oder deren Verschlüsselung zum Ziel haben. Daten werden häufig ausgetauscht und mit anderen Daten verknüpft, müssen interpretiert und schließlich präsentiert werden. Die beispielhaft in der Abbildung dargestellten Verarbeitungsformen können alle Bestandteile eines DV-Systems sein. Handelt es sich dabei um ein Geoinformationssystem, so erfolgt die Gliederung der DV-Arten in der Regel nach dem EVAP-Modell. Dabei werden die verschiedenen Tätigkeiten und Aufgaben, dem typischen Arbeitsablauf in einem GIS

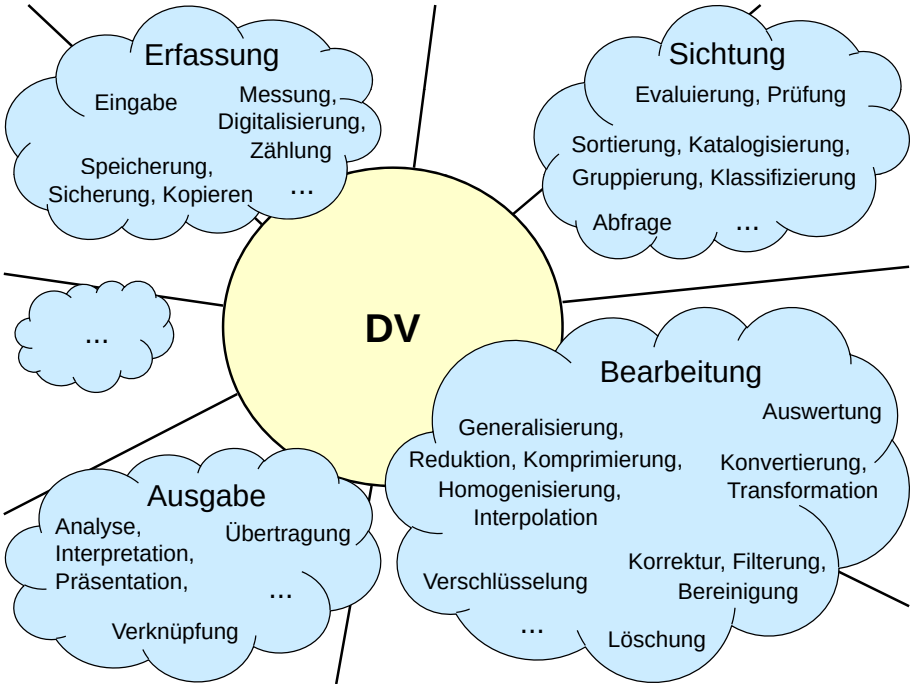


Abbildung 2.1: Darstellung von typischen Arten der Datenverarbeitung

folgend, nach den wesentlichen Bereichen **E**rfassung, **V**erwaltung, **A**nalyse und **P**räsentation gruppiert (Bill und Zehner [2001]).

Daneben sind noch einige weitere Gruppierungsmöglichkeiten denkbar, entscheidend hinsichtlich der Modellierung von Datenqualität ist jedoch die Einteilung in die folgenden drei Gruppen:

Qualitätsneutrale DV: Diese Arten der Datenverarbeitung nehmen keinen Einfluss auf die Qualität der Daten. In der Einteilung von Wiltschko [2004] entspricht diese Gruppe der *Verwendung* von Daten, allgemein umfasst sie jedoch wesentlich mehr Arten der DV. Durch sortieren und katalogisieren, abfragen oder verschlüsseln sowie präsentieren von Daten, ändert sich deren Qualität ebenfalls nicht.

Datenneutrale, jedoch qualitätsändernde DV: Diese Gruppe umfasst alle DV-Arten, bei denen die eigentlichen Daten nicht manipuliert werden, sich jedoch die Qualität der Daten durch die Maßnahme ändert. Dies kann beispielsweise die Speicherung oder redundante Sicherung von Daten sein, welche die Verfügbarkeit der Daten ändert. Auch die Überprüfung von Daten ändert deren Zuverlässigkeit grundlegend, ohne die Daten zu verändern.

Daten- und damit auch qualitätsändernde DV: In diese Gruppe fallen schließlich alle weiteren DV-Arten, die eine Änderung in Art oder Menge der Daten bewirken und damit automatisch auch eine Änderung der Qualität nach sich ziehen. Beispiele sind die Komprimierung, Korrektur oder das Löschen von Daten.

Eine über diese drei Gruppen hinausgehende, detailliertere Einteilung der DV-Arten wird an dieser Stelle nicht für notwendig erachtet. Im Kapitel 3.1 wird das von Wiltschko [2004] entwickelte Verfahren zur Analyse der Informationsqualität vorgestellt. Dieses beruht auf der zu Beginn des Kapitels dargestellten Einteilung in sechs Applikationsgruppen. Für alle weiteren Überlegungen wird jedoch bis auf Weiteres die Einteilung in die eben genannten drei DV-Arten herangezogen.

2.4 Einführung in Qualitätsmanagementsysteme

„*Ein Weg zum systematischen und erfolgreichen Führen und Betreiben einer Organisation kann die Einführung und Aufrechterhaltung eines Managementsystems sein*“(DIN EN ISO 9000 [2005]). Neben anderen Managementdisziplinen spielt dabei das Qualitätsmanagement eine wichtige Rolle. Wesentliches Ziel der Einführung und Pflege eines Qualitätsmanagementsystems (QM-Systems) ist dabei laut ISO die Erhöhung der Kundenzufriedenheit und damit verbunden der Erfolg einer Organisation. Die ISO-9000-Familie beschreibt dabei nur allgemein Mindestanforderungen an moderne QM-Systeme und es wird ein prozessorientierter Ansatz eines QM-Systems vorgeschlagen (DIN EN ISO 9001 [2008]). Konkrete Anforderungen an die Qualität von Produkten stellt sie hingegen nicht, stattdessen wird auf behördliche Vorgaben, Produktnormen und Vertragsvereinbarungen verwiesen.

Ganz allgemein wird ein QM-System in der DIN EN ISO 9000 [2005] als „*Managementsystem zum Leiten und Lenken einer Organisation bezüglich der Qualität*“definiert. Diese Definition betont die ganzheitliche Sicht und umfasst sämtliche Tätigkeiten, die die Qualität innerhalb und außerhalb einer Organisation tangieren. Es existieren eine Vielzahl an unterschiedlichen Modellen, Systemen und Methoden im Bereich des QM, „die gerade demjenigen, der sich in die Thematik einarbeiten möchte, die Orientierung nicht leicht macht“(Zollondz [2002]). Daher hat Zollondz ein Urmodell, das als kleinster, gemeinsamer Nenner für QM-Systeme zu verstehen ist, aufgestellt. Dieses Urmodell enthält die folgenden sieben QM-Elemente, die implizit Bestandteil jedes QM-Systems sind und daher mit **Conditio-Sine-Qua-Non-QM-Modell** (frei übersetzt: Bedingungen, ohne die kein QM-Modell Bestand hat) bezeichnet werden:

- Kunden
- Mitarbeiter
- Management
- Ressourcen
- Prozesse
- Messung & Analyse
- Verbesserung

Insbesondere die Gewichtung der einzelnen Bestandteile untereinander sowie die verwendeten Methoden und die abweichenden Zielvorgaben unterscheiden die einzelnen QM-Systeme.

Im Folgenden werden neben dem von der ISO vorgeschlagenen Management, welches als Mindestanforderung für die Zertifizierung eines QM-Systems angesehen werden kann, zwei weitere ganzheitliche Systeme vorgestellt und deren wesentliche Grundsätze kurz erläutert. Eine ausführliche Darstellung, Beurteilung und Gegenüberstellung der einzelnen Systeme kann hier aus Platzgründen nicht erfolgen, stattdessen wird auf die weiterführenden Quellen verwiesen. Neben den aufgeführten gibt es noch weitere QM-Systeme, die in der Praxis eine Rolle spielen. In dieser Arbeit soll gezeigt werden, dass das vorgestellte Konzept zum Management von Datenqualität in ein ganzheitliches, unternehmerisches Qualitätsmanagement eingebettet werden kann.

Als weiterführende Literatur zu verschiedenen QM-Systemen können u. a. Kamiske und Brauer [2008] und Zollondz [2002] empfohlen werden. Toutenburg und Knöfel [2008] beschreiben sehr ausführlich die 6σ-Methode (vgl. Abschnitt 2.4.3) und stellen diese anderen Managementsystemen gegenüber; in Rothlauf [2001] findet sich eine umfassende Darstellung des Total Quality Management (TQM), welches in Abschnitt 2.4.2 kurz vorgestellt wird.

2.4.1 Prozessorientierter Ansatz nach ISO 9000ff

Im Jahr 1987 erstmals erstellt, wurde die ISO 9000 mehrmals überarbeitet und steht nun in der vorläufig letzten Fassung als DIN EN ISO 9000 [2005] zur Verfügung. Sie ist als Einführung in das Gebiet des Qualitätsmanagements zu verstehen, beschreibt Grundlagen und erläutert Begriffe im Umfeld der Qualitätsmanagementsysteme (Kamiske und Brauer [2008]).

Mit der DIN EN ISO 9001 [2008] werden allgemeine Grundsätze und Mindestanforderungen an ein modernes Qualitätsmanagement einer Organisation gestellt. Diese Grundsätze sind allgemein anerkannt und genormt, daher ist eine Zertifizierung nach ISO 9000 möglich. Neben der Funktion als Checkliste bzw. Leitlinie bei der Einführung eines QM-Systems in einer Organisation können die gestellten Mindestanforderungen wiederum als ein mögliches QM-System verstanden werden. Das grundlegende Modell eines modernen prozessorientierten Qualitätsmanagementsystems, bei dem die Kundenforderungen im Mittelpunkt stehen, ist in der Abbildung 2.2 zu sehen.

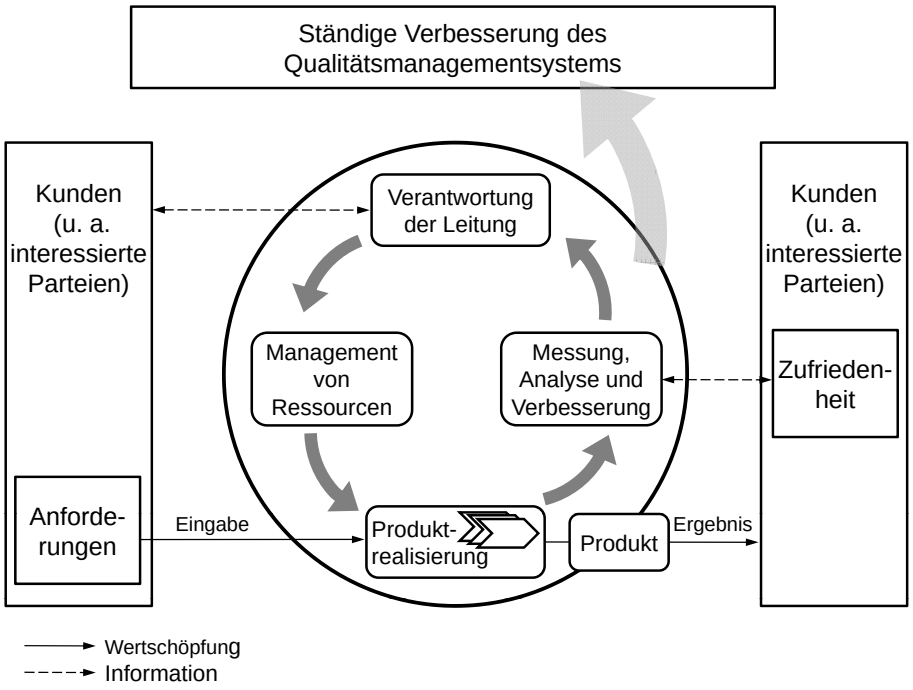


Abbildung 2.2: Prozessorientiertes QM-System nach DIN EN ISO 9000 [2005]

Zentraler Antrieb für das Streben nach ständiger Verbesserung ist die Zufriedenstellung der Kunden. Dazu ist die Erfassung und Eingabe der Kundenanforderungen in den Prozess der Produktrealisierung erforderlich. Neben der Kundenorientierung wurden sieben weitere Grundsätze formuliert, die eine Steigerung der Leistungsfähigkeit zum Ziel haben. Dabei spielt neben der Kundenzufriedenheit auch die Identifikation der Mitarbeiter mit der Organisation eine wesentliche Rolle. Grundsätzlich ist die Einbeziehung aller direkt oder indirekt beteiligten Personen anzustreben.

Die vier zentralen Bestandteile des Zyklus in Abbildung 2.2 sollen die Prozessverknüpfungen verdeutlichen und sind wie folgt zu verstehen:

Verantwortung der Leitung: Das Management muss die Qualitätspolitik festlegen und die Ziele sowie die Bedeutung des QM-Systems an die Mitarbeiter vermitteln. Sie muss eine regelmäßige Bewertung des QM-Systems bzw. des Stands der Umsetzung durchführen, dokumentieren, vermitteln und ggf. Maßnahmen ergreifen.

Management der Ressourcen: Das Management muss insbesondere die personellen und finanziellen Ressourcen zur Einführung und zur Pflege eines QM-Systems bereitstellen.

Produktrealisierung: Die Produktrealisierung muss im Einklang mit den Qualitätszielen und den Anforderungen der Kunden erfolgen.

Messung, Analyse und Verbesserung: Mit Hilfe interner Audits kann eine Überprüfung der Zielvorgaben erfolgen. Dabei müssen sowohl die Prozesse selbst als auch die Produkte auf die Einhaltung der Forderungen hin überprüft werden. Festgestellte Fehler oder nicht erfüllte Forderungen müssen behoben, bzw. Verbesserungsmaßnahmen entwickelt werden, die im Rahmen der nachfolgenden Zyklen umgesetzt werden können.

Da aus den in der Praxis gewonnen Erfahrungen neben den einzelnen Prozessen - und damit den Produkten und/oder Dienstleistungen - auch die Prozessverknüpfungen innerhalb der Organisation überprüft, beurteilt und ggf. verbessert werden können, wird auch das QM-System ständig verbessert.

Eine detaillierte Darstellung und Dokumentation aller in der Organisation vorhandenen Prozesse ist ebenfalls zur Zertifizierung nach ISO erforderlich. Dabei wird ausdrücklich auf die Möglichkeit der Darstellung einzelner Prozesse auf Grundlage des PDCA-Zyklus hingewiesen. Die PDCA-Methode besteht aus den sich immer wiederholenden Tätigkeiten **P**lan-**D**o-**C**heck-**A**ct und wird auf Deming [1982] zurückgeführt und daher auch mit Deming-Zyklus bezeichnet. Damit werden die ständige Planung neuer Verbesserungsmaßnahmen (z. B. durch neue oder verbesserte Prozesse), deren exemplarische Umsetzung und Überprüfung sowie, bei erfolgreichen Tests, die finale Einführung der Maßnahmen als ständig wiederkehrenden Tätigkeiten zur kontinuierlichen Verbesserung der Qualität in den Vordergrund gestellt.

2.4.2 Total Quality Management - TQM

Das TQM gilt als die umfassendste Strategie des Qualitätsmanagements einer Organisation. Das langfristige Ziel ist dabei die nachhaltige Einführung der Qualität als übergeordnetes Element in Unternehmenspolitik und -kultur und damit das **Leben** dieser Strategie durch sämtliche Mitarbeiter quer durch alle Bereichen des Unternehmens (Kamiske und Brauer [2008]). Die Abbildung 2.3 zeigt sehr einfach die grundlegende Strategie, basierend auf drei gleichwertigen Bestandteilen.

Wesentliches Merkmal dieses Managementsystems ist die Involvierung aller Beteiligten. Damit sind neben den Mitarbeitern auf allen Hierarchieebenen auch Kunden, Lieferanten und die Umwelt gemeint. Innerhalb des Unternehmens wird die Qualität als Aufgabe aller Mitarbeiter gesehen und die notwendigen Freiräume zur Erfüllung dieser Aufgaben werden bereitgestellt (z. B. in Form von Teambesprechungen und Qualitätszirkeln). Voraussetzung dafür ist eine entsprechende Aus und Weiterbildung der Mitarbeiter sowie die Anerkennung guter Leistungen und Ideen.

Der Begriff *Qualität* wird im TQM mit dem Erreichen der folgenden drei globalen Unternehmensziele gleichgesetzt,

- dem langfristigen und nachhaltigen Unternehmenserfolg,
- dem Nutzen für die Mitglieder der Organisation und
- dem Nutzen für die Gesellschaft.

Alle Anforderungen, die nach der DIN EN ISO 9001 [2008] an ein modernes (zertifizierbares) und umfassendes Qualitätsmanagementsystem gestellt werden, sind notwendige, jedoch nicht hinreichende Bedingungen zum Erreichen dieser Ziele. Das TQM stellt darüber hinaus noch höhere Ansprüche an alle Beteiligten in einem Unternehmen. Neben dem hier vorgestellten, eigenständigen TQM wird der Begriff des öfteren in seiner eigentlichen Bedeutung als „*umfassendes Qualitätsmanagement*“verwendet

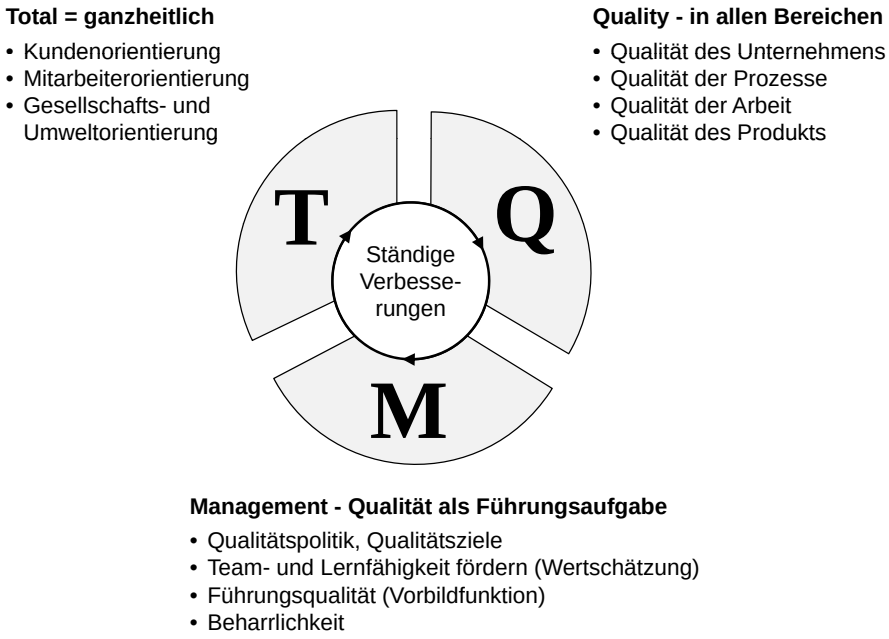


Abbildung 2.3: Grundlegende Strategie des TQM nach Kamiske und Brauer [2008]

und steht damit auch stellvertretend für jedes ganzheitliches, organisationsweites QM-System.

Eine sehr detaillierte Darstellung des TQM in Theorie und Praxis findet sich unter anderem in Rothlauf [2001].

2.4.3 Six-Sigma-Methode

Bei der Six-Sigma-Methode (oder auch 6σ-M.) liegt die oberste Priorität auf der Null-Fehler-Theorie für alle Prozesse. Hauptziel ist es auch hier, Kundenbedürfnisse vollständig und profitabel zu erfüllen, dabei wird insbesondere Wert auf die Messbarkeit von Prozessen gelegt (Toutenburg und Knöfel [2008]). Der Ausdruck „6σ“stammt dabei aus der Wahrscheinlichkeitsrechnung und gibt den Bereich an, in dem

$$\Phi(6) - \Phi(-6) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot \left(\int_{-\infty}^{6} -e^{-\frac{z^2}{2}} dz - \int_{-\infty}^{-6} -e^{-\frac{z^2}{2}} dz \right) = 99.99999990\,\% \tag{2.1}$$

der Realisierungen einer standardnormalverteilten Zufallsvariablen liegen. Dabei bezeichnet 1σ die Standardabweichung der normalverteilten Zufallsvariable und damit das Intervall um den Erwartungswert μ, der $68.3\,\%$ der Realisierungen enthält. Allerdings ist bei der Angabe des σ-Levels bereits eine erfahrungsgemäß bei realen Prozessen einzukalkulierende, langfristige Drift des Erwartungswertes μ enthalten. Diese Drift kann nach Evans [1975] im Durchschnitt die Größenordnung von etwa $\pm 1.5\sigma$ bei realen Prozessen erreichen[3]. Diese Größenordnung hat sich eingebürgert und wird in nahezu allen weiteren Quellen verwendet (z. B. Kamiske und Brauer [2008] oder Tennant [2001]). Daraus ergeben sich z. B. beim 6σ-

[3]Eine konkrete Angabe der über die Lebensdauer beliebiger Prozesse zu erwartenden Drift ist naturgemäß sehr schwer und hängt von vielen Faktoren ab. Daher hat sich Evans eher willkürlich auf Basis einiger empirischer Untersuchungen auf den Faktor 1,5 festgelegt, der seither im Zusammenhang mit 6σ stets zugrunde gelegt wird.

Niveau für die Verteilungsfunktion der Standardnormalverteilung ($\sigma = 1$) die Grenzen $z = -4.5$ und $z = 7.5$. Hier wurde willkürlich eine Drift nach rechts und damit eine Vergrößerung des Mittelwertes angenommen, wie sie beispielsweise durch Abnutzungserscheinungen an Maschinenteilen und der damit verbundenen Vergrößerung von Spaltmaßen oder Spielen entstehen kann. Mit diesen neuen Grenzen ergibt sich die Wahrscheinlichkeit von

$$\Phi(7.5) - \Phi(-4.5) = 99.99966\,\% \tag{2.2}$$

mit der die Realisierungen innerhalb der Grenzen liegen. In gleicher Weise ergeben sich weitere gängige σ-Niveaus, die in der folgenden Tabelle 2.2 dargestellt sind. Es wurde hier zur besseren Veranschaulichung zusätzlich die maximal zu erwartende Fehlerquote pro eine Million Einheiten (ppm - parts per million) angegeben:

Tabelle 2.2: Gängige Sigma-Niveaus mit den entsprechenden Fehlerraten

σ-Niveau	Ausschuss [ppm]	Fehlerfrei [%]
1	690 000	31.0
2	308 537	69.2
3	66 807	93.32
4	6 210	99.379
5	230	99.977
6	3.4	99.99966

Das 6σ-Niveau erscheint beispielsweise für die Fertigung von Mobiltelefonen eine vollkommen ausreichende Anforderung zu sein, betrachtet man jedoch die Flugbewegungen auf dem Rhein-Main-Flughafen, so ist eine 6σ-Forderung unzulänglich. Im Jahre 2007 wurden nahezu 500 000 Flugbewegungen registriert (Frankfurter Rundschau [2008]), d. h. es müsste mit $1 - 2$ Abstürzen pro Jahr gerechnet werden. Dieses Beispiel zeigt, dass das Sigma-Niveau an das entsprechende Problem realistisch angepasst werden muss und der Begriff *Six Sigma* daher vielmehr symbolisch für eine (angemessene) Minimierung der Fehlerraten in allen Prozessen zu verstehen ist.

Die Methode stammt ursprünglich aus der Fertigungsindustrie, bei der viele einzelne Bauteile zu einem Produkt zusammengesetzt werden müssen. Wenn an den gesamten Produktionsprozess und damit an das Endprodukt ein σ-Level von vier als Anforderung gestellt wird, was einem maximalen Ausschuss von $0.621\,\%$ oder $6210\,\mathrm{ppm}$ entspricht, dann müssen die Anforderungen an die einzelnen Bestandteile bzw. Teilprozesse noch wesentlich höher sein. Die Fehlerwahrscheinlichkeiten aller Bestandteile multiplizieren sich zur Wahrscheinlichkeit, mit der ein fehlerhaftes Endprodukt entsteht. Geht man vereinfacht von unkorrelierten Einzelprozessen und der Tatsache aus, dass jeder Defekt eines Einzelteils zum Defekt des gesamten Produkts führt, dann resultiert daraus (unter der Annahme eines 6σ-Niveaus für alle Einzelteile) bereits bei $i = 19$ Teilprodukten eine Wahrscheinlichkeit von

$$p_{\mathrm{ges}} = p_i^{19} = 0.9999966^{19} = 0.9936 = 99.36\,\% \tag{2.3}$$

für ein fehlerfreies Produkt. Dies entspricht etwa dem $4\,\sigma$-Niveau und damit bereits einem maximalen fehlerhaften Anteil von $6210\,\mathrm{ppm}$. Daraus erklärt sich die zunächst völlig unrealistische Forderung nach 6σ oder $3.4\,\mathrm{ppm}$ für die Teilprozesse bzw. Teilprodukte und damit nach nahezu fehlerfreien Prozessen.

In der Praxis erfolgt die grundlegende Verbesserung von Prozessen bis hin zu quasi fehlerfreien Prozessen im Sinne des Six-Sigma meist nach dem DMAIC-Prinzip (Define, Measure, Analyse, Improve, Control), mit dem bei Bedarf einzelne Prozesse im Rahmen von kurzen Projekten (es sollte sich nur um wenige Monate handeln) optimiert werden. Diese bedarfsorientierte und fokussierte Herangehensweise ist ein wesentliches Unterscheidungsmerkmal zu anderen QM-Systemen. Bei der DMAIC-Methode werden nach (Toutenburg und Knöfel [2008]) die folgenden fünf Phasen unterschieden :

Define: Projektauswahl, Problem- und Zieldefinition

Measure: Definition und Ermittlung geeigneter Kennzahlen und Parameter im Ist-Zustand

Analyze: Analyse des Problems, Auswertung der Kennzahlen; Identifikation von Ursachen

Improve: Entwicklung und Implementierung von Verbesserungsmaßnahmen

Control: Evaluierung der Wirksamkeit der Maßnahmen

Neben der DMAIC-Methode wurden weitere spezielle Verbesserungsprozesse entwickelt, unter anderem existiert die DMADV-Methode (**D**efine-**M**easure-**A**nalyze-**D**esign-**V**erify), die der Entwicklung und Einführung neuer Prozesse dient.

2.4.4 Zusammenfassung QM-Systeme

Es wurden exemplarisch drei weit verbreitete moderne QM-Systeme vorgestellt, die die Relevanz des Themas *Qualität* gezeigt haben. Heute lässt sich eine Organisation kaum langfristig erfolgreich führen, ohne eine umfassende und bis in alle Unternehmensbereiche hineinreichende Auseinandersetzung mit dem Thema Qualität. Der Schritt, weg von der reinen Betrachtung und Beachtung der Produktqualität hin zu der Einbeziehung der Qualität der Arbeitsbedingungen der Mitarbeiter, der Qualität der Lieferantenbeziehungen und der Lebensqualität aller Menschen im Umfeld der Organisation, ist die Grundlage in allen vorgestellten QM-Systemen. Die Systeme unterscheiden sich im Wesentlichen in den gelegten Schwerpunkten und der Herangehensweise, mit der die Ziele erreicht werden sollen. Tabelle 2.3 stellt die drei vorgestellten Systeme kurz gegenüber.

Tabelle 2.3: Vergleich der drei vorgestellten Qualitätsmanagementsysteme

System	Ziel / Aufgabe	Herangehensweise
ISO 9000	Minimalanforderungen für ganzheitliche QM-Systeme	Qualitätsverantwortliche werden bestellt; ständige Verbesserung (z. B. PDCA)
TQM	Qualität im Unternehmen „leben“ Umwelt und Mitmenschen einbeziehen	Alle Mitarbeiter einbeziehen und zum Mitdenken anregen; ständige Verbesserung (z. B. PDCA)
6σ	Null Fehler in allen Unternehmensbereichen	Einzelne Projekte werden nur bei Bedarf definiert (DMAIC oder DMADV)

Die Motivation für alle drei QM-Systeme ist der langfristige Unternehmenserfolg, daher haben alle drei Systeme die Kundenorientierung als ein Leitmotiv. Das Management der Produktqualität stellt aus

diesem Grund in allen Qualitätsmanagementsystemen eine zentrale Aufgabe dar, da sie eine der wesentlichen Grundlagen für zufriedene Kunden ist. Damit kann das in Abschnitt 5 vorgestellte Konzept zum Management von Datenqualität als ein wichtiger Bestandteil in der praktischen Umsetzung jedes umfassenden QM-Systems eingesetzt werden. Die Schnittstellen und Verknüpfungspunkte sind nur zu identifizieren und entsprechend anzupassen.

Insbesondere kann das Konzept zur Optimierung von Prozessen der Datenverarbeitung herangezogen werden. Dabei spielt zunächst die detaillierte Beschreibung der an den Prozessen beteiligten Datenarten mit Hilfe des vorgestellten Qualitätsmodells eine wichtige Rolle. Mit Hilfe geeigneter Verfahren müssen die Prozesse zunächst analysiert und geeignete Messmethoden für die Qualitätsparameter, die als Kenngrößen dienen, entwickelt werden. Das Konzept schlägt neben dem Qualitätsmodell geeignete qualitative Methoden vor und bietet mit dem in dieser Arbeit im Kapitel 3.2.4 vorgestellten und im Kapitel 4 näher untersuchten Verfahren auch die Möglichkeit der Qualitätsfortpflanzung in Prozessen. Damit wird beispielsweise die simulatorische Beurteilung von potenziell geeigneten Maßnahmen zur Qualitätsverbesserung bereits vor deren Umsetzung ermöglicht.

3 Darstellungs- und Fortpflanzungsverfahren für Datenqualität

Für die Beschreibung der Datenqualität in technischen Systemen sind geeignete Darstellungsmethoden und Verfahren zu deren Modellierung bzw. Fortpflanzung in Prozessen der Datenverarbeitung erforderlich. Es wird zunächst ein bestehendes Verfahren zur Analyse, Darstellung und Modellierung von Informationsqualität in Prozessen vorgestellt, welches im Rahmen der Forschungstätigkeiten von Wiltschko [2004] am IAGB entstanden ist. Das Verfahren wird im Detail erläutert und auf seine Eignung zur Abbildung und Modellierung aller sechs Qualitätsmerkmale des Modells, welches bereits im Kapitel 2.2 vorgestellt wurde, untersucht. Insbesondere die Möglichkeiten zur Beschreibung der Qualität auf Parameterebene wird geprüft. Schließlich werden die herausgearbeiteten Einschränkungen nachvollziehbar dargelegt, die unter anderem zu der vorliegenden Arbeit motiviert haben.

Zur Behebung der aufgezeigten Grenzen des bestehenden Verfahrens wird im Anschluss nach weiteren adaptierbaren Verfahren zur Ergänzung oder als Ersatz des vorgestellten Verfahrens gesucht, die sich bereits in anderen Fachdisziplinen zur Beschreibung komplexer Prozesse oder der Systemanalyse bewährt haben. Die dabei recherchierten und als potenziell geeignet eingestuften Verfahren werden -soweit es für eine objektive Beurteilung der Eignung erforderlich ist- erläutert und nachfolgend gegenübergestellt und beurteilt.

3.1 Ein Verfahren zur Analyse der Informationsqualität

Am IAGB wurde von Wiltschko [2004] ein Verfahren entwickelt, welches die übersichtliche Darstellung und die Berechnung von Änderungen der Qualität in Prozessen ("*Qualitätsflüsse*") ermöglicht. Das Verfahren erfordert die genaue Kenntnis des Aufbaus und Verhaltens des Daten verarbeitenden Systems, welches untersucht werden soll. Die Darstellung der Informationsflüsse in Form eines Informationsflussdiagramms, als erster Bestandteil des Verfahrens, entstand aus den beiden allgemein bekannten Standardverfahren

- Ereignisablaufanalyse (DIN 25419 [1985]) und
- Fehlerbaumanalyse[1] (DIN 25424 Teil 1 [1981] und DIN 25424 Teil 2 [1990]),

die bereits in den 1980er Jahren ursprünglich zur Beurteilung der Betriebssicherheit kerntechnischer Anlagen entwickelt wurden. Der zweite wesentliche Bestandteil des Verfahrens stellt das Rechenverfahren[2] zur quantitativen Bestimmung der Informationsqualität dar. Dieses Rechenverfahren basiert weitgehend auf Boolescher Algebra (VDI 4008 Blatt 2 [1998]).

Die Definition des Qualitätsmodells im vorigen Kapitel 2.2 ermöglicht die Beschreibung verschiedenartiger Daten auf der Grundlage von sechs festgelegten, inhärenten Qualitätsmerkmalen. Bei Betrachtung eines beliebigen DV-Prozesses können nun alle Eingangs- und Ausgangsdaten hinreichend detailliert mit geeigneten Qualitätsparametern beschrieben werden. Dabei müssen die Eingangs- und Ausgangsdatenarten nicht notwendigerweise dieselben sein. Je nach Art und Komplexität des DV-Prozesses sind diese in der Regel sowohl in ihrer Art als auch in der Anzahl unterschiedlich. Um die Veränderungen und

[1] Die Fehlerbaumanalyse wird auch als Fehlzustandsbaumanalyse bzw. *en.*: Fault Tree Analysis - FTA bezeichnet.

[2] Es handelt sich hier faktisch um ein Fortpflanzungsverfahren für Datenqualität. Wiltschko spricht in seiner Arbeit jedoch immer von „Rechenverfahren“, daher wird diese Bezeichnung bei der Darstellung des Verfahrens konsequent beibehalten.

Übergänge in der Datenqualität durch Prozesse hindurch modellieren zu können, hat Wiltschko [2004] ein Analyseverfahren entwickelt. Das Verfahren besteht aus zwei Komponenten, der

- grafischen Darstellung der Informationsflüsse im Informationsflussdiagramm und der
- rechnerischen Bewertung der Informationsqualität.

Das Analyseverfahren orientiert sich an den standardisierten Verfahren der Zuverlässigkeitsanalyse und basiert auf Boolescher Algebra. Die beiden Bestandteile werden im Folgenden im Detail vorgestellt und erläutert. Die Anwendung des Verfahrens erfordert die genaue Kenntnis der Prozesse und damit auch der Anwendung der Daten, daher wird das Verfahren an einfachen und konkreten Beispielen erläutert.

3.1.1 Das Informationsflussdiagramm

Bei dem Informationsflussdiagramm handelt es sich um

> *„ein grafisches Verfahren zur Darstellung des Informationsflusses innerhalb von informationsverarbeitenden Systemen"(Wiltschko [2004]).*

Damit wird die einfache und übersichtliche Beschreibung von funktional genau identifizierten Systemen oder Systembestandteilen zur Informationsverarbeitung ermöglicht. Ziel ist insbesondere die grafische Darstellung und Verknüpfung aller Systemkomponenten und Informationsquellen, die einen Einfluss auf die Informationsqualität haben. Das Diagramm soll im Unterschied zur Ereignisablaufanalyse und der Fehlerbaumanalyse alle Informationsflüsse sowie die Qualität beeinflussenden Bestandteile eines Systems vollständig abbilden. Als grafische Grundelemente stehen unter anderem die **Information**, dargestellt in abgerundeten Textboxen, und die **Applikation**, die mit einer eckigen Textbox symbolisiert werden, zur Verfügung. Diese können mit sogenannten **Wirkungslinien** entsprechend der Informationsflüsse miteinander verknüpft werden. Alle möglichen grafischen Symbole sind in der Tabelle 3.2 wiedergegeben. Zunächst werden jedoch die Grundzüge der Booleschen Algebra dargestellt, wie sie auch in VDI 4008 Blatt 2 [1998] zu finden sind, da sie für die weiteren Ausführungen eine wesentliche Rolle spielen.

Ausgehend von der Schaltalgebra werden für jede Systemkomponente bzw. deren Zustands- oder Schaltvariable X_i nur die beiden disjunkten Zustände

- $X_i = 1$ (Komponente ist funktionsfähig)
- $X_i = 0$ (Komponente ist ausgefallen)

zugelassen. Damit wird das System sehr vereinfacht, was jedoch für viele Anwendungen, bei denen beispielsweise nur Risikoabschätzungen während der Definitions- und Entwicklungsphase vorgenommen werden sollen, völlig ausreichend ist. In Abhängigkeit der Zustände der Komponenten X_i, kann nun die Funktionsfähigkeit des Systems φ ebenfalls binär beschrieben werden. Es ergeben sich die beiden möglichen Systemzustände

- $\varphi(X_1, X_2, ..., X_n) = 1$ das System ist funktionsfähig oder
- $\varphi(X_1, X_2, ..., X_n) = 0$ das System ist ausgefallen.

Die Verknüpfung der einzelnen Komponenten zur Abbildung eines gesamten Systems erfolgt mit Hilfe der Booleschen Operatoren: Konjunktion (logisches UND), Disjunktion (logisches ODER) und Negation (NICHT), die in der Tabelle 3.1 in Form von Wahrheitstafeln dargestellt sind.

Tabelle 3.1: Wahrheitstafeln der Booleschen Operatoren

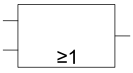
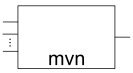
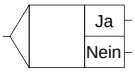
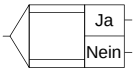
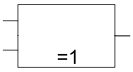
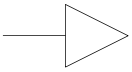
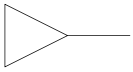
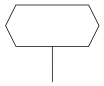
Zustand Komponenten		Konjunktion	Disjunktion	Negation
X_1	X_2	$X_1 \wedge X_2$	$X_1 \vee X_2$	$\bar{X_1}$
0	0	0	0	1
0	1	0	1	1
1	0	0	1	0
1	1	1	1	0

Neben den drei Operationen der Booleschen Algebra werden noch eine Reihe von weiteren grafischen Symbolen beschrieben, die zur Darstellung von Prozessen notwendig sind. Dazu gehören Verzweigungen von Informationsflüssen, deren Unterbrechung, der exklusiv ODER-Operator sowie Übertragungssymbole zur Beschreibung der Abhängigkeiten zwischen den Qualitätsparametern. Die Verzweigung des Informationsflusses in nur zwei Zweige wird dabei in Anlehnung an die DIN 25419 [1985] mit „**einfacher Verzweigung**“ bezeichnet. Daneben existiert dort auch eine Mehrfachverzweigung, die jedoch von Wiltschko nicht übernommen wurde. Analog wurde die zweite Möglichkeit, den Informationsfluss aufzuteilen, mit „einfacher Kontrolle“bezeichnet. Eine Verzweigung nach der Kontrolle eines Qualitätsmerkmals in mehrere Zweige wird in diesem Rechenverfahren ebenfalls nicht berücksichtigt. Der Unterschied zwischen der einfachen Verzweigung und Kontrolle besteht darin, dass für die Verzweigung nur die von Wiltschko eingeführte Vollständige Verfügbarkeit als Kombination der beiden Merkmale als Kriterium betrachtet wird (Daten vollständig und verfügbar: Ja/Nein). Bei der einfachen Kontrolle hingegen werden die Daten auf die Erfüllung eines der weiteren Qualitätsmerkmale hin untersucht. Dabei können konkrete Grenzwerte für die Entscheidung festgelegt werden. Die Tabelle 3.2 gibt einen kompakten Überblick über alle notwendigen grafischen Symbole zur Gestaltung eines Informationsflussdiagramms.

In der Regel verläuft der Informationsfluss im Diagramm von links nach rechts und von oben nach unten. Dies vereinfacht in Verbindung mit den teilweise eingefügten Richtungspfeilen die Lesbarkeit der Diagramme. Aufgrund des engen Zusammenhangs der beiden Merkmale: Verfügbarkeit und Vollständigkeit - beide beschreiben die Existenz der Information - fasst Wiltschko beide zunächst zur Vollständigen Verfügbarkeit (VV) zusammen. In späteren Veröffentlichungen wurden beide Merkmale jedoch wieder getrennt betrachtet, daher werden auch hier Verfügbarkeit und Vollständigkeit als eigenständige Merkmale behandelt (Wiltschko und Kaufmann [2005] oder Wiltschko und Möhlenbrink [2005]).

Zum besseren Verständnis, wie die Anwendung der vorgestellten grafischen Symbole im Informationsflussdiagramm erfolgen kann, zeigt die Abbildung 3.1 ein Beispiel aus der Geodäsie. Bei dem abzubildenden Prozess handelt es sich um die Berechnung von Positionslösungen auf einem mobilen, geodätischen GPS-Empfänger. Der Empfänger verarbeitet die Daten der empfangenen GPS-Satelliten (1) zusammen mit geeigneten Korrekturdaten (2) einer Referenzstation oder eines Referenznetzes zu Positionen (5). In dem vorliegenden Beispiel ist die Bereitstellung von Korrekturdaten mit Hilfe des Informationsflussdiagramms detaillierter modelliert. Durch eine einfache Verzweigung wird der Empfang geeigneter Korrekturdaten und damit deren Verfügbarkeit überprüft (2). Sofern passende Korrekturdaten empfangen wurden, kann auf dem Empfänger die Position berechnet werden. Andernfalls verzweigt sich der Informationsfluss in den „*Nein*“-Zweig und es tritt ein anderer Systemzustand ein, durch den eine weitere Applikation initiiert wird. Es muss nun geprüft werden, ob eine Prädiktion der Korrekturdaten möglich ist (7). Die Korrekturdaten, die in vorherigen Epochen gespeichert wurden, dienen hierbei als Eingangsinformation für eine mögliche Prädiktion der Daten in die nahe Zukunft. Ist dies möglich, dann

Tabelle 3.2: Grafische Symbole des Rechenverfahrens (nach Wiltschko [2004])

Symbol	Systemkomponente	Beschreibung
	Information mit Wirkungslinie	Information und Datenobjekt, die als Eingangsinformation, Zwischenergebnis oder Ausgangsinformation im Informationsfluss auftreten
	Systemkomponente mit Wirkungslinie	Darstellung technischer Komponenten (z. B. Sensoren als Datenlieferant) als Eingangs- bzw. Einflussgröße; Wirkungslinie zum Verknüpfen von Informationen und Systemkomponenten mit Applikationen
	Kommentar	Darstellung von Kommentaren im Informationsflussdiagramm
	UND-Verknüpfung	Applikation, bei der die Eingangsinformationen mit UND verknüpft sind. Die Anzahl der Eingänge ist beliebig
	ODER-Verknüpfung	Applikation, bei der die Eingangsinformationen mit ODER verknüpft sind. Die Anzahl der Eingänge ist beliebig
	mvn-Verknüpfung	Applikation, bei der von n Eingängen m funktionsfähige Eingänge gefordert sind. Es handelt sich hierbei um ein einfaches mvn-System (Majoritätsredundanz)
	einfache Verzweigung	Einfache Kontrolle der Verfügbarkeit der Information zur Verzweigung des Informationsflusses in zwei mögliche disjunkte Systemzustände
	einfache Kontrolle	Einfache Kontrolle eines Qualitätsmerkmals (nicht der Verfügbarkeit) zur Verzweigung des Informationsflusses in zwei disjunkte Systemzustände
	exklusive ODER-Verknüpfung	Verwendung zur Verknüpfung disjunkter Zustände; das Symbol wird nur im Anschluss an einfache Verzweigung und einfache Kontrolle benutzt
	Übertrag-Ausgang	Zur Unterbrechung des Informationsflusses an einer bestimmten Stelle
	Übertrag-Eingang	Zur Fortführung des Informationsflusses an einer bestimmten Stelle; tritt immer in Kombination mit dem Übertrag-Ausgang Symbol auf
	Übergang	Zur Darstellung des Einflusses anderer Qualitätsmerkmale auf die Qualität an einer Stelle

kann mit den für die auszuwertende Epoche prädizierten Korrekturdaten eine differentielle GPS-Position berechnet werden (9). Die beiden disjunkten Zustände „*empfangene Korrekturdaten*“und „*prädizierte Korrekturdaten*“werden mit der Exklusiven-ODER-Verknüpfung wieder vereinigt (8). Falls keine Prädiktion durchführbar ist, dann kann auch keine differentielle GPS-Lösung berechnet werden. Dies wird dann mit dem Kommentarfeld (10) im Informationsflussdiagramm dargestellt.

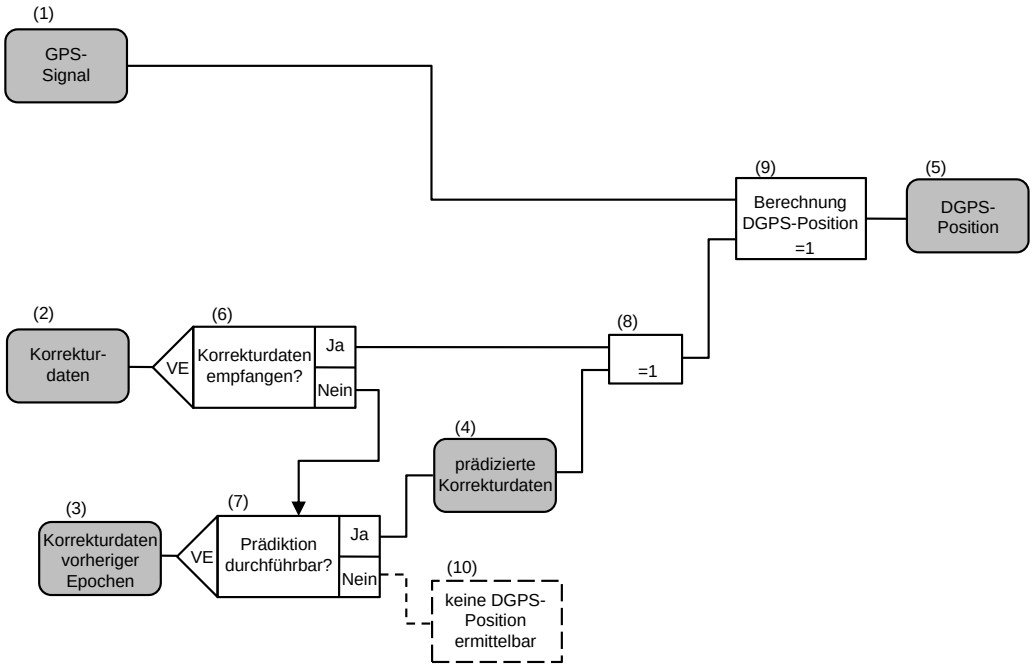


Abbildung 3.1: Anwendungsbeispiel für das Informationsflussdiagramm (nach Wiltschko [2004])

Wiltschko weist insbesondere auf die Möglichkeit der modularen Strukturierung von Prozessen mit Hilfe des Verfahrens hin sowie auf die Verwendung der vorgestellten Übertragsymbole zur Unterbrechung bzw. Fortführung der Darstellung komplexerer Systeme in Folgediagrammen. Beides dient der übersichtlicheren Darstellung komplexer Systeme, in dem der Gesamtprozess in einzelne Unterprozesse und diese wiederum in beliebig viele Teilprozesse gegliedert dargestellt werden können.

3.1.2 Rechenverfahren auf Basis Boolescher Algebra

Den zweiten Teil des von Wiltschko [2004] vorgeschlagenen Analyseverfahrens stellt ein Rechenverfahren zur Bewertung der Informationsqualität dar. Das Rechenverfahren setzt zunächst die Beschreibung der Qualität in Form von Merkmalserfüllungsgraden voraus, die in der Einheit [%] beschrieben werden können. Damit wird von der detaillierteren und wesentlich spezifischeren Betrachtung mit Hilfe von Qualitätsparametern zum Teil wieder abgerückt. Der Grund für diesen Schritt liegt in der Natur des verwendeten Rechenverfahrens. Das Rechenverfahren basiert auf binärer Schaltlogik und der Booleschen Algebra, bei denen die betrachteten Variablen mit 0 oder 1 nur zwei mögliche Zustände annehmen können. Dies war jedoch für die Anwendung auf Qualitätsmerkmale noch nicht ausreichend. Durch die Verallgemeinerung der binären Variablen auf den Wertebereich [0,1] und dem Übergang von den logischen Schaltoperationen auf die entsprechenden algebraischen Rechenregeln ist die Behandlung von Qualitätsmerkmalen grundsätzlich möglich (VDI 4008 Blatt 2 [1998]). Durch die Vorgabe des zulässigen Wertebereichs und die Einschränkung auf jeweils nur eine Variable je Merkmal ist die differenziertere

Behandlung eines einzelnen Merkmals durch mehrere Parameter jedoch nicht möglich. Jedes Merkmal kann nur durch eine Variable, den jeweiligen Merkmalserfüllungsgrad, repräsentiert werden. Die Genauigkeit nimmt hierbei eine Sonderstellung ein, da sie ausdrücklich nur mit dem bekannten Verfahren der Kovarianzfortpflanzung auf Parameterebene behandelt wird. Daher besteht keine Beschränkung hinsichtlich der Wertebereiche oder Anzahl der die Genauigkeit beschreibenden Parameter. Allerdings können durch die vollständig getrennte Betrachtung der Genauigkeit auch keine Abhängigkeiten zu den anderen Merkmalen dargestellt werden. Diese sind aber aus Sicht des Autors in einigen Fällen vorhanden, es kann beispielsweise aus der Genauigkeit mit Hilfe eines Schwellwertes auch immer ein Korrektheitsparameter abgeleitet werden.

Bevor eine konsistente und nachvollziehbare Darstellung des Rechenverfahrens erfolgen kann, müssen zunächst einmal die Abhängigkeiten zwischen den Qualitätsmerkmalen und deren Eigenschaften diskutiert werden. Wiltschko hat die folgenden Eigenschaften und wesentlichen Beziehungen zwischen den verschiedenen Merkmalen identifiziert:

- Die Verfügbarkeit stellt das wichtigste Merkmal innerhalb des Analyseverfahrens dar und hat keinen direkten Einfluss auf die anderen Qualitätsmerkmale. Eine fehlende Redundanz oder Update-Information kann jedoch einen indirekten Einfluss auf die Korrektheit oder Aktualität haben.
- Die Applikation „einfache Kontrolle“hat für alle Merkmale einen Einfluss auf die Verfügbarkeit.
- Die Aktualität wird für neue Ausgangsinformationen, die in Applikationen erzeugt wurden, immer auf $100\,\%$ gesetzt und ist damit am Ausgang für neue Informationen per se immer erfüllt. Mangelnde Aktualität einer Eingangsinformation hat unter Umständen Auswirkungen auf die Korrektheit der Ausgangsinformationen.
- Die Genauigkeit hat keinen direkten Einfluss auf andere Qualitätsmerkmale und wird ihrerseits auch von keinem Merkmal direkt beeinflusst. Daher kann die Genauigkeit separat behandelt werden.
- Die Korrektheit hat keinen direkten Einfluss auf andere Qualitätsmerkmale.
- Die Konsistenz und Aktualität beeinflussen bei informationsverarbeitenden Applikationen die Verfügbarkeit sowie die Korrektheit.

Daneben existieren weitere Abhängigkeiten, die jedoch nur unzureichend allgemeingültig formuliert werden können. Diese müssen jeweils anhand der konkret zu beschreibenden Prozesse erarbeitet werden.

Aufbauend auf dem Qualitätsmodell sowie den vorgestellten Abhängigkeiten zwischen den Merkmalen hat Wiltschko zur Wahrung der Übersichtlichkeit eine sequentielle Auswertung der Qualitätsmerkmale in den folgenden sechs Schritten vorgeschlagen:

1. **Auswertung der Konsistenz (KO)** an den Stellen im Informationsflussdiagramm, an denen Informationen von einer Systemkomponente in eine andere übergehen und bei denen aufgrund einer Informationsverarbeitung Auswirkungen auf die Verfügbarkeit und die Korrektheit zu erwarten sind.
2. **Auswertung der Aktualität (AK)** an den Stellen im Informationsflussdiagramm, an denen erfasste, berechnete oder genutzte Informationen eingehen, deren Entstehung eine gewisse Zeit zurückliegt und somit die Gefahr einer Verletzung der Aktualität besteht. Der Schwerpunkt der Betrachtung liegt bei Applikationen, bei denen die eingehenden Informationen zu neuen Informationen verarbeitet werden.

3. Auswertung von Verfügbarkeit (VE) und Vollständigkeit (VO) über das gesamte Informationsflussdiagramm unter Einbeziehung der Einflüsse der Konsistenz (aus Schritt 1) auf die VE. Da beide Merkmale sehr eng miteinander verwandt sind, erfolgt die Auswertung in einem Schritt. Wiltschko hatte beide Merkmale aus diesem Grund, wie bereits erwähnt, zunächst zur sogenannten *Vollständigen Verfügbarkeit* zusammengefasst.

4. Auswertung der Korrektheit (KR) über das gesamte Informationsflussdiagramm unter Einbeziehung der Einflüsse der Konsistenz und der Aktualität (aus Schritt 1 und 2) auf die Korrektheit.

5. Auswertung der Genauigkeit (GE) erfolgt mit den bekannten Methoden der Kovarianzfortpflanzung und Ausgleichungsrechnung. Die eingehenden und berechneten Standardabweichungen können in der selben Art und Weise wie die Wahrscheinlichkeitswerte der Qualitätsmerkmale im Informationsflussdiagramm protokolliert werden.

6. Zusammenstellung der Ergebnisse in einem Qualitätstupel, entsprechend dem festgelegten Satz von sechs Qualitätsmerkmalen. Dabei können im Bedarfsfall einzelne, für die Qualitätsbetrachtung irrelevante bzw. unbehandelte Qualitätsmerkmale ausgeklammert werden.

Von dieser strengen Vorgehensweise kann oder muss sogar im Einzelfall abgewichen werden. Beispielsweise wenn in einer einfachen Kontrolle die Korrektheit oder Genauigkeit Prüfkriterium ist, dann kann Schritt drei erst später stattfinden.

Im Folgenden sind die einzelnen Berechnungsvorschriften zur Bestimmung der Merkmalserfüllungsgrade für die wichtigsten Verknüpfungen und Verzweigungen dargestellt, die bereits in der Übersicht der grafischen Symbole (vgl. Tabelle 3.2) eingeführt wurden. Eine wesentliche Voraussetzung zur Behandlung der Merkmale mit dem Berechnungsverfahren ist nach Wiltschko [2004] die stochastische Unabhängigkeit der Eingangsinformationen hinsichtlich ihrer Informationsqualität. Diese Unabhängigkeit wurde bereits bei der Booleschen Algebra vorausgesetzt, auf Basis derer das Verfahren entwickelt wurde. Zur vereinfachten Darstellung der Erfüllungsgrade verschiedener Merkmale für Ein- und Ausgangsinformationen wurden von Wiltschko die in der Tabelle 3.3 dargestellten Abkürzungen eingeführt.

Tabelle 3.3: Abkürzungen im Informationsflussdiagramm

$\mathrm{QM}(i)$ für $1 \leq i \leq 6$	Darstellung der sechs Qualitätsmerkmale in der Reihenfolge VE, VO, KO, AK, KR, GE
$\mathrm{E}^{\mathrm{QM}(i)} \equiv p(E, QM(i))$	Wahrscheinlichkeit mit der die Eingangsinformation E das betrachtete Qualitätsmerkmal $\mathrm{QM}(i)$ erfüllt
$\mathrm{A}^{\mathrm{QM}(i)} \equiv p(\mathrm{A}, \mathrm{QM}(i))$	Wahrscheinlichkeit mit der die Ausgangsinformation A das betrachtete Qualitätsmerkmal $\mathrm{QM}(i)$ erfüllt
$\mathrm{Z}_{\mathrm{Ja}}^{\mathrm{QM}(i)} \equiv p(\mathrm{Z}_{\mathrm{Ja}}, \mathrm{QM}(i))$	Wahrscheinlichkeit für die Fortsetzung des Informationsflusses im Ja-Zweig
$\mathrm{Z}_{\mathrm{Nein}}^{\mathrm{QM}(i)} \equiv p(\mathrm{Z}_{\mathrm{Nein}}, \mathrm{QM}(i))$	Wahrscheinlichkeit für die Fortsetzung des Informationsflusses im Nein-Zweig
$0.9_4 = 0.9999$	Verkürzte Schreibweise für die Anzahl der Wiederholungen der vorhergehenden Ziffer

UND-Verknüpfung

Es handelt sich um die am häufigsten auftretende Art der Verknüpfung von Informationen. Alle Eingänge sind zur Generierung der Ausgangsdaten erforderlich. Diese Applikationsform stellt immer eine Verarbeitung von Informationen zu einer neuen Information dar. Dementsprechend berechnet sich die Wahrscheinlichkeit zur Erfüllung der Qualitätsmerkmale $\mathrm{A}^{\mathrm{QM}(q)}$, mit Ausnahme der Genauigkeit, aus den Erfüllungswahrscheinlichkeiten der n Eingangsinformationen durch Multiplikation der einzelnen Erfüllungswahrscheinlichkeiten $\mathrm{E}_i^{\mathrm{QM}(q)}$.

$$\mathrm{A}^{\mathrm{QM}(q)} = \prod_{1=i}^{n} \mathrm{E}_i^{\mathrm{QM}(q)} \qquad \text{für } 1 \leq q \leq 5 \tag{3.1}$$

Die Aktualität der generierten Ausgangsinformationen bezieht sich laut Definition immer auf den Generierungszeitpunkt. Daher ist die Aktualität der Ausgangsinformation unmittelbar nach der Generierung mit $\mathrm{A}^{\mathrm{AK}} = 100\,\%$ gegeben. Ein möglicher Mangel der Aktualität der Eingangsinformationen hat Einfluss auf die Verfügbarkeit und die Korrektheit. Veraltete Eingangsinformationen verringern in der Regel die Korrektheit der Ausgangsinformation und können auch die Verfügbarkeit der Information herabsetzen, wenn aufgrund veralteter Information keine Berechnung von Ausgangsinformationen mehr durchgeführt wird. Diese Einflüsse müssen bei der Darstellung einer Applikation im Informationsflussdiagramm mit Hilfe des Übergangssymbols modelliert werden.

ODER-Verknüpfung

Die Oder-Verknüpfung wird insbesondere zur Beschreibung von Redundanzen verwendet. Bei Ausfall einer Informationsquelle wird auf eine andere Quelle, mit in der Regel gleichwertiger Information, zurückgegriffen. Dabei wird die Entscheidung über die Verwendung einer alternativen Eingangsinformation nur aufgrund der Verfügbarkeit der Information getroffen. Eine Überprüfung nach einem Kriterium wie z. B. Genauigkeit oder Vollständigkeit ist in dieser Verknüpfung nicht enthalten. Gegebenenfalls wird eine derartige Prüfung als Kontrolle im Informationsflussdiagramm vorangestellt. Die Wahrscheinlichkeit der Verfügbarkeit der Ausgangsinformation $\mathrm{A}^{\mathrm{QM}(1)}$ berechnet sich bei der Verknüpfung redundanter Information aus deren Verfügbarkeit $\mathrm{E}_i^{\mathrm{VE}}$ als

$$\mathrm{A}^{\mathrm{QM}(1)} = \mathrm{A}^{\mathrm{VE}} = 1 - \prod_{i=1}^{n} \left(1 - \mathrm{E}_i^{\mathrm{VE}}\right) \qquad \text{mit } 1 \leq i \leq n \text{ unabh., redundante Eingänge.} \tag{3.2}$$

Die weiteren Qualitätsmerkmale bleiben bei dieser Verknüpfungsart zunächst unverändert. Da das Vorhandensein der Ausgangsinformation $\mathrm{A}^{\mathrm{QM}(q)}$ jedoch unmittelbar von der Verfügbarkeit der Eingangsdaten $\mathrm{E}_i^{\mathrm{QM}(q)}$ abhängt, können die Erfüllungswahrscheinlichkeiten der weiteren Qualitätsmerkmale mit Hilfe des gewichteten, arithmetischen Mittels der Wahrscheinlichkeiten $\mathrm{E}_i^{\mathrm{QM}(q)}$ der n Eingangsinformationen berechnet werden.

$$\mathrm{A}^{\mathrm{QM}(q)} = \sum_{i=1}^{n} \left(\frac{\mathrm{E}_i^{\mathrm{VE}}}{\sum_{j=1}^{n} \mathrm{E}_j^{\mathrm{VE}}} \cdot \mathrm{E}_i^{\mathrm{QM}(q)} \right) \qquad \text{für } 2 \leq q \leq 5 \text{ und } 1 \leq i \leq n \text{ unabh., red. Eing.} \tag{3.3}$$

mvn-Verknüpfung

Die mvn-Verknüpfung ist zur Behandlung von Majoritätsredundanzen notwendig. Dabei ist die Funktionstüchtigkeit von m aus n Eingängen erforderlich bzw. die Wahrscheinlichkeiten für die Merkmalser-

füllungen der Ausgangsinformation sind abhängig von der Verfügbarkeit und den Merkmalserfüllungsgraden von mindestens m aus n Eingangsinformationen.

Diese Verknüpfung soll hier nur der Vollständigkeit halber erwähnt werden. Da sie in den weiteren Ausführungen und Folgerungen keine Rolle spielt, wird auf eine detailliertere Darstellung und die Angabe der Rechenvorschriften verzichtet. Gegebenenfalls können die Details in Wiltschko [2004] nachgelesen werden.

Einfache Verzweigung

Die Einfache Verzweigung (zur Herkunft des Begriffs, vgl. 3.1.1) teilt den Informationsfluss auf Grundlage eines übergeordneten Kriteriums auf. Dies kann zum einen die Verfügbarkeit einer Information sein, andererseits kann auch ein bestimmter Informationsinhalt oder es können verschiedene Systemzustände (Schaltzustände) zu einer Verzweigung führen. Wird die Entscheidung aufgrund der Verfügbarkeit getroffen, so kann der Nein-Zweig keine Ausgangsdaten generieren, da die Information in diesem Systemzustand nicht vorhanden ist. Alle weiteren Qualitätsmerkmale werden mit der Applikation *Einfache Kontrolle* behandelt, da sie für die Entscheidung einen Kontrollmechanismus zur Prüfung der jeweiligen Merkmalserfüllung benötigen.

Die Wahrscheinlichkeiten der Verfügbarkeit der Information an den beiden Ausgängen A^{VE} berechnen sich in diesem Fall einfach aus der Wahrscheinlichkeit der Eingangsinformation E^{VE}. Dies entspricht ebenfalls der Wahrscheinlichkeit p, mit der der jeweilige Systemzustand eintritt.

$$\mathrm{A}^{\mathrm{VE}}_{\mathrm{Ja}} = \mathrm{E}^{\mathrm{VE}} = p_{\mathrm{Ja}} \qquad \text{im Ja-Zweig} \tag{3.4}$$

$$\mathrm{A}^{\mathrm{VE}}_{\mathrm{Nein}} = 1 - \mathrm{E}^{\mathrm{VE}} = p_{\mathrm{Nein}} \qquad \text{im Nein-Zweig} \tag{3.5}$$

Die Verzweigung aufgrund der Verfügbarkeit hat Auswirkungen auf die weiteren Merkmalserfüllungsgrade. Die Erfüllungswahrscheinlichkeiten der Ausgangsinformation im Ja-Zweig für die Merkmale VO, KO, AK und KR berechnen sich daher wie folgt:

$$\mathrm{A}^{\mathrm{QM}(q)}_{\mathrm{Ja}} = \mathrm{E}^{\mathrm{VE}} \cdot \mathrm{E}^{\mathrm{QM}(q)} \qquad \text{für } 2 \leq q \leq 5 \tag{3.6}$$

Im Nein-Zweig ist der Informationsfluss unterbrochen, daher kann auch keine Qualität der Information angegeben werden. Zur Verdeutlichung der Unterbrechung des Informationsflusses im Nein-Zweig wird dieser im Symbol grau hinterlegt.

$$\mathrm{A}^{\mathrm{QM}(q)}_{\mathrm{Nein}} = - \qquad \text{für } 2 \leq q \leq 6 \tag{3.7}$$

Wichtig ist bei der Verzweigung die Trennung von Merkmalserfüllungsgraden ($\mathrm{A}^{\mathrm{QM}(i)}_{\mathrm{Ja}}$ / $\mathrm{A}^{\mathrm{QM}(i)}_{\mathrm{Nein}}$) und Wahrscheinlichkeiten von Systemzuständen (p_{Ja} / p_{Nein}). Die Wahrscheinlichkeit, mit welcher der Systemzustand „*VE nicht erfüllt*“(Nein-Zweig) eintritt, ist für die weiteren Berechnungen im Informationsflussdiagramm wichtig. Zum Beispiel, wenn eine Verletzung der Verfügbarkeit eine redundante Informationsbeschaffung initiiert und die disjunkten Informationsflüsse zu einem späteren Zeitpunkt wieder vereinigt werden sollen. Durch das Mitführen der Systemzustands-Wahrscheinlichkeiten wird gewährleistet, dass die Verfügbarkeit nach der Vereinigung nicht auf über $100\,\%$ anwachsen kann.

Einfache Kontrolle

Im Gegensatz zur *Einfachen Verzweigung* werden bei der *Einfachen Kontrolle* Qualitätsmerkmale zur Entscheidung herangezogen. Es werden Schwellwerte für eines der Qualitätsmerkmale definiert, anhand derer die Entscheidung getroffen wird. Die Kontrolle kann grundsätzlich auf allen Qualitätsmerkmalen basieren, mit Ausnahme der Verfügbarkeit, deren Prüfung bereits mit Hilfe der Einfachen Verzweigung abgebildet werden kann.

Die Erfüllungswahrscheinlichkeit des kontrollierten Merkmals steigt aufgrund der Kontrolle und Aussortierung von Informationen, die den Anforderungen nicht genügen. Im Idealfall würde die Erfüllungswahrscheinlichkeit nach der Kontrolle auf $100\,\%$ ansteigen, allerdings ist in der Regel kein Prüfmechanismus in der Lage, sämtliche unzureichenden Informationen auszusortieren. Der Wirkungsgrad (Kontrollgüte) der Kontrollmechanismen kann stark variieren und muss zur Modellierung der Datenqualität abgeschätzt werden. Im Weiteren wird der Wirkungsgrad eines Prüfmechanismus mit $\mathrm{K_{Ja}}$ bezeichnet. $\mathrm{K_{Ja}} = 90\,\%$ bedeutet damit, dass $90\,\%$ der (hinsichtlich des zu prüfenden QM) fehlerhaften Informationen als solche erkannt und eliminiert werden. Es verbleiben damit in diesem Fall $10\,\%$ der falschen Daten unerkannt im Datenfluss. Der Fall, dass korrekte Informationen fälschlicherweise aussortiert werden, wird dabei ausgeschlossen. Die Verfügbarkeit der Information wird durch eine Einfache Kontrolle stets vermindert oder bleibt bestenfalls gleich, da die Anzahl der Informationen verkleinert wird. Die Verfügbarkeit der Ausgangsinformation ergibt sich somit aus der Verfügbarkeit der Eingangsinformation und verkleinert sich um den Anteil, der aufgrund des geprüften Kriteriums eliminiert wird. Die Wahrscheinlichkeit

$$p_{\mathrm{elim}}^{\mathrm{QM}(q)} = (1 - \mathrm{E}^{\mathrm{QM}(q)}) \cdot \mathrm{K}_{\mathrm{Ja}}^{\mathrm{QM}(q)} \qquad \text{für } 2 \leq q \leq 5 \tag{3.8}$$

stellt den Anteil dar, der das zu prüfende Kriterium nicht erfüllt und daher eliminiert wird. Dabei wird berücksichtigt, dass nur der als Kontrollgüte angegebene Anteil $\mathrm{K}_{\mathrm{Ja}}^{\mathrm{QM}(q)}$ und nicht alle fehlerhaften Informationen eliminiert werden.

Am Ja-Zweig, das heißt im Falle der geprüften und hinsichtlich des Qualitätsmerkmals für ausreichend empfundenen Informationen, ergibt sich die Verfügbarkeit $\mathrm{A}_{\mathrm{Ja}}^{\mathrm{VE}}$ durch Multiplikation der Eingangsverfügbarkeit E^{VE} mit der Wahrscheinlichkeit, dass der Kontrollmechanismus greift $(\mathrm{K}_{\mathrm{Ja}}^{\mathrm{QM}(q)} \cdot \mathrm{E}^{\mathrm{QM}(q)})$. Zusätzlich muss der Anteil $\mathrm{K}_{\mathrm{Nein}}^{\mathrm{QM}(q)}$ berücksichtigt werden, bei dem der Kontrollmechanismus versagt. Dieser führt ebenfalls zum Ja-Zweig, da damit der Anteil beschrieben wird, der den Anforderungen nicht genügt, jedoch nicht als solcher erkannt wird.

$$\mathrm{A}_{\mathrm{Ja}}^{\mathrm{VE}} = \mathrm{E}^{\mathrm{VE}} \cdot \left(1 - p_{\mathrm{elim}}^{\mathrm{QM}(q)}\right) = \mathrm{E}^{\mathrm{VE}} \cdot \left(1 - (1 - \mathrm{E}^{\mathrm{QM}(q)}) \cdot \mathrm{K}_{\mathrm{Ja}}^{\mathrm{QM}(q)}\right) = ...$$

$$\mathrm{A}_{\mathrm{Ja}}^{\mathrm{VE}} = \mathrm{E}^{\mathrm{VE}} \cdot \left(\mathrm{K}_{\mathrm{Ja}}^{\mathrm{QM}(q)} \cdot \mathrm{E}^{\mathrm{QM}(q)} + \mathrm{K}_{\mathrm{Nein}}^{\mathrm{QM}(q)}\right) = (p_{\mathrm{Ja}}) \qquad \text{für } 2 \leq q \leq 5 \tag{3.9}$$

Durch Negation ergibt sich daraus auch die Verfügbarkeit $\mathrm{A}_{\mathrm{Nein}}^{\mathrm{VE}}$ einer Information am Nein-Zweig:

$$\mathrm{A}_{\mathrm{Nein}}^{\mathrm{VE}} = 1 - \mathrm{E}^{\mathrm{VE}} \cdot \left(\mathrm{K}_{\mathrm{Ja}}^{\mathrm{QM}(q)} \cdot \mathrm{E}^{\mathrm{QM}(q)} + \mathrm{K}_{\mathrm{Nein}}^{\mathrm{QM}(q)}\right) = (p_{\mathrm{Nein}}) \qquad \text{für } 2 \leq q \leq 5 \tag{3.10}$$

Die Verfügbarkeit einer Information an den Ausgängen der Applikation beschreibt das Aufteilungsverhältnis des Datenflusses, das durch den Prüfmechanismus bedingt ist. Daraus ergibt sich die Erfüllungswahrscheinlichkeit $\mathrm{A}^{\mathrm{QM}(q)}$ der Qualitätsmerkmale, die nicht als Prüfkriterium herangezogen wurden, durch Multiplikation mit der Wahrscheinlichkeit p des betrachteten Zweiges.

$$\mathrm{A}_{\mathrm{Ja}}^{\mathrm{QM}(q)} = (p_{\mathrm{Ja}}) \cdot \mathrm{E}^{\mathrm{QM}(q)} \qquad \text{für } 2 \leq q \leq 5 \text{ ohne zu prüfendes QM} \tag{3.11}$$

$$\mathrm{A}_{\mathrm{Nein}}^{\mathrm{QM}(q)} = (p_{\mathrm{Nein}}) \cdot \mathrm{E}^{\mathrm{QM}(q)} \qquad \text{für } 2 \leq q \leq 5 \text{ ohne zu prüfendes QM} \tag{3.12}$$

Die durch eine Kontrolle erhöhte Erfüllungswahrscheinlichkeit des untersuchten Merkmals im Ja-Zweig berechnet sich durch Multiplikation der Wahrscheinlichkeit des Ja-Zweiges p_{Ja} mit der Kontrollgüte $\mathrm{K}_{\mathrm{Ja}}^{\mathrm{QM}(q)}$. Hinzu kommt zusätzlich der Anteil, bei dem der Kontrollmechanismus versagt hat.

$$\mathrm{A}_{\mathrm{Ja}}^{\mathrm{QM}(q)} = (p_{\mathrm{Ja}}) \cdot \left(\mathrm{K}_{\mathrm{Nein}}^{\mathrm{QM}(q)} \cdot \mathrm{E}^{\mathrm{QM}(q)} + \mathrm{K}_{\mathrm{Ja}}^{\mathrm{QM}(q)}\right) \qquad \text{für kontrolliertes QM } q \tag{3.13}$$

$$\mathrm{A}_{\mathrm{Nein}}^{\mathrm{QM}(q)} = 1 - \mathrm{A}_{\mathrm{Ja}}^{\mathrm{QM}(q)} \qquad \text{für kontrolliertes QM } q \tag{3.14}$$

XODER-Verknüpfung

Bei der Exklusiv-Oder Applikation schließen sich die beiden oder auch mehrere Eingangsinformationen gegenseitig aus. Daher können die Erfüllungswahrscheinlichkeiten aller Qualitätsmerkmale der Ausgangsinformation als Summe der Erfüllungswahrscheinlichkeiten aller Eingangsinformationen $\mathrm{E}_i^{\mathrm{QM}(q)}$ berechnet werden.

$$\mathrm{A}^{\mathrm{QM}(q)} = \sum_{i=1}^{n} \mathrm{E}_i^{\mathrm{QM}(q)} \qquad \text{für } 1 \leq q \leq 5 \text{ und } 1 \leq i \leq n \text{ disjunkte Eingänge} \tag{3.15}$$

Zur Veranschaulichung der Vorgehensweise bei der Anwendung der Methodik wird ein einfaches Beispiel herangezogen, an dem die Auswertung jedes Merkmals separat erfolgt. Aus der mit einem Tachymeter gemessenen Strecke sowie der Richtung zu einem Reflektor sollen die lokalen, kartesischen Koordinaten des markierten Punkts berechnet werden. Es handelt sich dabei um die Transformation von polaren in kartesische Koordinaten. Die Bewertung der Qualität erfolgt nach den sechs Schritten, wie sie zu Beginn des Kapitels 3.1.2 vorgestellt wurden:

1. Auswertung der Konsistenz: Die Untersuchung der Datenkonsistenz erfolgt in der Regel immer dann, wenn Daten einer Verarbeitung zugeführt werden. Das ist im Beispiel die Zuführung der gemessenen Elemente *Strecke* und *Richtung zum Reflektor* zur Verarbeitung. Die Konsistenzprüfung wird gedanklich dem Verarbeitungsprozess vorgeschaltet, wie in Abbildung 3.2 links oben für die Streckenmessung im Teil a) dargestellt. Laut Wiltschko tritt die Inkonsistenz einer Information erst mit der Übernahme in das Subsystem zur Informationsverarbeitung auf (diese Auffassung wird hier zunächst nicht in Frage gestellt, eine umfassende Beurteilung des gesamten Rechenverfahrens erfolgt im Kapitel 3.1.4). Die Konsistenz der beiden Eingangsinformationen sei in $10^{-4} = 0.1\,‰$ nicht erfüllt (z. B. negative Strecke s, nicht-numerischer Wert für die Richtung r). Daraus kann die Konsistenzrate der Strecke s aus der zunächst einmal konsistenten Strecke s' mit

$$p(s, \mathrm{KO}) = 1.0 \cdot 0.9_4 = 0.9_4 \tag{3.16}$$

berechnet werden. Für die Konsistenz der Richtungsmessung $p(r, \mathrm{KO})$ wird im Folgenden derselbe Wert angenommen.

2. Auswertung der Aktualität: Es wird angenommen, dass die Strecke und Richtung unmittelbar vor der Berechnung der kartesischen Koordinaten gemessen wurden. Ansonsten müsste die Art des signalisierten Objektes zur Beurteilung der AK berücksichtigt werden. Bei Beobachtung einer fahrenden Baumaschine wird die AK sehr viel schneller verletzt als bei der Koordinatenbestimmung eines Grenzpunktes. Zur Beurteilung wäre daher die genaue Kenntnis der Anwendung ebenso erforderlich, wie geeignete Qualitätsparameter zur Beurteilung der Aktualität. Dies könnten zum Beispiel Zeitstempel oder Zeitspannen sein, aus denen auf das Alter der Daten geschlossen

werden kann. Daraus können Rückschlüsse auf den Merkmalserfüllungsgrad der Aktualität gezogen werden. Der Einfachheit halber wird hier jedoch auf eine Auswertung der Aktualität verzichtet.

3. Auswertung von Verfügbarkeit und Vollständigkeit: Die technische Verfügbarkeit beider notwendiger Eingangsinformationen bedingt die Verfügbarkeit der Ausgangsinformation. Aus statistischer Sicht ergibt sich daher bei einer angenommenen VE beider Messwerte von $p(s) = p(r) = 0,9_38$ gemäß dem Rechenverfahren eine Verfügbarkeit der kartesischen Koordinaten von

$$p(\mathbf{x}_{\mathrm{Aus}}, \mathrm{VE}) = p(s) \cdot p(r) = 0,9_38 \cdot 0,9_38 = 0,9_36. \tag{3.17}$$

Die Verfügbarkeit der Eingangsinformationen ist essentiell für die Generierung von Ausgangsinformation. Da im Falle einer nicht verfügbaren Eingangsinformation keine Ausgangsinformation berechenbar ist, wird die Verfügbarkeit nicht zur Berechnung der weiteren Merkmale benötigt.
Die Berechnung der Vollständigkeit jedoch erfordert laut Wiltschko die Berücksichtigung des Einflusses der Konsistenz. Daher wird die Vollständigkeitsrate der Ausgangsinformation wie folgt ermittelt:

$$p(\mathbf{x}_{\mathrm{Aus}}, \mathrm{VO}) = p(s, \mathrm{KO}) \cdot p(s, \mathrm{VO}) \cdot p(r, \mathrm{KO}) \cdot p(r, \mathrm{VO})$$

$$p(\mathbf{x}_{\mathrm{Aus}}, \mathrm{VO}) = 0,9_4 \cdot 0,9_3 \cdot 0,9_4 \cdot 0,9_3 = 0,9_278 \tag{3.18}$$

Dabei wurde jeweils für beide Eingangsinformationen eine Vollständigkeit von $0,9_3$ angenommen.

4. Auswertung der Korrektheit: Die Auswertung der Korrektheit erfolgt analog zur Auswertung der Vollständigkeit. Die Merkmalserfüllung der Ausgangsinformation berechnet sich ebenfalls unter Berücksichtigung der Konsistenz der Eingangsinformationen.

$$p(\mathbf{x}_{\mathrm{Aus}}, \mathrm{KR}) = p(s, \mathrm{KO}) \cdot p(s, \mathrm{KR}) \cdot p(r, \mathrm{KO}) \cdot p(r, \mathrm{KR})$$

$$p(\mathbf{x}_{\mathrm{Aus}}, \mathrm{KR}) = 0,9_4 \cdot 0,9_26 \cdot 0,9_4 \cdot 0,9_26 = 0,9_218 \tag{3.19}$$

Hier wurde für die Korrektheit der Eingänge für beide Eingangsgrößen jeweils mit $0,9_26$ angenommen.

5. Auswertung der Genauigkeit: Wie bereits erläutert, nimmt die Genauigkeit eine Sonderstellung ein. Diese wird nicht wie die anderen Merkmale auf Merkmalsebene mit Erfüllungsgraden behandelt, sondern es erfolgt eine explizite Beschreibung mit Genauigkeitsmaßen auf Parameterebene. In der Regel werden hier Standardabweichungen angegeben, die nach den Regeln der Kovarianzfortpflanzung in den Prozessen behandelt werden. Dazu muss der funktionale Zusammenhang in der Applikation hinreichend bekannt sein.
Die Strecken- bzw. Richtungsmessungen seien mit den empirischen Standardabweichungen von $S_s = 7\,\mathrm{mm}$ und $S_r = 1\,\mathrm{mgon}$ gegeben und unkorreliert ($S_{sr} = 0$). Der funktionale Zusammenhang ergibt sich über die Winkelfunktionen.

$$\mathbf{F}(\mathbf{s}, \mathbf{r}) = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} s \cdot \cos r \\ s \cdot \sin r \end{bmatrix} \tag{3.20}$$

Mit Hilfe der klassischen Kovarianzfortpflanzung lassen sich die Standardabweichungen der Ausgangsinformationen abschätzen. Mit Hilfe der Matrizenrechnung können beide funktionalen Zusammenhänge gleichzeitig behandelt werden und es ist gegebenenfalls auch die Berücksichtigung von Kovarianzen zwischen den Eingangsgrößen möglich. Das Kovarianzfortpflanzungsgesetz lautet:

$$\mathbf{\Sigma_{xx}} = \mathbf{F} \cdot \mathbf{\Sigma_{ll}} \cdot \mathbf{F^T} \tag{3.21}$$

mit $\mathbf{\Sigma_{ll}}$ als Varianz-/Kovarianzmatrix der Eingangsgrößen

$$\mathbf{\Sigma_{ll}} = \begin{bmatrix} S_s^2 & S_{sr} \\ S_{sr} & S_r^2 \end{bmatrix} = \begin{bmatrix} 49\,\mathrm{mm}^2 & 0 \\ 0 & 1\,\mathrm{mgon}^2 \end{bmatrix} \tag{3.22}$$

und der Jacobi-Matrix $\mathbf{F}$, die aus den partiellen Ableitungen der funktionalen Zusammenhänge nach den fehlerbehafteten Größen aufgebaut ist:

$$\mathbf{F} = \begin{bmatrix} \frac{\partial F_1(s,r)}{\partial s} & \frac{\partial F_1(s,t)}{\partial r} \\ \frac{\partial F_2(s,r)}{\partial s} & \frac{\partial F_2(s,t)}{\partial r} \end{bmatrix} = \begin{bmatrix} \cos r & -s \cdot \sin r \\ \sin r & s \cdot \cos r \end{bmatrix} \tag{3.23}$$

Unter Annahme der gemessenen Eingangsgrößen von Strecke $s = 50\,\mathrm{m}$ und Richtung $r = 50\,\mathrm{gon}$ ergibt sich nach dem Kovarianzfortpflanzungsgesetz die folgende Varianz-Kovarianzmatrix der gesuchten Ausgangsgrößen $\mathbf{\Sigma_{xx}}$:

$$\mathbf{\Sigma_{xx}} = \begin{bmatrix} S_x^2 & S_{xy} \\ S_{xy} & S_y^2 \end{bmatrix} = \begin{bmatrix} 0.25 \cdot 10^{-4}\,\mathrm{m}^2 & 0.24 \cdot 10^{-4}\,\mathrm{m}^2 \\ 0.24 \cdot 10^{-4}\,\mathrm{m}^2 & 0.25 \cdot 10^{-4}\,\mathrm{m}^2 \end{bmatrix} \tag{3.24}$$

Damit ergeben sich die Standardabweichungen der Ausgangsinformationen in diesem Fall zu

$$S_x = S_y = \sqrt{0,25 \cdot 10^{-4}\,\mathrm{m}^2} = 5,0\,\mathrm{mm} \tag{3.25}$$

Die Kovarianz steht ebenfalls zur weiteren Verwendung zur Verfügung.

6. Zusammenfassung der Ergebnisse: Im letzten Schritt werden die Ergebnisse der Auswertungen in einem Qualitätstupel zusammengefasst. In diesem Fall umfasst das Qualitätstupel $\mathbf{Q}_F$ die Merkmale: Verfügbarkeit, Vollständigkeit, Korrektheit sowie zwei Parameter für die Genauigkeit der kartesischen Koordinaten x und y. Die Konsistenz der Ausgangsinformation ist mit 1 anzunehmen, da nur im Falle konsistenter Eingangswerte für s und r auch Koordinaten mit Hilfe der Rechenvorschrift als Ausgangswerte erzeugt werden können. Damit ergibt sich das Ergebnis als

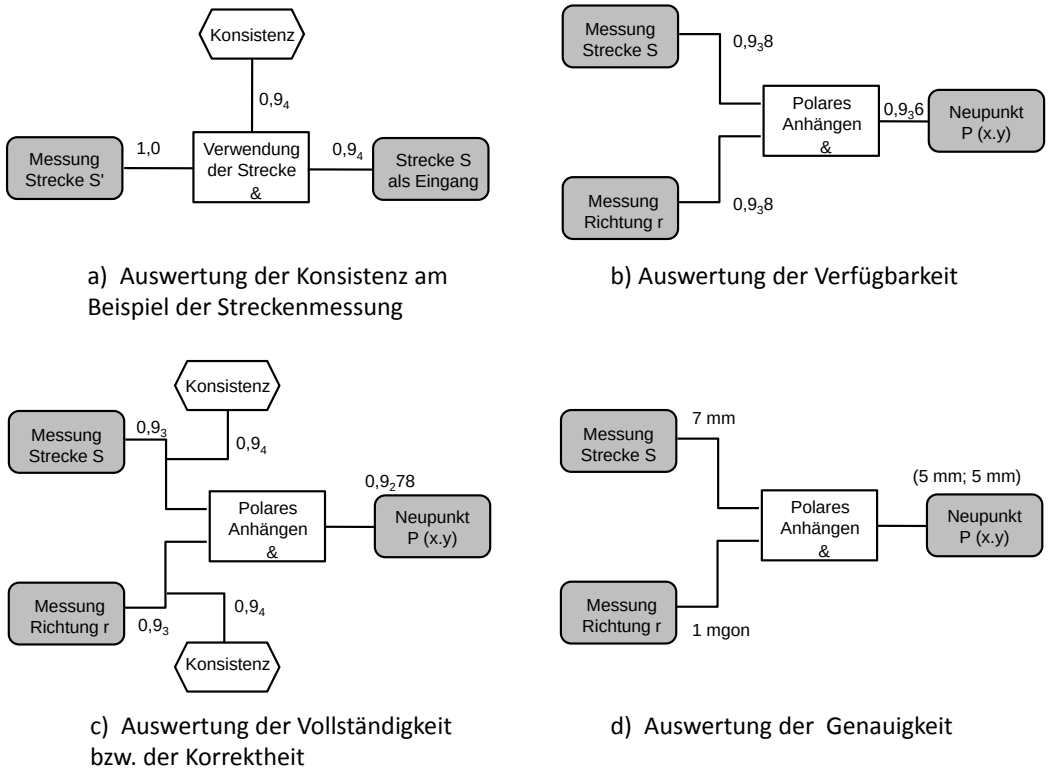
$$\begin{aligned} \mathbf{Q_F} = (p(\mathbf{x}_{\mathrm{Aus}}, \mathrm{VE}); p(\mathbf{x}_{\mathrm{Aus}}, VO); p(\mathbf{x}_{\mathrm{Aus}}, \mathrm{KO}); p(\mathbf{x}_{\mathrm{Aus}}, \mathrm{KR}); S_x; S_y) = \\ (0.9_36; 0.9_278; 1.0; 0,9_218; 5.0\,\mathrm{mm}; 5.0\,\mathrm{mm}). \end{aligned} \tag{3.26}$$

Die beiden Ausgangsinformationen werden dabei aufgrund der Gleichartigkeit hinsichtlich VE, VO und KR als ein Ergebnis behandelt, daher genügt hier jeweils ein Merkmalserfüllungsgrad. Unterscheiden sich die Ausgangsinformationen in ihrer Entstehung wesentlich, insbesondere in den erforderlichen Eingangsinformationen, so ist eine getrennte Betrachtung aller Qualitätsmerkmale erforderlich. Hier genügt eine getrennte Angabe der Genauigkeit, die jedoch in diesem einfachen Beispiel mit der gemessenen Richtung $r = 50\,\mathrm{gon}$ ebenfalls zum selben Ergebnis für die Standardabweichungen beider Koordinatenwerte führt.

In der Abbildung 3.2 sind die einzelnen Auswerteschritte mit Hilfe des Informationsflussdiagramms dargestellt. Die berechneten Ausgangsinformationen können zusammen mit ihrer Qualitätsbeschreibung nun ihrerseits wieder als Eingangsinformationen für weitere Teilprozesse dienen.

3.1.3 Kovarianzfortpflanzung

Die Kovarianzfortpflanzung, wie sie bereits im vorigen Abschnitt 3.1.2 angewendet wurde, ist gut zur Abbildung der Genauigkeit in Prozessen geeignet. Sofern eine analytische Berechnung der Standardabweichungen erfolgen soll, muss dazu der funktionale Zusammenhang zwischen Eingangs- und Ausgangsinformationen bekannt sein. Damit ist die Bestimmung der partiellen Ableitungen und damit die



a) Auswertung der Konsistenz am Beispiel der Streckenmessung

b) Auswertung der Verfügbarkeit

c) Auswertung der Vollständigkeit bzw. der Korrektheit

d) Auswertung der Genauigkeit

Abbildung 3.2: Nach Merkmalen getrennte grafische Darstellung der Qualitätsbewertung nach dem Rechenverfahren von Wiltschko

Berechnung der Varianz-Kovarianzmatrix der Ausgangsgrößen möglich. Bei sehr komplexen, partiellen Ableitungen können die partiellen Ableitungen nach den fehlerbehafteten Größen auch mit Hilfe des numerischen Differenzierens berechnet werden. Dies muss jedoch bei nicht-linearen Zusammenhängen und variierenden Standardabweichungen der Eingangsgrößen immer von Neuem erfolgen, da die Gradienten jeweils nur in einem sehr kleinen Bereich der Funktion gültig sind.

Sind der funktionale Zusammenhang bzw. die resultierenden partiellen Ableitungen sehr komplex, so kann die Varianzfortpflanzung durch eine Monte-Carlo-Simulation (vgl. auch 3.2.3) ersetzt werden. Um eine genaue Kenntnis über die Verteilung der Ausgangsgrößen zu erhalten, muss jedoch eine sehr große Anzahl von Beispielen simuliert werden.

Auf eine Herleitung der Grundlagen der Kovarianzfortpflanzung wird hier verzichtet, diese kann der einschlägigen Literatur entnommen werden (z. B. in Höpcke [1980] oder Sachs und Hedderich [2006]).

3.1.4 Grenzen des bestehenden Verfahrens

Nach der ausführlichen Beschreibung des von Wiltschko entwickelten Verfahrens zur Darstellung und Bewertung der Informationsqualität im vorhergehenden Kapitel 3.1.2, werden in diesem Abschnitt bestehende Einsatzgrenzen der Vorgehensweise herausgearbeitet. Dabei werden insbesondere die Anwendungsbereiche sowie die praktische Umsetzbarkeit kritisch betrachtet.

Die Betrachtung der Qualität von Informationen baut in dem Konzept von Wiltschko auf dem zunächst entwickelten Qualitätsmodell mit seinen sechs inhärenten Merkmalen auf. Dabei wird insbesondere die

flexible Anpassung des Modells an unterschiedlichste Datenarten mit Hilfe einer nicht begrenzten Anzahl von Qualitätsparametern hervorgehoben. Zu jedem Qualitätsmerkmal können grundsätzlich beliebig viele Parameter zur Konkretisierung der Merkmale definiert werden. Damit ist eine sehr differenzierte und problemangepasste Beschreibung der Qualitätsaspekte möglich. Allerdings ist es mit dem Rechenverfahren nicht möglich, diese differenzierte Beschreibung auch in das Informationsflussdiagramm zur Prozessdarstellung aufzunehmen. Insbesondere bietet das Rechenverfahren auf Basis Boolescher Algebra nicht die Möglichkeit, verschiedenartige Qualitätsparameter zu behandeln. Die Berechnung wird hier sogar explizit auf die Merkmalsebene beschränkt. Es ist lediglich eine Behandlung aller Merkmale mit jeweils einem im Intervall $[0, 1]$ liegenden sogenannten Merkmalserfüllungsgrad möglich. Wie aus den verschiedenen Parametern zur Beschreibung eines Merkmals ein einzelner derartiger Merkmalserfüllungsgrad abgeleitet werden kann, ist in der Literatur nicht beschrieben. Dieses ist aus Sicht des Autors auch funktional nicht möglich. Es stellt sich daher auch die Frage, aus welchem Grunde eine aufwändige Erarbeitung einer problemangepassten Menge von Qualitätsparametern überhaupt erforderlich ist, wenn deren Behandlung im Informationsflussdiagramm nicht möglich ist. Diese Einschränkung wird daher als ein wesentlicher Nachteil des Verfahrens im Hinblick auf die Verwendbarkeit über die konzeptionelle Planung von Prozessen hinaus angesehen.

Die vereinfachte Darstellung jeglicher Art von Datenverarbeitung mit der UND-Verknüpfung erscheint nicht immer sinnvoll. Insbesondere bei komplexeren Verarbeitungsschritten, die verschiedenste Arten von Daten in unterschiedlichen zeitlichen Abständen und räumlichen Ausdehnungen erfordern, vermag die vereinfachte UND-Verknüpfung die realen Verhältnisse nicht korrekt abzubilden. Die einfache Multiplikation der Merkmalserfüllungsgrade aller Eingänge ist bei sehr unterschiedlichen Eingangsdaten eine grobe Vereinfachung der tatsächlichen Verhältnisse. Zur Verdeutlichung soll hier, ohne näher auf die Zusammenhänge einzugehen, die Generierung der Längs- und Querabweichung beim polaren Anhängen von Neupunkten dienen. Dieses Anwendungsbeispiel wird im Kapitel 4.1 nochmals aufgegriffen.

Zur Ermittlung der Längs- und Querabweichung eines tachymetrisch bestimmten Neupunktes sind in dem einfachen Beispiel die Eingangsparameter

- gemessene Strecke,
- atmosphärische Korrektur der Streckenmessung (ppm),
- Richtungsmessgenauigkeit des Instruments sowie
- der entfernungsunabhängige Anteil der Streckenmessgenauigkeit des Instruments

notwendig. Alle einzelnen Eingangsinformationen sind zur Generierung der Ausgangsinformation erforderlich, daher kann die Berechnung der Verfügbarkeit nach dem Rechenverfahren von Wiltschko erfolgen. Alle weiteren Qualitätsmerkmale können jedoch aufgrund der sehr unterschiedlichen Eingangsgrößen nicht mit einer einfachen Multiplikation der Merkmalserfüllungsgrade ermittelt werden. Hier müsste zumindest eine geeignete Gewichtung der verschiedenen Eingänge erfolgen, um der Komplexität des Prozesses, insbesondere der Verschiedenartigkeit der Eingangsdaten, Rechnung zu tragen. Dabei stellt sich jedoch unmittelbar das Problem der Bestimmung der Gewichte, mit denen die Qualität der einzelnen Eingangsgrößen berücksichtigt werden müsste.

Des Weiteren ist zur Aufstellung eines Informationsflussdiagramms und zur Berechnung der Qualitätsmerkmalserfüllungsgrade die genaue Kenntnis des zu beschreibenden Prozesses oder Teilprozesses erforderlich. Die Abbildung der Datenflüsse mit den zur Verfügung stehenden Symbolen (vgl. 3.2), als Vorarbeit zur Berechnung der Qualität, erfordert ein hohes Maß an Abstraktionsvermögen. Dies gilt insbesondere bei komplexeren Datenverarbeitungsprozessen wie beispielsweise der Mobilfunkortung, die im Kapitel 4.2 noch zur Veranschaulichung herangezogen wird. Sind die Teilprozesse nicht hinreichend

bekannt, so kann auch keine detaillierte Modellierung der Datenqualität mit dem Informationsflussdiagramm erfolgen. Die Modellierung ist dann auf eine makroskopische Sicht beschränkt, bei der die Anwendung der im Kapitel 3.1.2 vorgestellten Rechenvorschriften nur schwer oder gar nicht möglich ist. Die Verwendung der UND-Verknüpfung zur Darstellung der Qualitätszusammenhänge eines nicht genauer bekannten Prozesses der Datenverarbeitung scheitert zunächst an der Gewichtung der einzelnen Informationseingänge. Es ist allenfalls eine empirische Gewichtung der verschiedenen Eingangsgrößen möglich. Dies ist jedoch sehr aufwändig, insbesondere da stets die bereits angesprochene Beschränkung auf die Merkmalsebene verbleibt.

Schließlich sind zur Modellierung von Qualität mit Hilfe des Rechenverfahrens meist Annahmen für die Merkmalserfüllungsgrade der Eingangsgrößen erforderlich. Diese können bei einem realen Prozess in Echtzeit nicht zur Verfügung gestellt werden. Daher bleibt das Verfahren auf die Planungsphase begrenzt.

In der folgenden Aufstellung sind die aufgezeigten Nachteile des vorgestellten Rechenverfahrens nochmals kurz und übersichtlich dargestellt:

- Die Reduzierung auf Merkmalserfüllungsgrade, Qualitätsparameter können nicht abgebildet werden. Damit ist die erforderliche Abstraktion des zu beschreibenden Problems eine der größten Hürden bei der Anwendung des Rechenverfahrens.

- Die UND-Verknüpfung stellt eine unzulässige Vereinfachung komplexer Prozesse dar. Eine Gleichgewichtung aller Eingänge ist in der Regel nicht sinnvoll.

- Eine sehr detaillierte Kenntnis des zu beschreibenden Prozesses ist zwingend erforderlich.

- Das Verfahren ist nur in der Planung einsetzbar, die Abbildung der Datenqualität eines realen Prozesses in Echzeit ist nicht möglich.

Das Rechenverfahren ist damit nur eingeschränkt und auf einer höheren Abstraktionsebene einsetzbar. Insbesondere kann damit eine Abschätzung der Qualitätsverhältnisse in der Planungs- und Entwicklungsphase von Datenverarbeitungsprozessen erfolgen. Die Behandlung der Verfügbarkeit mit Boolescher Algebra erscheint sinnvoll, da hier die Beschreibung auf Merkmalsebene mit Hilfe einer Verfügbarkeitsrate in Prozent möglich ist. Die detaillierte Abbildung der Datenqualität in beliebigen, komplexen Datenverarbeitungsprozessen hingegen ist nicht möglich. Insbesondere die Darstellung der Abbildung von Datenqualität von Eingangs- auf Ausgangsdaten auf Parameterebene gelingt damit nicht. Lediglich die Genauigkeitsparameter können mit Hilfe der bereits seit langer Zeit bekannten Methode der Kovarianzfortpflanzung exakt abgebildet werden. Sind die partiellen Ableitungen des funktionalen Modells schwer berechenbar, so kann auf numerisches Differenzieren zurückgegriffen werden. Allerdings setzt die Anwendung der Kovarianzfortpflanzung einen mit Formeln geschlossen beschreibbaren, funktionalen Zusammenhang zwischen Ein- und Ausgangsdaten voraus. Die Abbildung eines subtileren, nicht exakt formulierbaren Zusammenhangs gelingt damit jedoch nicht.

Aus diesen Gründen ist es erforderlich, nach weiteren Methoden und Verfahren zu suchen, die eine durchgängige und konkrete Abschätzung von Qualitätsflüssen auf Parameterebene ermöglichen. Insbesondere die Abbildung von subtileren, nicht funktional darstellbaren Zusammenhängen zwischen Ein- und Ausgangsdatenqualität, wie sie zum Beispiel bei der Prognose von ökonomischen Zeitreihen auftreten, soll damit erfolgen können. In den folgenden Abschnitten werden in Frage kommende Methoden und Verfahren beschrieben und deren Eignung zur Lösung der gestellten Aufgabe wird beurteilt. Insbesondere wird dabei nach einer Möglichkeit zur einheitlichen Behandlung aller Qualitätsmerkmale gesucht.

3.2 Alternative Darstellungs- und Fortpflanzungsverfahren

Im vorigen Abschnitt 3.1.2 wurde das von Wiltschko [2004] entwickelte Rechenverfahren vorgestellt und an einem einfachen Beispiel angewendet. Die Grenzen des Verfahrens wurden identifiziert und ausführlich erläutert. In diesem Abschnitt werden die Ergebnisse der Recherche nach weiteren geeigneten Verfahren, die zumindest als Grundlage zur grafischen Abbildung und Modellierung von Qualität in Prozessen dienen könnten, vorgestellt. Dabei orientiert sich die Ausführlichkeit der Darstellung der Verfahren an deren subjektiv beurteilten Eignung zur umfangreichen Beschreibung von Qualität in Prozessen. Ziel ist es, dem Leser die Auswahl des aus Sicht des Autors am Besten geeigneten Verfahrens nachvollziehbar darzulegen und nicht, die angesprochenen Verfahren in ihrer Gänze darzustellen.

Schließlich erfolgt eine kritische Beurteilung der Verfahren und die begründete Auswahl des offensichtlich am besten geeigneten Verfahrens zur Beschreibung von Qualität in Prozessen.

3.2.1 Methoden und Verfahren der Zuverlässigkeitsanalyse

Unter dem Begriff *technische Zuverlässigkeit* oder *Funktionszuverlässigkeit* versteht man im allgemeinen Sprachgebrauch die Eigenschaft einer technischen Einheit, unter den gegebenen Bedingungen seinen Zweck über einen definierten Zeitraum zu erfüllen bzw. fehlerfrei zu funktionieren (VDI 4003 [2007]). Laut Pieruschka [1963] liegt es in der Natur der Dinge, dass jedes technische Erzeugnis irgendwann versagen wird. Wann es versagt ist lediglich eine Frage der Zeit. Daher ist es allgemein üblich, die Zuverlässigkeit von technischen Systemen als *durchschnittliche Zeit zwischen zwei Ausfällen* (Mean time between failure -MTBF), als *Anzahl Ausfälle pro Zeitraum* oder in Form der *prognostizierten Lebensdauer* anzugeben.

Die Notwendigkeit von Zuverlässigkeitsanalysen wurde bereits in den 40er Jahren des letzen Jahrhunderts erkannt. In der Luft- und Raumfahrt waren und sind realistische Abschätzung von technischen Systemen zwingend zur bestmöglichen Vermeidung von Risiken für Leib und Leben notwendig. Die Kerntechnik mit ihren äußerst schwer kalkulierbaren Risiken gab weiteren Anlass für intensive Forschungen, um die damit verbundenen Risiken für die Bevölkerung in Grenzen zu halten bzw. überhaupt quantifizieren zu können. In den folgenden Jahrzehnten wurde die Notwendigkeit der Zuverlässigkeitsabschätzung auf nahezu alle technischen Produkte ausgedehnt. Neben den Sicherheitsaspekten spielt hier mehr und mehr auch die Kundenzufriedenheit, als ein wesentlicher Wettbewerbsfaktor, eine wichtige Rolle für den nachhaltigen wirtschaftlichen Erfolg eines Unternehmens.

Im Rahmen einer Reihe von VDI-Richtlinien wurden die im Laufe der Zeit entwickelten Methoden zur Behandlung der Zuverlässigkeit übersichtlich in dem VDI-Handbuch *Technische Zuverlässigkeit* zusammengestellt und damit als Quasi-Standard auf nationaler Ebene eingeführt. Hierbei wurde die ursprüngliche Definition der Zuverlässigkeit konkretisiert und auf die Einflussfaktoren der Verfügbarkeit wie folgt ausgedehnt (VDI 4001 Blatt 2 [2006]):

> *„[Zuverlässigkeit ist ein] zusammenfassender Ausdruck zur Beschreibung der Verfügbarkeit und ihrer Einflussfaktoren Funktionsfähigkeit, Instandhaltbarkeit und Instandhaltungsbereitschaft.“*

Das Handbuch ist ausdrücklich unbeschränkt auf technische Produkte aller Art und in allen Projektphasen anwendbar und soll helfen, technische Probleme, Kosten und Risiken hinreichend gut beherrschbar zu machen.

Nach Meyna [1982] wird zwischen Zuverlässigkeitsanalyseverfahren auf Grundlage Boolescher und nicht-Boolescher Modellbildung differenziert. Zusätzlich werden bei den Verfahren auf Grundlage der Booleschen Modellbildung zwei Arten von sicherheitstechnischen Analyseverfahren unterschieden. Zum einen sind dies die induktiven Analysen, bei denen ein unerwünschtes Ereignis vorgegeben wird und daraufhin alle möglichen Fehlerfolgen gesucht und eingehend betrachtet werden (z. B. Fehlermöglichkeits- und -einflussanalyse oder Ereignisablaufanalyse). Zum anderen die deduktiven Analysen, bei denen der Ablauf umgekehrt ist. Es wird ein Fehler vorgegeben und nachfolgend nach allen möglichen Ursachen, die zu diesem Fehler führen können, gesucht (z. B. Ursache-Wirkungs-Diagramm oder Fehlerbaumanalyse).

In der Richtlinie VDI 4003 [2007] werden nur *analytische* und *experimentelle* Methoden und Verfahren unterschieden. Die Abbildung 3.3 gibt einen Überblick über einige ausgewählte Verfahren der Zuverlässigkeitsanalyse. Sofern vorhanden, sind jeweils die zugehörigen, aktuellen Normen und Richtlinien mit angegeben.

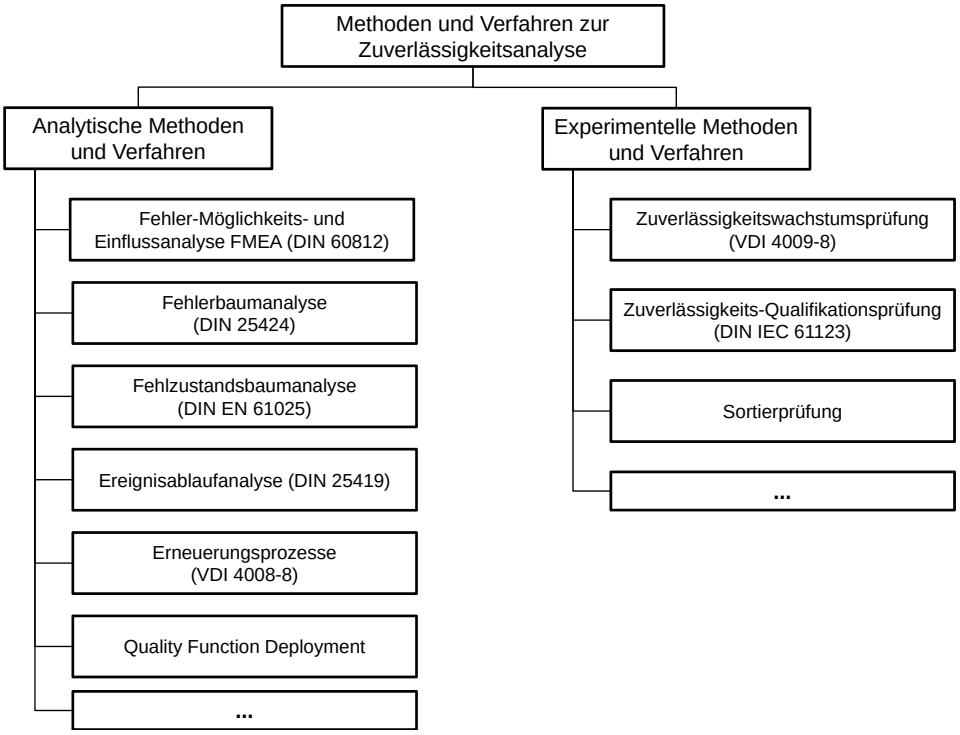


Abbildung 3.3: Übersicht über einige gängige Verfahren der Zuverlässigkeitsanalyse, die in der Richtlinie VDI 4003 [2007] aufgeführt sind.

Die experimentellen Methoden und Verfahren (vgl. Abbildung 3.3, rechte Seite) dienen der Feststellung der Anzahl der Ausfälle in einem Produktionssystem über die Zeit sowie deren Verteilung. Dazu werden in der Regel explizit geeignete Stichproben geprüft. Diese Verfahren sind für die Beschreibung der Qualitätsmerkmale und -parametern von Daten nicht geeignet und werden daher auch nicht näher erläutert. Bei den analytischen Methoden zur Zuverlässigkeitsanalyse sind einige wenige Verfahren dabei, die für diesen Zweck geeignet sein könnten. Daher werden im Folgenden die, aus Sicht des Autors, am erfolgversprechendsten Verfahren kurz vorgestellt und hinsichtlich ihrer Erweiterbarkeit zur Beschreibung von Datenqualität beurteilt.

Die in den Kapiteln 3.2.2 und 3.2.3 separat dargestellten Petri-Netze und die Methodik der Monte-Carlo-Simulation werden in der Richtlinie VDI 4003 [2007] ebenfalls zu den analytischen Methoden der Zuverlässigkeitsanalyse gezählt. Insbesondere aus geodätischer Sicht geht das Potenzial dieser Methoden jedoch weit über die Behandlung der Zuverlässigkeit hinaus, daher erfolgt hier eine Darstellung dieser Verfahren in eigenständigen Kapiteln.

Im Folgenden wird neben den beiden erwähnten Verfahren, die etwas später betrachtet werden, zusätzlich lediglich ein Verfahren aus dem Bereich der Zuverlässigkeitsanalyse exemplarisch dargestellt. Alle weiteren Verfahren sind entweder zu speziell auf einen Aspekt der Zuverlässigkeit ausgerichtet (z. B. die Erneuerungsprozesse, deren Fokus nur auf der Berechnung der zeitlichen Aspekte von Systemen liegt, wie beispielsweise der Lebensdauer) oder die nicht genügend Potenzial zur Erweiterung bieten, wie z. B. das Quality Function Deployment. Die Ereignisablaufanalyse sowie die Fehlerbaumanalyse wurden von Wiltschko bereits zur Entwicklung seines Rechenverfahrens herangezogen und werden daher nicht mehr weiter betrachtet. Die Fehlerbaumanalyse bietet gegenüber der Fehlermöglichkeits- und -einflussanalyse den Vorteil, dass Ausfallkombinationen bzw. Wechselwirkungen zwischen Subsystemen dargestellt werden können. Allerdings basiert das Verfahren, wie das von Wiltschko entwickelte Verfahren, auf Boolescher Algebra und ist daher ebenfalls nicht zur Abbildung komplexer und/oder unbekannter DV-Prozesse auf Parameterebene geeignet.

Fehlermöglichkeits- und -einflussanalyse (FMEA)

Die Fehlermöglichkeits- und -einflussanalyse (FMEA[3]) wird meist als Teammethode eingesetzt, um ein System auf seine Zuverlässigkeit hin zu analysieren. Es handelt sich dabei um ein textbasiertes, induktives Verfahren zur umfassenden Untersuchung der Ausfallarten aller Baueinheiten eines Systems und deren Auswirkungen (Effekte) auf das System. Die FMEA wurde im Raumfahrtprogramm der NASA in den 1960er Jahren entwickelt und findet heute in fast allen Bereichen der Industrie Anwendung (Müller und Tietjen [2000]). Erst mit der DIN 25448 [1990] und später mit der DIN EN 60812 [2006] wurde sie schließlich in die Normenfamilie aufgenommen.

Ausfallkombinationen können mit dieser Methode nicht behandelt werden, die Untersuchung wird jeweils immer nur für eine vorgegebene Ausfallart durchgeführt. Wesentliches Ergebnis der Methode ist eine umfassende Dokumentation von möglichen Problemen, die im untersuchten Subsystem in einem Bauteil auftreten können sowie deren Auswirkungen auf das Gesamtsystem. Das Verfahren wird oft um die Beurteilung der Kritizität (*en.*: criticality) zur FMECA erweitert. Durch die Beurteilung der Kritizität jeder Ausfallart hinsichtlich **Auftreten** (A), **Bedeutung** (B) und **Erkennbarkeit/Behebbarkeit** (E) jeweils auf einer Skala von 1 bis 10 kann zusätzlich jeder Fehlermöglichkeit eine sogenannte **Risikoprioritätszahl (RPZ)** auf einer Skala von 1 bis 1000 zugeordnet werden. Dies erlaubt trotz der stark nicht-linearen Skala eine einfache quantitative Beurteilung verschiedener Risiken und die Bestimmung einer Priorisierungsliste für die Dringlichkeit von Gegenmaßnahmen. Daneben kann die Auswirkung einer Qualitätssicherungsmaßnahme anhand der Senkung der RPZ nachgewiesen werden (vgl. Abschnitt 5.2.6). Dabei ist es ratsam, die Beurteilungstabellen für Auftreten, Erkennbarkeit und Behebbarkeit dem zu beurteilenden System anzupassen, um den unterschiedlichen, zu beurteilenden Systemen gerecht zu werden.

[3] Es existieren in der Literatur eine Reihe von Beschreibungen der Abkürzung FMEA. Gleichbedeutend sind die folgenden Ausdrücke: Fehlermöglichkeits- und -einflussanalyse; Fehlerzustandsart- und -auswirkungsanalyse; Auswirkungsanalyse; Ausfalleffektanalyse; Failure Mode and Effects Analysis

Nach Müller und Tietjen [2000] ist das Ziel der systematisierten Herangehensweise die Beantwortung folgender Fragen:

- **Wo** könnte ein Fehler auftreten?
- **Wie** würde sich der Fehler äußern bzw. wie tritt der Fehler auf?
- **Was** für eine Fehlerfolge könnte sich einstellen?
- **Warum** kann der Fehler oder die Fehlerfolge eintreten?
- **Welche** Auswirkungen auf das Gesamtsystem können auftreten?

Die Frage nach der Auswirkung auf das Gesamtsystem kann mit Hinblick auf die Beschreibung und Modellierung von Datenqualität um die Fragen

- Auf **welche** Qualitätsmerkmale wirkt sich der Fehler aus?
- **Wie** wirkt sich der Fehler qualitativ auf die Qualitätsparameter aus?
- **Wie** kann diesen Qualitätsmängeln begegnet werden?

erweitert werden.

Tabelle 3.4: Formblatt zur Durchführung einer FMECA mit Berücksichtigung der Qualität

lfd. Nr.	Bauteil	Funktion	potenzielle Ausfallart	mögliche Ursachen	lokale Auswirkung	Auswirkung auf System	Auswirkung auf Qualität	Erkennung	vorsorgliche Gegenmaß-nahmen	A/B/E RPZ
	1.1	...								
	...									

Tabelle 3.4 zeigt ein angepasstes Formblatt zur Dokumentation der Ergebnisse einer FMECA. In der letzen Spalte werden die drei geschätzten Komponenten und die daraus berechnete RPZ dokumentiert.

Trotz der Bestimmung von Zahlenwerten, wird die FMECA zu den qualitativen Verfahren gerechnet, da die RPZ lediglich einen Indikator zur Bewertung der Fehlermöglichkeiten darstellt. Die Skala verläuft stark nicht-linear und viele der theoretisch 1000 verschiedenen Abstufungen können schon rein rechnerisch nicht erreicht werden. Zwischen den RPZ 900 (z. B. mit A=10, B=10 und E=9 erreichbar) und 1000 (für A=B=E=10) gibt es keine Zwischenstufen, zwischen 1 und 10 hingegen, sind alle Zahlen möglich.

Die quantitative Beschreibung von Qualitätsmerkmalen und -parametern in den Formblättern ist grundsätzlich möglich. Allerdings bietet die Methode FMECA keine Verfahren, mit denen die Qualitätsparameterwerte konkret berechnet werden können. des Weiteren ist eine Beschreibung von Qualität in realen Prozessen in Echtzeit nicht möglich. Daher kann eine Modellierung der Datenqualität, wie sie in dieser Arbeit angestrebt wird, allenfalls durch Ergänzung der FMECA mit einem weiteren Verfahren erfolgen. Zur qualitativen Analyse von Systemen und zur Evaluierung von Qualitätssicherungs- und -verbesserungsmaßnahmen ist dieses Verfahren jedoch gut geeignet. Daher wird die FMECA im Kapitel 5 als möglicher Bestandteil eines Qualitätsmanagementkonzeptes nochmals aufgegriffen.

3.2.2 Petri-Netze (PN)

Petri-Netze sind sehr gut zur grafischen Darstellung von dynamischen Vorgängen geeignet. Insbesondere können nebenläufige, d. h. nicht in kausalem Zusammenhang stehende Prozesse dargestellt werden. Seit 2008 existiert eine VDI-Richtlinie, in der die Grundlagen der PN, deren Geschichte bis in die 1960er Jahre zurück reicht, sehr detailliert und allgemeingültig dargestellt sind (VDI 4008 Blatt 4 [2008]). Die formale Modellierung von Prozessen erfolgt dabei im Kern mit sogenannten Stellen oder Plätzen und Transitionen (*en.*: Places and Transitions), die durch gerichtete Kanten (*en.*: Edges) zu einem Netzwerk verbunden sind. Die Übergänge zwischen verschiedenen Systemzuständen werden mit Hilfe von Kantengewichten und Marken (*en.*: Token) modelliert. Die Angaben der maximalen Anzahl der Marken symbolisieren dabei die Kapazitäten an den jeweiligen Stellen im Netz. In der ursprünglichen Theorie der PN sind die Marken nicht voneinander unterscheidbar (anonyme Marken) und die Kantengewichte sind immer mit 1 gegeben. In „farbigen“Stellen-Transitionsnetzen können die Stellen jedoch als verschiedene, in der Regel zählbare Objekte bzw. Informationen (z. B. Hammer, Nagel, Koordinate) oder als Zustand (z. B. Nagel in die Wand eingeschlagen, Punkt ist angezielt) verstanden werden. Transitionen stellen Prozesse oder Vorgänge (z. B. Nagel einschlagen, Auto lackieren, Position berechnen) dar.

Mit PN lassen sich insbesondere Eigenschaften wie Nebenläufigkeit von Prozessen und deren Synchronisierung modellieren sowie gegenseitige Ausschlüsse von Prozessen oder Nutzungsbegrenzungen darstellen. Damit ist diese Methodik gut zur Darstellung von Arbeitsabläufen (*en.*: Work flow management - WfM), Organisationsstrukturen oder Systemsteuerungen mit bedingten Schaltzuständen geeignet.

Die Modellierung der funktionalen Eigenschaften eines Petri-Netzes erfolgt mit Hilfe von Kapazitäten, die die maximal mögliche Anzahl von Marken an einer Stelle angibt, sowie durch die Angabe von Mindestanzahlen von Marken für einen Schaltvorgang für jede weiterführende Kante, den sogenannten Kantengewichten. Daneben können zur Modellierung von Wenn-Dann-Bedingungen Kommunikations- und Sperrkanten eingesetzt werden. Dabei handelt es sich um binäre Plätze, die belegt bzw. frei sein müssen, damit eine Transition schaltbereit ist. Durch geschickte Anordnung der zur Verfügung stehenden Netzsymbole gelingt die Modellierung logischer Operatoren wie UND-, ODER- sowie Exklusiv-ODER-Verknüpfungen. In der VDI 4008 Blatt 4 [2008] sind die Netzanordnungen im Detail dargestellt.

In der Abbildung 3.4 ist ein einfaches Netz dargestellt, welches das Potenzial von Petri-Netzen bereits andeutet. Neben der Darstellung der wesentlichen Grundelemente der Petri-Netze sind auch begrenzte Kapazitäten (es wurden nur 20 Nägel gekauft) und die Synchronisierung von Nebenläufigkeiten bzw. die Lösung von Konflikten (nur ein Hammer steht für beide Personen zur Verfügung) in dem Beispiel verdeutlicht. Mit Hilfe von Kommunikationskanten (rot dargestellt) wird ein Konflikt zwischen den beiden Personen aufgelöst, die je einen Hammer für die Aufgabe *„Nagel in die Wand schlagen“* benötigen. Die Kantengewichte, die auch als Kosten bezeichnet werden, sind hier alle mit 1 gegeben.

Der Ablauf der Teilprozesse in dem Prozess *„Bilder aufhängen“* erfolgt in den folgenden Schritten:

- Eine Packung Nägel mit 20 Stück wird gekauft
- Beide Personen benötigen je eine Marke vom Typ *Nagel* bzw. *Hammer*, letztere wird nicht verbraucht, sondern nur verwendet (oder verbraucht und gleich darauf wieder erzeugt)
- Durch die Einführung der Kommunikationskanten und den zwei Stellen (in rot dargestellt) wird der Konflikt um den Hammer gelöst
- Person 1 kann mit der Arbeit beginnen, da die rote Marke im Urzustand auf dem oberen Platz liegt
- Nachdem auf der Stelle *„eingeschlagenen Nagel“*bei Person 1 eine Marke entstanden ist, ist die Transition *„Nagel einschlagen“* für Person 2 schaltbereit usw.

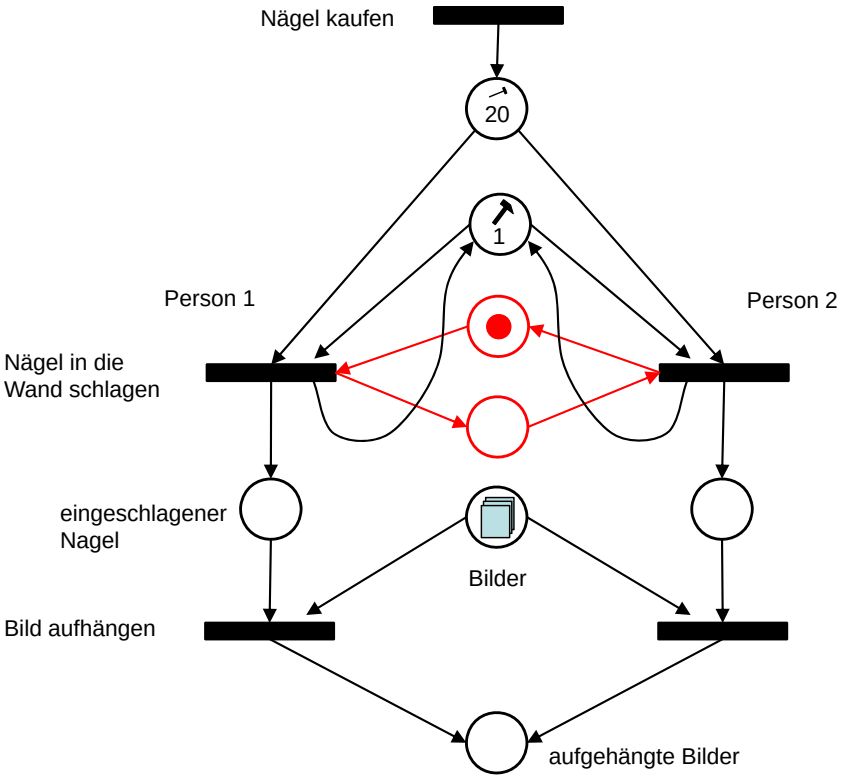


Abbildung 3.4: Einfaches Petri-Netz zur Darstellung des Vorgangs „*Bild aufhängen*“

- Mit den Marken *„eingeschlagener Nagel“* und *„Bild“* kann Person 1 bereits eine Marke des Typs *„aufgehängtes Bild“* erzeugen
- usw.

Der Prozess läuft so lange von oben nach unten durch, bis entweder die Nägel oder die Bilder verbraucht wurden.

Durch *Einfärben* der Marken kann eine Unterscheidung von Marken auf einem Platz erfolgen. Dies erhöht die Flexibilität und vereinfacht die Darstellung. In sogenannten fluiden Petri-Netzen ist es zusätzlich möglich, neben deterministischen Marken (z. B. Personen oder Teile) auch kontinuierliche Marken (z. B. die Temperatur) und damit zeitkontinuierliches Verhalten zu beschreiben (Horton u. a. [1998]).

Zur Anwendung der Petri-Netze für die Analyse der Zuverlässigkeit von Systemen ist meist eine temporale Erweiterung erforderlich. Die Transitionen werden erweitert um eine Zeitspanne T, die als zusätzliche Bedingung erfüllt sein muss, bevor die Transition schalten kann. Da die Schaltvorgänge in Petri-Netzen (Löschen und Erzeugen von Marken) per Definition ohne Zeitverzögerung erfolgen, wird dazu eine zusätzliche Stelle eingeführt, auf der die Marke vorübergehend *geparkt* werden kann. Neben der deterministischen Zeitbewertung von Petri-Netzen kann durch Einführung von Schaltwahrscheinlichkeiten oder Ausfallraten auch eine stochastische Zeitbewertung erfolgen. Damit gelingt bereits eine wesentlich realistischere Beschreibung der Vorgänge in Systemen.

Ein ähnlicher Ansatz zur Erweiterung der Modellierungsmöglichkeiten mit Petri-Netzen ist die Einführung von sogenannten Bewertungsfunktionen für die Stellen und Transitionen. Damit kann ebenfalls der binäre Charakter wie zum Beispiel

- Transition schaltet / schaltet nicht oder
- die maximale Anzahl an Marken auf einem Platz ist erreicht / nicht erreicht

durch unscharfe Beschreibungen ersetzt werden. Die durch die sog. Bewertungsfunktionen erweiterten PN werden als Fuzzy-Petri-Netze bezeichnet (z. B. Cardoso u. a. [1998]).

Herkömmliche Petri-Netze sind markenbasiert, daher können damit zunächst nur abzählbare Eigenschaften behandelt werden. Die Darstellung von Systemen mit herkömmlichen Petri-Netzen bleibt daher auf die ganzheitliche Betrachtung einzelner Produktbestandteile sowie deren Mengen und deren Verwendung zur Bildung weiterer Produkte beschränkt. Damit kann grundsätzlich nur das Qualitätsmerkmal Verfügbarkeit abgebildet werden. Mit den kurz angedeuteten Erweiterungen der klassischen Petri-Netze scheint die Beschreibung weiterer Qualitätsmerkmale - insbesondere von Wahrscheinlichkeiten - neben der Verfügbarkeit grundsätzlich möglich. Mit Hilfe der temporalen Erweiterung von Petri-Netzen ist eine Abbildung der Aktualität denkbar. Die parallele Behandlung verschiedener Qualitätsmerkmale ist jedoch bereits bei einfachen Systemen mit einem sehr großen Aufwand verbunden. Insbesondere die Definition der Bewertungsfunktionen und die Anordnung von Stellen und Transitionen zur Abbildung des Systems erfordert eine genaue Kenntnis der funktionalen Zusammenhänge in den Prozessen.

Die Verwendbarkeit von PN zur Beschreibung von Qualitätsparametern in Prozessen wird derzeit im Rahmen des 2009 gestarteten DFG-Projektes EQuip[4] untersucht (vgl. Berkhahn u. a. [2010]). Das Projekt beschäftigt sich insbesondere mit der Qualitätssicherung ingenieurgeodätischer Prozesse im Bauwesen und läuft vorraussichtlich noch bis 2013. Eine wesentliche Rolle spielt dabei die Beschreibung und Modellierung der Qualität dieser Prozesse, die mit Prädikat/Transitionsnetzen erfolgt. Es wurden bereits erste Erfolge erzielt, die in Kürze veröffentlicht werden sollen (Schweitzer und Schwieger [2011]).

3.2.3 Monte-Carlo-Simulation

Die Monte-Carlo-Simulation (MC-Simulation) ist ein numerisches Verfahren, das bereits in den 1940er-Jahren im Rahmen der Forschung an der Atombombe in Los Alamos von Metropolis und Ulam [1949] entwickelt wurde. Der Name geht auf das Spielkasino im Fürstentum Monaco zurück, in dem mit dem Roulette eine der einfachsten Zufallsgeneratoren im Einsatz ist (Sobol [1991]). Die MC-Simulation dient der Lösung mathematischer Probleme mit Hilfe der simulatorischen Bestimmung geeigneter Zufallsvariablen. Dabei wird nach Stopp [1973]

> *„[...] ein zum realen Objekt oder Vorgang analoges Wahrscheinlichkeitsexperiment konstruiert und [...] genügend oft durchgeführt.“*

Die Anzahl der Durchführungen bedingt dabei die Genauigkeit der Approximation. Die Methode ist insbesondere für Prozesse geeignet, die unbekannt sind oder nur schwer analytisch dargestellt werden können. Allerdings müssen für eine passende Abbildung des vorliegenden Problems in einem Wahrscheinlichkeitsexperiment die Verteilungen der eingehenden Zufallsvariablen hinreichend genau bekannt sein. Die Ergebnisse des Zufallsexperiments ergeben die bis dahin unbekannte Verteilung der gesuchten Zufallsvariablen. Neben der Schätzung des Erwartungswertes μ können weitere Parameter der Verteilung, wie die Varianz σ^2, die Schiefe oder der Exzess, ermittelt werden.

Ein sehr einfaches und anschauliches Beispiel für den Einsatz einer Monte-Carlo-Simulation ist die Annäherung der Zahl π und damit die Bestimmung eines Integrals im Intervall [0,1]. Die Zahl π ist unter anderem definiert als die Fläche des Einheitskreises mit dem Radius 1. Damit ist die Fläche F unter dem

[4] EQuip steht für: Effizienzoptimierung und Qualitätssicherung ingenieurgeodätischer Prozesse im Bauwesen.

Einheitskreis im ersten Quadranten mit $F = \frac{\pi}{4}$ gegeben. Um das erste Viertel des Einheitskreises wird, wie in Abbildung 3.5 dargestellt, ein Quadrat mit der Kantenlänge 1 gelegt. Es werden nun n Punkte bestehend aus je zwei gleichverteilten Zufallszahlen für die Koordinaten x und y im Intervall [0,1] generiert, die in der Abbildung als blaue Punkte erkennbar sind. Durch Auszählen kann nun die Anzahl m der innerhalb des Kreissegmentes liegenden Punkte und damit die Fläche $F = \frac{m}{n}$ ermittelt werden. Die gesuchte Fläche F wird dabei als normalverteilte Zufallsvariable der MC-Simulation betrachtet, deren Erwartungswert μ sich dem wahren Wert von $\frac{\pi}{4}$ mit steigender Anzahl n der Realisierungen des Experiments beliebig genau immer weiter annähert.

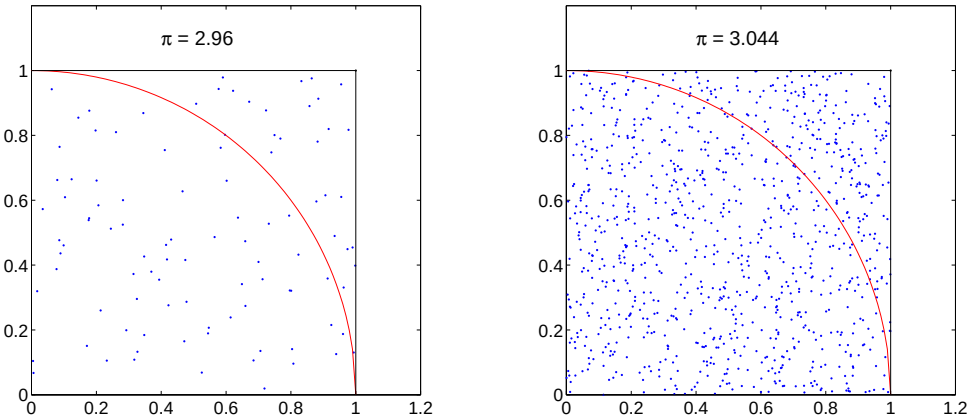


Abbildung 3.5: Annäherung der Zahl π mit Hilfe der MC-Simulation mit $n = 100$ (linke Grafik) und $n = 1000$ (rechte Grafik) Zufallszahlen (in Anlehnung an Ambrosius [2008])

Bei $n = 100$ zufälligen Punkten, ergibt sich nach dem Auszählen zunächst mit $\mu \approx 3$ nur eine sehr grobe Näherung der Konstanten π (vgl. 3.5, linke Grafik). Bei mehrmaligem Durchführen der 100 Experimente ergibt sich eine Streuung von etwa ± 0.5. Bei $n = 1000$ ergibt sich bereits die richtige Vorkommastelle (vgl. 3.5, rechtes Bild) und bei $n = 10000$ ist mit dem Ergebnis 3.159 die Kreiszahl π bereits schon bis zur ersten Nachkommastelle angenähert. Bei einer Million Experimenten stimmt der Erwartungswert von F mit $\mu = 3.141$ bis auf etwa zwei Nachkommastellen mit dem wahren Wert π überein.

An diesem Beispiel ist bereits der wesentliche Nachteil des Verfahrens erkennbar. Um einen hohen Grad der Approximation zu erlangen, steigt die Anzahl der erforderlichen Zufallsexperimente und damit der Rechenaufwand und die Rechenzeit sehr schnell an. Um die Genauigkeit zu verzehnfachen, d. h. um eine Nachkommastelle zu steigern, muss nach Hengartner und Theodorescu [1978] die Anzahl der Experimente verhundertfacht werden.

In der Richtlinie VDI 4008 Blatt 6 [1999] wird gezeigt, dass sich das Problem bei sinkender Eintrittswahrscheinlichkeit für die zu untersuchenden Ereignisse noch verschärft. So sind „*[...] zur Untersuchung eines System mit einer Ausfallwahrscheinlichkeit von* 10^{-6} *[...]* 10^{8} *Spiele (Experimente) erforderlich, um mit einer Wahrscheinlichkeit von* $68\,\%$ *(1* σ*) eine Angabe mit einem relativen Fehler von* $> 10\,\%$ *zu gewährleisten*“.

Im Hinblick auf die Modellierung von Datenqualität ist die MC-Simulation insbesondere zur Bestimmung der Genauigkeit einer Zufallsgröße geeignet. Dazu kann aus der ermittelten Verteilung und deren ggf. geschätzten Erwartungswert über die Verbesserungen (bei geschätztem Erwartungswert) oder die zufälligen Abweichungen (bei bekanntem Erwartungswert) die Varianz und damit die Standardabweichung

einer gesuchten Zufallsvariablen ermittelt werden. Die MC-Simulation kann eine Kovarianzfortpflanzung ersetzen (vgl. 3.1.3). Der Vorteil liegt insbesondere darin, dass die Bestimmung der unter Umständen sehr komplexen partiellen Ableitungen nach den fehlerbehafteten Größen entfällt. Daneben kann die MC-Simulation sogar bei unbekannten Zusammenhängen zwischen Eingangs- und Ausgangsdaten eingesetzt werden. Bei Verwendung großer Stichprobenumfänge kann die empirische Bestimmung der exakten Lösung beliebig angenähert werden.

Für die Beschreibung weiterer Qualitätsmerkmale von Daten wie der Verfügbarkeit, Aktualität, Vollständigkeit oder Konsistenz ist die MC-Simulation jedoch nicht geeignet. Vollständige und verfügbare Daten werden bei der MC-Simulation bereits als gegeben vorausgesetzt. Lediglich die Korrektheit statistischer Tests kann über die Bestimmung der Fehler 1. und 2. Art aus der Verteilung beurteilt werden. Beide Kenngrößen geben Aufschluss über die Zuverlässigkeit einer Testentscheidung.

3.2.4 Künstliche neuronale Netze - KNN

Bei künstlichen neuronalen Netzen[5] (KNN), im Englischen mit *artificial neural networks (ANN)* bezeichnet, handelt es sich um eine Methodik, die insbesondere zur Lösung von Problemen entwickelt wurde, bei denen klassische Kombinatorik an ihre Grenzen stößt. Vorbild ist dabei das Neuronennetz des Menschen, bestehend aus Gehirn, Rückenmark sowie den Nervenbahnen. Das Nervensystem des Menschen besteht aus ca. 10^{11} Neuronen, die bei einem erwachsenen Menschen jeweils über durchschnittlich 10^4 Synapsen mit benachbarten Neuronen vernetzt sind (Zell [1997]). Dadurch entsteht eine sehr hohe Komplexität und es wird eine enorm effiziente, parallele Informationsverarbeitung ermöglicht, die der klassischen von-Neuman-Rechnerarchitektur und auch den heutigen Parallelrechnern mit mehreren CPUs nach wie vor weit überlegen ist.

Wesentliche Eigenschaften von KNN

Mit einer sehr vereinfachten Nachbildung der menschlichen Neuronenstruktur in Form von KNN kann man sich die folgenden Eigenschaften des menschlichen Gehirns zu Nutze machen:

- Parallele Informationsverarbeitung,
- Lernfähigkeit und Anwendung des Erlernten,
- Generalisierungs- und Assoziationsfähigkeit; Lösung von Problemklassen,
- hohe Fehlertoleranz (z. B. gegenüber verrauschter Eingangsgrößen) sowie der
- sehr schnellen Informationsverarbeitung nach der Lernphase.

Die Forschung auf dem Gebiet der KNN reicht in die frühen 1940er-Jahre zurück. Mit einem sehr einfachen Modell einer Nervenzelle versuchten McCulloch und Pitts [1943] die Funktionsweise des menschlichen Gehirns besser zu verstehen. Damit begann die erste Phase intensiver Forschungen auf dem Gebiet der KNN, die mit einer Veröffentlichung von Minsky und Papert [1969] nahezu zum Erliegen kam. Minsky und Papert erbrachten den Nachweis, dass die bis dahin verwendeten einfachen Neuronalen Netze, die sogenannten Perzeptrone, nicht in der Lage sind, die XOR-Funktion und damit eine ganze Klasse von Problemen nachzubilden. Daraus wurde zu Unrecht der Schluss gezogen, dass diese Einschränkung auf alle Arten von KNN zutreffen würde und die Forschungstätigkeiten auf diesem Gebiet kamen ins

[5] Oft wird auch nur von „neuronalen Netzen“oder „neural networks“gesprochen; ob jeweils natürliche oder künstliche Netze gemeint sind, ergibt sich in der Regel aus dem Zusammenhang.

Stocken. Nur eine Handvoll Forscher waren in den 1970er-Jahren auf diesem Gebiet tätig, unter ihnen Kohonen, Anderson und Hopfield, bevor Mitte der 1980er-Jahre mit fortschreitender Entwicklung leistungsfähiger Computer ein bis heute ungebremster Boom auf dem Gebiet der KNN einsetzte.

Anwendungspektrum für KNN

KNN kommen heute wegen ihrer Flexibilität in einem sehr breiten Spektrum zur Lösung der unterschiedlichsten Anwendungen zum Einsatz. Insbesondere in den Ingenieur- und Naturwissenschaften sowie den Wirtschaftswissenschaften spielen KNN eine wichtige Rolle. Zur Vermittlung eines kleinen Eindrucks der Vielseitigkeit von KNN sind im Folgenden einige ausgewählte Beispiele aus verschiedenen Disziplinen bzw. Problemklassen dargestellt, in denen KNN heute in der Praxis eingesetzt werden:

- Mustererkennung: Erkennung von Text oder Sprache, Identifikation von Gesicht, Stimme, etc.,
- Maschinen- und Anlagenbau: Regelung von Anlagen,
- Literaturwissenschaft: Analyse und Erkennung von Schreibstilen,
- Wirtschaftswissenschaften: Prognose von Kursentwicklungen,
- Hydrologie: Hochwasservorhersage,
- Qualitätskontrolle: Beurteilung von Schweißnähten; Qualitätsanalyse anhand Motorengeräusch,
- Geodäsie: Signalprognose bei GPS, Überwachung von Rutschhängen; Bildverarbeitung,
- Immobilienwirtschaft: Ermittlung von Immobilienpreisen,
- Luft- und Raumfahrt: Weiterentwicklung des Autopiloten, Instrumentenlandesysteme,
- Medizintechnik: Diagnostik, Analyse von EEG- oder MRT-Bildern.

Eine sehr umfangreiche Zusammenstellung von Anwendungsbereichen in diversen Disziplinen findet sich beispielsweise in Hagan u. a. [1996]. Zell [1997] stellt einige sehr detailliert beschriebene Anwendungsbeispiele aus den Bereichen Mustererkennung, Prognose sowie der Regelungstechnik dar. Eine umfangreiche Sammlung an Veröffentlichungen in jüngster Zeit zur Anwendung von KNN in der Geodäsie wurde in Frank [2010] zusammengestellt. Aus der langen Liste von Beispielen lässt sich ablesen, wie vielseitig KNN tatsächlich sind. Die wesentliche Aufgabe der KNN in den genannten Anwendungen ist der Ersatz von Expertenwissen und das Erkennen und Erlernen komplexer Zusammenhänge. Es muss eine Verknüpfung zwischen den Eingangsgrößen des zu beschreibenden Systems (Bilder, Sprache, Signal, Kurswerte, Messwerte, etc.) und dem Ergebnis, der Systemantwort (Zahl, Schriftsteller, Kursentwicklung, Hochwasserereignis, Verschiebung, etc.) gefunden werden. In der Regel eignen sich KNN genau dann, wenn die Zusammenhänge zwischen Ein- und Ausgang nicht klar definierbar bzw. formal darstellbar sind. Zum Beispiel ist der Zusammenhang zwischen einem Bild einer Schweißnaht und dem Urteil des Experten („Naht ist ok“oder „Nacharbeit erforderlich“) kaum in einer exakten Abbildungsgleichung darstellbar. Dasselbe gilt für verrauschte Eingangsgrößen, wie beispielsweise bei der Gesichtserkennung. Die Eingangsgröße *Bild des Gesichts der zu identifizierenden Person* ist von diversen Faktoren wie Beleuchtung, Gesundheitszustand, Aufnahmewinkel, Gesichtsausdruck, Bekleidung (Sonnenbrille, Mütze, usw.), Bart, und vielen weiteren abhängig. Es ergeben sich daher nahezu unendlich viele Varianten der Eingangsgröße, die alle auf dieselbe Ausgangsgröße -die zu identifizierenden Person- abgebildet werden sollen. Für derartige Probleme, bei denen die konventionelle Numerik an ihre Grenzen stößt, stellen KNN einen möglichen Ausweg dar. Es gibt heute eine große Anzahl mehr oder weniger spezieller Netzarten. Nach ihrer Topologie lassen sich im Wesentlichen drei Arten von Netzen unterscheiden,

- die klassischen Feed-Forward-Netze,
- Netze mit Rückkopplungen und
- vollständig verbundene Netze.

In den ersten beiden Netzvarianten wird das Netz in verschiedenen Schichten angeordnet, die jeweils mehrere Neuronen zusammenfassen. Dabei sind bei den Feed-Forward-Netzen, wie der Name bereits andeutet, nur vorwärtsgerichtete Verknüpfungen zwischen Neuronen zulässig. Dies setzt für das gesamte Netz eine definierte Fließrichtung voraus, die auch bei Netzen mit Rückkopplungen vorgegeben sein muss. Dort sind allerdings einzelne Rückkopplungen in die jeweils vorhergehende Neuronenschicht zulässig. In vollständig verbundenen Netzen hingegen gibt es keine vorgegebene Fließrichtung.

Des Weiteren kann eine Unterscheidung der Netze unter anderem nach dem eingesetzten Lernverfahren, nach ihrer Dimensionierung und nicht zuletzt nach ihren Erfindern erfolgen. Aus funktionaler Sicht können KNN des Weiteren in statische und dynamische Netze unterteilt werden[6]. In dynamischen Netzen werden zeitliche Verzögerungen modelliert und der Netzausgang berücksichtigt neben den aktuellen Eingangswerten auch die vorhergehenden Ausgänge und/oder Eingangswerte. In dieser Arbeit werden jedoch nur die klassischen Feed-Forward-Netze näher betrachtet und für die Modellierung von Qualität in Prozessen herangezogen. Dieser Netztyp findet sich in der Literatur, und dabei insbesondere bei praktischen Problemlösungen mit KNN, mit Abstand am häufigsten. Aufgrund ihres Potenzials sind diese Netze nach dem aktuellen Kenntnisstand auch für den hier verfolgten Zweck gut geeignet. Es soll in dieser Arbeit insbesondere der Nachweis erbracht werden, dass KNN grundsätzlich für die Modellierung von Datenqualität geeignet sind. Daher genügt hier die Anwendung von relativ einfach zu handhabenden KNN. Weitergehende Ansätze werden im Ausblick aufgezeigt. Im Folgenden wird die Funktionsweise der KNN, insbesondere der Feed-Forward-Netze, in ihren Grundzügen dargestellt und erläutert.

Funktionsweise Künstlicher Neuronaler Netze

Seit den 1980er Jahren wurden eine Vielzahl neuer Netzarten entwickelt, die einzelne Nachteile der klassischen KNN wettmachen sollten oder auf spezielle Problemklassen besser zugeschnitten sind. In allen folgenden Ausführungen liegt der Fokus jedoch auf den klassischen **Multilayer Feed-Forward-Netzen - MLFFN**. Bei diesen Netzen existiert eine definierte Fließrichtung der Informationsverarbeitung innerhalb des Netzes von den Eingangswerten hin zu den Ausgangswerten. Es wird keine Schicht (Neuronen sind innerhalb der Netze in der Regel in Schichten (*en.:* layern) angeordnet) übersprungen und es gibt keine direkten oder indirekten Rückkopplungen.

Eine Problemklasse, die sich mit Hilfe eines klassischen MLFFN lösen lässt, kann ganz allgemein wie in der Abbildung 3.6 als Black Box dargestellt werden. Mit dem KNN soll der unbekannte Zusammenhang zwischen den n Eingangs- und m Ausgangsgrößen, der sich rein mathematisch als Abbildung der Form

$$f : \mathbb{R}^n \rightarrow \mathbb{R}^m \tag{3.27}$$

beschreiben lässt, modelliert werden. Nicht nur jeder funktionale Zusammenhang, der eine endliche Anzahl von Unstetigkeiten aufweist, kann mit KNN beschrieben werden (Hornik u. a. [1989]), sondern grundsätzlich gelingt auch die Darstellung andersartiger, subtilerer Zusammenhänge zwischen den Ein- und Ausgängen. Zum Beispiel wird niemand bestreiten, dass die Aussicht, die umliegende Infrastruktur

[6] Die Bezeichnungen „statische“und „dynamische Netze“werden hier analog wie in Matlab® [2010] verwendet. Mit statischen Netzen sind in diesem Zusammenhang solche ohne Rückkopplungen und Verzögerungen in der Eingabe der Beispiele gemeint.

oder der gesellschaftliche Ruf einer Gegend einen signifikanten Einfluss auf die Immobilienpreise haben werden, allerdings ist dieser Zusammenhang nur sehr schwer in einer Formel darstellbar. Das ein Problem nicht funktional beschreibbar ist, kann unterschiedliche Ursachen haben. Entweder ist zu wenig über den Zusammenhang der Größen bekannt, der Zusammenhang unterliegt der Geheimhaltung oder die Beschreibung gelingt nicht aufgrund der Komplexität oder Subtilität des Zusammenhangs.

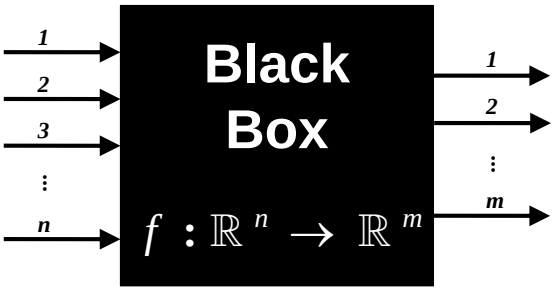


Abbildung 3.6: Black Box als Darstellung der mit KNN lösbaren Problemklasse

Das Erlernen des durch die Black Box symbolisierten, unbekannten Systemverhaltens durch das KNN kann auf drei unterschiedliche Arten erfolgen. Beim **überwachten Lernen** ist eine ausreichend große Trainingsmenge **P** mit Eingangs- und Ausgangsvektoren **p** und **t** vorhanden. Diese Art des Trainings entspricht in der Regel nicht dem biologischen Lernen, ist jedoch relativ einfach umzusetzen und wird sehr häufig verwendet. Daneben besteht die Möglichkeit, das Netz nur durch **bestärkendes Lernen** zu trainieren. Dabei wird nur eine binäre Information über den Systemausgang bereitgestellt. Zu jedem Eingangsdatensatz ist nicht der zugehörige Ausgangswert, sondern nur die Information „Ausgang korrekt/nicht korrekt“bekannt. Diese Art des Lernens ist vergleichbar mit einer Tierdressur. Richtiges Verhalten wird mit einer Belohnung honoriert. Schließlich gibt es eine Klasse von Problemen, bei denen keine Information über die Eingangs-Ausgangs-Beziehung der Lernbeispiele existieren, die Ausgangswerte sind unbekannt. Die Eingangstrainingsmenge kann nur auf Ähnlichkeiten hin untersucht werden und dadurch kann eine Clusterung der Eingangswerte erlernt werden. Diese Art des Lernens entspricht am Besten dem biologischen Vorbild und wird als **unüberwachtes Lernen** bezeichnet. Peters [2008] ist es gelungen, mit dieser Art des Lernens ein KNN zur Vorhersage des sehr selten eintretenden Falls einer Überschwemmung trainieren zu können. Aufgrund der enorm kleinen Trainingsmenge von Hochwasserereignissen kam für diese Anwendung weder ein überwachtes noch ein bestärkendes Lernen in Frage. Allerdings muss hier erwähnt werden, dass die sogenannten selbstorganisierenden Merkmalskarten, die ohne Lernbeispiele bzw. Ausgangsdaten trainiert werden können, lediglich in der Lage sind, eine intelligente Clusterung der Eingangsdaten vorzunehmen. Damit sind diese Netze zur Modellierung von Datenqualität nicht geeignet.

Zum besseren Verständnis der Funktionsweise der KNN und des Begriffs „Lernen“in diesem Zusammenhang, wird im Folgenden zunächst einmal der Aufbau und die Funktionsweise von natürlichen Nervenzellen und ihren künstlichen Pendants sowie deren Vernetzung dargestellt. Die Art der Darstellung sowie die Notation entspricht weitgehend derjenigen, die in Hagan u. a. [1996] verwendet wird. Diese Notation kommt auch in der Neural Network Toolbox der Matlab-Software zum Einsatz.

Die Abbildung 3.7 zeigt im linken Teil den schematischen Aufbau einer Nervenzelle und im rechten Teil deren vereinfachte mathematische Nachbildung. Im biologischen Sinne versteht man unter einer

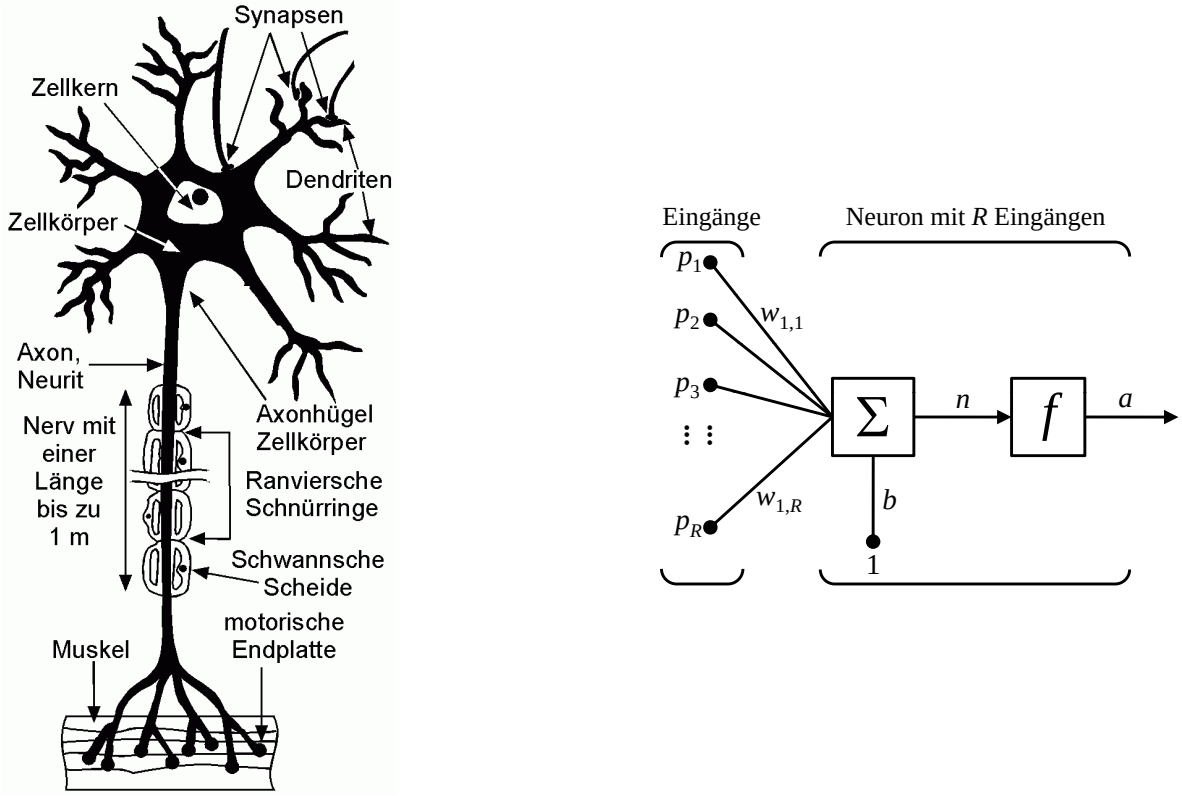


Abbildung 3.7: Gegenüberstellung einer Nervenzelle (Völz [1999]) und eines künstlichen Neurons (nach Hagan u. a. [1996])

Nervenzelle eine auf die Informationsverarbeitung spezialisierte Zelle. Dabei werden verschiedene Unterarten wie Haupt-, Inter-, Sensor- oder -wie hier abgebildet- Motoneuronen nach ihren Aufgaben unterschieden. Die Motoneuronen leiten eine Reaktion auf einen über Sensorneuronen empfangenen Reiz in die entsprechenden Muskeln weiter, um zum Beispiel die Hand von der versehentlich berührten, heißen Herdplatte wegzuziehen. Vereinfacht dargestellt besteht eine Nervenzelle aus dem Zellkörper, der den Zellkern und damit die Erbinformation enthält sowie den baumartig verzweigten Dendriten, die als Rezeptoren für Synapsen benachbarter Nervenzellen dienen (vgl. Abbildung 3.7, links). Ein einzelnes Neuron ist in der Regel auf diese Art und Weise mit tausenden weiteren Zellen verbunden. Des Weiteren verfügt jedes Neuron über ein Axon, welches den Reizleiter bzw. Nerv darstellt und das in einer Verzweigung von Synapsen endet. Diese Nerven können bis zu $1\,\mathrm{m}$ Länge erreichen, im Gegensatz zu den Dendriten, die lediglich eine Ausdehnung von wenigen $\mu\mathrm{m}$ bis mm haben (Völz [1999]).

Die Information, die im Neuronennetz „fließt“, besteht innerhalb der Nervenzelle aus elektrischen Impulsen, die durch fortwährend angeregte, kurze Depolarisierungen (Dauer ca. $2.4\,\mathrm{ms}$) mit einer Geschwindigkeit von bis zu $100\frac{\mathrm{m}}{\mathrm{s}}$ wellenartig entlang der Zellmembran des Axoms bis zur Synapse wandert (Zell [1997]). An der Synapse erfolgt die Informationsweiterleitung biochemisch auf die Dendrite der benachbarten Nervenzelle oder direkt in eine Muskelzelle. Damit ist die Schaltgeschwindigkeit eines Neurons mit ca. $10^{-3}\,\mathrm{s}$ viel langsamer als die Taktfrequenz eines herkömmlichen Computers, bei dem heute Schaltzeiten jedes Transistors von weit weniger als $10^{-9}\,\mathrm{s}$ üblich sind. Dies wird jedoch durch die massive Parallelität der Signalverarbeitung und -weiterleitung in Neuronennetzen weit mehr als nur kompensiert (Zell [1997]). Ob und wann ein Neuron „feuert“, d. h. Impulse weiterleitet, wird durch Schwellwerte geregelt, die erst überschritten werden müssen, damit ein Aktivierungspotenzial die Zellreaktion in Gang setzten kann. Dies hängt entscheidend von der Aufgabe des Neurons und der Gewichtung der eintreffenden Reize und damit der Stärke der einzelnen Synapsen anderer Neuronen ab.

Die Nachbildung eines Neurons ist auf der rechten Seite der Abbildung 3.7 dargestellt. Vergleichbar mit dem Zellkörper eines Neurons, besteht der Hauptteil des künstlichen Abbildes aus der mit $\sum$ symbolisierten Übertragungsfunktion, in der alle eintreffenden Signale zusammengefasst werden und der Aktivierungsfunktion, die der Berechnung des Neuronenausgangs dient. Die Aktivierungsfunktion wird in der Regel mit f bezeichnet. Die Eingangswerte p_i werden jeweils mit den Gewichten $w_{1,i}$ gewichtet, wobei der erste Index das jeweilige Neuron bezeichnet und der zweite den jeweiligen Eingangswert. In der Abbildung besteht das Netz zunächst nur aus einem Neuron, daher ist der erste Index hier immer 1. Der Pfeil mit dem Ausgabewert a entspricht dem Nerv im biologischen Vorbild der Nervenzelle. Der Schwellwert b (*en.*: bias) ist vergleichbar mit den Eingängen, jedoch gibt es immer nur einen Schwellwert je Neuron und der Eingangswert ist immer $= 1$. Der Schwellwert bietet in KNN eine weitere Möglichkeit der Modellierung, kann aber rein rechnerisch auch durch einen weiteren Eingabewert p ersetzt werden. Dies vereinfacht die Berechnung eines KNN erheblich, da die separate Behandlung der Schwellwerte entfällt. Gelegentlich findet sich in der Literatur zwischen der Transferfunktion und dem Netzausgang noch eine sogenannte *Ausgabefunktion*. Da diese Ausgabefunktion in der Realität meist aus der Identitätsfunktion besteht, wird hier auf die Darstellung dieser zusätzlichen Möglichkeit der Netzmodellierung zu Gunsten der Übersichtlichkeit verzichtet.

Der Ausgabewert a ergibt sich als Funktionswert der Aktivierungsfunktion f mit der Netzeingabe n, die sich mit Hilfe der Übertragungsfunktion $\sum$ als gewichtete Summe der R Eingänge und dem Schwellwert b mit

$$n = \sum_{i=1}^{R} (w_{i,1} \cdot p_i) + b = \mathbf{W} \cdot \mathbf{p} + b \tag{3.28}$$

berechnet. Dabei beinhaltet die Gewichtsmatrix $\mathbf{W}$ hier nur einen Vektor mit den Gewichten des R-dimensionalen Eingangsvektors $\mathbf{p}$ des Neurons 1:

$$\mathbf{W} = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,R} \end{bmatrix} \qquad \mathbf{p} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_R \end{bmatrix} \tag{3.29}$$

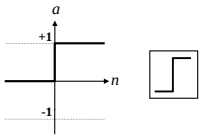
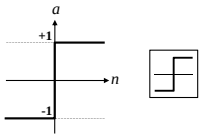
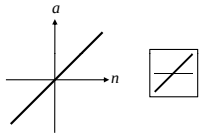
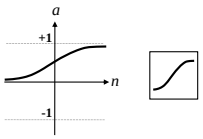
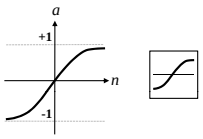
In Matrixschreibweise lässt sich damit die Berechnung des Ausgangswertes des Neurons wie folgt darstellen:

$$a = f(n) = f(\mathbf{W} \cdot \mathbf{p} + b) \tag{3.30}$$

Die Transferfunktion f kann dabei je nach gewünschtem Verhalten des Netzes sehr unterschiedlich gewählt werden. Die Art des gewünschten Ausgangswertes schränkt die Wahl der Transferfunktion bereits ein. Wird beispielsweise ein binärer Ausgang mit den Funktionswerten 0 oder 1 gefordert, so kann eine Schwellwert-Transferfunktion (*en.*: hard limit transfer function - hardlim) geeignet sein. Die folgende Tabelle 3.5 gibt einen kleinen Überblick über häufig verwendete Transferfunktionen und deren Charakteristik. In der letzten Spalte ist jeweils die Bezeichnung der Funktion in Matlab angegeben.

Die logarithmische Sigmoidfunktion (vgl. Tabelle 3.5) kommt insbesondere in mehrschichtigen Netzen zum Einsatz, da der häufig verwendete Backpropagation-Lernalgorithmus eine differenzierbare Funktion erfordert. Neben diesen am häufigsten eingesetzten Transferfunktionen existieren weitere, gängige Funktionen, die bei speziellen Problemstellungen eine Rolle spielen. Grundsätzlich können auch neue Transferfunktionen eingeführt werden, die mit Hinblick auf den verwendeten Lernalgorithmus ggf. stetig differenzierbar sein müssen.

Tabelle 3.5: Typische Transferfunktionen (Hagan u. a. [1996] und Matlab® [2010])

Graph und Symbol	Name	Abbildung f	Funktion in Matlab
	Schwellwert-funktion	$a = 0 \,\forall\, n < 0$ $a = 1 \,\forall\, n \leq 0$	hardlim
	Symmetrische Schwellwert-funktion	$a = -1 \,\forall\, n < 0$ $a = +1 \,\forall\, n \leq 0$	hardlims
	Linearfunktion	$a = n$	purelin
	Logarithmische Sigmoidfunktion	$a = \frac{1}{1+e^{-n}}$	logsig
	Tangens Hyperbolicus	$a = \frac{e^{n}-e^{-n}}{e^{n}+e^{-n}}$	tansig

Zur Lösung komplexerer Fragestellungen mit KNN ist die Verknüpfung der einzelnen Neuronen zu Netzen erforderlich. Dabei kann das Netz in zwei Dimensionen erweitert werden. Einerseits können mehrere Neuronen in einer Schicht parallel angeordnet werden und andererseits ist die Verwendung mehrerer Schichten üblich. Intuitiv ist klar, dass mit der Komplexität der Aufgabe auch die Netzgeometrie komplexer werden muss. Der Übersichtlichkeit halber werden in einem komplexeren Netz nicht mehr alle Verbindungen dargestellt, sondern es erfolgt eine vektorielle Zusammenfassung mit der Angabe der Vektor- und Matrixdimensionen. In der Abbildung 3.8 ist die vereinfachte Darstellung eines KNN bestehend aus einer Schicht veranschaulicht.

In der Regel werden in einer Schicht identische Transferfunktionen eingesetzt. Dies ist jedoch nicht zwingend erforderlich. Die Dimension der Vektoren b und a ist $S \times 1$. Sie hängt von der Anzahl S der Neuronen in der jeweiligen Schicht ab. Die Dimension des Eingabevektors p ist $R \times 1$, daher ergibt sich die Dimension der Gewichtsmatrix

$$\mathbf{W} = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,R} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,R} \\ \vdots & \vdots & & \vdots \\ w_{S,1} & w_{S,2} & \cdots & w_{S,R} \end{bmatrix} \tag{3.31}$$

zu $S \times R$. Damit lässt sich der Ausgabevektor $\mathbf{a}$, ähnlich wie in der Formel 3.30, durch Ersetzen des skalaren Schwellwertes b und der Funktion f durch die Vektoren $\mathbf{b}$ bzw. $\mathbf{f}$ berechnen:

$$\mathbf{a} = \mathbf{f}(\mathbf{n}) = \mathbf{f}(\mathbf{W} \cdot \mathbf{p} + \mathbf{b}) \tag{3.32}$$

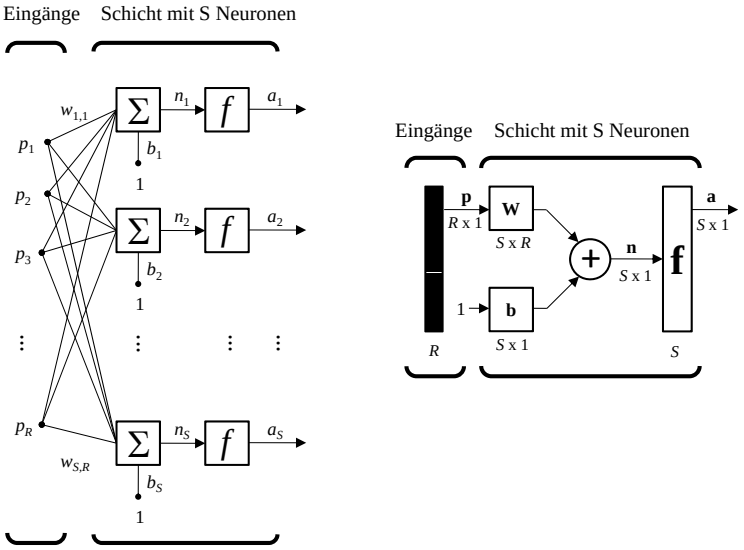


Abbildung 3.8: KNN mit einer Schicht: Ausführliche und verkürzte Darstellung

Wie bereits in der Einleitung erwähnt wurde, können KNN mit nur einem Layer nicht alle Klassen von Funktionen darstellen (vgl. 3.2.4). Daher werden in der Regel Netze modelliert, die aus mehreren aufeinander folgenden Schichten bestehen. Durch die oben eingeführten Symbole ist eine übersichtliche Darstellung derartiger Netze trotz unüberschaubar vieler Verknüpfungen möglich. Jede Schicht verfügt über einen eigenen Schwellwert-Vektor $\mathbf{b}$ sowie eine Funktionsmatrix $\mathbf{f}$ und eine Gewichtsmatrix $\mathbf{W}$. Der jeweilige Eingangsvektor ist entweder der R-dimensionale Netzeingang $\mathbf{p}$, wenn es sich um die erste Schicht handelt, oder der Ausgangsvektor $\mathbf{a}$ der jeweils vorhergehenden Schicht. Im zweiten Fall entspricht die Dimension der Anzahl S von Neuronen der vorangegangenen Schicht. Entsprechend dem Vorschlag in Hagan u. a. [1996] wird die Zugehörigkeit der Größen zu einem bestimmten Layer mit Hilfe eines hochgestellten Indexes -der Nummer des Layers- verdeutlicht. Damit lassen sich auch mehrschichtige Netze übersichtlich darstellen, wie in Abbildung 3.9 an einem dreischichtigen Netz verdeutlicht wird.

Die letzte Schicht von Neuronen wird als *Ausgabeschicht (en.: output layer)* bezeichnet, alle weiteren Schichten als *verdeckte Schichten*[7] *(en.: hidden layers)* (Hagan u. a. [1996]). Das in der Abbildung 3.9 dargestellte Netz verfügt damit über zwei verdeckte Schichten, Schicht eins und Schicht zwei. Ein derartig dimensioniertes KNN wird oft auch mit „$s^1 - s^2 - s^3$−Netz“bezeichnet, wobei s^i für die Anzahl der Neuronen in der i-ten Schicht steht. Damit kann die Dimension des Netzes bereits aus der Bezeichnung abgeleitet werden.

Die Berechnung des Ausgangs eines mehrschichtigen Netzes kann durch Schachtelung der Formel 3.32 durchgeführt werden. Dabei stellen die Vektoren $\mathbf{p}$, $\mathbf{a}^1$ und $\mathbf{a}^2$ die Eingangsvektoren der ersten,

[7] In der Literatur wird entgegen der hier dargestellten Definition gelegentlich auch die Ausgabeschicht zu den verdeckten Schichten hinzugezählt.

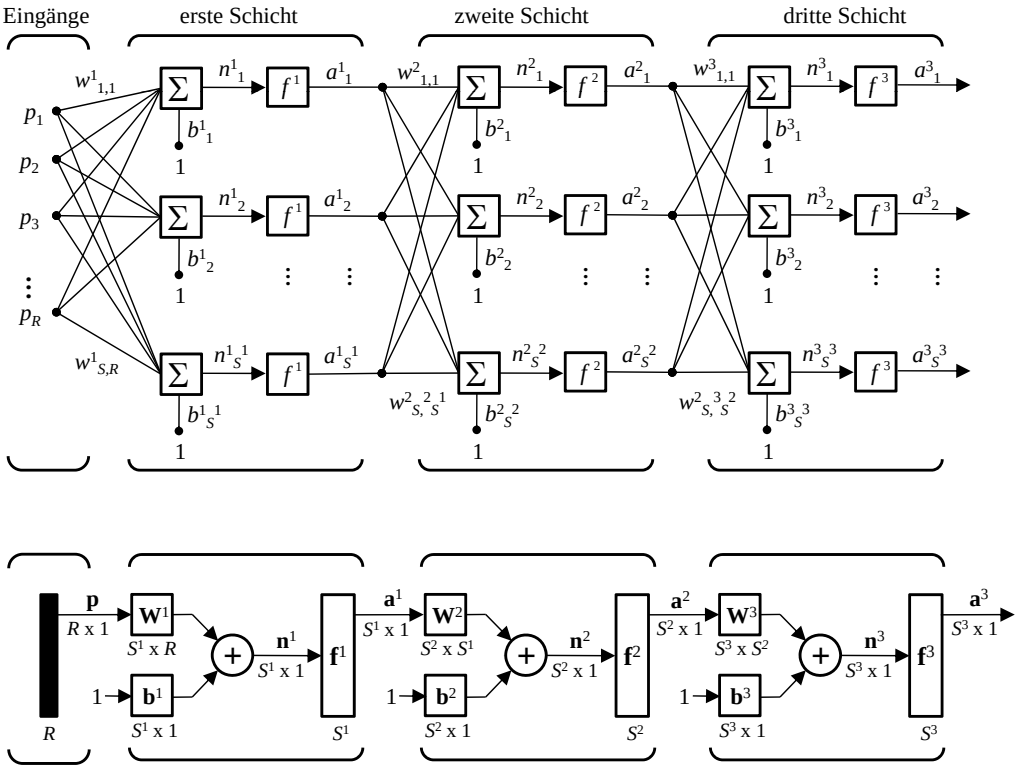


Abbildung 3.9: KNN mit drei Schichten: Ausführliche und verkürzte Darstellung

zweiten und dritten Neuronenschicht dar. Der Ausgangsvektor des gesamten Systems ist hier $\mathbf{a}^3$. Der Ausgabevektor des vorgestellten Beispielnetzes mit zwei verdeckten Schichten ergibt sich daher zu

$$\mathbf{a^3} = \mathbf{f^3}(\mathbf{W^3} \cdot (\mathbf{f^2}(\mathbf{W^2} \cdot (\mathbf{f^1}(\mathbf{W^1} \cdot \mathbf{p} + \mathbf{b^1}) + \mathbf{b^2}) + \mathbf{b^3}). \tag{3.33}$$

Der Begriff „*Lernen*“im Zusammenhang mit KNN wird mit der Bestimmung der Gewichtsmatrizen $\mathbf{W}^i$ und der Schwellwert-Vektoren $\mathbf{b}^i$ aller Schichten gleichgesetzt. Dabei gibt es grundsätzlich beliebig viele Möglichkeiten, wie die bestimmten Gewichte und Schwellwerte in ein und dem selben Netz belegt werden können. In der Regel hängt das von den gewählten Startwerten ab, so dass sich bei jedem neuen Training des Netzes ein unterschiedlicher Endzustand hinsichtlich der Netzparameter einstellt. Daneben ist natürlich auch die Änderung der Netzdimension oder die Verwendung anderer Transferfunktionen möglich, um die Netzcharakteristik zu ändern. Das biologische Vorbild lernt in sehr ähnlicher Art und Weise durch die Hemmung oder Verstärkung von Synapsen und durch die Knüpfung von neuen oder Löschung alter Verbindungen (dies entspricht der Gewichtung im KNN). Im Folgenden wird das überwachte Lernen näher erläutert, da es sich dabei um die am häufigsten verwendete Trainingsform für KNN handelt und diese auch für das Training der im Kapitel 4 verwendeten MLFF-Netze notwendig ist.

Kommt das überwachte Lernen zur Anwendung, so lernt das KNN die Lösung eines Problems mit Hilfe von Beispielen aus der Problemklasse und ist bei korrekter Lernphase anschließend in der Lage, im Rahmen der Intervalle, in denen die Lernbeispiele lagen, zu abstrahieren und ähnliche Probleme ebenfalls zu lösen. Dies entspricht genau dem biologischen Lernen aus Erfahrung: Ein Kind lernt, dass Feuer heiß ist und nicht angefasst werden darf. Dazu genügt eine schmerzhafte Erfahrung mit einem Feuer, daraufhin kann jegliche Art von Feuer (sofern es ähnlich aussieht, d. h. im Sinne der KNN in dem Intervall der Lernbeispiele liegt) als Gefahr erkannt werden.

Bei mehrschichtigen künstlichen Netzen ist die Bestimmung der Gewichte und Schwellwerte nicht trivial. Insbesondere besteht dabei die Gefahr des „*Überlernens (en.: Overlearning)*“. D. h. es werden nur die Beispieldatensätze erlernt und nicht das zugrunde liegende Schema. Wird das Netz im Anschluss an das Training mit unbekannten Eingängen konfrontiert, ist keine Abstraktion möglich und das Netz liefert grob falsche Ausgangswerte. Dieser Gefahr muss begegnet werden, indem im Anschluss an ein Training immer ein unabhängiger Testdatensatz aus dem vorgegebenen Intervall evaluiert wird.

Die Wahl des Trainingsalgorithmus hängt von der Art des zu trainierenden Netzes, der Rechenkapazität und den verfügbaren Lernbeispielen ab. Die meisten der zum Trainieren von Feed-Forward-Netzen eingesetzten Lern- oder Trainingsalgorithmen basieren auf dem Backpropagation-Algorithmus, der im Zusammenhang mit KNN erstmals 1986 von Rummelhard und McClelland [1986] vorgestellt wurde. Damit ist es möglich, mehrschichtige vorwärts gerichtete Netze erfolgreich zu trainieren. Im Kern handelt es sich bei (Error-)Backpropagation (*de.:* Fehlerrückführung) um ein Gradientenabstiegsverfahren zur Suche des globalen Minimums im mehrdimensionalen Fehlerraum. Die Fehler werden nach jedem Durchlauf aus dem Vergleich der vorgegebenen Soll-Ausgangswerte mit den vom Netz berechneten Ist-Werten bestimmt. Somit wird die Dimension des Fehlerraumes durch die Anzahl der Ausgangswerte, bzw. der Dimension von $\mathbf{a}$ vorgegeben. Mit dem Backpropagation-Algorithmus wird die Fehlerquadratsumme minimiert, in dem die freien Parameter, d. h. die Gewichte des Netzes um einen Bruchteil des negativen Gradienten der Fehlerfunktion schrittweise verbessert werden. Der Bruchteil wird dabei bestimmt durch die Lernrate η[8], welche die Schrittweite bzw. die Geschwindigkeit zur Suche des globalen Minimums der fehlerzeigenden Figur vorgibt. Eine zu kleine Lernrate kann zum Hängenbleiben in einem lokalen Minimum der fehlerbeschreibenden Figur führen, eine zu große Schrittweite führt u. U. zum Überspringen des globalen Minimums. Die Lernrate spielt auch in der praktischen Umsetzung von KNN mit Hilfe der Matlab-Software eine wichtige Rolle und kann manuell geändert werden. Es wird so lange iteriert, bis eine weitere Anpassung der Gewichte zu keiner wesentlichen Verkleinerung der Fehlerquadratsumme mehr führt. Dieser Grenzwert kann in der Regel auch als Parameter in der Software angegeben werden.

Im Weiteren von besonderer Bedeutung ist der nach seinen Erfindern benannte Levenberg-Marquardt Algorithmus (LM-Algorithmus), der eine rechenoptimierte und damit schnellere Weiterentwicklung des ursprünglichen Backpropagation-Algorithmus darstellt. Es handelt sich bei dem Verfahren um eine Variation des Newton-Verfahrens zur Annäherung von Nullstellen in nichtlinearen Funktionen. Ein Iterationsschritt zur Verbesserung der Gewichte in einem Netz mit nur einer verdeckten Neuronenschicht kann im Algorithmus von Levenberg und Marquardt wie folgt dargestellt werden:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - [\mathbf{J}^T\mathbf{J} + \mu\mathbf{I}]^{-1}\mathbf{J}^T\mathbf{e} \tag{3.34}$$

Wobei $\mathbf{x}_k$ und $\mathbf{x}_{k+1}$ die Gewichtsvektoren der k-ten und (k+1)-ten Iteration darstellen. Mit $\mathbf{J}$ wird die Jacobimatrix mit den partiellen Ableitungen der Fehlerfunktion nach den fehlerbehafteten Gewichten, mit $\mathbf{I}$ die Einheitsmatrix und mit $\mathbf{e}$ der Fehlervektor bezeichnet. Der skalare Parameter μ spielt eine wichtige Rolle in dem Trainingsalgorithmus von Levenberg und Marquardt. Für $\mu = 0$ entartet der LM-Algorithmus zum klassischen Newton-Verfahren zur Nullstellensuche bzw. Minima-Suche in Funktionen. Für große Werte von μ stellt der LM-Algorithmus ein Gradientenabstiegsverfahren mit kleinen Schrittweiten dar. Damit kann die Eigenschaft des Algorithmus mit Hilfe des Parameters μ der Aufgabe besser angepasst werden.

In Hagan u. a. [1996] ist eine ausführliche Beschreibung des Backpropagation-Algorithmus sowie der Erweiterung, die von Levenberg und Marquardt entwickelt wurde, zu finden.

[8] In Hagan u. a. [1996] wird die Lernrate mit dem griechischen Buchstaben α bezeichnet. In Zell [1997] hingegen bezeichnet α den Glättungskoeffizienten im Momentum-Term.

3.2.5 Weitere Verfahren

Neben den vorgestellten Analyseverfahren existiert eine Vielzahl weiterer Verfahren zur Systemanalyse, die jedoch auf den ersten Blick kein Erweiterungspotenzial zur Behandlung diverser Qualitätsmerkmale und -parameter bieten und daher im Rahmen dieser Arbeit nicht weiter untersucht wurden. Dies gilt unter anderem für die bekannten Verfahren wie

- Ursache-Wirkungs-Diagramm
- weitere Verfahren aus dem Bereich der Zuverlässigkeitsanalyse wie z. B. die Gefährdungsanalyse,
- evolutionäre Algorithmen,
- Fuzzy-basierte Verfahren zur Beschreibung unscharfer Zustände oder
- naturanaloge Verfahren aus dem Bereich des Soft Computing wie z. B. Boolesche Netze und Zellularautomaten,

die in der Literatur zu finden sind. Diese Verfahren sind (zumindest in ihrer Grundform) allenfalls zur Modellierung einzelner Qualitätsaspekte wie der Verfügbarkeit oder Aktualität geeignet. Eine Abdeckung aller Qualitätsmerkmale des in Kapitel 2.2 vorgestellten Modells ist hingegen nur schwer möglich. Eine explizite Beschreibung auf Ebene der Qualitätsparameter erscheint mit vertretbarem Aufwand derzeit nicht praktikabel.

3.3 Gegenüberstellung der Verfahren

Nach der Vorstellung ausgewählter Methoden und Verfahren in den vorigen Abschnitten erfolgt nun die Auswahl des, nach Abwägung aller Vor- und Nachteile, am besten zur Darstellung von Qualitätsmerkmalen und -parametern geeigneten Verfahrens. In der folgenden Tabelle 3.6 werden die vorgestellten Methoden und Verfahren nochmals in Stichworten kurz mit ihren wesentlichen Eigenschaften dargestellt. In der letzen Spalte wurden die einzelnen Vor- und Nachteile hinsichtlich einer Erweiterung von der ausschließlichen Beurteilung der Zuverlässigkeit und damit in erster Linie der Verfügbarkeit, hin zu einer differenzierteren Qualitätsbeschreibung und -beurteilung aller bereits in Tabelle 2.1 vorgestellten Merkmale, aufgeführt. Insbesondere soll das gewählte Verfahren neben der Beschreibung von prozentualen Werten, wie es bereits das vorgestellte Rechenverfahren auf Basis Boolescher Algebra ermöglicht (vgl. Abschnitt 3.1.2), auch beliebige Qualitätsparameterwerte wie zum Beispiel Standardabweichungen, metrische Größen, oder Ähnliches behandeln können. Zum besseren Vergleich sind in der ersten Zeile nochmals die wesentlichen Vor- und Nachteile des von Wiltschko entwickelten Verfahrens dargestellt.

Nach eingehender Betrachtung der wesentlichen Vor- und Nachteile der verschiedenen Verfahren zeigt sich, dass die Methoden aus dem Bereich der Zuverlässigkeitsanalyse in ihrer Grundform nicht die Möglichkeiten bieten, beliebige Qualitätsparameter von einzelnen Daten zu behandeln. Eine Erweiterung erscheint nur bei den Petri-Netzen möglich, da diese über ein entsprechendes Potenzial und die notwendige Flexibilität verfügen. Einige vielversprechende Ansätze der PN wurden im Abschnitt 3.2.2 aufgezeigt. Alle weiteren betrachteten gängigen Verfahren der Zuverlässigkeitsanalyse sind in erster Linie nur für ihren ursprünglichen Zweck geeignet. Eine Behandlung von Qualitätsparametern für Datengruppen in Prozent ist teilweise möglich, die Erweiterung auf beliebige Qualitätsparameter hingegen erscheint, insbesondere auch für einzelne Daten, nur schwer realisierbar. Detaillierte Untersuchungen bleiben weiteren Arbeiten vorbehalten.

Für eine Systembetrachtung im Hinblick auf ein ganzheitliches Qualitätsmanagement hingegen, erscheint der Einsatz eines mehr qualitativen Verfahrens sinnvoll. Beispielsweise bietet die FME(C)A die

Tabelle 3.6: Gegenüberstellung ausgewählter Verfahren zur Darstellung und Fortpflanzung von Datenqualität in Datenverarbeitungsprozessen

Verfahren	Referenznorm	Eignung für beliebige Q-Merkmale und Parameter
Rechenverfahren von Wiltschko	abgeleitet aus: DIN 25419 (1985) und DIN 25424 (1981/1990)	+ Exakte Darstellung einfacher Prozesse möglich − Funktionaler Zusammenhang muss bekannt sein − Qualität meist nicht auf Parameterebene modellierbar; Ausnahme: GE, sonst nur Merkmalerfüllungsgrade möglich − Anwendung in Echtzeit nicht möglich
FME(C)A	DIN EN 60812 (2006)	+ Gut zur qualitativen Analyse von Prozessen geeignet + FMECA gut zum Nachweis der Wirksamkeit von QS-Maßnahmen geeignet − Qualität nicht quantitativ modellierbar − Keine Darstellung von Fehlerkombinationen möglich − Anwendung in Echtzeit nicht möglich
Petri-Netze PN	VDI 4008 Blatt 4 (2008)	− Funktionaler Zusammenhang muss bekannt sein + Begrenzte Kapazitäten und Konflikte modellierbar + Erweiterungen klassischer PN hat Potenzial − Dynamische Abläufe darstellbar − Modellierung der Qualität wird schnell komplex + Anwendung in Echtzeit möglich
Monte-Carlo-Simulation	VDI 4008 Blatt4 (2008)	+ Funktionaler Zusammenhang muss nicht bekannt sein + Zuverlässigkeit und Genauigkeit aus Verteilung ableitbar − Verteilung, Erwartungswert und Varianz der Eingangsvariablen muss bekannt sein − Erweiterung auf weitere Q-Merkmale ist nicht möglich − Schwache Konvergenz des Verfahrens + Fehlertoleranz ggü. verrauschten Eingangsgrößen − Anwendung in Echtzeit nur bedingt möglich
künstliche neuronale Netze KNN	keine; diverse Literatur	+ Funktionaler Zusammenhang muss nicht bekannt sein + Grundsätzlich alle Q-Merkmale und Parameter darstellbar + Sehr breites Einsatzspektrum und extrem flexible Methode − Viele Lernbeispiele erforderlich − Netzmodellierung ist nicht trivial + Fehlertoleranz ggü. verrauschten Eingangsgrößen + Anwendung in Echtzeit möglich

Möglichkeit der Analyse von Fehlermöglichkeiten von Systemteilen, die explizit Folgen für einzelne Qualitätsmerkmale und -parameter haben. Dazu muss lediglich der Fokus der Betrachtungen auf der Beeinflussung der Qualität gelegt werden bzw. die Betrachtung darf nicht bei der Dokumentation der einzelnen identifizierten Fehlermöglichkeiten liegen, sondern vielmehr auf deren Auswirkung auf einzelne Qualitätsparameter. Mögliche Gegenmaßnahmen unterstützen dann die ständige Verbesserung der Produktqualität. Daher wird die FME(C)A als möglicher Bestandteil eines umfassenden Qualitätsmanagementkonzeptes für Daten, welches im Kapitel 5 vorgestellt wird, wieder aufgegriffen. Für eine flexible Modellierung von Datenqualität ist diese Methode jedoch nicht geeignet.

Die Künstlichen Neuronalen Netze erscheinen insbesondere wegen ihrer großen Flexibilität und breit gefächerten Einsatzmöglichkeit als Mittel der Wahl. Die häufig auftretende mangelnde Kenntnis der exakten Zusammenhänge zwischen der Qualität des Systemeingangs und dem Systemausgang macht KNN sehr attraktiv, da die Abbildung unbekannter Zusammenhänge eine der Kernkompetenzen dieses Verfahrens darstellt. Des Weiteren werden grundsätzlich keine Formatanforderungen an die Art der Eingangs- und Ausgangsdaten gestellt. Dies gewährleistet die freie Wählbarkeit der zu simulierenden Qualitätsparameter. Die Anwendung von KNN erscheint vielversprechend, da hier im Gegensatz zu den Petri-Netzen keine umfassende Erweiterung des Verfahrens erforderlich ist (vgl. Kapitel 3.4 und 4). Aus messtechnischer Sicht erscheint die Fehlertoleranz enorm wichtig, da es sich bei den behandelten Daten oft um stochastische Größen handelt, die ein mehr oder weniger starkes (Mess-)Rauschen aufweisen.

In welchem Maße die KNN tatsächlich zur Beschreibung von Qualität in Prozessen geeignet sind, soll durch die im Kapitel 4 detailliert untersuchten Anwendungsbeispiele aus verschiedenen Anwendungsgebieten geklärt werden. Im folgenden Abschnitt wird zunächst kurz dargestellt, welche Art von KNN zur Abbildung von Datenqualität in Frage kommt und wie die Adaption dieser universellen Methode auf die hier gestellte Aufgabe, der Modellierung von Datenqualität in Prozessen, erfolgen kann.

3.4 Adaption der künstlichen neuronalen Netze zur Beschreibung von Qualität in Prozessen

Künstliche neuronale Netze sollen einen Zusammenhang zwischen den Eingangsdaten und Ausgangsdaten eines beliebigen, Daten verarbeitenden Systems oder Prozesses herstellen. Genauer gesagt, sollen insbesondere die Beschreibungen der Datenqualität aufeinander abgebildet werden, d. h. es sollen alle, durch die Art des Prozesses vorhandenen Abhängigkeiten zwischen verschiedenen Qualitätsparametern identifiziert und modelliert werden. Dabei sind im Allgemeinen sowohl die Eingangs- und Ausgangsdatenart oder -arten als auch die Realisierung ihrer Qualitätsmodelle, d. h. die erforderlichen Qualitätsparameter unterschiedlich.

Aufgrund der definierten Problemklasse, der möglichst genauen Abbildung der Qualitätsparameter (QP) der Datenart A auf die Qualitätsparameter der Datenart B, wie in Abbildung 3.10 schematisch dargestellt, wird ein Netztyp gewählt, der durch überwachtes Lernen trainiert werden kann. In der zu untersuchenden Anwendung besteht im Allgemeinen die Möglichkeit, für das Training eines Netzes die erforderliche Anzahl Beispiele zur Verfügung zu stellen. Im Rahmen dieser Arbeit wird zunächst das Vorhandensein des Daten verarbeitenden Systems vorausgesetzt, dessen Auswirkung auf die Qualität der Daten mit Hilfe der KNN analysiert werden soll. Der Einsatz in der Planungsphase eines Prozesses wird hier nicht betrachtet, da dieser Fall für den Nachweis der grundsätzlichen Eignung der KNN für die Qualitätsbeschreibung nicht relevant ist. Bei der Prozessplanung kann ein Training mit simulierten Beispielen anstelle empirisch bestimmter erfolgen, ansonsten ist die Aufgabenstellung identisch.

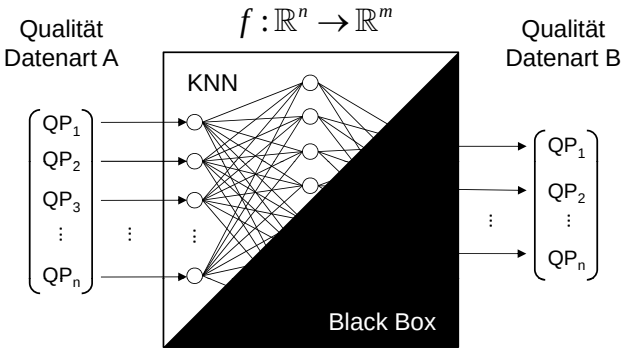


Abbildung 3.10: Modellierung der Qualitätsübergänge mittels KNN in einem im Detail unbekannten DV-System

Zur Lösung der gestellten Aufgabe werden Multilayer Feed-Forward-Netze (MLFFN) eingesetzt, da diese klassische Art von Neuronalen Netzen alle erforderlichen Eigenschaften zur Verfügung stellt und zur Abbildung jedes funktionalen Zusammenhangs geeignet ist, der eine endliche Anzahl von Unstetigkeiten aufweist. MLFFN wurden in der Geodäsie bereits in vielen Anwendungen erfolgreich eingesetzt. In Karabork u. a. [2008] dienen MLFFN der Höheninterpolation in digitalen Geländemodellen. Leandro u. a. [2007] prädizieren GPS-Beobachtungen auf der L2-Frequenz aus den Beobachtungen eines Einfrequenzempfängers und Heine [1999] setzt MLFFN zur Beschreibung von Deformationsprozessen erfolgreich ein. Eine umfangreiche Sammlung von Veröffentlichungen aus dem Bereich der Geodäsie findet sich in Frank [2010]. Die hier untersuchte Anwendung erfordert keine Rückkopplungen innerhalb des Netzes, wie es beispielsweise bei Optimierungs- oder Regelungsaufgaben der Fall ist.

Für viele Anwendungen genügen MLFFN mit einer verdeckten Schicht bereits den Anforderungen. Daher wird auch hier zur Simulation von Datenqualität im ersten Schritt nur mit einer verdeckten Schicht gearbeitet. Gegebenenfalls kann bei steigender Komplexität des zu beschreibenden Problems neben der Anzahl der verdeckten Neuronen auch die Anzahl der verdeckten Schichten auf zwei oder drei erhöht werden. Dies erhöht jedoch bei schichtweise vollständig verbundenen Netzen, wie sie hier zum Einsatz kommen, den Rechenaufwand zur Identifizierung der Gewichte ebenso wie die Anzahl erforderlicher Lernbeispiele enorm. Daher ist eine noch weitere Erhöhung der Schichtanzahl meist nicht mehr sinnvoll.

Die Dimensionierung der verdeckten Schicht(en) ist nicht trivial. Poddig und Sidorovitch [2001] bezeichnen die Spezifikation der Netzwerkarchitektur als „*[] den am schwierigsten zu handhabenden Freiheitsgrad bei der Modellentwicklung []*“. Es finden sich einige wenige Hinweise und Faustformeln in der Literatur, die einen Zusammenhang zwischen der Anzahl der Eingangsgrößen und der Anzahl verdeckter Neuronen in den verdeckten Schichten beschreiben. In wie weit die Anzahl der Ausgangsneuronen eine Rolle bei der Dimensionierung der verdeckten Schicht spielt, ist im Detail noch unklar. Bei den im folgenden Kapitel 4 durchgeführten Untersuchungen wurden die beiden folgenden Faustformeln berücksichtigt:

- Otto [1995] gibt für die Anzahl h der verdeckten Neuronen eine Obergrenze an, die von der Anzahl der Trainingsbeispiele Q sowie der Anzahl Eingangs- und Ausgangswerte (n und m) des Netzes abhängt. Die Anzahl verdeckter Neuronen in der verdeckten Schicht wird dabei auf

$$h = \frac{Q}{5 \cdot (n + m)} \tag{3.35}$$

  begrenzt.

- In Kinnebrock [1992] wird hingegen eine Faustformel für Netze mit einer verdeckten Schicht angegeben, in der nur die Anzahl der Eingangsgrößen berücksichtigt wird und h mit

$$h = 2 \cdot n + 1 \tag{3.36}$$

  ermittelt werden kann.

Grundsätzlich handelt es sich bei diesen Formeln um Faustformeln, die nur als Anhaltspunkt dienen können. Die Feinabstimmung der Parameter, insbesondere die Dimensionierung des Netzes und die zu wählende Anzahl der Trainingsbeispiele, muss im Anschluss experimentell erfolgen. Oft ist die Anzahl der Trainingsbeispiele jedoch durch die Rahmenbedingungen des zu beschreibenden Problems bereits gegeben, beispielsweise wenn empirische Daten die Grundlage bieten und keine weiteren Messungen möglich sind. In diesem Fall bleibt als einziger variabler Parameter -sofern die Anzahl der verdeckten Schichten nicht variiert wird- die Dimension der verdeckten Schicht. Es gibt erste Ansätze, die Dimensionierung von KNN exakt rechnerisch zu ermitteln. Allerdings sind diese bislang nur für sehr klein dimensionierte Netze praktikabel und versagen noch bei größeren, mehrschichtigen Netzen. Neuner und Kutterer [2010] haben dies am Beispiel der Modellselektion in der ingenieurgeodätischen Deformationsanalyse gezeigt. Daher werden in der hier untersuchten praktischen Anwendung die in der Literatur zu findenden Faustformeln für eine erste Dimensionierung der Netze herangezogen. Die weiteren Anpassungen erfolgen herkömmlich nach dem Prinzip Versuch und Irrtum, wie im folgenden Abschnitt etwas näher erläutert wird.

Die Dimensionierung des einzusetzenden KNN muss mit Sorgfalt erfolgen. Wird das Netz zu klein dimensioniert, so ist das Netz u. U. nicht in der Lage einen komplexeren Zusammenhang zu erlernen. Dieses als *Underfitting* bezeichnete Phänomen ist vergleichbar mit dem Versuch, eine unbekannte Funktion höherer Ordnung mit einem Polynom zweiten Grades zu approximieren. Umgekehrt kann ein zu groß dimensioniertes Netz zu einem *Overfitting* führen, d. h. das Netz lernt nur die Trainingsbeispiele, jedoch nicht die Abbildung selbst und ist damit nicht fähig zu generalisieren. Grundsätzlich sollte daher mit kleineren Netzdimensionen gestartet werden, die dann langsam vergrößert werden. Ein Underfitting des Netzes kann leicht daran erkannt werden, dass die Fehler zwischen Soll und Ist der Netzausgabe nicht gegen Null gehen und damit die Netzgewichte nicht iterieren. Ein Overfitting ist mit Hilfe eines Testdatensatzes nachweisbar, der nicht während der Trainingsphase verwendet wurde. Ein beginnendes Overfitting des Netzes zeigt sich, wenn der Fehler in den Trainingsdaten mit steigender Anzahl Iterationen weiter sinkt, während der Fehler in den Validierungsdaten bereits wieder steigt. An diesem Punkt ist das Netz bestmöglich trainiert. In Otto [1995] wird in diesem Zusammenhang von *Übergeneralisierung* gesprochen, wobei ganz allgemein das Nachbilden der Störungen bzw. des Rauschens des zu beschreibenden, nicht linearen Systems gemeint sind. Bildlich kann man sich Overfitting als Oszillation der bestangepassten (mehrdimensionalen) Kurve zwischen den Stützpunkten, mit denen das Netz trainiert wurde, vorstellen (Hagan u. a. [1996]).

Für das Trainieren von Netzen mit einer oder mehreren verdeckten Schichten wird ein optimierter Lernalgorithmus auf Basis des Backpropagation-Verfahrens eingesetzt. Aufgrund der schnellen Iteration des Lernprozesses, wird in den Beispielen im Kapitel 4 ausschließlich der LM-Algorithmus gewählt, der bereits am Ende des Kapitel 3.2.4 kurz erläutert wurde.

Die Auswahl der Transferfunktionen muss problemangepasst erfolgen und hängt entscheidend von der Art der gewünschten Ausgangsqualitätsparameter und insbesondere von dessen Wertebereich ab. Die Auswahl wird jedoch etwas eingeschränkt, da der Backpropagation-Algorithmus zwingend die Verwendung von stetig differenzierbaren Transferfunktionen erfordert. Die purelin-Funktion schränkt den Wertebereich des Ausgangs nicht ein und ist daher universell einsetzbar. Es kann jedoch von Vorteil sein, bei rein binären Ausgangsgrößen eine nach oben und unten beschränke Transferfunktion, wie z. B. die

logsig- oder tansig-Funktion zu wählen (vgl. Tabelle 3.5). In der verdeckten Schicht (oder den verdeckten Schichten) ermöglicht die Wahl einer sich asymptotisch verhaltenden Funktion die Behandlung von in ihrem Betrag sehr unterschiedlichen Größen. Damit wird das Netz auch sensibel gegenüber kleinen Größen bei gleichzeitiger Verarbeitung von großen Zahlen, wie es beispielsweise bei der Behandlung von Messgrößen und deren Standardabweichungen häufig auftritt. Die mit einem Tachymeter gemessene Strecke kann im Bereich von Kilometern liegen, während die Standardabweichung der selben Streckenmessung im Bereich von Millimetern liegt. Der Größenunterschied der Eingangswerte kann damit in diesem Beispiel im Bereich von 10^{-6} liegen und u. U. zu numerischen Problemen führen.

Grundsätzlich können in jeder Schicht und für jedes Neuron unterschiedliche Transferfunktionen zum Einsatz kommen. Üblicherweise wird jedoch je Schicht nur eine Transferfunktion eingesetzt und es kommen die vorgestellten Standardfunktionen zum Einsatz. Da hier keine Gründe für die Definition neuer Transferfunktionen bestanden, wurden die üblicherweise eingesetzten Funktionen

- Linearfunktion,
- Logarithmische Sigmoidfunktion sowie der
- Tangens Hyperbolicus

insbesondere wegen ihrer Differenzierbarkeit gewählt.

Der Lernalgorithmus nach Levenberg und Marquardt kann auch mit der Verwendung des Momentum-Terms kombiniert werden. Die Verbesserung der Gewichte und Schwellwerte nach jeder Iteration kann mit und ohne dessen Berücksichtigung erfolgen. Der Momentum-Term besteht aus einem Faktor zwischen 0 und 1, der mit der letzten Matrix der Verbesserungen multipliziert und zu der im aktuellen Iterationsschritt berechneten Verbesserungsmatrix addiert wird. Der Momentum-Term sorgt somit für ein längeres „Gedächtnis“bei der schrittweisen Verbesserung der Gewichte. Durch das Aufsummieren der Verbesserungen werden einerseits flache Plateaus in der Fehlerfläche wieder (schneller) verlassen. Andererseits wird ein Oszillieren in engen Schluchten gedämpft und ein erneutes Herausspringen ggf. verhindert, da die Verbesserungen, bedingt durch das alternierende Vorzeichen, an den Rändern der Schlucht verkleinert werden (Zell [1997]). Daher wird in den folgenden Untersuchungen stets der Momentum-Term berücksichtigt.

Für die Performance- oder Leistungsfunktion, mit der aktuelle Restfehler der Anpassung in jeder Epoche aus den einzelnen Ist-Soll-Fehlern ermittelt wird, stehen grundsätzlich mehrere Möglichkeiten zur Verfügung. Neben dem mittleren quadratischen Fehler (*en.*: mean square error, MSE) kann die Beurteilung auch anhand der Summe der Fehlerquadrate (*en.*: square sum error, SSE) oder der Wurzel aus dem mittleren quadratischen Fehler (*en.*: root mean square, RMS) erfolgen. Die Unterschiede sind für die hier untersuchte Anwendung nicht von Bedeutung, daher wurde eher willkürlich der MSE-Wert zur Beurteilung herangezogen.

Die Modellierung und der Test der MLFFN erfolgt mit der Matlab-Software. Diese bietet eine mächtige Toolbox zur Modellierung von KNN. Neben einer Vielzahl an Funktionen, die damit über die Kommandozeile zur Verfügung stehen, bietet die Neural Network Toolbox eine grafische Benutzeroberfläche (*en.*: graphical user interface, GUI) mit deren Hilfe viele gängige Netzarten modelliert, validiert und angewendet werden können. Das GUI *nntool* bietet einen großen Funktionsumfang und erleichtert die Modellierung von KNN mit Matlab enorm. Daher wird das GUI in allen weiteren Untersuchungen im Rahmen dieser Arbeit eingesetzt.

Im Anhang 6 erfolgt eine kurze Vorstellung der Funktionsweise des verwendeten GUI der Matlab-Software. Insbesondere werden die zur Modellierung der Netze und zur Optimierung der Lernphase relevanten Einstellungen und Parameter erläutert. Aus Platzgründen kann hier lediglich ein kleiner Überblick über die Möglichkeiten und die Bedienung des GUI vermittelt werden. Dem interessierten Leser sei das umfangreiche Handbuch zur Toolbox empfohlen, welches online abrufbar ist (Matlab® [2010]).

4 Anwendung von KNN zur Fortpflanzung der Datenqualität an Beispielen

Im Abschnitt 3.3 wurden die Künstlichen Neuronalen Netze als grundsätzlich am Besten geeignet zur Modellierung und Simulation von Datenqualität in Prozessen bewertet. Der praktische Nachweis für deren Eignung wird in diesem Kapitel anhand geeigneter Beispiele erbracht. Zunächst dient ein einfaches Beispiel aus der geodätischen Messpraxis zur Veranschaulichung der Vorgehensweise für den Einsatz der KNN. Dabei liegt der Fokus auf der Modellierung einzelner Qualitätsmerkmale. Abschließend wird dann ein komplexeres Beispiel aus dem Bereich der Verkehrstelematik mit realen Daten untersucht.

Zur Modellierung und Simulation von Datenqualität in DV-Prozessen wurden, entsprechend der Begründung im vorigen Kapitel 3.4, aus der mittlerweile relativ großen Zahl altbekannter und neuer Netztypen die klassischen Feed-Forward-Netze gewählt, bei denen keine Rekursionen oder Rückkopplungen und auch keine Abkürzungen (Überspringen einzelner Schichten) möglich sind. Im Einzelnen konnten damit die folgenden wesentlichen Parameter bei der Anwendung von KNN in den in dieser Arbeit dargestellten Beispielen vorab festgelegt werden:

- Netzart: Feed-Forward-Netze; keine Rekursionen oder Rückkopplungen und keine Abkürzungen
- Netzmächtigkeit: maximal 2-3 Schichten, d. h. 1-2 verdeckte Schichten
- Dimensionen der Schichten: Anzahl der Neuronen in den verdeckten Schichten orientiert sich zunächst an Faustformeln von Otto [1995] und Kinnebrock [1992]
- Nur statische Netze, Reihenfolge der Trainingsbeispiele ist beliebig; keine Zeitverzögerungen
- Wahl einer identischen Transferfunktion für alle Neuronen einer Schicht
- Lernalgorithmus: Levenberg-Marquardt Backpropagation mit Momentum-Term
- Linearfunktion, Tangens Hyperbolicus oder Sigmoidfunktion als stetig differenzierbare Transferfunktionen
- Einsatz der Matlab-Toolbox nntool zur Modellierung der KNN
- Verwendung der programmseitig vorgeschlagenen Startwerte für Gewichte und Schwellwerte

Der Einsatz der KNN in den folgenden Beispielen orientiert sich im Wesentlichen an den obigen Festlegungen. Die Suche nach der am Besten geeigneten Netzkonfiguration erfolgt in der Regel empirisch. Eine ausführlichere Darstellung der Anpassung von KNN auf die vorliegende Fragestellung findet sich im Abschnitt 3.4. Dort werden auch die definierbaren Abbruchkriterien für den Lernalgorithmus näher erläutert.

4.1 Beispiel aus der klassischen Geodäsie

Eine typische Aufgabe in der Geodäsie ist die Berechnung von Standardabweichungen des Neupunktes in Längs- und Querrichtung (l und q) bei der polaren Punktbestimmung aus den beiden Messgrößen

horizontale Richtung r und Strecke s (Abbildung 4.1). Es handelt sich dabei um die auf die Zielrichtung bezogenen Standardabweichungen des Neupunktes N, die aus der physikalisch begrenzten Messgenauigkeit des Tachymeters resultieren. Die beiden Parameter l und q definieren die Halbachsen der Fehlerellipse um den Neupunkt und damit den Bereich, in dem der Neupunkt mit einer statistischen Wahrscheinlichkeit von $68\,\%$ (dies entspricht dem 1σ-Bereich) zu liegen kommt. Dabei wird von einer Gauß´schen Normalverteilung der beiden Messgrößen Richtung und Strecke und damit auch der beiden Zielgrößen ausgegangen.

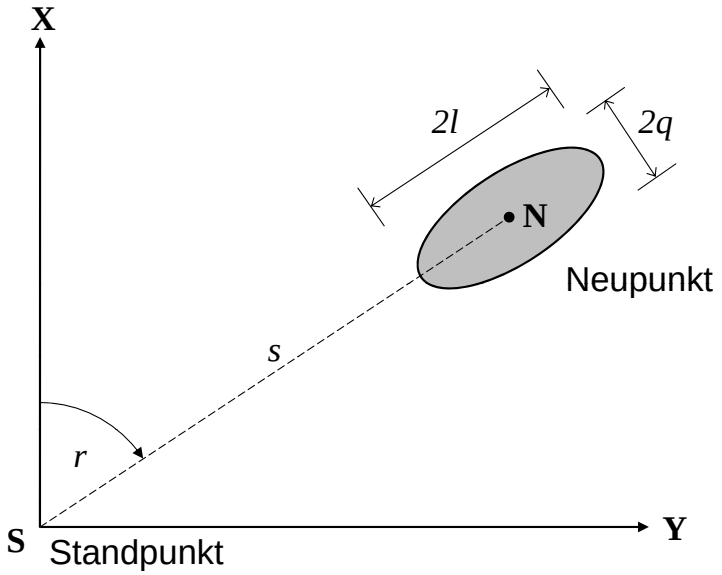


Abbildung 4.1: Neupunktbestimmung durch polares Anhängen an einen bekannten Standpunkt (lokales Koordinatensystem)

Zur Untersuchung der Möglichkeiten von KNN zur Modellierung und Simulation von Datenqualität wurde zunächst dieser einfache DV-Prozess gewählt. Der Prozess ist funktional bekannt, damit ist es möglich, nahezu beliebig viele Lernbeispiele zu erzeugen. Um das Potenzial der KNN im Detail zu untersuchen, liegt der Fokus hier auf der Betrachtung einzelner Qualitätsmerkmale.

Aus Vereinfachungsgründen werden im Folgenden die Begriffe *Längsabweichung* und *Querabweichung* synonym für die Standardabweichungen des Neupunktes in Längsrichtung und in Querrichtung, bezogen auf die Richtung vom Standpunkt zum Neupunkt, verwendet.

4.1.1 Modellierung der Genauigkeit

Zunächst wird mit Hilfe eines Feed-Forward-Netzes (FFN) das Qualitätsmerkmal Genauigkeit modelliert. Ausgangswerte sind daher nur die beiden Zielgrößen l und q, bei denen es sich um Genauigkeitsparameter handelt. Zu deren Bestimmung werden vier Eingangsgrößen benötigt. Zum einen sind dies neben der originären Streckenmessung s auch die Standardabweichung σ_s der Streckenmessung und der ppm-Wert als entfernungsabhängiger Anteil der Streckenmessgenauigkeit. Zum anderen wird auch die Genauigkeit der Richtungsmessung in Form der Standardabweichung σ_r benötigt. Mit Ausnahme der gemessenen Strecke können alle Eingangsgrößen dem Merkmal Genauigkeit zugeordnet werden. Die absolute Richtung, in der der Neupunkt im lokalen Koordinatensystem liegt, hat in diesem einfachen Beispiel keinen Einfluss auf Längs- und Querabweichung. In der Tabelle 4.1 sind alle Eingangs- und

Ausgangswerte nochmals übersichtlich und kompakt dargestellt. In der ersten und vierten Spalte wird das übergeordnete Qualitätsmerkmal der Eingangs- bzw. Ausgangsgröße gemäß den Abkürzungen in Tabelle 2.1 eingetragen. Handelt es sich um eine Größe, die nur zur Bestimmung der Ausgangswerte erforderlich ist (hier im Beispiel die gemessene Strecke), so wurde dies in der QM-Spalte mit „-“vermerkt.

Tabelle 4.1: Eingangs- und Ausgangswerte beim Polaren Anhängen; nur GE berücksichtigt

Eingang			Ausgang		
QM	Parameter		QM	Parameter	
GE	σ_s	Standardabweichung Strecke	GE	l	Längsabweichung Neupunkt
GE	σ_r	Standardabweichung Richtung	GE	q	Querabweichung Neupunkt
GE	ppm	Atmosphärische Korrektur			
-	s	gemessene Strecke			
$n = 4$ Eingangsparameter			$m = 2$ Ausgangsparameter		

Der funktionale Zusammenhang ist in diesem Beispiel bekannt. Längs- und Querabweichung können mit

$$l = \sqrt{\sigma_s^2 + (\mathrm{ppm} \cdot s[\mathrm{km}])^2} \quad und \quad q = \sigma_r[\mathrm{rad}] \cdot s \tag{4.1}$$

mit hoher Genauigkeit berechnet werden. Aus der Faustformel von Kinnebrock [1992] ergibt sich die Anzahl der verdeckten Neuronen zu $h = 2 \cdot n + 1 = 9$. Nach der Faustregel von Otto [1995] sind dazu mindestens 270 Trainingsbeispiele erforderlich, daher wurden $Q = 450$ Lernbeispiele berechnet und an das GUI übergeben. Nach der Standardeinstellung des KNN-Simulators wird diese Menge in $60\,\%$ und damit 270 Trainingsbeispiele und jeweils $20\,\%$ bzw. 90 Validierungs- und Testbeispiele unterteilt. Damit werden beide Empfehlungen eingehalten (vgl. Formeln 3.35 und 3.36). Als Transferfunktion des Netzes wurde den ersten praktischen Erfahrungen entsprechend in der verdeckten Schicht der Tangens Hyperbolicus (tansig) und in der Ausgabeschicht die Linearfunktion (purelin) gewählt. Die Netzarchitektur kann somit, wie in Abbildung 4.2 zu sehen ist, dargestellt werden.

Wie bereits erwähnt, wird die in den KNN-Simulator eingespeiste Menge an Lernbeispielen zufällig mit den Anteilen $60\,\%/20\,\%/20\,\%$ in Trainingsmenge, Validierungsmenge und Testmenge eingeteilt. Das Netz wird nur mit der Trainingsmenge trainiert und nach jedem kompletten Durchlauf wird die aktuelle Netzgüte mit Hilfe der Validierungsmenge geprüft. Dabei wird der mittlere quadratische Fehler (MSE) über alle Validierungsbeispiele zur Beurteilung herangezogen. Der MSE berechnet sich als Mittelwert aus den quadrierten Verbesserungsvektoren $\mathbf{v}(k)$ aller Q Lernbeispiele mit

$$\mathrm{MSE} = \frac{1}{Q}\sum_{k=1}^{Q}\mathbf{v}(k)^T\mathbf{v}(k) = \frac{1}{Q}\sum_{k=1}^{Q}(\mathbf{t}(k) - \mathbf{a}(k))^T(\mathbf{t}(k) - \mathbf{a}(k))[-]. \tag{4.2}$$

Wobei $\mathbf{t}(k)$ den Soll-Ausgabevektor des k-ten Beispiels (*en.*: target) des Netzes und $\mathbf{a}(k)$ den vom KNN tatsächlich berechneten Ausgang darstellt.

Der MSE hat im Allgemeinen keine Einheit, da die Ausgangswerte meist unterschiedliche Einheiten haben. Es werden lediglich die Zahlenwerte zur Berechnung herangezogen. Daher ist der MSE nur relativ zu betrachten und lässt im Wesentlichen qualitative Aussagen über den aktuellen Trainingszustand und die Eignung des Netzes zu. Steigt der MSE wieder an, oder sinkt unter einen definierten Schwellwert, so

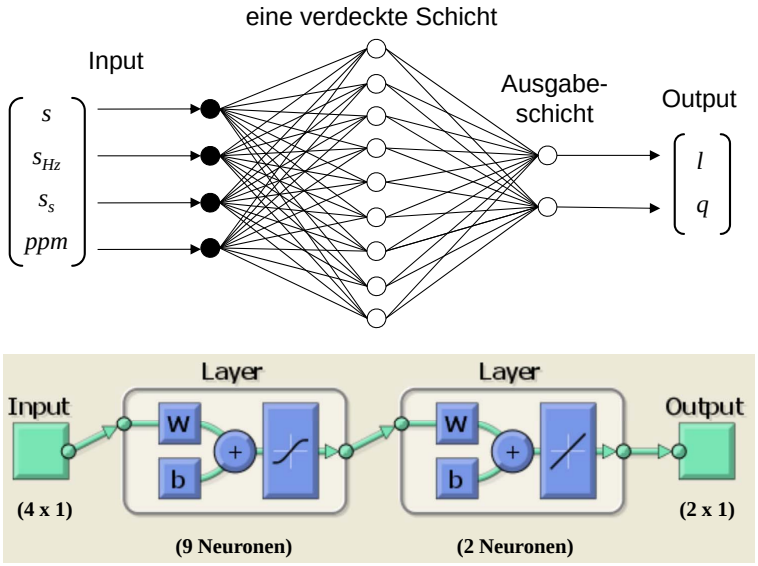


Abbildung 4.2: Netz zur Modellierung der Genauigkeitsparameter beim Polaren Anhängen

wird das Training beendet. An der erreichten Größenordnung kann abgeschätzt werden, ob das Netz für das trainierte Problem geeignet ist. Ist der MSE nicht zufriedenstellend klein (im Vergleich zu anderen Netzvarianten oder Erfahrungswerten), muss zunächst geprüft werden, ob bei der Suche des globalen Minimums, möglicherweise aufgrund falscher Startwerte für die Gewichte und Schwellwerte, versehentlich ein lokales Minimum gefunden wurde. Liefert die mehrmalige Neuinitialisierung und nachfolgendes Trainieren des Netzes keine besseren Ergebnisse, so ist das Netz falsch dimensioniert oder die Netzart ungeeignet für die gestellte Aufgabe.

Schließlich wird abschließend die Leistungsfähigkeit und insbesondere die Abstrahierfähigkeit des trainierten Netzes mit Hilfe der Testdatenmenge überprüft. Mit Hilfe der Testbeispiele ist eine quantitative Aussage über die Güte und Genauigkeit des trainierten Netzes möglich. Die Testbeispiele wurden dem Netz bislang nicht eingespeist, können ergo auch nicht „auswendig gelernt“worden sein. Wichtig ist dabei jedoch, dass die Testdaten nur aus den zuvor trainierten Datenintervallen und damit dem Eingaberaum der Funktion stammen. KNN sind generell nur sehr bedingt zur Extrapolation der Eingangswerte geeignet.

Für das hier ausgewählte Netz zur Modellierung der Genauigkeitsparameter beim polaren Anhängen, ist ein schnelles Iterieren zu beobachten (vgl. Abbildung 4.3). Die MSE-Werte der Trainingsdatenmenge sind in blau dargestellt (unterste Kurve bei Epoche 1000), für die Validierungsdatenmenge wurde grün (mittlere Kurve bei Epoche 1000) und für die MSE-Werte der Testdatenmenge rot (oberste Kurve bei Epoche 1000) gewählt. Die MSE-Werte sinken bereits nach etwa 50 Iterationen auf unter $1 \cdot 10^{-8}$. Damit scheint das Netz bereits zur Lösung der gestellten Aufgabe geeignet. Aufgrund der sehr niedrig gesetzten Abbruchkriterien für den Abstiegsgradienten für die Änderung der Gewichte ($< 1 \cdot 10^{-10}$) und 0 für den MSE, durchläuft das Training in diesem Beispiel mit 1000 die definierte Maximalzahl an Iterationen. Der MSE sinkt dabei auf ca. $1 \cdot 10^{-10}$ ab. Die Rechenzeit zum Trainieren des Netzes ist aufgrund der relativ kleinen Netzdimension mit 15 s moderat.

Zusätzlich wurden für eine von der verwendeten Software Matlab unabhängige Überprüfung der Netzgüte 15 weitere Testbeispiele frei aus dem gesamten Eingaberaum gewählt, für den das Netz trainiert wurde. Dabei wurden jeweils alle Eingangswerte im Rahmen der trainierten Intervalle variiert, um einen

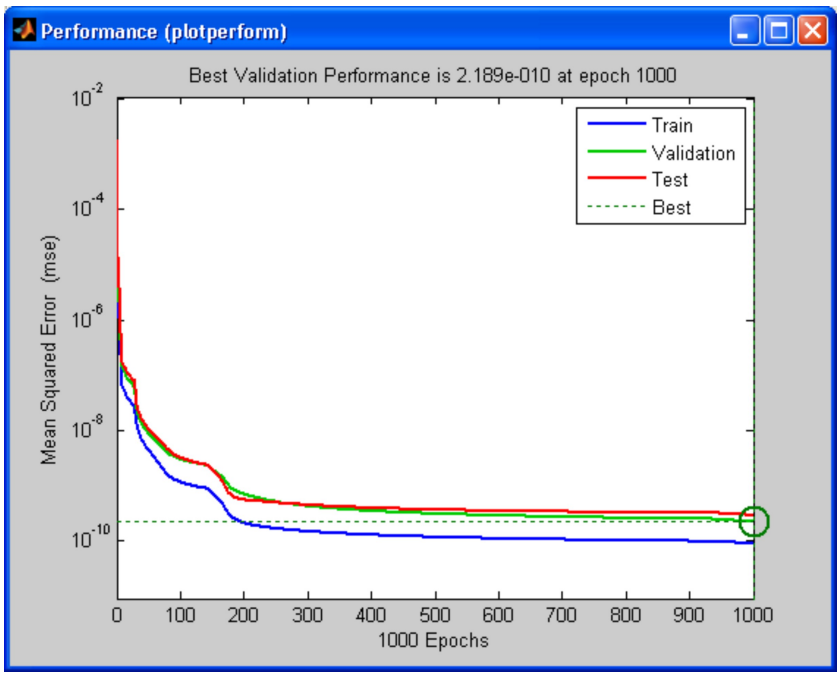


Abbildung 4.3: Darstellung des MSE während der Trainingsphase des KNN; Erläuterung der Legende: *Train:* Beispiele, die zum Trainieren des Netzes herangezogen werden; *Validation:* Beispiele, die nach jeder Iteration zur Beurteilung des Trainingsstands dienen; *Test:* Beispiele, die erst nach Abschluss des Trainings zum Test herangezogen werden

möglichst repräsentativen Eindruck der Leistungsfähigkeit des Netzes zu erhalten. Wie der Vergleich der Sollwerte mit den von dem trainierten Netz generierten Werten in der Abbildung 4.4 zeigt, sind die Restfehler der mit dem Netz ermittelten Längs- und Querabweichungen gegenüber den exakt berechneten Werten kleiner als $1 \cdot 10^{-4}\,\mathrm{m}$ bzw. $0.1\,\mathrm{mm}$. Damit ist die Genauigkeit wesentlich besser als die erreichbare Koordinatengenauigkeit mit einem genauen Tachymeter. Das getestete KNN ist gut zur Modellierung der Genauigkeitsparameter geeignet und die Faustformeln zur Dimensionierung des Netzes sind für diese Aufgabenstellung sinnvoll.

Die Ergebnisse der von der Trainingsdatenmenge unabhängigen Testdaten können daher die guten Ergebnisse bestätigen, die von Matlab mit dem für den Test herangezogene Anteil der Trainingsdatenmenge berechnet wurden. Feed-Forward-Netze mit lediglich einer verdeckten Schicht sind demnach geeignet, die Genauigkeitsparameter beim polaren Anhängen zu modellieren. Das trainierte Netz ist nun auch zur Simulation verschiedener Situationen geeignet. Es kann beispielsweise die Verwendung eines speziellen Instrumentes untersucht und die damit erreichbare Genauigkeit abgeschätzt werden.

Im nächsten Schritt wurde untersucht, ob weitere Qualitätsmerkmale mit einem derartigen neuronalen Netz ebenfalls behandelt werden können. Um die Eignung im Detail beurteilen zu können, wurde zunächst die Modellierung von anderen Merkmalen getestet, ohne das in der ersten Variante des Beispiels behandelte Genauigkeitsmerkmal explizit zu berücksichtigen.

4.1.2 Modellierung der Verfügbarkeit

Die Verfügbarkeit wird hier als Eigenschaft jeder einzelnen Datenart bzw. Eingangs- und Ausgangsgröße verstanden. Dies macht aufgrund der im Allgemeinen unterschiedlichen Datenquellen Sinn, da ein

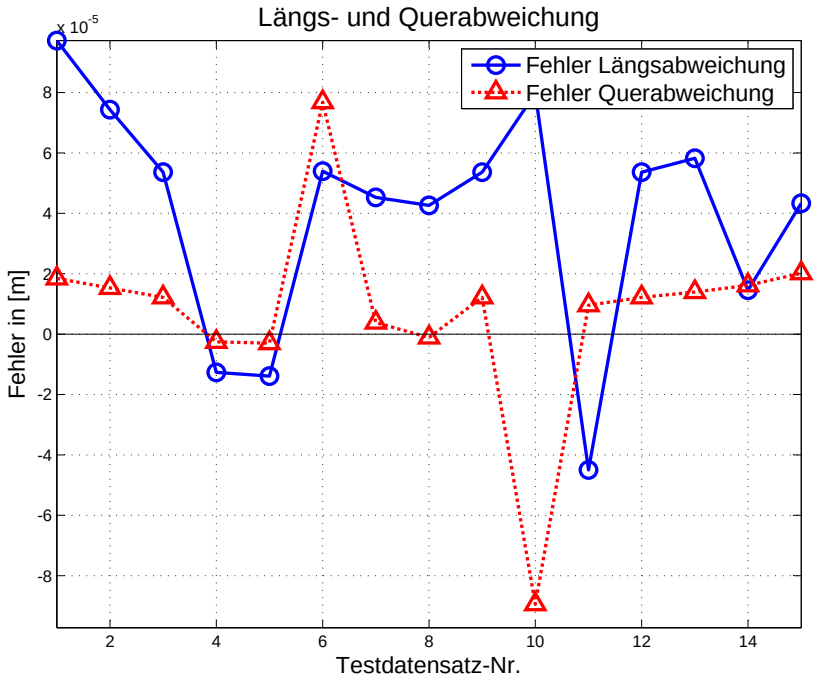


Abbildung 4.4: Restfehler der Längs- und Querabweichungen der Testdaten bei Modellierung der GE (KNN [9−2])

Ausfall einer Datenquelle nicht automatisch den Ausfall aller Eingangsdaten und damit eines ganzen Eingangsdatensatzes zur Folge haben muss. Gehören verschiedene Datenarten allerdings einem zusammengehörigen Datensatz an, da diese z. B. aus dem selben vorhergehenden DV-Prozess stammen, so kann nur die Verfügbarkeit des gesamten Datensatzes als Einheit beschrieben werden. Fehlt ein einzelner Wert in einem derartigen Datensatz, so wird dies durch die Vollständigkeit beschrieben (vgl. Abschnitt 4.1.3). Der Datensatz ist in diesem Fall verfügbar aber nicht vollständig. Im Folgenden wird jedoch zunächst davon ausgegangen, dass alle Eingangs- und Ausgangsdaten getrennt behandelt werden können. Es werden daher keine Datensätze als logisch zusammengehörende Einheiten betrachtet. Grundsätzlich können die beiden Ausgangswerte Längs- und Querfehler (l und q) auch als ein Datensatz betrachtet werden, da sie im selben DV-Prozess entstehen. Da hier jedoch die Untersuchung der Modellierungsmöglichkeiten von einzelnen Qualitätsmerkmalen im Vordergrund steht, wird auf die Betrachtung von Übergängen zwischen verschiedenen Merkmalen verzichtet. In dem vorliegenden Fall würde sich ein einzelner, nicht verfügbarer Ausgangswert direkt in der nicht erfüllten Vollständigkeit des entstandenen Ausgangsdatensatzes zeigen.

Zur Modellierung der Verfügbarkeit wird der einfache Ansatz aus dem vorigen Kapitel 4.1.1 verwendet, bei dem aus den vier Eingangswerten Strecke, Standardabweichung der Streckenmessung, entfernungsabhängiger atmosphärischer Streckenfehler sowie der Standardabweichung der Richtungsmessung, die beiden Genauigkeitsparameter Längs- und Querabweichung berechnet werden. Tabelle 4.3 gibt einen Überblick über die Ein- und Ausgangsdatenarten.

Die Modellierung der Verfügbarkeit der einzelnen Eingangsdaten kann in KNN grundsätzlich auf verschiedene Arten geschehen. Dabei stellt sich das Problem, nicht verfügbare Daten oder Datenbestandteile im Ein- und Ausgangsvektor darzustellen, damit ein KNN für die Abbildung der Verfügbarkeit trainiert werden kann. Der einfachste Ansatz ist die Darstellung nicht verfügbarer Daten mit einem nicht-numerischen Symbol wie beispielsweise *NaN*. Allerdings führt das unter Umständen zu Problemen bei der

praktischen Implementierung der KNN. Auch die Darstellung nicht verfügbarer Daten durch betragsmäßig sehr hohe Werte (z. B. 99999 oder -99999) ist grundsätzlich möglich, allerdings kann dies zum einen zu numerischen Problemen durch unter Umständen stark inhomogene Größenordnungen in den unterschiedlichen Datenarten führen. Wie im Weiteren noch dargestellt wird, ist zum anderen zusätzlich eine Normierung der Eingangs- und Ausgangsdaten erforderlich, so dass aus diesen Gründen diese Möglichkeit zur Darstellung von mangelnder Verfügbarkeit nicht weiter in Betracht gezogen wurde.

Die symbolische Darstellung nicht verfügbarer Daten durch eine „0“ist möglich. Dabei kann jedoch die Gefahr bestehen, dass sehr kleine und damit gute Werte z. B. für ermittelte Standardabweichungen mit der *exakten* 0 verwechselt werden. Durch die hohe Rechengenauigkeit PC-basierter Programme bzw. Programmiersprachen, ist die Verwechslung einer exakten 0 mit verfügbaren Eingangsgrößen nahe 0 unwahrscheinlich. Da beim Einsatz neuronaler Netze an eine Aufgabe jedoch Näherungsverfahren Verwendung finden, ist die Verwechslungsgefahr bei den Ausgangsdaten wesentlich größer. Das KNN wird daher nur in der Lage sein, Ergebnisse nahe 0 zu produzieren, jedoch keine exakte Null. Die Verwendung einer Treppenfunktion als Transferfunktion würde die Erzeugung einer „echten“Null ermöglichen, dies ist jedoch aufgrund der bereits angesprochenen Stetigkeitsforderung des Lernalgorithmus an die Transferfunktionen nicht möglich. Dennoch wird dieser Ansatz, nicht verfügbare Daten mit 0 darzustellen, mit Hinblick auf die hohe Rechengenauigkeit der eingesetzten Matlab-Software als vielversprechend bewertet und im Weiteren verfolgt.

Tabelle 4.2: Möglichkeiten zur Darstellung von nicht verfügbaren Werten in KNN

Symbolik	Beurteilung der Eignung in KNN
nicht-numerische Symbole	Schwer zu implementieren
99999/-99999	Numerische Probleme wahrscheinlich
0 (im Datenfeld)	Möglich und sinnvoll
Flag 0/1 (neues Feld)	Möglich aber aufwändig

Die zweite praktikable Möglichkeit besteht in der Erweiterung der Ein- und Ausgangsvektoren um Indikatoren, die zur Darstellung zusätzlicher Eigenschaften für jeden einzelnen Datenbestandteil geeignet sind. Da die Verfügbarkeit jedoch auch am Vorhandensein des jeweiligen Wertes erkennbar ist, wird diese Methode zu deren Modellierung nicht eingesetzt. Für die Darstellung von inkonsistenten oder nicht korrekten Daten hingegen erscheint diese Methode gut geeignet zu sein, auch wenn sie aufgrund der Verdopplung der Dimensionen für Eingangs- und Ausgangsvektoren weitaus komplexere KNN erfordert. Dieser Ansatz wird im Abschnitt 4.1.3 weiter untersucht.

In Tabelle 4.2 sind die Möglichkeiten zur Darstellung nicht verfügbarer Daten nochmals zusammengefasst. Es kommen hier nur die beiden letzten Ansätze in Frage, da die restlichen entweder zu numerischen Problemen führen, oder die Umsetzung in der Implementierung der KNN sehr aufwändig ist. Die Modellierung der Verfügbarkeit erfolgt aus praktischen Gründen jedoch ohne zusätzliche Indikatoren direkt durch die Definition des Datenwertes 0 als *„Datum nicht verfügbar“*. Die Eingangs- und Ausgangswerte, die mit Ausnahme der gemessenen Strecke Genauigkeitsparameter darstellen, „transportieren“somit zusätzlich die Verfügbarkeit. Trotz der Konzentration auf die Modellierung der Verfügbarkeit wird in diesem Beispiel daher automatisch auch die Genauigkeit modelliert. Ergibt sich für die Ausgangswerte $l \geq \varepsilon$ bzw. $q \geq \varepsilon$ mit $\varepsilon > 0$ als kleiner, jedoch plausibler numerischer Wert, so sind Längs- bzw. Querabweichung verfügbar. Der numerische Wert quantifiziert damit den Genauigkeitsparameter.

Damit bleibt die Dimension der Eingangs- und Ausgangsvektoren mit $n = 4$ bzw. $m = 2$ gleich wie bei der Abbildung der Genauigkeit in Abschnitt 4.1.1. Tabelle 4.3 gibt einen kompakten Überblick über die Netzein- und Ausgänge.

Tabelle 4.3: Eingangs- und Ausgangswerte beim Polaren Anhängen; nur VE explizit berücksichtigt

Eingang			Ausgang		
QM	Parameter		QM	Parameter	
GE/VE	σ_s	Standardabweichung Strecke	GE/VE	l	Längsabweichung Neupunkt
GE/VE	σ_r	Standardabweichung Richtung	GE/VE	q	Querabweichung Neupunkt
GE/VE	ppm	Atmosphärische Korrektur			
VE	s	gemessene Strecke			
$n = 4$ Eingangsparameter			$m = 2$ Ausgangsparameter		

Wie empirische Untersuchungen ergaben, ist es aufgrund der unterschiedlichen auftretenden Größenordnungen von Eingangs- und Ausgangsgrößen sinnvoll, diese zunächst zu normieren. In dem vorliegenden Beispiel unterscheiden sich die Eingangsgrößen, betrachtet man die Standardabweichung der Richtungsmessung in der Einheit Radiant und die gemessene Strecke in Metern, etwa um den Faktor 10^9. Damit sind numerische Probleme ohne Normierung sehr wahrscheinlich. Wie in der Tabelle 4.4 in der untersten Zeile zum Vergleich exemplarisch dargestellt, werden beim Trainieren des Netzes sehr kleine Werte für den MSE erreicht. Allerdings kann das Netz die Testdaten nur um Faktor 10 (bzw. Faktor 6 für die nicht verfügbaren Werte) schlechter abbilden als dasselbe Netz, welches mit normierten Werten trainiert wurde. Offensichtlich beugt die „Homogenisierung der Beobachtungen“, wie die Normierung im Zusammenhang mit der Ausgleichungsrechnung bezeichnet wird (vgl. Benning [2010]), dem Auswendiglernen der Trainingsbeispiele vor. Daher wird in allen weiteren Untersuchungen zunächst eine Normierung der Eingangswerte vorgenommen.

Zunächst wurde die gleiche Netzarchitektur $[9 - 2]$ wie für die Modellierung der Genauigkeit in Abschnitt 4.1.1 gewählt. In den 500 Lernbeispielen wurden ca. $30\,\%$ der Eingangswerte als *nicht verfügbar* durch Nullen ersetzt. Dabei wurde absichtlich ein derart großer und unrealistischer Anteil gewählt, um das Netz auch mit diesem -in der Realität- meist selten eintreffenden Fall von nicht verfügbaren Daten oder Datenbestandteilen zu konfrontieren. In den nach erfolgreicher Lernphase eingesetzten Testdaten wurde der Anteil mit $5\,\%$ wesentlich geringer und damit realistischer gewählt, zeitgleich wurde die Anzahl der Testbeispiele auf 50 erhöht.

Wie sich bereits bei den ersten Versuchen herausgestellt hat, muss hier von der Faustformel zur Berechnung der Neuronen im Falle einer verdeckten Schicht (vgl. Formel 3.36) Abstand genommen werden. Die Verwendung von lediglich 9 Neuronen in der verborgenen Schicht führt hier nicht zum Ziel. Bei mehrfacher Wiederholung des Versuchs bricht die Berechnung der Netzgewichte jedes Mal bereits nach wenigen Iterationen aufgrund des großen Validierungsfehlers ab. Die MSE-Werte erreichen nur Werte von einigen 10^{-3}, eine Bestimmung der Längs- und Querabweichungen auf Millimeter sowie die Modellierung deren Verfügbarkeit (insbesondere die Unterscheidung kleiner Werte von echten Nullen) ist nicht möglich. Daher wurden verschiedene Netzarchitekturen getestet, um die optimale Dimension des Netzes für diese Aufgabe zu finden. In Tabelle 4.4 sind die Ergebnisse einer stufenweisen Vergrößerung des KNN gegenübergestellt.

Grundlage für das Netztraining waren jeweils 5000 Lernbeispiele. In der gesamten Eingangsmatrix der Dimension 4×5000 wurden wieder $30\,\%$ als nicht verfügbar deklariert und daher durch Nullen ersetzt. Die Testdaten, mit denen eine Beurteilung der trainierten Netze durchgeführt wurde, bestanden für alle Varianten aus den identischen 50 Beispielen mit $5\,\%$ nicht verfügbaren Werten. Es wurden verschiedene Varianten nicht verfügbarer Eingangsdaten unterschieden, die in der Abbildung 4.5 als Verzweigungsbaum dargestellt sind. Eine nicht verfügbare Streckenmessung bedingt nicht verfügbare Längs- und Querabweichungen, eine nicht verfügbare Standardabweichung der Richtungsmessung hingegen beeinträchtigt die Längsabweichung nicht. Ebenso kann die Querabweichung mit hinreichender Genauigkeit bestimmt werden, ohne dass die Standardabweichung für die Strecke oder der ppm-Wert verfügbar sind. Es ergeben sich somit vier unterschiedliche Ergebnisklassen, die auf verschiedene Wege erreicht werden können. Alle Fälle mussten bei der Generierung der Soll-Ergebnisvektoren berücksichtigt werden.

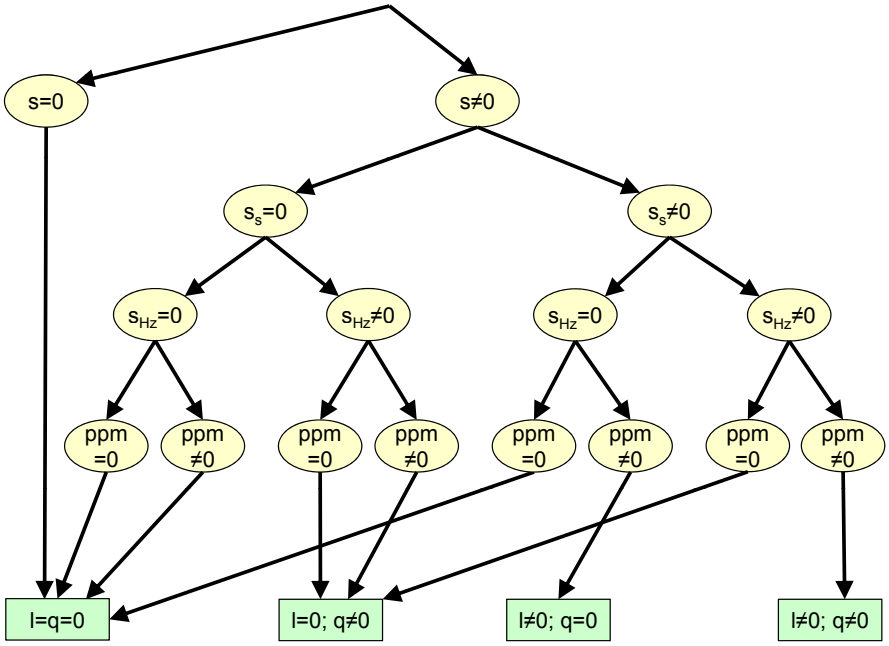


Abbildung 4.5: Fallbaum für nicht verfügbare Eingangsdaten beim Polaren Anhängen

Die Rechenzeit war in allen dargestellten Netzkonfigurationen kleiner als zwei Minuten und kann daher beim Vergleich vernachlässigt werden. Die Anzahl der Iterationen hängt stark von den zufällig gewählten Startwerten für die Gewichte und die Schwellwerte ab. Es lässt sich allerdings feststellen, dass komplexere Netze mehr Iterationen benötigen, dafür aber langsam und stetig iterieren. Netze mit kleinerer Neuronenzahl in der oder den beiden verdeckten Schichten tendieren zu stärkeren Schwankungen während der Iterationsphase. Daher passiert es gelegentlich, dass die Anpassung der Gewichte sehr früh aufgrund des maximal zulässigen Validierungsfehlers abbricht. Eine erneute Anpassung der Gewichte mit neuen Startwerten behebt dieses Problem jedoch schnell, sofern das Netz grundsätzlich in der Lage ist, die gestellte Aufgabe zu lösen. Bei den durchgeführten Tests wurde die Iteration meist nach etwa $200 - 400$ Iterationen aufgrund eines Validierungsfehlers gestoppt. Lediglich bei der $[15 - 15 - 2]$ Netzkonfiguration wurde die maximale Anzahl von 1000 Iterationen erreicht. Wie das mehrmalige Trainieren einzelner Netze gezeigt hat, ist die Anzahl der Iterationen bis zum Abbruch sehr unterschiedlich und daher wenig aussagekräftig.

Tabelle 4.4: Variation der Netzdimension eines KNN zur expliziten Modellierung der VE

Netz-architektur	MSE	theoretischer Wertebereich für l und q [mm]	GE v_l/v_q[mm] SOLL − IST	VE v_l/v_q[mm] SOLL = 0 − IST
$9-2$	$4\cdot10^{-4}$	$1 \le l \le 14$ $4\cdot10^{-3} \le q \le 141$ falls VE verletzt: $l = 0$ und/oder $q = 0$	< 1.6	< 0.6
$15-2$	$1\cdot10^{-4}$		< 0.7	< 0.2
$20-2$	$2\cdot10^{-5}$		< 0.5	< 0.08
$30-2$	$1\cdot10^{-5}$		< 0.5	< 0.08
$50-2$	$2\cdot10^{-5}$		< 0.7	< 0.5
$9-9-2$	$6\cdot10^{-7}$		< 0.3	< 0.07
$15-15-2$	$1\cdot10^{-6}$		< 0.08	< 0.03
Zum Vergleich: Ohne normierte Ein- und Ausgangswerte trainiertes KNN				
$15-15-2$	$5\cdot10^{-9}$		< 0.5	< 0.3

In der ersten Spalte der Tabelle 4.4 ist jeweils die Netzdimension zu erkennen. Es wurde zunächst mit der laut Formel 3.36 erforderlichen Anzahl von neun Neuronen bei vier Eingangswerten begonnen. Die Anzahl der Neuronen wurde dann schrittweise erhöht und zuletzt wurden auch zwei Netze mit jeweils zwei verdeckten Schichten getestet. Dabei wurde stets darauf geachtet, die Komplexität der Netze so gering wie möglich zu halten, um einem Überlernen des Netzes vorzubeugen. Neben dem MSE aus der letzten Epoche als erste Information über den Grad der Restfehlerminimierung ist jeweils die maximale Verbesserung v_l und v_q der Parameter Längs- und Querabweichungen angegeben. In der vorletzten Spalte handelt es sich um die Verbesserungen der verfügbaren Längs- und Querabweichungen, in der letzten Spalte werden die Verbesserungen zu den Netzausgängen bei nicht verfügbaren Längs- oder Querabweichungen und damit die Differenz zum Sollwert 0 angegeben.

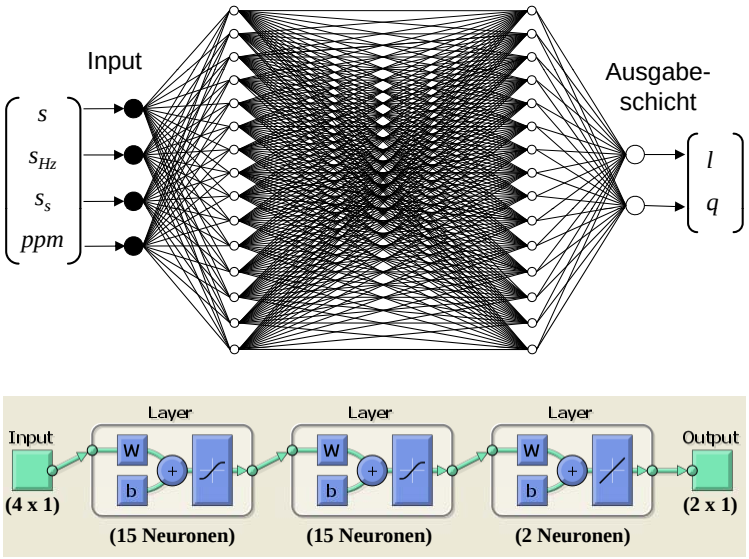


Abbildung 4.6: KNN welches gut zur gleichzeitigen Modellierung von GE und VE geeignet ist.

Wie sich herausgestellt hat, sinkt der MSE zunächst bei Steigerung der Anzahl verdeckter Neuronen bis auf 30 und beginnt dann wieder zu steigen. Die niedrigsten Werte werden in den Netzen mit zwei verborgenen Schichten erreicht. Die Verbesserungen zu den tatsächlichen verfügbaren Längs- und Querabweichungen sinken nur langsam mit steigender Netzkomplexität. Die beiden getesteten Netze mit drei Schichten können den wahren Wert am Besten annähern. Dasselbe gilt für den Fall nicht verfügbarer Längs- und Querabweichungen, wenngleich diese durch die zweischichtigen Netze mit 20 oder 30 verdeckten Neuronen ebenfalls gut abgebildet werden können. Zur Beurteilung der erzielten Genauigkeiten sind in der mittleren Spalte die theoretischen Wertebereiche für Längs- und Querabweichungen angegeben. Ist die Verfügbarkeit verletzt, so liegt der Sollwert bei 0. Aus der Tabelle wird ersichtlich, dass die Verwendung zweier verdeckter Schichten mit moderater Anzahl Neuronen die kleinsten Restfehler liefert. Die besten Ergebnisse ergaben sich beim Einsatz des dreischichtigen Netzes $[15 - 15 - 2]$ wie es in der Abbildung 4.6 dargestellt ist. Der etwa doppelt so große MSE im Vergleich zur Variante mit jeweils nur 9 Neuronen in den verdeckten Schichten ist dabei immer noch vernachlässigbar klein. Die grafische Darstellung der Netzstruktur lässt erahnen, welche Intelligenz in diesem relativ kleinen Netz bereits steckt. Andererseits wird sofort klar, dass mit der schnell steigenden Anzahl zu bestimmender Parameter bei Vergrößerung des Netzes auch die Anzahl der erforderlichen Lernbeispiele erhöht werden muss. Zwangsläufig geht damit auch eine wesentliche Erhöhung der Anzahl erforderlicher Iterationen und damit der Trainingsdauer einher.

In Abbildung 4.7 sind die Restfehler bei der Prozessierung der Testdaten mit dem 3-schichtigen Netz $[15 - 15 - 2]$ dargestellt. Die Datensätze, in denen Längs- und/oder Querfehler nicht verfügbar sind, wurden mit einem grün ausgefüllten Quadrat auf der X-Achse markiert. Um die Abbildbarkeit der Verfügbarkeit genauer beurteilen zu können, wurden in Abbildung 4.8 nur Restfehler der nicht verfügbaren Datensätze dargestellt. Dabei ist erkennbar, dass das Netz in diesem Beispiel nicht verfügbare Längsabweichungen wesentlich besser abbilden kann ($v_l < |0.005\,\mathrm{mm}|$) als nicht verfügbare Querabweichungen ($v_q < |-0.027\,\mathrm{mm}|$).

Problematisch ist jedoch die Erkennbarkeit von „echten“Nullen in den Ausgangsdaten, wenn die entsprechenden Ausgangswerte, die sich bei verfügbaren Daten ergeben, sehr klein sind. Dies ist hier bei der Querabweichung bei sehr genauen Geräten und kurzen Entfernungen rein rechnerisch der Fall. Hier kann es zur Verwechslung nicht verfügbarer Ausgangswerte mit kleinen (jedoch verfügbaren) Ergebnissen kommen. Die Querabweichung bei einer Entfernung von $1.5\,\mathrm{m}$ verbunden mit einem sehr genauen Instrument ($\sigma_r = 0.15\,\mathrm{mgon}$) ergibt rein rechnerisch eine Querabweichung von $0.04\,\mathrm{mm}$. Allerdings ist diese Genauigkeit sowohl bei der manuellen Anzielung als auch bei einer automatischen Zielerfassung auf sehr kurze Distanzen nicht möglich. In der Literatur wird für eine automatische Anzielung eines Robottachymeters eine erreichbare 3D-Punktgenauigkeit von etwa $1\,\mathrm{mm}$ angegeben (z. B. Leica [2009] oder Trimble [2010]). Nimmt man dabei gleiche Komponenten in allen drei Koordinatenrichtungen an (Konfidenz-Kugel), so kann in der minimal messbaren Zielentfernung von $1.5\,\mathrm{m}$ etwa von einer minimalen Längs- und Querabweichung von $0,6\,\mathrm{mm}$ ausgegangen werden. Damit ist eine Verwechslung von nicht verfügbaren Werten für Längs- und Querfehler mit tatsächlich berechneten, kleinen Werten in der Praxis nicht möglich. Im dargestellten Beispiel werden die Nullen für nicht verfügbare Werte mit einer Genauigkeit von weniger als $0.03\,\mathrm{mm}$ bestimmt (vgl. Abbildung 4.8). Voraussetzung für eine klare Unterscheidung nicht verfügbarer Werte mit kleinen Werten ist jedoch eine genaue Berücksichtigung der realen Verhältnisse. In dem hier dargestellten Beispiel *polares Anhängen* wird die Anzielgenauigkeit nicht berücksichtigt, daher kommt es bei kurzen Entfernungen zu unrealistischen Genauigkeiten und damit zu dem Problem der Trennung von GE und VE. Eine sorgfältige Planung der Implementierung eines Problems in einem KNN ist daher anzustreben, um zufriedenstellende Ergebnisse erzielen zu können.

Bislang wurde die Verfügbarkeit einzelner unabhängiger Datenarten betrachtet und modelliert. Es macht in vielen Anwendungsfällen jedoch Sinn, beispielsweise die Eingangsdaten (oder einen Teil da-

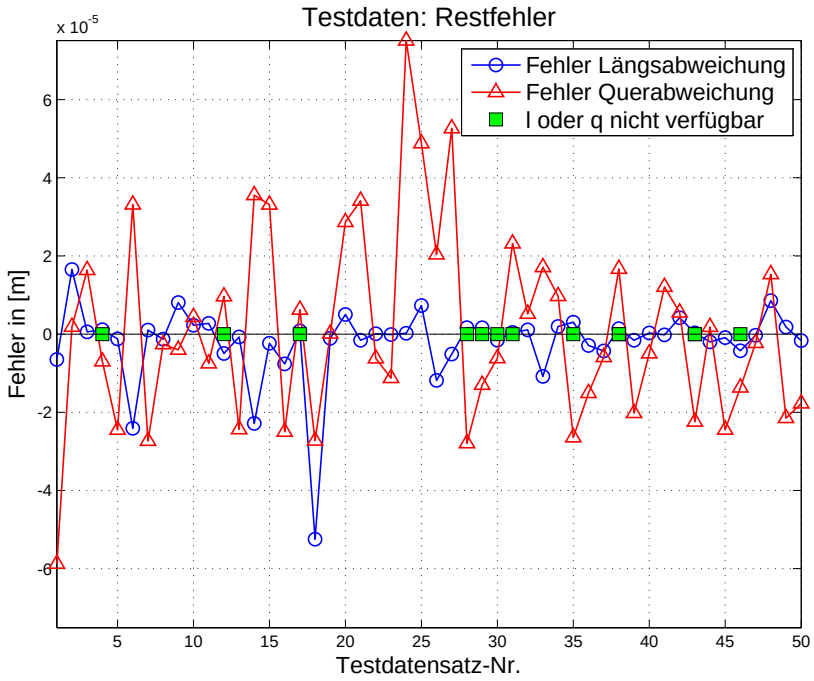


Abbildung 4.7: Restfehler der Längs- und Querabweichungen der Testdaten bei Modellierung von VE und GE (KNN $[15-15-2]$)

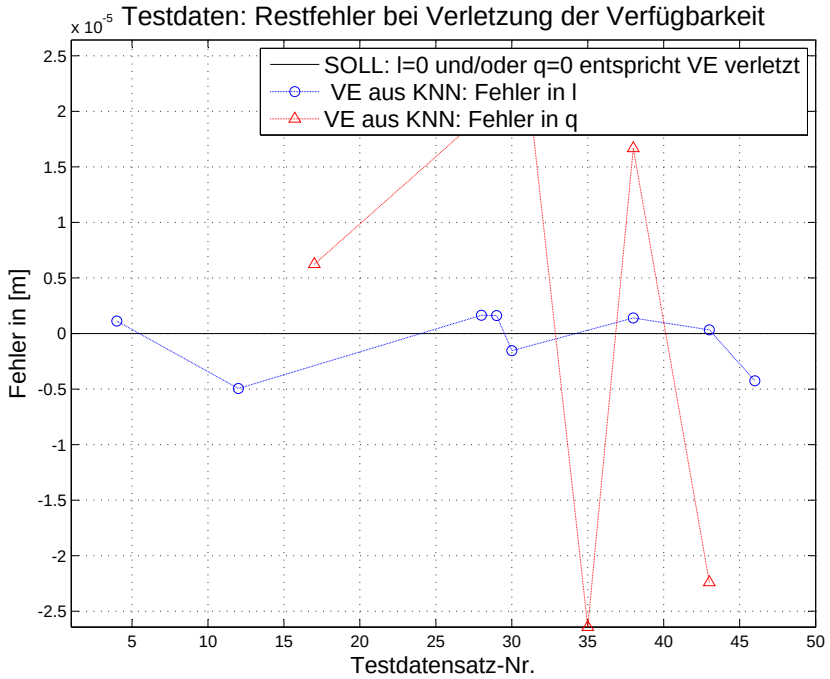


Abbildung 4.8: Restfehler für Längs- und Querabweichung des KNN $[15-15-2]$ bei verletzter Verfügbarkeit

von) als zusammengehörige Einheit und damit als Datensatz zu betrachten. Liefert ein einzelnes Messsystem mehrere Messwerte, so wird deren Verfügbarkeit nur auf Datensätze bezogen betrachtet. D. h. ein Datensatz ist verfügbar oder nicht verfügbar. Als Beispiel soll eine Inertialmesseinheit dienen, die

in der Regel drei Drehraten und drei Beschleunigungswerte zu einem Zeitpunkt liefert. Damit besteht der Datensatz aus sechs einzelnen Messgrößen, die zeitgleich ausgegeben werden. Ist das System z. B. in Folge eines Stromausfalls oder Softwareabsturzes nicht verfügbar, so werden auch keine Datensätze ausgegeben. Die Verfügbarkeit wird nur auf gesamte Datensätze bezogen und das Fehlen einzelner Daten in einem Datensatz wird in diesem Fall per Definition als Mangel in der Vollständigkeit betrachtet. Mit der Modellierung der Vollständigkeit beschäftigt sich der folgende Abschnitt 4.1.3.

4.1.3 Modellierung der Vollständigkeit

Wie in dem vorigen Abschnitt 4.1.2 bereits angedeutet wurde, hängen Verfügbarkeit und Vollständigkeit sehr eng zusammen. Die Unterscheidung kann erst getroffen werden, wenn klar definiert wurde, welche Eingangs- und Ausgangsgrößen logisch zu Datensätzen zusammengehören. Die Vollständigkeit lässt sich nicht für einzelne Daten, sondern immer nur für Datensätze angeben. Daher genügt für deren Modellierung die Ergänzung der Eingangs- und Ausgangsdaten um einen Parameter je Datensatzart, der die Vollständigkeit des jeweiligen Datensatzes markiert. Es bietet sich die Verwendung der binären Werte $1/0$ für „Datensatz vollständig“oder „Datensatz nicht vollständig“an. Damit kann beim Auftreten verschiedener Eingangsdatensätze, z. B. Messwerte aus unterschiedlichen Sensoren, auch eine gleichzeitige Modellierung von VE und VO erfolgen. Im folgenden Abschnitt werden jedoch alle Eingangsgrößen als ein zusammengehöriger Datensatz betrachtet.

Als Grundlage der folgenden Untersuchung dient wieder der Ansatz, der bereits bei der Modellierung der Genauigkeits- und Verfügbarkeitsparameter verwendet wurde. Dabei werden alle zusammengehörenden Eingangsdaten und alle Ausgangsdaten im Folgenden als Datensätze betrachtet. Zur Beschreibung deren Vollständigkeit werden Eingangs- und Ausgangsvektor um jeweils einen binären Parameter erweitert. Es werden wieder 5000 Lernbeispiele generiert, die ca. $30\,\%$ nicht verfügbare Einzelwerte enthalten, was aufgrund der Zusammenfassung zu Datensätzen einen Mangel in der Vollständigkeit des betreffenden Datensatzes darstellt. Wieder werden nach dem selben Schema wie bei der Untersuchung der Verfügbarkeit 50 unabhängige Testbeispiele mit $5\,\%$ nicht verfügbaren Einzelwerten generiert. In der Tabelle 4.5 sind die Ein- und Ausgangsdaten nochmals übersichtlich dargestellt.

Tabelle 4.5: Eingangs- und Ausgangswerte beim Polaren Anhängen; nur VO explizit berücksichtigt

Eingang			Ausgang		
QM	Parameter		QM	Parameter	
GE	σ_s	Standardabweichung Strecke	GE	l	Längsabweichung Neupunkt
GE	σ_r	Standardabweichung Richtung	GE	q	Querabweichung Neupunkt
GE	ppm	Atmosphärische Korrektur	VO	$\mathrm{VO_A}$	Vollständigkeit Ausgangsdaten
-	s	gemessene Strecke			
VO	$\mathrm{VO_E}$	Vollständigkeit Eingangsdaten			
$n = 5$ Eingangsparameter			$m = 3$ Ausgangsparameter		

Aufgrund der Erweiterung um zwei binäre Parameter werden nun mit Hilfe des KNN fünf Eingangs- auf drei Ausgangswerte abgebildet. Aus der bisherigen Erfahrung heraus muss die Netzdimension durch die Erweiterung aller Voraussicht nach ebenfalls vergrößert werden. Zum Vergleich mit dem Netz, welches zur Abbildung der Genauigkeit im Abschnitt 4.1.1 hinreichend war, wird zunächst jedoch ebenfalls

dieses einfache Netz eingesetzt und die Dimension dann schrittweise erhöht, um die beste Netzkonfiguration zu finden. Neben den Varianten 9, 15 und 30 Neuronen in einer verdeckten Schicht werden zwei Netze mit zwei verdeckten Schichten mit jeweils 9 und 15 verdeckte Neuronen in beiden Schichten getestet (vgl. Tabelle 4.6). Zusätzlich zu dem jeweils erzielten MSE wurden die Rechendauer sowie die Verbesserungen der Ausgangswerte für jede getestete Netzdimension gegenübergestellt. Dabei wurden zum einen wieder die beiden Genauigkeitsparameter Längs- und Querfehler mit den Sollwerten verglichen. Zum anderen wurde die Abweichung des vom Netz bestimmten Vollständigkeitsparameters in der letzten Spalte dargestellt. Dabei handelt es sich um die Differenz zwischen dem binären Sollwert $1/0$ für Datensatz vollständig/nicht vollständig und dem vom Netz generierten Istwert, der die Vollständigkeit des Ausgangsdatensatzes bestehend aus Längs- und Querfehler beschreibt.

Tabelle 4.6: Variation der Netzdimension eines KNN zur expliziten Modellierung der VO

Netz-architektur	MSE	Dauer [min:s]	GE $v_l/v_q[\mathrm{mm}]$ SOLL − IST	VO [-] SOLL − IST
$9-3$	$7\cdot10^{-4}$	$1:13$	<3	$<8\cdot10^{-3}$
$15-3$	$3\cdot10^{-4}$	$0:40$	<4	$<2\cdot10^{-2}$
$30-3$	$6\cdot10^{-6}$	$4:18$	<0.3	$<9\cdot10^{-4}$
$9-9-3$	$2\cdot10^{-4}$	$2:35$	<2	$<2\cdot10^{-2}$
$15-15-3$	$5\cdot10^{-7}$	$6:48$	<0.1	$<4\cdot10^{-5}$

Wie in der Tabelle 4.6 zu sehen ist, überschreiten die Verbesserungen des die Vollständigkeit beschreibenden Parameters in keinem Fall 0.02. Damit kann die Vollständigkeit bereits mit einem sehr einfachen KNN mit einer verdeckten Schicht mit neun Neuronen zufriedenstellend abgebildet werden. Wird hingegen auch die Verbesserung von Längs- und Querfehler betrachtet, so können diese nur mit den Netzen der Dimension $[30-3]$ und $[15-15-3]$ mit hinreichender Genauigkeit abgebildet werden. Die maximalen Abweichungen erreichen in diesen beiden Fällen maximal 1 bis 3 Zehntel-Millimeter. Aufgrund des etwa um Faktor 10 kleineren MSE-Wertes und der wesentlich besseren Modellierung der Vollständigkeit fällt hier die Auswahl erneut auf das dreischichtige Netz mit jeweils 15 Neuronen in den beiden verdeckten Schichten. Die Rechenzeit ist mit knapp 6 Minuten zwar etwas höher als bei dem zweischichtigen Netz, bei dem bis zum Abbruch etwas mehr als 4 Minuten erforderlich waren, sie ist aber durchaus noch akzeptabel. Insbesondere wenn man die große Streuung der Anzahl Iterationen und damit verbunden die Spanne der Rechenzeiten betrachtet, die sich bei mehrmaligem Wiederholen der Lernphase mit unterschiedlichen Startwerten ergaben. Diese erreichten Größenordnungen von wenigen Minuten, je nachdem ob und wie schnell die Validierungsprüfung zum Abbruch der Iteration geführt hat. Die Streuung der Testergebnisse für die verschiedenen trainierten Netze war bei allen Wiederholungen gering.

In Abbildung 4.9 sind die Restfehler aller Ausgangsparameter für alle 50 Testbeispiele dargestellt. Bis auf zwei Ausnahmen bewegen sich die mit dem KNN bestimmten Längs- und Querfehler im Bereich $\pm0.06\,\mathrm{mm}$. Die beiden Ausreißer in der Querabweichung bleiben jedoch kleiner als $0.1\,\mathrm{mm}$ und damit innerhalb der erforderlichen Genauigkeit. Der in grün dargestellte Fehler des Parameters, der die Vollständigkeit des Ausgangsdatensatzes beschreibt, bleibt in allen Fällen unter ±0.04. Diese Modellierungsgenauigkeit ist weit mehr als ausreichend, da hier lediglich eine klare Zuordnung zu den beiden binären Werten 0 und 1 für „Datensatz vollständig“oder „Datensatz nicht vollständig“erforderlich ist.

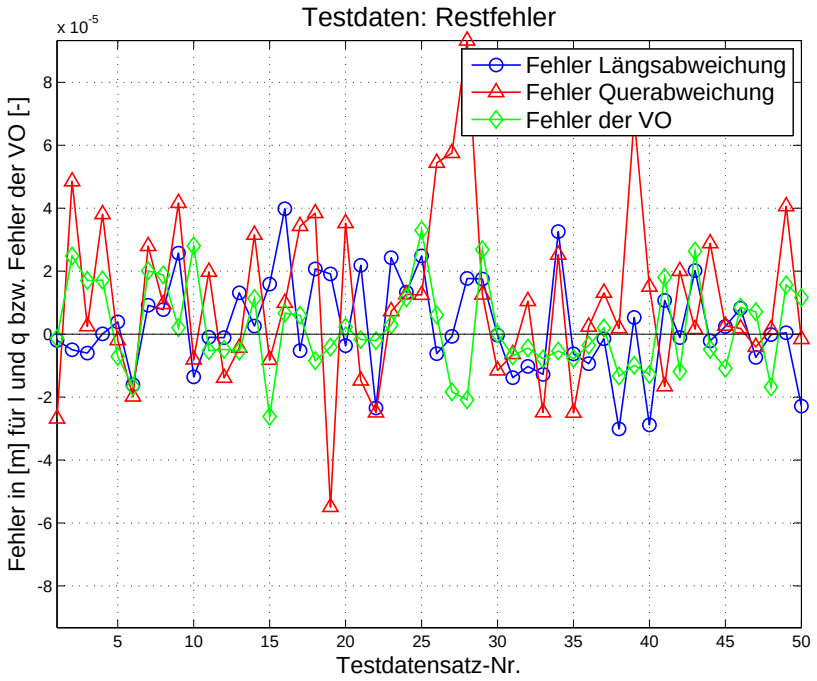


Abbildung 4.9: Restfehler für Längs- und Querabweichung sowie für den Parameter der Vollständigkeit des KNN $[15 - 15 - 3]$

Mit diesem Beispiel konnte gezeigt werden, dass eine Modellierung von Qualitätsparametern, die die Vollständigkeit von Datensätzen beschreiben, mit KNN grundsätzlich relativ einfach erfolgen kann. Die Dimension der Eingangs- und Ausgangsvektoren steigt zur Modellierung der Vollständigkeit für jeden zu beurteilenden Datensatz in den Eingangs- und Ausgangsdaten um einen Parameter an. Damit muss unter Umständen auch die Dimension des KNN sowie die Anzahl der Lernbeispiele erhöht werden.

4.1.4 Modellierung der Konsistenz

Laut Definition beschreibt die Konsistenz das Ausmaß der Übereinstimmung der Information mit dem Informationsmodell (Tabelle 2.1). Damit kann die Konsistenz nur beurteilt werden, wenn ein Daten- oder Informationsmodell vorliegt. In diesem Modell müssen sowohl die zulässigen Wertebereiche der einzelnen Eingangs- und Ausgangsparameter festgelegt als auch deren Datentyp definiert werden.

Zur Generierung der Beispieldaten mussten daher zunächst für jede Eingangsdatenart zwei ineinander liegende Wertebereiche definiert werden. Es wurde jeweils ein zulässiger Wertebereich und ein kleineres, innerhalb der zulässigen Werte liegendes Intervall zur Eingrenzung konsistenter Daten definiert. Die Intervalle wurden dabei eher willkürlich in realistischer Größenordnung gewählt um das prinzipielle Vorgehen zu testen. Liegt ein zufällig aus dem großen Intervall gewählter Wert auch innerhalb des kleineren, so ist der Wert konsistent, andernfalls ist er und damit auch der gesamte Datensatz inkonsistent. Wird der Datensatz als inkonsistent erkannt, so erfolgt keine weitere Berechnung. Die definierten Intervalle für die vier Eingangswerte sind in der Tabelle 4.7 zusammengefasst.

Zum prinzipiellen Nachweis der Modellierbarkeit der Konsistenz mit KNN dient erneut das Beispiel des polaren Anhängens. Um das Modell möglichst einfach zu halten, wird wie bei der Darstellung der Vollständigkeit nur ein binärer Parameter für die Konsistenz eingeführt. Das heißt, es werden die

Tabelle 4.7: Zulässige und konsistente Wertebereiche für die Eingangswerte

Eingangsparameter	Wertebereich zulässig	Wertebereich konsistent
Standardabweichung Strecke	$0.7\,\mathrm{mm} \leq \sigma_s \leq 9.0\,\mathrm{mm}$	$1.0\,\mathrm{mm} \leq \sigma_s \leq 7.0\,\mathrm{mm}$
Standardabweichung Richtung	$0.12\,\mathrm{mgon} \leq \sigma_{\mathrm{Hz}} \leq 4.0\,\mathrm{mgon}$	$0.15\,\mathrm{mgon} \leq \sigma_{\mathrm{Hz}} \leq 3.0\,\mathrm{mgon}$
Atmosphärische Korrektur	$0 \leq \mathrm{ppm} \leq 5$	$1 \leq \mathrm{ppm} \leq 4$
gemessene Strecke	$-10\,\mathrm{m} \leq s \leq 3100\,\mathrm{m}$	$1.5\,\mathrm{m} \leq s \leq 3000\,\mathrm{m}$

Eingangs- und Ausgangsdaten wieder als zusammengehöriger Datensatz betrachtet, deren Konsistenz genau dann gegeben ist, wenn alle Bestandteile des Datensatzes konsistent sind. Wie in den vorherigen Versuchen zur Modellierung der Vollständigkeit werden auch in diesem Fall wieder 5000 Lernbeispiele generiert und das trainierte Netz im Anschluss mit 50 zufälligen Beispielen aus dem zulässigen Wertebereich getestet. Damit ergeben sich insgesamt die Eingangs- und Ausgangsdaten, wie sie in der Tabelle 4.8 dargestellt sind.

Tabelle 4.8: Eingangs- und Ausgangswerte beim Polaren Anhängen; nur KO explizit berücksichtigt

Eingang			Ausgang		
QM	Parameter		QM	Parameter	
GE	σ_s	Standardabweichung Strecke	GE/VE	l	Längsabweichung Neupunkt
GE	σ_r	Standardabweichung Richtung	GE/VE	q	Querabweichung Neupunkt
GE	ppm	Atmosphärische Korrektur	KO	$\mathrm{KO_E}$	Konsistenz Eingangsdaten
-	s	gemessene Strecke			
KO	$\mathrm{KO_E}$	Konsistenz Eingangsdaten			
$n = 5$ Eingangsparameter			$m = 3$ Ausgangsparameter		

Es wird hier nicht unterschieden, welcher Eingangswert oder welche Eingangswerte im Datensatz zu einer Inkonsistenz geführt hat. Daher wird bei inkonsistentem Eingangsdatensatz kein Ergebnis berechnet, das Ergebnis ist somit nicht verfügbar. Die nicht verfügbaren Ausgangswerte werden, wie im Abschnitt 4.1.2 bereits getestet, mit „echten“Nullen dargestellt. Die beiden Ausgangsparameter l und q stellen daher sowohl die Genauigkeit als auch die Verfügbarkeit dar. Der zur Beschreibung der Konsistenz eingeführte Parameter am Eingang wird hier zusätzlich auch in den Ausgangsdatensatz übertragen. Damit ist auch in den Ausgangsdaten erkennbar, weshalb der Datensatz nicht berechnet werden konnte und damit l und q im Datensatz nicht verfügbar sind. Grundsätzlich ist dieser Parameter als Bestandteil der Ausgangsdaten jedoch nicht unbedingt erforderlich, da dieselbe Information bereits im Eingangsdatensatz enthalten ist.

Die Trainingsdaten enthalten mit $58\,\%$ sehr viele inkonsistente Datensätze. Nach den bisherigen Erfahrungen, die unter anderem bei der Modellierung der Verfügbarkeit gemacht wurden, ist es sinnvoll, für das Training des Netzes einen unrealistisch hohen Anteil kritischer Werte des zu modellierenden Parameters zu verwenden. Dadurch werden die Fälle, die in der Realität selten auftreten, hinreichend

trainiert. In dem zufälligen Testdatensatz, mit dem das trainierte Netz im Anschluss getestet wird, liegt der Anteil der nicht konsistenten Datensätze bei ca. $55\,\%$. Diese Anteile ergaben sich durch die Größe der zulässigen Datenintervalle und der innerhalb dieser liegenden kleineren Intervalle, die die konsistenten Daten markieren.

Aufgrund der sehr ähnlichen Konfiguration mit fünf Eingangs- und drei Ausgangswerten, wie sie bereits bei der Modellierung der Vollständigkeit vorlag, orientiert sich die Dimensionierung des KNN an den dort gemachten Erfahrungen (vgl. Tabelle 4.6). Die besten Ergebnisse konnten bei der Untersuchung der Vollständigkeit mit zwei verdeckten Schichten mit jeweils 15 Neuronen gemacht werden. Daher wird auch hier die Netzkonfiguration $[15 - 15 - 3]$ gewählt.

Das Netz iteriert relativ schnell, bereits nach ca. 100 Epochen sinkt der MSE auf unter 10^{-6} ab und erreicht nach Durchlauf der maximalen Anzahl von 1000 Iterationen etwa $1 \cdot 10^{-8}$ (vgl. Tabelle 4.9). Die Lernphase dauert aufgrund der relativ großen Netzdimension und Anzahl der Lernbeispiele knapp sechs Minuten.

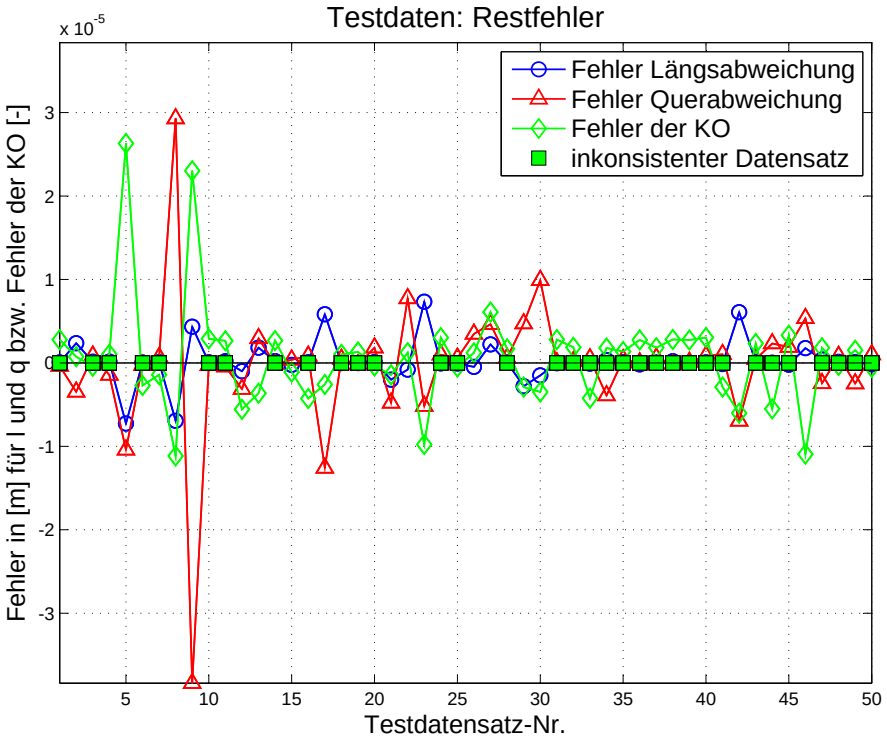


Abbildung 4.10: Restfehler für Längs- und Querabweichung sowie für den Parameter der Konsistenz des KNN $[15 - 15 - 3]$

Die Abweichungen in den Längs- und Querfehlern bewegen sich im Bereich von wenigen $10^{-2}\,\mathrm{mm}$ und sind damit ein bis zwei Zehnerpotenzen kleiner als real vorkommende Längs- und Querfehler bei der polaren Punktaufnahme mit Tachymetern. Damit ist eine Verwechslung von kleinen Längs- und Querfehlern mit „echten“Nullen, die nicht verfügbare Ausgangswerte repräsentieren sollen, sehr unwahrscheinlich. In der Abbildung 4.10 sind sämtliche inkonsistente Datensätze mit einem grünen Quadrat gekennzeichnet. Wie in der Darstellung erkennbar ist, treten bei inkonsistenten Datensätzen wesentlich kleinere Abweichungen zu den Sollwerten auf, es gibt jedoch einzelne Ausnahmen.

Wie die Ergebnisse bestätigen, ist die Dimension des getesteten KNN ausreichend, um die Konsistenz der Daten in diesem Beispiel hinreichend genau zu modellieren. Möglicherweise ist bereits ein einfacheres Netz dazu in der Lage. Auf eine weitergehende Untersuchung und den Vergleich verschiedener Netzdimensionen, wie er z. B. in Kapitel 4.1.3 durchgeführt wurde, wird jedoch verzichtet, da hier lediglich der Nachweis erbracht werden soll, dass KNN die Modellierung der Konsistenz prinzipiell leisten können.

Tabelle 4.9: Ergebnis der expliziten Modellierung der KO mit einem KNN $[15-15-3]$

Netz-architektur	MSE	Dauer $[\min:\mathrm{s}]$	GE $v_l/v_q[\mathrm{mm}]$ SOLL − IST	KO [-] SOLL − IST
$15-15-3$	$1,4\cdot10^{-8}$	$5:40$	<0.04	$<3\cdot10^{-5}$

Inkonsistenzen in den Datensätzen, die durch falsche Datentypen wie z. B. alphanumerische Zeichenketten anstatt Zahlenwerten auftreten können, müssen im voraus abgefangen werden. Wie bereits in der Tabelle 4.2 dargestellt wurde, erlaubt die Software zur Berechnung der KNN nur numerische Eingangs- und Ausgangswerte. Alle anderen Datentypen führen zu einem Fehler und das KNN kann nicht trainiert werden.

4.1.5 Modellierung der Aktualität

Laut der Definition der Aktualität in Kapitel 2.2, werden unter diesem Qualitätsmerkmal alle Parameter zusammengefasst, die das Ausmaß der Übereinstimmung mit der sich zeitlich ändernden Realität beschreiben. Dazu gehören relative Angaben, wie z. B. die Dauer eines Prozessschritts oder Angaben auf der absoluten Zeitskala, wie z. B. Zeit- oder Datumsstempel. Die Beurteilung der Erfüllung der Aktualität kann nur erfolgen, wenn von Nutzerseite überprüfbare Anforderungen bestehen und formuliert werden. Gibt es vom Nutzer für einzelne Anwendungen beispielsweise eine Vorgabe, wie alt die Daten sein dürfen, so kann jeder Datensatz hinsichtlich der Aktualität anhand seines Alters beurteilt werden. Das Alter der Daten muss aus den mitgeführten Qualitätsparametern ermittelbar sein, d. h. das Alter zum Beurteilungszeitpunkt kann entweder mit Hilfe eines Zeitstempels (des Entstehungszeitpunktes der Eingangsinformationen) und der Prozessierungsdauer oder aus der Differenz geeigneter Zeitstempel berechnet werden.

Die Beurteilung der Aktualität mit statischen KNN ist in einfacher Art und Weise möglich. Ist der jeweilige Schwellwert für die Beurteilung einer Zeitspanne oder eines Zeitpunktes bekannt, so kann die Entscheidung „zeitliche Anforderung erfüllt/nicht erfüllt“im Netz mitmodelliert werden. Die Modellierung der Aktualität stellt dabei nur eine einfache Abfrage zu einem festen Zeitpunkt dar. Sollen zeitvariante Größen mit dem Netz modelliert werden, so sind dynamische Netze erforderlich, in denen die zeitliche Reihenfolge der Eingangs- und Ausgangsdaten berücksichtigt wird. Typische Anwendungsgebiete dynamischer Netze sind beispielsweise die Zeitreihenanalyse, die betriebswirtschaftliche Marktprognose oder die Steuerung autonomer Fahrzeuge (Poddig und Sidorovitch [2001]), Zell [1997]). Da diese Netzart im Rahmen dieser Arbeit nicht weiter untersucht wird, werden sich zeitlich ändernde Größen bei der Untersuchung der Anwendbarkeit von KNN zur Modellierung von Qualität in Prozessen ausgeschlossen. Die triviale Beurteilung der Aktualität einzelner Datensätze anhand der Anforderung wird daher auch nicht weiter bei den Untersuchungen berücksichtigt. Nach den bisher gemachten Erfahrungen mit der Anwendung von KNN ist diese einfache Art der Modellierung von Aktualität aber möglich.

4.1.6 Modellierung der Korrektheit

Die Korrektheit von Daten beschreibt das Ausmaß der Übereinstimmung der Informationen mit der Realität. Dabei wird die Aktualität explizit vorausgesetzt und damit eine mögliche zeitliche Änderung der Qualität der Daten nicht als mangelnde Korrektheit definiert (vgl. Tabelle 2.1).

Der Parameter der Korrektheit kann dabei entweder ein aus der Genauigkeit abgeleiteter Parameter sein oder es handelt sich um eine von der Genauigkeit unabhängige und eigenständige Eigenschaft der Daten. Zum Beispiel kann eine Straßenkante als Bestandteil der berechneten Trajektorie eines Fahrzeugs auf dem digitalen Straßennetz lediglich korrekt oder nicht korrekt sein, Zwischenstufen sind hier nicht zugelassen. Im Gegensatz dazu ist die Angabe der Korrektheit einer mit GPS ermittelten einzelnen Fahrzeugposition nicht ohne weiteres möglich. Dazu ist zunächst die Vorgabe einer Grenze für die Standardabweichung der Position erforderlich, anhand derer die Entscheidung „korrekt“oder „nicht korrekt“erfolgen kann. Diese Grenze kann nur aus einer externen Quelle stammen, da eine vom zu beschreibenden Prozess unabhängige Referenz notwendig ist. Meist sind unabhängige Messungen mit genaueren Messmethoden erforderlich, um entsprechende Sollwerte als Referenzdaten zu generieren.

Zur Modellierung der Korrektheit mit Hilfe eines KNN, muss der Eingangsvektor um weitere binäre Werte erweitert werden. Dies können einerseits aus Parametern der Genauigkeit abgeleitete Korrektheitsparameter sein oder aber neue Parameter, die sich nur binär darstellen lassen. Funktional betrachtet ist die Aufgabe des KNN jedoch vergleichbar mit der Erweiterung um einen binären Parameter, der die Vollständigkeit des Datensatzes oder dessen Konsistenz beschreibt. Bei der Betrachtung der Konsistenz müssen ebenfalls Referenzdaten in Form von Datenintervallen vorhanden sein, um die Beurteilung der Konsistenz durchzuführen. In diesem Fall handelt es sich jedoch in der Regel um intern definierte Intervalle, zu deren Festlegung keine weiteren, externen Messungen erforderlich sind. Daher wird an dieser Stelle auf den vorigen Abschnitt 4.1.4 verwiesen und auf die Umsetzung in dem Beispiel verzichtet.

4.1.7 Zusammenfassung

Der Einsatz künstlicher neuronaler Netze konnte anhand dieses Beispiels aus dem geodätischen Umfeld untersucht werden. Wie gezeigt wurde, sind die KNN in der Lage, die Genauigkeitsparameter Längs- und Querabweichung in zufriedenstellender Genauigkeit aus den Eingangsparametern vorherzusagen. Auch die zusätzliche Abbildung der Verfügbarkeit von Daten als weiterer Qualitätsparameter können die künstlichen neuronalen Netze leisten. Dabei wurden nicht verfügbare Daten durch den Wert „0“dargestellt. Wie sich in den Untersuchungen gezeigt hat, ist das KNN in der Lage, diese glatte 0 von kleinen Werten nahe 0 zuverlässig zu unterscheiden, sofern für die Messgenauigkeit nur realistische Werte zugelassen wurden. In anderen Fällen kann dies jedoch zu Problemen führen und ist daher ggf. von Fall zu Fall näher zu untersuchen.

Werden einzelne Datenarten zu Datensätzen zusammengefasst, so kann deren Vollständigkeit beschrieben werden. Es wurde anhand simulierter Daten, bei denen vereinzelt die Vollständigkeit der Datensätze verletzt war, gezeigt, dass deren Modellierung durch Einführung eines binären Parameters je Datensatz grundsätzlich möglich ist. Das Netz erlernt zuverlässig die Bedeutung der zusätzlichen Parameter in den Eingangs- und Ausgangsdaten und kann diese auch vorhersagen. In der Regel muss dazu jedoch die Netzdimension aufgrund der zusätzlichen Parameter entsprechend vergrößert werden, ebenso wie die Anzahl an Trainingsbeispielen. Die Trainingsbeispiele müssen den Fall unvollständiger Datensätze in ausreichender Zahl beinhalten, andernfalls ist das Netz nicht in der Lage diese Fälle zu erlernen. Wie

gezeigt wurde, ist dazu auch die künstliche Erhöhung dieses Anteils weit über einen real vorkommenden Anteil hinaus legitim.

Durch die Hinzunahme eines weiteren Parameters, konnte auch die Datenkonsistenz mittels KNN abgebildet werden. Dazu war jedoch für deren Festlegung zunächst die Definition von zulässigen und vorkommenden Wertebereichen für alle Eingangsgrößen erforderlich. Auf den trivialen Fall, dass nicht zulässige Zeichen als Werte in den Daten vorkommen und so für eine Inkonsistenz sorgen, wurde hier verzichtet. Da die Realisierung der KNN mit Matlab keine nicht-numerischen Werte zulässt, müsste dieser Fall zunächst durch eine Abfrage aller Eingangswerte überprüft werden. Die Abbildung derartiger Inkonsistenzen in den Daten mit Hilfe der KNN-Toolbox von Matlab ist nicht direkt möglich.

Aufgrund der Beschränkung auf statische Netze wurde die Modellierbarkeit von Parametern, die die Aktualität beschreiben, im Rahmen dieser Arbeit nicht untersucht. Die Beispiele wurden dem KNN in Form einer Matrix zugeführt, die in beliebiger zeitlicher Reihenfolge abgearbeitet werden konnte.

Die Darstellung von Parametern, die die Korrektheit der Daten beschreiben, kann mit binären Parametern erfolgen. Handelt es sich dabei um aus der Genauigkeit abgeleitete Parameter, so ist das Vorgehen vergleichbar mit der Modellierung der Datenkonsistenz. Es wird ein Grenzwert festgelegt, ab dem die Daten als korrekt beurteilt werden dürfen. Stellt der Korrektheitsparameter einen eigenständigen Parameter dar, muss die Entscheidung für den Parameterwert 1 (korrekt) oder 0 (nicht korrekt) anhand anderer Referenzdaten erfolgen. In jedem Fall sind Referenzinformationen zur Beurteilung der Korrektheit erforderlich. Da diese in der Regel nicht vorliegen, wurde auf eine weitere, allgemeine Untersuchung verzichtet und vielmehr auf das Vorgehen zur Behandlung der Konsistenz verwiesen.

Das erste Beispiel zur Untersuchung des Potenzials von KNN ist theoretischer Natur und daher nur bedingt auf die Realität verallgemeinerbar. Insbesondere die komplexen Zusammenhänge zwischen den unterschiedlichen Qualitätsparametern können mit simulierten Daten nur bedingt real dargestellt werden. Daher wird in dem folgenden realen Beispiel die Praxistauglichkeit der KNN zur Modellierung von Qualitätsparametern und zur Abbildung deren komplexen Zusammenhänge näher untersucht.

4.2 Beispiel: Floating Phone Data

Im Folgenden wird die Eignung der künstlichen neuronalen Netze zur Modellierung von Datenqualität in Prozessen anhand realer Daten gezeigt. Es standen hierfür Daten aus dem im Laufe dieser Arbeit bereits mehrfach erwähnten Projekt Do-iT im Bereich der Verkehrstelematik zur Verfügung. Das vom BMWi im Rahmen der Initiative *Verkehrsmanagement 2010* geförderte Projekt wurde im Zeitraum $2005 - 2009$ am Institut der Anwendungen der Geodäsie im Bauwesen bearbeitet. Weitere Projektpartner waren die DDG Gesellschaft für Verkehrsdaten mbH, das Innenministerium Baden-Württemberg, die Städte Karlsruhe und Stuttgart sowie der Lehrstuhl für Verkehrsplanung und Verkehrsleittechnik der Universität Stuttgart. Fokus des Projektes war die Untersuchung von Möglichkeiten zur Generierung von Ortungsinformationen von Straßenverkehrsteilnehmern aus anonymisierten Mobilfunkdaten und die Bewertung des Potenzials, dieser sogenannten Floating Phone Data (FPD) in verkehrstechnischen Anwendungen zu nutzen. Diese FPD-Trajektorien beschreiben den im digitalen Straßennetz zurückgelegten Weg einzelner Verkehrsteilnehmer mit zeitlichem Bezug und spiegeln somit grundsätzlich das Verkehrsgeschehen im Straßennetz wider.

Zum besseren Verständnis der Datenprozessierung erfolgt im nächsten Abschnitt eine kurze Vorstellung des Projektumfelds und der Vorgehensweisen zur Generierung von FPD-Trajektorien. Es werden

die beiden entwickelten Ansätze vorgestellt und, soweit für die weiteren Untersuchungen notwendig, in groben Zügen erläutert.

4.2.1 Generierung von FPD-Trajektorien

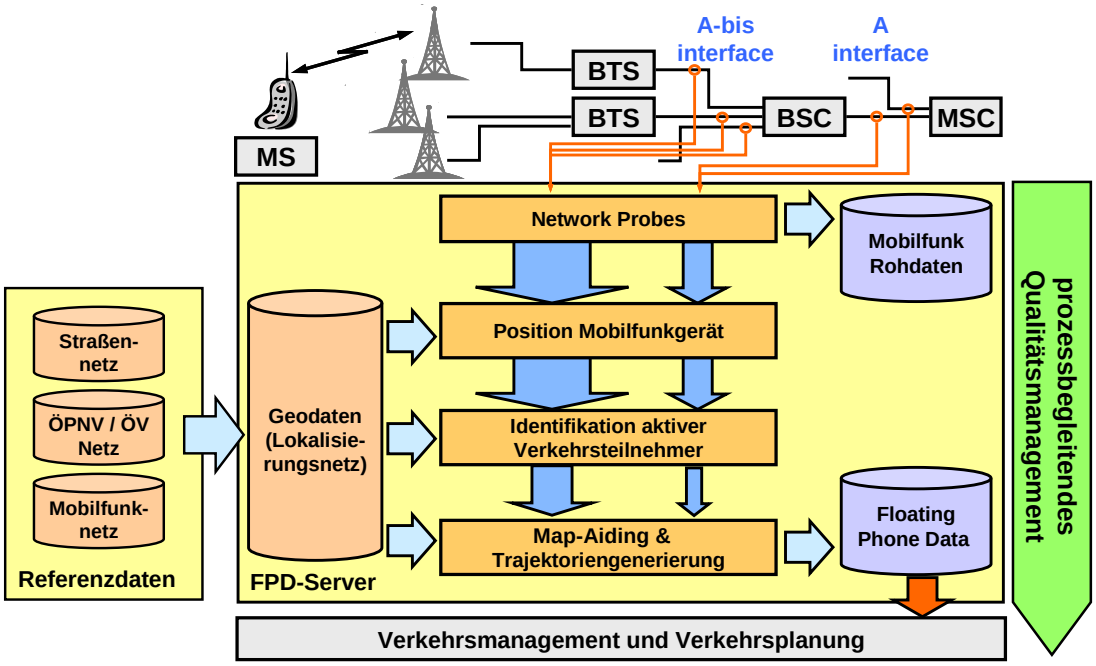


Abbildung 4.11: Ablaufschema zur Erzeugung von FPD (Wiltschko u. a. [2007])

Das Projektgebiet erstreckte sich über das Autobahnviereck Stuttgart, Karlsruhe, Walldorf und Weinsberg und beinhaltete zusätzlich die Ballungsgebiete Karlsruhe und Stuttgart. In dem gesamten Gebiet standen Daten eines Mobilfunkanbieters zur Verfügung. Dabei handelte es sich um die standardmäßig im GSM-Netz anfallenden Daten, die an zwei unterschiedlichen Schnittstellen A und A-bis im Mobilfunknetz abgegriffen werden können. Im oberen Bereich der Abbildung 4.11 ist die GSM-Netzinfrastruktur angedeutet. Die Mobilstation (MS) kommuniziert über Funkzellen (durch Antennenmasten angedeutet) mit der bedienenden Basisstation (*en.*: Base Transceiver Station, BTS). Eine Basisstation ist in der Regel für die Versorgung von ein bis -im Fall von Sektorantennen- sechs Funkzellen verantwortlich. Im Rahmen des Projekts traten jedoch nur Sektorantennen mit maximal 3 Funkzellen auf. Mehrere dieser BTS werden ihrerseits über den sogenannten A-bis-Link (daher die Schnittstellenbezeichnung A-bis) einem Base Station Controller (BSC) angeschlossen. Mehrere BSC sind ihrerseits über den A-Link mit einer digitalen Vermittlungsstelle, dem Mobile-services Switching Centre (MSC) verbunden.

Je nach Schnittstelle unterscheiden sich die abgegriffenen Daten in ihrem Umfang wesentlich (Abbildung 4.11). Die A-Daten[1] beinhalten zunächst nur Informationen, die zur stetigen Anbindung eines eingeschalteten Mobiltelefon an die Netzinfrastruktur erforderlich sind. Dies sind im Wesentlichen die letzte und die aktuelle LAC[2] sowie die Zelle, in die beim letzten LAC-Wechsel gewechselt wurde. Dieses Minimum an Kommunikation zwischen mobilem Gerät und der Infrastruktur ist unter anderem notwendig, um ein bewegtes Gerät an benachbarte Zellen, LACs oder auch andere Netzanbieter weiterzureichen

[1] Entsprechend der Schnittstelle des Datenabgriffs, werden die Daten als A- bzw. A-bis-Daten bezeichnet.
[2] LAC steht für Location Area Code und bezeichnet eine administrative Organisationseinheit in der GSM-Netzarchitektur.

und somit die Bereitschaft zum Aufbau eines Telefonats jederzeit sicherzustellen. Ist das mobile Gerät jedoch aktiv, so beinhalten die A-Daten während dieser Zeit lückenlos alle Funkzellen in denen das Gerät angemeldet war.

Im Gegensatz dazu entstehen A-bis-Daten ausschließlich bei Telefongesprächen oder aktiven Datenverbindungen. Die A-bis-Daten beinhalten neben den Informationen der A-Daten, die jederzeit zur Verfügung stehen, unter anderem auch vom Endgerät gemessene Signalstärken von den bis zu sieben empfangenen Antennen mit der höchsten Signalstärke und den sogenannten TA-Wert[3] zur bedienenden Zelle. Diese Daten müssen während einer stehenden Verbindung, insbesondere bei schnell bewegten Mobilfunkteilnehmern, aktuell gehalten werden, um das Weiterreichen in benachbarte Zellen und LACs rechtzeitig durchführen zu können und stehen daher mit der Frequenz von 2 Hz zur Verfügung.

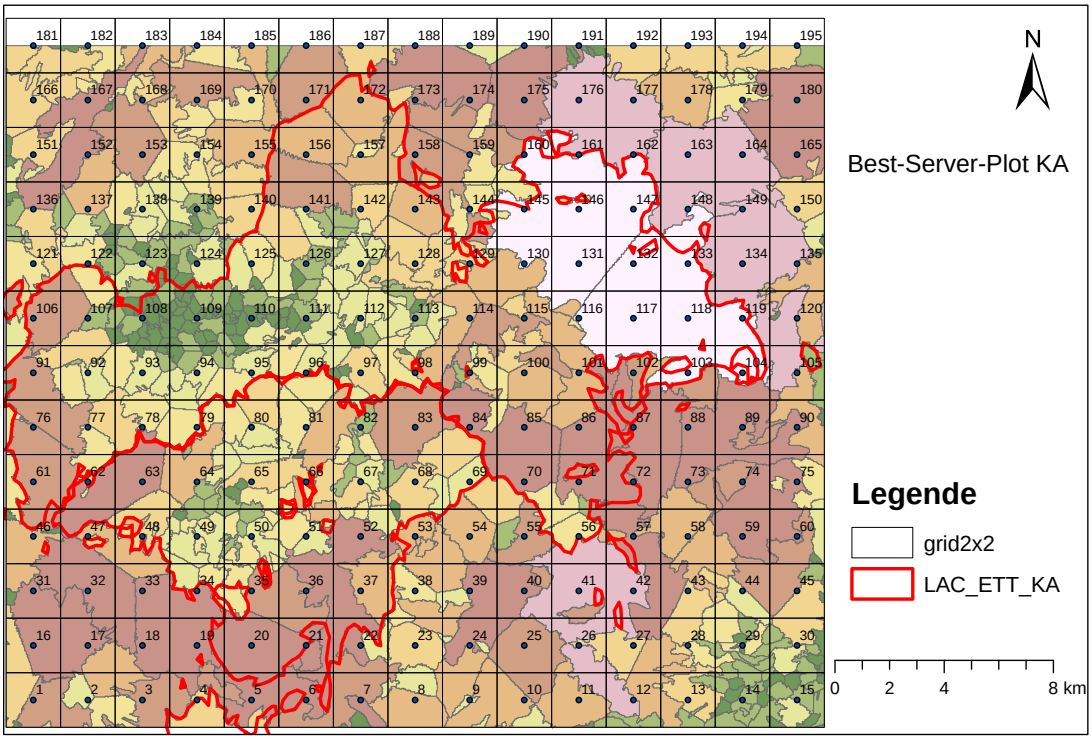


Abbildung 4.12: Best-Server-Plot im Bereich Karlsruhe mit den rot markierten LACs Karlsruhe und Ettlingen in denen A-bis-Daten zur Verfügung standen

Aus technischen Gründen standen die A-bis-Daten im Gegensatz zu den A-Daten nur im Großraum Karlsruhe, den LAC Karlsruhe und Ettlingen zur Verfügung. Die beiden LACs sind in der Abbildung 4.12 dick rot umrandet dargestellt. Die einzelnen Flächenpolygone stellen die Verteilung der am besten empfangbaren Antennen, den sogenannten Best-Server-Plot, dar. Die kleineren, grün dargestellten Funkzellen markieren dabei die eng bebauten Innenstadtbereiche von Karlsruhe (links, oberhalb der Kartenmitte) und Ettlingen (linkes unteres Viertel der Karte), in denen aufgrund der starken Abschattungen und der großen erforderlichen Netzkapazitäten eine sehr hohe Antennendichte erforderlich ist.

[3]TA steht für Timing Advance und dient der zeitlichen Synchronisation der im Netz zu übertragenden Daten des Teilnehmers. Der Wert wird aus der Signallaufzeit abgeleitet und stellt daher eine grobe Entfernungsangabe zwischen Antenne und Mobilstation (MS) dar.

Im Rahmen des Projektes Do-iT wurden zwei unterschiedliche Ansätze, basierend auf den beiden grundsätzlich zur Verfügung stehenden Datenarten, verfolgt. Beiden Ansätzen gemeinsam war das Vorgehen, wie es in dem Ablaufschema in Abbildung 4.11 angedeutet wird. Mit Hilfe der im GSM-Netz an unterschiedlichen Schnittstellen installierten Network Probes wurden die Mobilfunkrohdaten in Echtzeit abgegriffen und gespeichert. Daraus konnten mit den beiden verschiedenen Verfahren Positionen von Mobilstationen berechnet werden. Bereits vorher, jedoch insbesondere nachher, wurden aus den Informationsketten der Teilnehmer mit verschiedenen Verfahren die aktiven Individualverkehrsteilnehmer identifiziert. Insbesondere der ÖPNV und der schienengebundene Verkehr sowie Fußgänger und sich nicht bewegende Geräte mussten aus den Daten herausgefiltert werden. Anschließend wurden aus den Positionsfolgen der als aktiv eingestuften Straßenverkehrsteilnehmern, unter anderem mit Hilfe von Map-Aiding-Verfahren, Trajektorien im digitalen Straßennetz berechnet (Czommer [2000]). Diese Floating Phone Data standen schließlich zur Evaluierung in unterschiedlichen verkehrstechnischen Anwendungen der Projektpartner zur Verfügung. Anwendungen der Partner waren hier unter anderem die Verkehrslageermittlung in Echtzeit, die Steuerung von Wechselverkehrszeichen und die Ermittlung des Routenwahlverhaltens der Verkehrsteilnehmer.

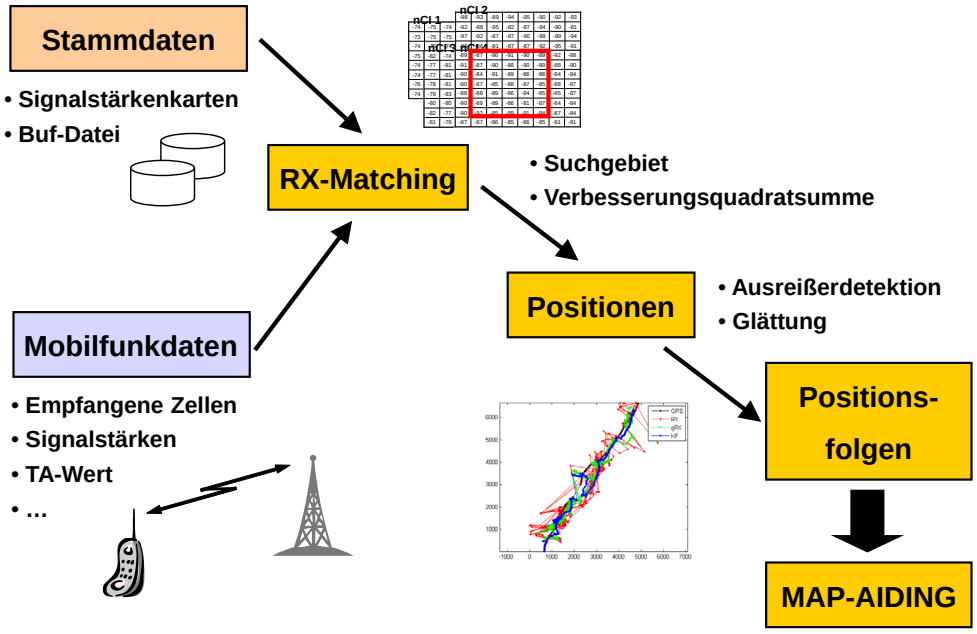


Abbildung 4.13: Ablauf der Mobilfunkortung aus Abis-Daten (Do-iT [2009b])

Auf Grundlage der A-Daten wurde ein zellbasierter Algorithmus entwickelt, mit dem grundsätzlich jede zeitlich und räumlich längere Informationskette eines sich bewegenden Teilnehmers ausgewertet werden kann, egal ob sich diese nur aus Zellwechseln eines passiven Telefons oder (auch) aus kurzen Gesprächen oder Kurznachrichten zusammensetzt. Die räumlich wenig dichten Informationen lassen allerdings naturgemäß auch nur eine relativ grobe Zuordnung des zurückgelegten Weges eines Teilnehmers zu. Im Gegensatz dazu stehen auf A-bis-Ebene mehr und sowohl räumlich als auch zeitlich viel dichtere Informationen eines Teilnehmers zur Verfügung, was eine genauere Zuordnung des zurückgelegten Weges im digitalen Straßennetz erlaubt. Der grobe Ablauf der Mobilfunkortung aus A-bis-Daten ist in der Abbildung 4.13 zu sehen. Eine erste genäherte Position des Teilnehmers kann aus der Antennenausrichtung und dem TA-Wert, der einen Hinweis zur Laufzeit des Signals und somit zur Entfernung zwischen Mobilstation und Antenne liefert, ermittelt werden (Do-iT [2009b]). Mit Hilfe der Signalstärkedifferenzen benachbarter Zellen kann in jeder Signalstärkenkarte (diese stammen aus der Netzplanung

und werden vom Mobilfunkbetreiber mit Hilfe von Ausbreitungsmodellen und digitalen Oberflächenmodellen für jede Antenne erstellt) eine räumliche Position abgeleitet werden. Durch die Minimierung der Verbesserungsquadratsumme der räumlichen Zuordnungen in den verschiedenen Karten kann die wahrscheinlichste Position der Mobilstation (MS) berechnet werden. Aus der daraus entstehenden Positionsfolge eines Teilnehmers, die zunächst auf Ausreißer untersucht und geglättet wird, kann schließlich mit Hilfe von Map-Aiding-Verfahren die Ermittlung einer FPD-Trajektorie im digitalen Straßennetz erfolgen. Eine detaillierte Beschreibung der beiden Verfahren ist in Ramm und Schwieger [2008] sowie in Do-iT [2008b] und Do-iT [2009b] zu finden.

4.2.2 Modellierung der Querabweichung

Um die Modellierung von Qualität für die Generierung von FPD-Trajektorien zu zeigen, wird aus mehreren Gründen die A-bis-Datenauswertung, die aufgrund der Datenverfügbarkeit lediglich im Raum Karlsruhe stattfinden konnte, herangezogen. Die Trajektoriengenerierung im Projekt wurde verstärkt auf A-bis-Ebene vorangetrieben und in mehreren Iterationsstufen bis zur Demonstratorphase am Projektende weiterentwickelt. Dies war insbesondere darin begründet, dass bis etwa ein Jahr vor Projektende die Bereitstellung der A-Daten in Echtzeit von Seiten des Netzbetreibers noch nicht gewährleistet werden konnte. Daher war aus Zeitgründen nach der Entwicklung der ersten Version des Algorithmus zur Auswertung der A-Daten, trotz des großen, identifizierten Potenzials, keine weitere Optimierung des Verfahrens mehr möglich. Des Weiteren standen für die Beurteilung der generierten Trajektorien aus A-Daten noch keine geeigneten Parameter zur Verfügung, die während der Demonstratorphase aufgezeichnet werden konnten. Somit besteht nicht die Möglichkeit, ein KNN zur Abbildung eines Qualitätszusammenhangs für die auf A-Daten basierende Auswertung zu trainieren.

Für die umfassende Beschreibung der Qualität von FPD-Trajektorien wurde ein Qualitätsmodell aufgestellt, wie es in Tabelle 5.1 auszugsweise dargestellt ist. Es wurde eine Reihe von Qualitätsparametern definiert, die während der Datenprozessierung oder bei der sich anschließenden Evaluierung der Trajektorien von Bedeutung waren. Unter anderem wurde zur internen Beurteilung des Algorithmus, basierend auf den A-bis-Daten, der interne Genauigkeitsparameter *Querabweichung* als Maß für die mittlere, orthogonale Abweichung einer Positionsfolge von der wahrscheinlichsten Route des Teilnehmers definiert. Der Parameter alleine sagt wenig über die Korrektheit der ermittelten Trajektorie aus, war jedoch sehr hilfreich bei der Optimierung des Auswertealgorithmus. Eine Verkleinerung der Querabweichung ist ein Indiz dafür, dass auch die Korrektheit der Trajektorie steigt, auch wenn es aufgrund des komplexen Zusammenspiels vieler Faktoren dafür keinen klaren formelmäßig darstellbaren Zusammenhang gibt. So konnte empirisch bestätigt werden, dass eine Querabweichung $< 250\,\mathrm{m}$ in Verbindung mit einer Trajektorienvollständigkeit (diese beschreibt den Anteil der berechneten Einzelpositionen eines Mobilfunkteilnehmers, die für die Trajektorienberechnung genutzt werden konnten) von $> 97\,\%$ zu einer Trajektorienkorrektheit vom Typ B (korrekt zugeordneter Anteil bezogen auf die gesamte berechnete Trajektorie) von über $70\,\%$ führt (vgl. Do-iT [2009b]). Eine externe Überprüfung der Korrektheit der Trajektorien anhand der zwei Korrektheitsparameter, der Zuordnungskorrektheit Typ A und Typ B (vgl. Tabelle 5.1) konnte jedoch nur in einigen wenigen Messfahrten erfolgen, bei denen während der Telefonate parallel GPS-Daten aufgezeichnet wurden. Aufgrund der relativ hohen Anforderungen an die Rohdaten (vgl. Do-iT [2009b]) wie

- eine Mindestgesprächsdauer von $1\,\mathrm{min}$,
- ein minimaler Abstand der ersten und letzten Position einer Folge von $2\,\mathrm{km}$
- eine Positionsfolge muss aus mindestens 15 Einzelpositionen bestehen und
- die Geschwindigkeit um die letzte Position zu erreichen muss $< 200\,\mathrm{km/h}$ sein,

und Schwierigkeiten bei der Zuordnung der Teilnehmer sowie einiger technischer Schwierigkeiten bei der Identifizierung der Testfahrten in den anonymisierten Rohdaten, standen bei Projektende insgesamt lediglich elf Trajektorien aus Testfahrten im Großraum Karlsruhe für die Evaluierung zur Verfügung (Do-iT [2008b]). Diese Testfahrten stellen zudem keine repräsentative Datengrundlage dar, da hauptsächlich spezielle Szenarien und Grenzfälle für die Untersuchungen gewählt wurden.

Die Abbildung 4.14 gibt einen Überblick über das Lokalisierungsnetz im Großraum Karlsruhe, in dem den telefonierenden Straßenverkehrsteilnehmern die wahrscheinlichsten Trajektorien zugeordnet werden müssen. Es ist deutlich der sehr dicht bebaute Innenstadtbereiche in Karlsruhe mit dem sich nach Norden öffnenden Schlossgarten, in dem kaum öffentlichen Straßen vorhanden sind, zu erkennen (Nordwestliches Viertel in der Karte). Von Norden nach Südwesten ist deutlich die Autobahn A5 und von Osten kommend die A8 erkennbar. Südlich von Karlsruhe in dem kleineren, rot umrandeten LAC liegt die Stadt Ettlingen. Aus technischen Gründen können Teilnehmer, die in die LACs Karlsruhe oder Ettlingen einfahren, erst nach einer Übergangszeit von bis zu wenigen Minuten das erste Mal geortet werden. Dies hängt mit der teilweise massiven Überlappung der Funkzellen und damit auch LACs zusammen, die den gesamten Datenverkehr abwickeln. Die bedienende Funkzelle wird zur Minimierung der Wechsel so lange wie möglich beibehalten. Daher kann ein Teilnehmer, der sich bereits einige Zeit innerhalb der nächsten LAC bewegt, immer noch über die alte LAC mit dem Mobilfunknetz verbunden sein. Seine A-bis-Daten stehen daher erst nach einiger Zeit, genauer: nach dem technischen Wechsel in die überwachte LAC zur Mobilfunkortung zur Verfügung. Auf den Autobahnen kann dies dazu führen, dass ein Fahrzeug erst nach mehreren Kilometern Fahrstrecke innerhalb der LAC lokalisiert werden kann.

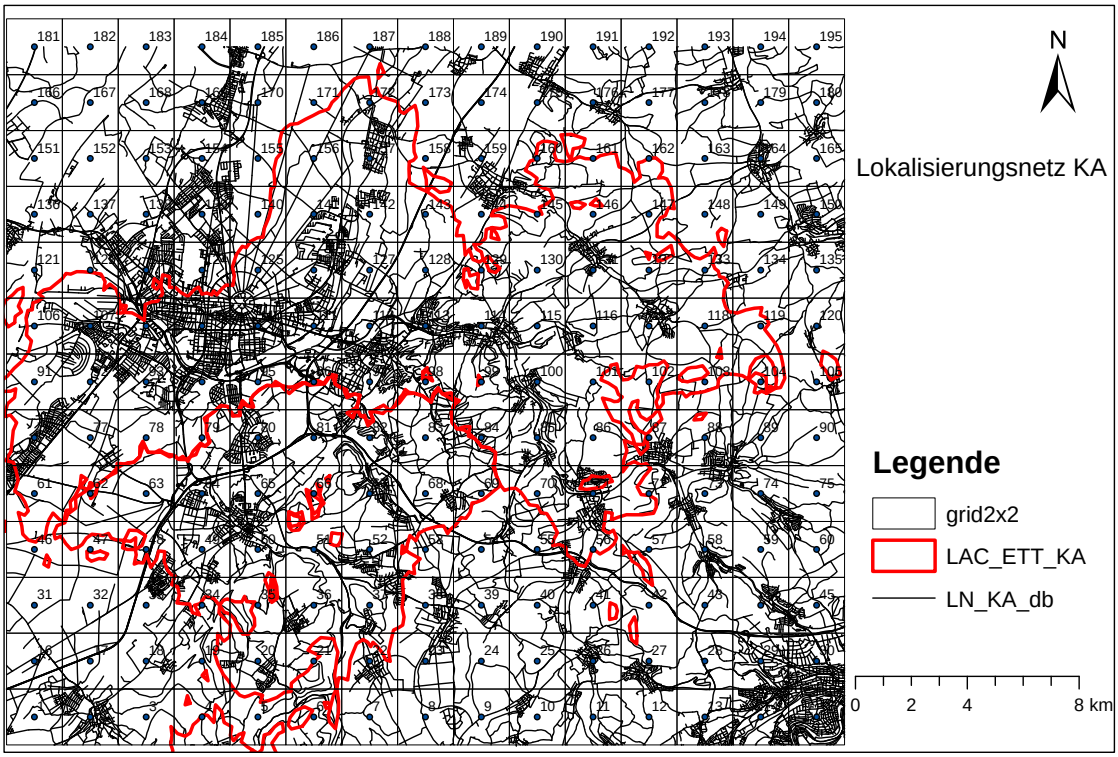


Abbildung 4.14: Digitales Straßennetz im Raum Karlsruhe mit den rot markierten LACs Karlsruhe und Ettlingen in denen Abis-Daten zur Verfügung standen

Für die Untersuchung der Modellierbarkeit der Querabweichung stehen eine Reihe von Eingangsgrößen zur Verfügung, die einerseits die Infrastruktur zum Zeitpunkt der Trajektorienberechnung widerspie-

geln und andererseits direkt aus den Berechnungen stammen. Es wurden aus der Projekterfahrung heraus die folgenden fünf Größen identifiziert, die mit hoher Wahrscheinlichkeit einen signifikanten Einfluss auf die Qualität der FPD-Trajektorien haben und aus dem Archiv derzeit auch noch zur Verfügung stehen. Dies sind

- die Trajektorienlänge in [m],
- die Trajektoriendauer in [s],
- die Mittlere Antennendichte in $[\mathrm{Ant./km^2}]$,
- die Mittlere Straßendichte $[\mathrm{RE}^4/\mathrm{km^2}]$ sowie
- die Anzahl der RE aus der die Trajektorie besteht.

Bei den Dichtewerten handelt es sich um externe Informationen, die als Eingangsdaten zur Verfügung stehen. Die weiteren Einflussparameter Trajektorienlänge, -dauer sowie die Anzahl der RE in der Trajektorie wurden mit dem Algorithmus aus den Mobilfunkrohdaten berechnet, es handelt sich daher um Ausgangsparameter des Prozesses. Damit kann die Modellierung der Qualitätszusammenhänge mit KNN unmittelbar der Trajektoriengenerierung nachgeschaltet erfolgen. Ist das für die Aufgabe geeignete KNN ausreichend trainiert, so können die Qualitätsinformationen auch nahezu in Echtzeit bereitgestellt werden. Damit ist eine Online-Überwachung des Prozesses mit Hilfe des Qualitätsparameters *Querabweichung* möglich.

Wie die Untersuchungen im ersten Anwendungsbeispiel gezeigt haben, ist eine Normierung der Eingangsgrößen vorteilhaft und verbessert die Resultate wesentlich (vgl. Kapitel 4.1). Daher erfolgt auch hier eine Normierung der Eingangsgrößen, bevor die Daten dem KNN zur Verfügung gestellt werden.

Tabelle 4.10: Übersicht über die zur Verfügung stehenden A-bis-Trajektorien mit einer Querabweichung kleiner $250\,\mathrm{m}$ aus dem Evaluierungszeitraum von Do-iT (23.-31. März 2009)

Tag (24 h)	Anzahl Trajektorien	Mittelwert der			
		Querabw. [m]	Länge [m]	Dauer [s]	Anzahl RE
Mo. 23.03.09	582	186	6455	240	39
Di. 24.03.09	527	186	6616	258	40
Mi. 25.03.09	628	180	6592	258	39
Do. 26.03.09	508	186	6770	288	41
Fr. 27.03.09	578	184	6290	243	38
Sa. 28.03.09	593	177	5233	185	34
So. 29.03.09	597	176	5330	174	33
Mo. 30.03.09	731	182	6170	220	38
Di. 31.03.09	699	184	6464	252	39

Die Beschreibung eines Zusammenhangs der genannten Einflussgrößen mit den Korrektheitsparametern vom Typ A und Typ B ist aufgrund fehlender Beispiele bzw. einer viel zu geringen Zahl an Beispielen

[4] RE steht für *en.*: Road Element und bezeichnet ein Straßenelement zwischen zwei Knotenpunkten in der digitalen Straßenkarte.

nicht möglich. Jedoch steht für alle im Evaluierungszeitraum berechneten Trajektorien die Querabweichung als Ausgangsparameter zur Verfügung. Daher wird im Folgenden untersucht, in welchem Maße die Querabweichung aus den obigen Einflussgrößen durch ein KNN prognostiziert werden kann. Insgesamt konnte auf die Daten von neun Tagen zurückgegriffen werden, die unmittelbar vor Projektende vom 23. bis zum 31. März 2009 aufgezeichnet und prozessiert wurden. Diese Daten stehen in Form von ASCII-files im Archiv zur Verfügung und wurden zum Trainieren der KNN herangezogen. Die Auswertung erfolgt tageweise (24 h, beginnend morgens um 3 Uhr) und es wurden jeweils 50 Trajektorien für die unabhängige Evaluierung der Matlab-Ergebnisse zurückgehalten. Die Tabelle 4.10 gibt einen Überblick über die an den unterschiedlichen Tagen zur Verfügung stehenden Daten. Dabei wurden bereits einzelne Trajektorien gelöscht, die offensichtlich fehlerhaft (Trajektoriendauer von $0\,\mathrm{s}$) oder zu kurz (nur aus 1 RE bestehend) waren oder doppelt vorkamen. Nach den ersten Erfahrungen wurden zusätzlich nur Trajektorien verwendet, die eine berechnete Querabweichung unter $250m$ aufwiesen. Größere Querabweichungen deuten auf eine nicht korrekt zugeordnete Trajektorien hin und werden daher von der Untersuchung ausgeschlossen (Do-iT [2009b]). Dieser Eindruck wird durch die in der letzten Zeile der Tabelle 4.12 dargestellten Ergebnisse des Trainings eines gut für die Aufgabe geeigneten KNN mit allen Trajektorien eines Tages bestätigt. Die Testergebnisse sind um Faktor $2-3$ schlechter als bei Beschränkung auf Trajektorien mit einer Querabweichung $< 250\,\mathrm{m}$.

Die Tabelle 4.10 zeigt eine Zusammenstellung aller berechneten Trajektorien aus dem Evaluierungszeitraum, die für die weiteren Untersuchungen in Frage kommen. Dabei beinhaltet die gesamte Anzahl an Trajektorien mit Querabweichungen unter $250\,\mathrm{m}$ auch jeweils die 50 Testbeispiele, mit denen die Netzgüte bei der tageweisen Untersuchung jeweils unabhängig geprüft wurde. Insgesamt ergibt sich ein sehr homogenes Bild über den untersuchten Zeitraum Ende März 2009. Die mittleren Querabweichungen schwanken an den Wochentagen nur um maximal $6\,\mathrm{m}$ oder ca. $3\,\%$, die Trajektorienlänge um $200\,\mathrm{m}$ bzw. ebenfalls $3\,\%$ und die mittlere Anzahl RE je Trajektorie um $3\,\mathrm{RE}$ bzw. weniger als $8\,\%$. Die Trajektoriendauer schwankt dagegen um bis zu $68\,\mathrm{s}$ oder ca. $30\,\%$, was maßgeblich auf zwei Tage zurückzuführen ist, an denen durchschnittlich kürzere (Mo. 30.03.09) bzw. längere Trajektorien (Do. 26.03.09) berechnet wurden. Der Auswertezeitraum ist allerdings zu kurz, um wöchentlich wiederkehrende Änderungen des Telefonierverhaltens der Verkehrsteilnehmer aufdecken zu können, daher kann über die Ursachen für diese Tatsachen nur spekuliert werden. Es fällt weiter auf, dass am Wochenende kürzere Telefongespräche im fließenden Verkehr geführt werden als an den Wochentagen und damit auch die mittlere Anzahl der RE je Trajektorie um 4 bis 8 und die Trajektorienlänge um teilweise über $1000\,\mathrm{m}$ sinkt. Damit lässt sich auch die kleinere Querabweichung an diesen beiden Tagen begründen. Tendenziell kann bei etwas kürzeren Punktfolgen besser zwischen richtiger und falscher Zuordnung unterschieden werden, somit ergibt sich eine eher kleinere Querabweichung bei den als korrekt beurteilten Trajektorien und sehr große Querabweichungen bei nicht korrekt zugeordneten Trajektorien (natürlich gilt dies nicht für sehr kurze Trajektorien, die generell schwer im digitalen Straßennetz zuzuordnen sind). Bei längeren Trajektorien hingegen „verschmieren“einzelne, falsch zugeordnete Abschnitte der berechneten Trajektorie die Querabweichung, so dass hier schlechter anhand der Querabweichung der gesamten Trajektorie zwischen 'richtig' und 'falsch' zugeordnet, unterschieden werden kann.

Die Modellierung der Datenqualität bei der Generierung von FPD-Trajektorien aus Abis-Daten beschränkt sich aus den bereits genannten Gründen auf den Ausgangsparameter *Querabweichung*. Die Antennen- und RE-Dichten als Eingangsgrößen werden jedoch noch differenzierter betrachtet. Im ersten Schritt wurde die Dichte der Antennenstandorte und die Straßendichte nur für Gitterquadrate mit einer Seitenlänge von $5\,\mathrm{km}$ berechnet. Die Dichte des dem ersten RE der Trajektorie am nächsten liegende Quadrat wurde für die gesamte Trajektorie übernommen. Diese Gittergröße und die Annahme nur einer Dichte über die gesamte Trajektorie von durchschnittlich $6\,\mathrm{km}$ hat sich jedoch als zu ungenau herausgestellt. Stattdessen wurden die Dichtewerte für ein quadratisches Raster von $2\,\mathrm{km}$ Seitenlänge berechnet, wie es in den Abbildungen 4.12 und 4.14 zu sehen ist. Die Dichte je Quadrat ist durch Auszählen der

Tabelle 4.11: Eingangs- und Ausgangsparameter für die Modellierung der Querabweichung mit KNN

Eingang			Ausgang		
QM	Parameter		QM	Parameter	
KR	$\mathrm{L_{FPD}}$	Trajektorienlänge [m]	GE	$\mathrm{QA_{FPD}}$	Querabweichung
KR	$\Delta\mathrm{T_{FPD}}$	Trajektoriendauer [s]			der Trajektorie
VE	$\Delta\rho_{\mathrm{Ant.}}$	Differenz der Antennendichten $[\mathrm{Ant./km^2}]$			
VE	$\bar{\rho}_{\mathrm{Ant.}}$	mittlere Antennendichte $[\mathrm{Ant./km^2}]$			
VE	$\Delta_{\rho_{\mathrm{RE}}}$	Differenz der RE-Dichten $[\mathrm{RE/km^2}]$			
VE	$\bar{\rho}_{\mathrm{RE}}$	mittlere RE-Dichte $[\mathrm{RE/km^2}]$			
-	$\mathrm{Anz_{RE}}$	Anzahl RE in der Trajektorie			
$n = 7$ Eingangsparameter			$m = 1$ Ausgangsparameter		

innenliegenden und angeschnittenen Zellen bzw. RE erfolgt. Über den metrischen Abstand der RE zum nächsten Gitterquadrat wurde jedem RE, welches Anfang oder Ende einer Trajektorie bildet, je ein Dichtewert für die Antennen- und die RE-Dichte zugewiesen. Aus diesen zugeordneten Dichtewerten wurden mit der Differenz und dem Mittelwert zwei neue Parameter berechnet. Es ergaben sich für den Einsatz der KNN daher die in der Tabelle 4.11 dargestellten Ein- und Ausgangsgrößen.

Im Anschluss an die Untersuchungen einzelner Netze und einzelner Tage mit den Dichtedifferenzen und -mittelwerten zur Berücksichtigung der Infrastrukturdichten wurde auch die Berechnung der Mittelwerte und Dichten aus den Maximal- und Minimalwerten für jede einzelne Trajektorie getestet. Der Verdacht lag nahe, dass diese Werte die Dichteverhältnisse entlang der Trajektorien besser repräsentieren, als die Dichten am ersten und letzten RE der Trajektorien. Dazu wurden die Trajektorien mit einer Querabweichung $< 250\,\mathrm{m}$ vom 26.03.09 herangezogen und die Ergebnisse, die sich für die am Besten geeignete Netzvariante [9-9-1] ergaben, wurden in der Tabelle 4.12 zusätzlich gegenübergestellt. Wie die Ergebnisse exemplarisch zeigen, liefert die Berechnung der Parameter aus dem ersten und letzten RE jedoch kleinere Querabweichungen. Daher wurde diese Methode weiter verfolgt.

Die aus der Infrastruktur abgeleiteten Parameter, welche die Antennen- und Straßennetzdichte beschreiben, werden im Qualitätsmodell für FPD-Trajektorien dem Qualitätsmerkmal Verfügbarkeit zugeordnet. Bei Trajektorienlänge und Trajektoriendauer hingegen handelt es sich um Parameter, die die Korrektheit der Trajektorien beschreiben. Die Anzahl RE in der berechneten Trajektorie ist abhängig von der RE-Dichte, der Straßenkategorie auf der sich der Teilnehmer bewegt hat sowie der Länge der Trajektorie. Daher ist eine klare Zuordnung dieses Parameters zu einem der sieben Qualitätsmerkmale nicht möglich. Dennoch ist zu erwarten, dass der Parameter in Verbindung mit den weiteren Faktoren einen signifikanten Einfluss auf die zu erwartende Querabweichung hat.

Um den Einfluss der einzelnen ausgewählten Eingangsparameter auf die Querabweichung vor der Implementierung in dem KNN grundsätzlich qualitativ nachweisen zu können, wurden die Einflussgrößen für jeden Tag gegen die Querabweichung geplottet. Die Plots der verschiedenen Tage unterscheiden sich

nur wenig, was nach der Gegenüberstellung der einzelnen Daten in Tabelle 4.10 auch nicht zu erwarten war. In den Abbildungen 4.15 sind exemplarisch die Anzahl der RE, aus denen sich die Trajektorien zusammensetzen (linke Grafik) und die metrische Trajektorienlänge (rechte Grafik) im Verhältnis zur Querabweichung für Donnerstag, den 26. März 2009 dargestellt.

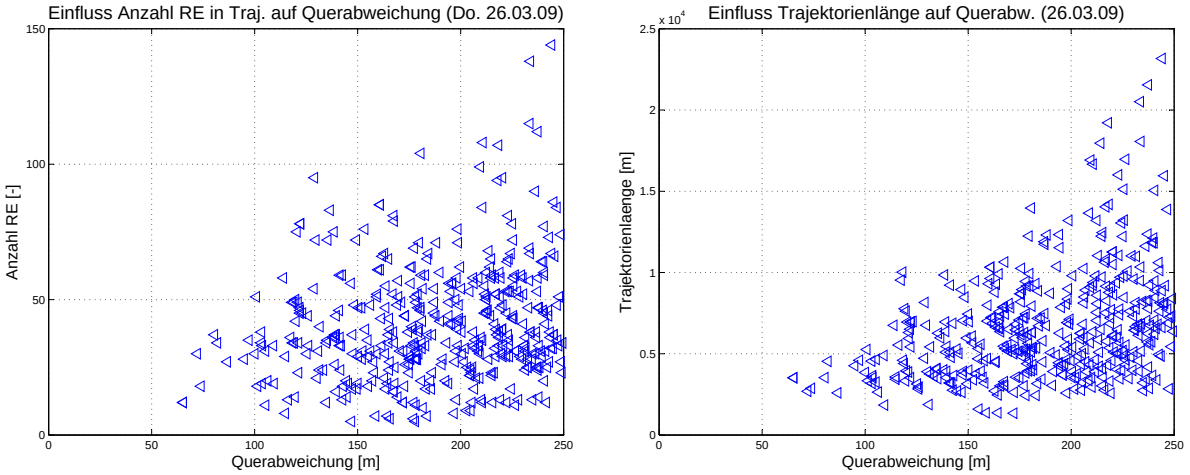


Abbildung 4.15: Einfluss der Anzahl RE je Trajektorie (links) sowie der Trajektorienlänge (rechts) auf die Querabweichung (Do. 26.03.09)

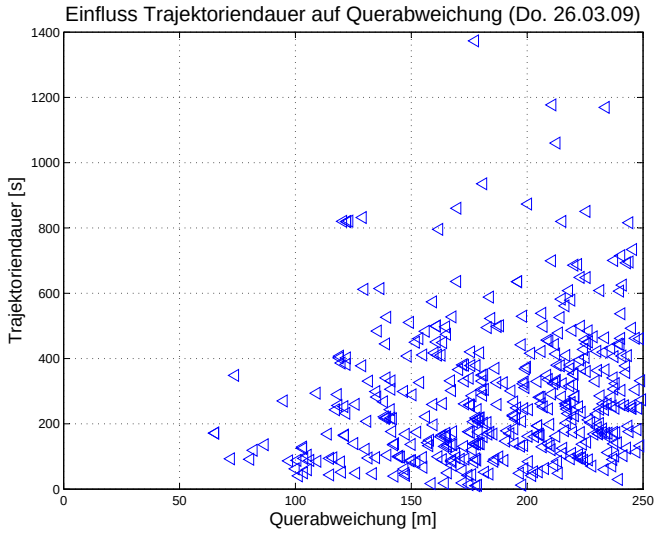


Abbildung 4.16: Einfluss der Trajektoriendauer auf die Querabweichung (Do. 26.03.09)

Die Trajektorienlänge und die Anzahl der RE hängen eng miteinander zusammen, je mehr RE eine Trajektorie enthält, desto länger wird sie in der Regel. Allerdings hängt die durchschnittliche Länge der RE von der Dichte der Straßeninfrastruktur ab. Je dichter das Straßennetz, desto kürzer sind die RE als Verbindungen zwischen Knotenpunkten. Wie der Abbildung 4.14 zu entnehmen ist, schwankt die Straßendichte im Gebiet, in dem Abis-Daten zur Verfügung standen, stark. Im südlichen und östlichen Bereich gibt es einige Gebiete mit dünner Infrastruktur, wohingegen im zentralen und nordwestlichen Bereich der beiden LACs ein sehr dichtes Straßennetz vorhanden ist. Daher ist es eher unerwartet, dass

beide Grafiken in Abbildung 4.15 dennoch eine sehr ähnliche Abhängigkeit für die Querabweichung zeigen. Die Querabweichung steigt tendenziell mit steigender Anzahl RE und steigender Trajektorienlänge an, daher ist eine Berücksichtigung beider Parameter im KNN sinnvoll.

Im Falle der Trajektoriendauer lässt sich ein ähnlicher Zusammenhang erahnen, wie Abbildung 4.16 repräsentativ für alle acht Evaluierungstage zeigt. Allerdings ist der Zusammenhang zwischen zeitlich längeren Trajektorien und einer Zunahme der Querabweichung wesentlich schwächer ausgeprägt. Dies liegt an der Tatsache, dass eine länger andauernde Trajektorie sowohl metrisch sehr kurz sein kann, wenn der Teilnehmer im Stadtverkehr langsam unterwegs ist, als auch sehr lang, wenn das Fahrzeug mit hoher Geschwindigkeit auf der Autobahn Karlsruhe tangiert hat. Da sich beide Effekte überlagern, lässt sich nicht direkt eine eindeutige Abhängigkeit zwischen der Dauer und der Querabweichung einer Trajektorie ableiten.

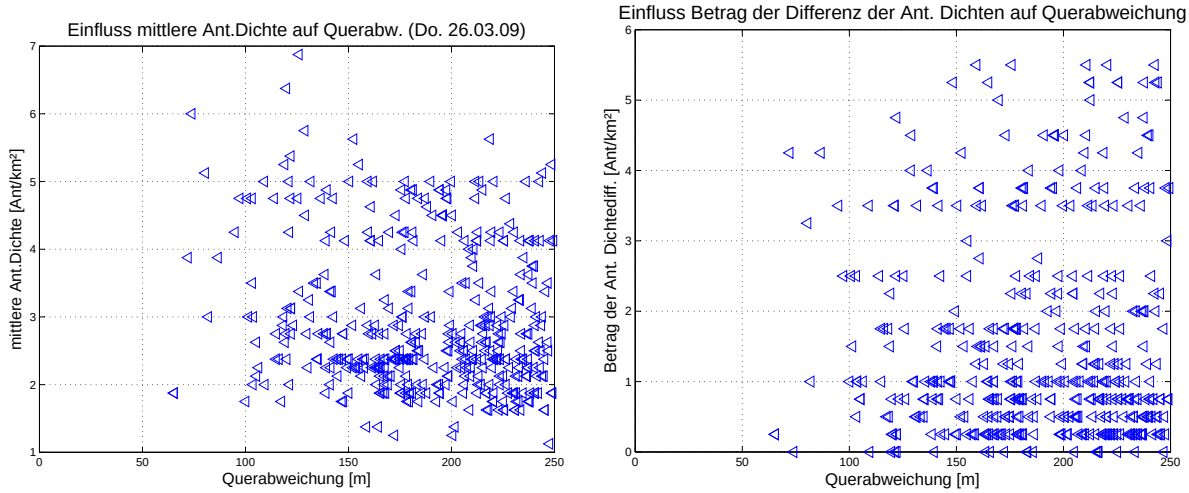


Abbildung 4.17: Einfluss der mittleren Antennendichte (links) und des Betrags der Differenz der Antennendichten (rechts) aus Ende - Anfang der Trajektorien auf die Querabweichung (Do. 26.03.09)

Abbildung 4.17 zeigt den Einfluss der Antennendichte auf die Querabweichung. Zum einen wurde aus den zugeordneten Dichten der ersten und letzten Road- Elemente jeder Trajektorie der Mittelwert gebildet und den Trajektorien zugeordnet. Zum anderen wurde aber auch der Einfluss der Dichteänderung entlang der Trajektorie vereinfacht durch die Differenz der zugeordneten Dichte am Ende und am Anfang jeder Trajektorie berücksichtigt. Dieser mögliche Einfluss würde durch die alleinige Berücksichtigung des Mittelwertes verloren gehen. Die linke Grafik in der Abbildung 4.17 lässt einen Zusammenhang zwischen höheren Antennendichten und kleineren Querabweichungen erahnen, allerdings ist der Einfluss nur sehr schwach ausgeprägt. Es überlagern sich zwei Einflüsse: Eine höhere Antennendichte ermöglicht eine bessere räumliche Zuordnung, allerdings wird die Antennendichte insbesondere dort erhöht, wo auch das Telefonieraufkommen und damit meist auch die Straßendichte höher ist. Dies wiederum führt zu einer schwierigeren Identifikation der korrekten Trajektorien und damit zu größeren mittleren Querabweichungen. Die rechte Grafik mit den Beträgen der Dichtedifferenzen zeigt eine Datenlücke bei einem Wert von 3 für den Betrag der Antennendichten. Diese ist aufgrund der Art der Dichteberechnung aus den Informationen des Netzbetreibers rein zufälliger Natur. In der rechten Grafik ist des Weiteren zu erkennen, dass die Dichtedifferenzen zwischen Ende und Anfang der Trajektorien meist kleiner als $\pm 2\,\mathrm{Ant./km^2}$ ist. Das heißt, die Antennendichte ist in dem untersuchten Gebiet relativ homogen bzw. die Trajektorien laufen meist durch Gebiete mit ähnlicher Antennendichte.

Da die Lage der Trajektorien im Mobilfunknetz zufällig ist, wurde zunächst keine Abhängigkeit der Querabweichung vom Vorzeichen der berechneten Dichtedifferenz erwartet. Die linke Grafik in Abbildung 4.18 bestätigt dies jedoch nicht. Es ist zwar eine gewisse Symmetrie zu einer horizontalen Spiegelachse erkennbar, allerdings liegt diese unter der 0-Achse, bei der die Differenz 0 beträgt. Die Grafiken der Dichtedifferenz an allen neun Tagen bestätigen dieses Ungleichgewicht zwischen positiven und negativen Dichtedifferenzen. Über den gesamten Zeitraum schwankt der Anteil positiver Differenzen, d. h. mit höherer Antennendichte am Ende der Trajektorie, zwischen $35\,\%$ und $38\,\%$ und beträgt im Mittel nur $36\,\%$ der Antennendifferenzen, die $\neq 0$ sind. Dieser Umstand ist in der Auswertestrategie begründet: Die Einzelpositionen eines Teilnehmers lassen sich in Bereichen mit kleineren Funkzellen systembedingt tendenziell genauer bestimmen. Da nur Positionen mit einer gewissen Mindestgenauigkeit der weiteren Auswertung zugeführt werden, überwiegt der Anteil an Trajektorien, die von Gebieten mit dichterer Antenneninfrastruktur in weniger dichte verlaufen. Der Zusammenhang zwischen der Antennendichte und der Querabweichung der Trajektorien muss somit vorzeichenrichtig berücksichtigt werden.

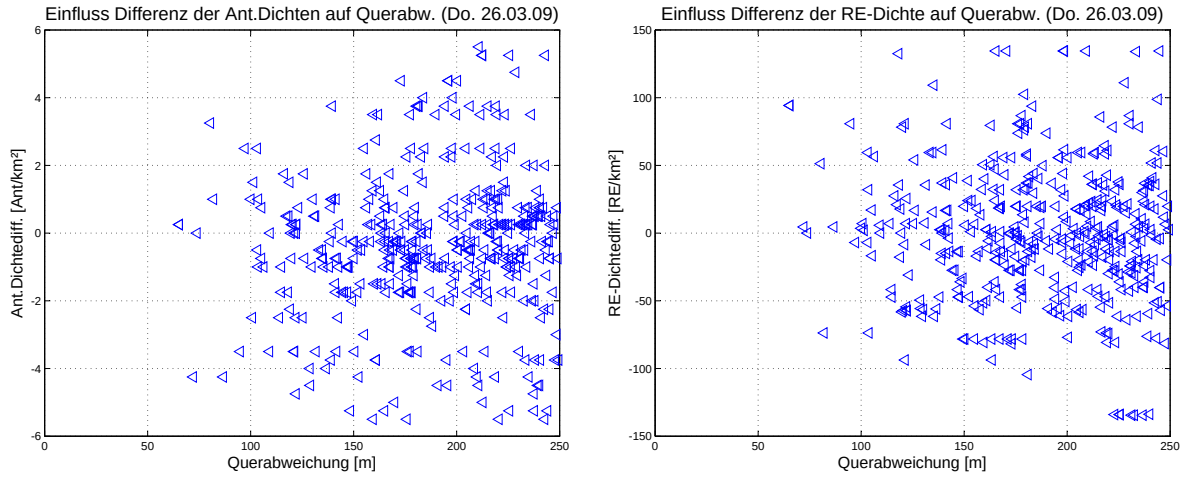


Abbildung 4.18: Einfluss der vorzeichenrichtig dargestellten Dichtedifferenzen aus Ende - Anfang der Trajektorien für die Antennendichte (links) bzw. die RE-Dichte (rechts) auf die Querabweichung (Do. 26.03.09)

Die Straßendichte im Bereich Karlsruhe liegt meist unter $50\,\mathrm{RE/km^2}$, wie in der linken Grafik der Abbildung 4.19 zu sehen ist. Die Abhängigkeit der Querabweichung von der Dichte des Straßennetzes ist nur schwach ausgeprägt. Eine Verschlechterung der Querabweichung bei steigender RE-Dichte ist erkennbar. Dies liegt an dem Umstand, dass sich hier ebenfalls zwei Effekte überlagern. Einerseits ist bei kürzeren Road-Elementen grundsätzlich eine genauere Zuordnung zwischen detektierter Position und RE möglich, andererseits liegt in diesen Bereichen allerdings auch eine hohe RE-Dichte vor, so dass die korrekte Zuordnung sehr schwer wird. Diese wird bei einem weitmaschigen Straßennetz im Außenbereich von Karlsruhe erleichtert, dort sind die RE allerdings sehr lang (mehrere Kilometer sind keine Seltenheit), was die genaue koordinatenmäßige Zuordnung ebenfalls schwer macht. Damit ergeben sich sowohl innerstädtisch, als auch im Randbereich teilweise große Querabweichungen. Die Berücksichtigung der RE-Dichte macht dennoch Sinn, da dieser Einfluss in Kombination mit den anderen Einflüssen interpretiert werden muss. Dies leistet das KNN automatisch, da die Trainingsdatenmenge derartige Beispiele enthält.

Die Dichtedifferenz wird in die Modellierung des KNN ebenfalls einbezogen, um auch im Falle der RE-Dichten mögliche Abhängigkeiten, die sich durch die Änderung der Dichte innerhalb einzelner Trajektorien ergeben, abzubilden. In der rechten Grafik der Abbildung 4.18 ist keine ungleiche Verteilung

der positiven und negativen Dichtedifferenzen zu erkennen. Trotz kleinerer, zur horizontalen 0-Achse unsymmetrischen Cluster, ist keine systematische Verschiebung der Spiegelachse zu sehen. Dies belegt auch die Streuung des Anteils positiver Differenzen über den Untersuchungszeitraum, der sich zwischen $45\,\%$ und $55\,\%$ bewegt und im Mittel $49\,\%$ erreicht. Es existiert jedoch eine eher schwach ausgeprägte Abhängigkeit der Querabweichung von der RE-Dichtedifferenz (vgl. Abbildung 4.18). Trotz keiner signifikanten Abweichung von der Symmetrie wird auch hier die vorzeichenrichtige Differenz zum Trainieren des Netzes verwendet.

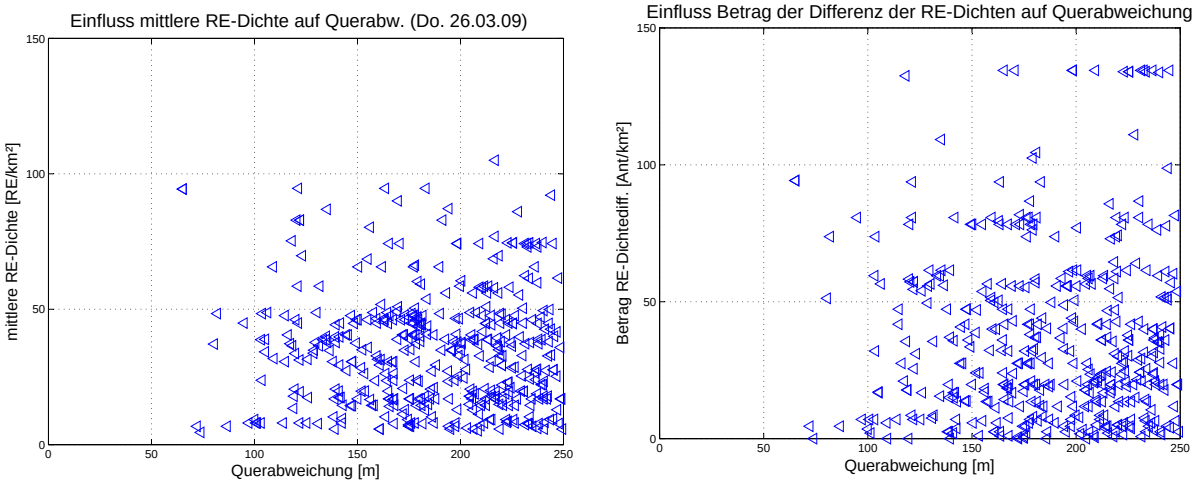


Abbildung 4.19: Einfluss der mittleren RE-Dichte (links) und des Betrags der Differenz (rechts) der RE-Dichten aus Ende - Anfang der Trajektorien auf die Querabweichung (Do. 26.03.09)

Für die Trajektorienvollständigkeit ($\mathrm{VO}_{\mathrm{FPD}}$ in Tabelle 5.1), die den prozentualen Anteil der zur Trajektorienberechnung verwendeten Positionen aus der gesamten Positionsfolge angibt, kann keine signifikante Abhängigkeit der Querabweichung festgestellt werden. Der Plot der Trajektorienvollständigkeit der Trajektorien eines Tages, zusammen mit den zugehörigen Querabweichungen, ergibt eine augenscheinlich zufällige Verteilung und zeigt keinerlei Trends. Daher wird diese Information nicht im Training der KNN berücksichtigt.

Für die Modellierung der Querabweichung aus den oben dargestellten sieben Einflussgrößen, die für die Evaluierungsphase von insgesamt neun Tagen Ende März 2009 zur Verfügung stehen, wurden wieder eine Reihe von möglichen Netzen getestet. Zur Orientierung wurden hier ebenfalls die Faustformeln von Otto und Kinnebrock herangezogen, nach denen im Falle einer verdeckten Neuronenschicht, diese aus $h = 2n + 1 = 15$ Neuronen bestehen soll (vgl. Formel 3.36). Laut Otto sind dazu etwa $Q = h \cdot 5 \cdot (n + m) = 600$ erforderlich. Wie in der Tabelle 4.10 zu sehen ist, steht diese Anzahl an Lernbeispielen nicht an jedem Tag zur Verfügung, insbesondere da für die unabhängige Überprüfung der Ergebnisse noch weitere 50 Trajektorien zurückgehalten werden. Dennoch wurden die einzelnen Tage zunächst getrennt behandelt und es wurden jeweils vier verschiedene Netzkonfigurationen getestet. Neben den beiden zweischichtigen Netzen mit 9 bzw. 15 verdeckten Neuronen wurden auch wieder zwei Netze mit zwei verdeckten Schichten und ebenfalls je 9 bzw. 15 verdeckten Neuronen untersucht. In der Tabelle 4.12 sind die Ergebnisse der vier Netzvarianten repräsentativ für alle neun Tage konsequenterweise erneut für den 26.03.2009 dargestellt. Neben den vier Netzvarianten wurde - wie bereits erwähnt - zusätzlich das [9-9-1]-Netz mit allen Trajektorien des Tages trainiert. Die um Faktor $2 - 3$ schlechteren Ergebnisse bestätigen nochmals das Vorgehen, alle Trajektorien mit einer berechneten Querabweichung $> 250\,\mathrm{m}$ nicht zu verwenden.

Tabelle 4.12: Variation der Netzdimension eines KNN zur Modellierung der Querabweichung (26.03.2009)

Netz-architektur	MSE	Maximale Abw. [m]	Standardabw. [m]	Mittelwert [m]
$9-1$	0.026	83.5	33.5	1.0
$15-1$	0.025	84.5	34.1	8.3
$9-9-1$	0.022	96.6	32.9	0.7
$15-15-1$	0.020	76.8	35.4	10.0
zum Vergleich: Dichtemittelwerte und -differenzen aus Extremalwerten				
$9-9-1$	0.021	83.0	40.4	6.0
zum Vergleich: Alle Trajektorien ohne Begrenzung auf $QA < 250\,\mathrm{m}$				
$9-9-1$	0.022	230.8	81.2	23.0

Die Abbildung 4.20 zeigt für einen Testdatensatz die Querabweichungen aller getesteten Netzvarianten übereinander geplottet. Zusätzlich wurden in der Grafik die Sollwerte für die Querabweichung gepunktet eingetragen. Diese Sollwerte stammen aus der regulären Datenprozessierung im Projekt Do-iT. Die 50 Trajektorien, die für den Test herangezogen wurden, stammen aus dem Tagesdatenbestand. Es handelt sich um die letzten 50 Trajektorien, die am 26.03.2009 berechnet wurden. Diese unterscheiden sich in Länge, Dauer oder in sonstigen Eigenschaften jedoch nicht von allen anderen Trajektorien, die an diesem Tag berechnet wurden und mit denen das Netz trainiert worden ist. Die resultierenden Fehler in der Vorhersage der Querabweichungen aller getesteten Netzvarianten sind in der Grafik 4.21 gegenübergestellt.

Trotz der größten maximalen Abweichung zu den Sollwerten für die Querabweichung von etwa $97\,\mathrm{m}$ (vgl. Tabelle 4.12), vermag das dreischichtige Netz mit je neun verdeckten Neuronen den Prozess am besten abzubilden (in den beiden Abbildung 4.20 und 4.21 jeweils gelb dargestellt). Die Standardabweichung, die sich aus den Verbesserungen der Querabweichung (SOLL-IST) ergibt, ist mit knapp $33\,\mathrm{m}$ am kleinsten und der Mittelwert aller 50 Abweichungen liegt nahe bei $0\,\mathrm{m}$. Die Netze haben alle innerhalb weniger Iterationen $(4-13)$ ihre bestmögliche Konfiguration erreicht. Der Abbruch wurde jeweils durch den *Validation Check* ausgelöst, d. h. die Ergebnisse haben sich nach weiteren Änderungen der Gewichte sechsmal hintereinander wieder verschlechtert, so dass bei weiteren Iterationen mit keiner weiteren Verbesserung mehr zu rechnen war.

Es ist zu erkennen, dass alle Netzkonfigurationen bei der Abbildung von extrem großen oder kleinen Querabweichungen Probleme haben, diese werden nur in wenigen Fällen korrekt abgebildet. Des Weiteren fällt auf, dass einige Trajektorien doppelt auftreten, wie z. B. die Trajektorien Nummer 1 und 2 oder 20 und 21. Grundsätzlich stellt dies im Datenbestand keine Inkonsistenz dar, da es prinzipiell möglich ist, dass zwei Personen im selben Fahrzeug gleichzeitig telefonieren, allerdings tritt dies hier ungewöhnlich häufig auf, so dass in diesem Fall eher ein Fehler in der Datenaufbereitung vermutet werden kann. Da diese doppelten Trajektorien prinzipiell nichts an der Aussagekraft der Untersuchungen ändern, werden diese jedoch nicht aussortiert. Zusätzlich lag der Fokus hier in der Nutzung des vorhandenen Datenmaterials als Beispiel zur Untersuchung der Anwendung von KNN zur Fortpflanzung von Datenqualität. Die Suche nach Fehlern im Datenmaterial und die Verbesserung der zugrunde liegenden Algorithmen war nicht die Aufgabe dieser Untersuchung.

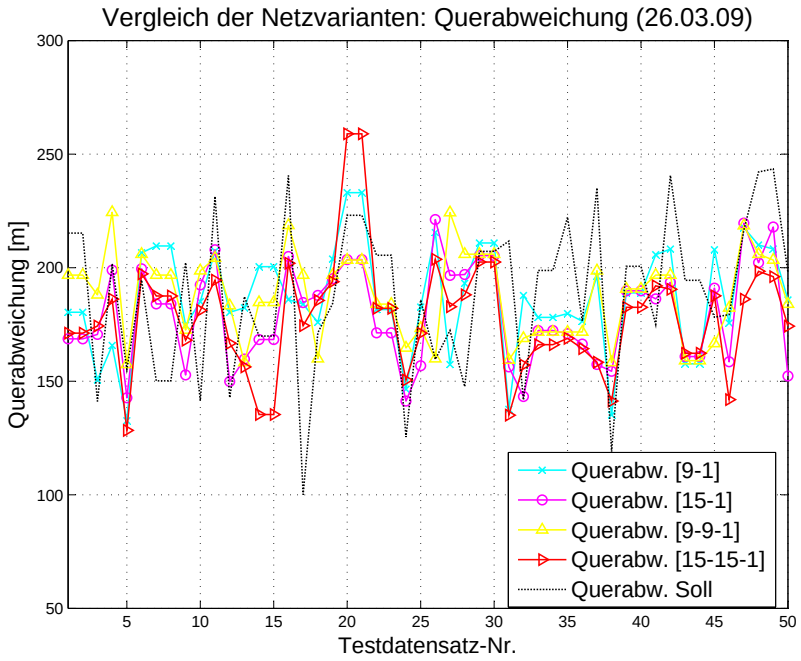


Abbildung 4.20: Modellierung der Querabweichung für die Testdaten vom Do. 26.03.09: Darstellung der berechneten Querabweichungen verschiedener Netzvarianten im Vergleich

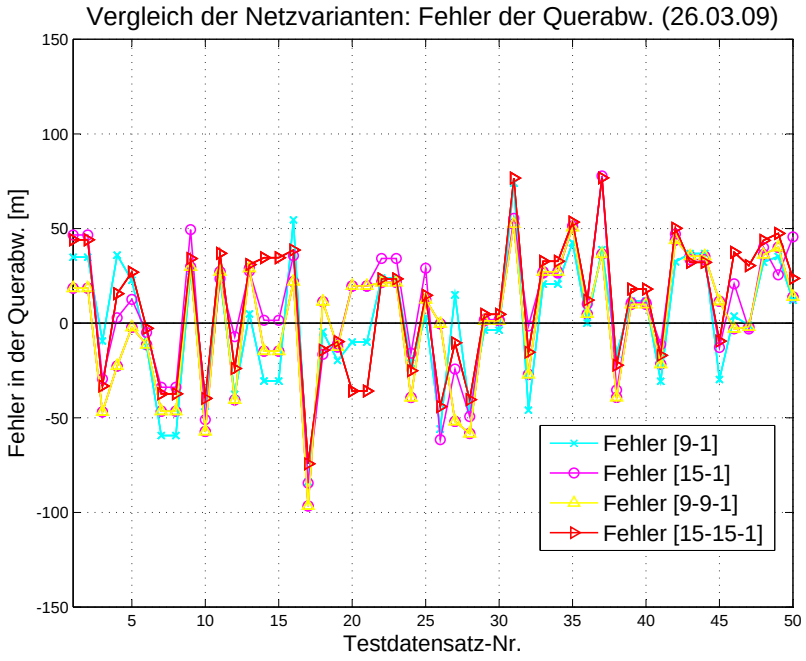


Abbildung 4.21: Modellierung der Querabweichung für die Testdaten vom Do. 26.03.09: Darstellung der Fehler in den Querabweichungen verschiedener Netzvarianten im Vergleich

Die Fehler in der Prognose der Querabweichungen bewegen sich weitgehend zwischen $\pm 50\,\mathrm{m}$, wobei die bereits erwähnten, schlecht abzubildenden Extremwerte der Querabweichung die Ausnahmen bilden.

Dies wird beim Vergleich der beiden Grafiken der Abbildung 4.20 und 4.21 deutlich. In der Abbildung 4.21 scheint im rechten Drittel eine ansteigende Tendenz der Fehler in der Querabweichung vorzuliegen. Wie der Vergleich aller neun Tage gezeigt hat, sind jedoch keine Tendenzen in den Abbildungen zu erkennen und aufgrund der Unabhängigkeit der verschiedenen berechneten Trajektorien auch nicht zu erwarten. Daher handelt es sich hier nur um eine zufällige Schwankung der letzten fünf Trajektorien mit dem selben Vorzeichen.

Eine Alternative zur Berechnung der Einflussgrößen, die die Infrastrukturdichte beschreiben, aus den zugehörigen Dichtewerten des jeweils ersten und letzten Straßenabschnitts der Trajektorien, wurde ebenfalls untersucht. Dabei wurden die Differenzen der Straßennetzdichte sowie der Antennendichte aus den jeweils entlang der Trajektorien auftretenden minimalen und maximalen Dichtewerten berechnet. Entsprechend wurden auch die mittleren Dichten aus diesen Extrema ermittelt. Trotz der scheinbar besseren Abbildung der Verhältnisse der Infrastruktur entlang der Trajektorie, bringt dieses Vorgehen keinen Vorteil. Wie die Gegenüberstellung für den 26.03.2009 in der Tabelle 4.12 zeigt, sinkt zwar die maximale Abweichung um fast $14\,\mathrm{m}$ im Vergleich zu der selben Netzkonfiguration, die mit den Dichten der Anfangs- und Endwerte der Trajektorien trainiert wurden, jedoch steigt die Standardabweichung um mehr als $7\,\mathrm{m}$ auf über $40\,\mathrm{m}$. Daher wird auch bei den nachfolgenden Untersuchungen weiterhin die Dichte der Anfangs- und Endstücke der Trajektorien verwendet.

Tabelle 4.13: Jeweils beste Netzkonfiguration aller Evaluierungstage zur Modellierung der Querabweichung mit KNN

Tag (24 h)	Netz-architektur	MSE	Maximale Abw. $[\mathrm{m}]$	Standardabw. $[\mathrm{m}]$
Mo. 23.03.09	$9-9-1$	0.024	102.5	43.4
Di. 24.03.09	$9-9-1$	0.024	77.8	39.1
Mi. 25.03.09	$9-9-1$	0.025	100.3	45.9
Do. 26.03.09	$9-9-1$	0.022	96.6	32.9
Fr. 27.03.09	$9-9-1$	0.021	75.6	40.3
Sa. 28.03.09	$9-9-1$	0.024	111.0	48.5
So. 29.03.09	$9-9-1$	0.029	104.5	37.4
Mo. 30.03.09	$9-9-1$	0.024	81.5	40.2
Di. 31.03.09	$15-15-1$	0.023	95.0	41.4
Mittelwerte		0.024	93.9	41.0

Insgesamt ergibt sich ein sehr homogenes Bild über alle Evaluierungstage und es hat sich bei den Untersuchungen -mit einer Ausnahme- immer die selbe Netzkonfiguration $[9-9-1]$ als Beste herausgestellt. Die Beurteilung der besten Konfiguration wurde dabei anhand der einfachen Standardabweichung, die als Wurzel der mittleren Verbesserungsquadratsumme über alle 50 Testbeispiele berechnet wurde, vorgenommen. Die in der vierten Spalte dargestellten Extremwerte wurden erst nachgeordnet beurteilt, da es sich meist lediglich um ein oder zwei Ausreißer handelt. Am letzten Tag, dem 31.03.09 war die Konfiguration $[15-15-1]$ noch besser zur Modellierung der Qualität geeignet. Insgesamt schwankt die Standardabweichung zwischen $33\,\mathrm{m}$ und $49\,\mathrm{m}$ und erreicht im Mittel etwa $41\,\mathrm{m}$. Der Betrag der maximalen Abweichung erreicht bis zu $111\,\mathrm{m}$. Damit kann die Reproduzierbarkeit der Methode in die-

sem Anwendungsbeispiel grundsätzlich bestätigt werden. Die Größenordnung der erzielten Ergebnisse erscheint plausibel, eine detailliertere Bewertung dieser findet sich im Abschnitt 4.2.3. Die erzielten Ergebnisse der jeweils besten Netzkonfiguration sind in der Tabelle 4.13 gegenübergestellt. Es fällt auf, dass die Daten vom 28. März, trotz kleinem MSE, aus nicht nachvollziehbarer Ursache wesentlich schlechter zum Trainieren der Netze geeignet sind, als die anderen Tage. Die maximale Abweichung sowie die Standardabweichung liegen mit $111\,\mathrm{m}$ und $48.5\,\mathrm{m}$ knapp $20\,\%$ über dem Durchschnitt. Daher wird dieser Tag bei der weiteren, gemeinsamen Auswertung des gesamten Evaluierungszeitraumes vernachlässigt.

Die Anzahl Iterationen bis zum Abbruch lagen in allen Fällen zwischen 3 und 15, was mit Rechenzeiten von durchweg unter $2\,\mathrm{s}$ verbunden war. Die mittlere Abweichung vom Sollwert jeweils aller 50 Lernbeispiele liegt an allen untersuchten Tagen für alle Netzkonfigurationen zwischen $+20\,\mathrm{m}$ und $-20\,\mathrm{m}$, was meist auf einzelne große Ausreißer zurückzuführen ist. Eine Tendenz zu stets zu groß oder zu klein vorhergesagten Querabweichungen ist nicht zu erkennen.

Im Anschluss wurden zwei Netze mit allen zur Verfügung stehenden Daten von acht Tagen untersucht. Aus den genannten Gründen wurden dabei die Daten vom 28.03.09 von der Auswertung ausgeschlossen. Aus der bisherigen Erfahrung heraus haben sich die Untersuchungen auf die beiden dreischichtigen Netzvarianten mit 9 bzw. 15 verdeckten Neuronen je Schicht beschränkt. Die Ergebnisse sind in der Tabelle 4.14 zusammengefasst.

Tabelle 4.14: Variation der Netzdimension eines KNN zur Modellierung der Querabweichung (gemeinsame Auswertung aller acht Tage)

Netz-architektur	MSE	Maximale Abw. [m]	Standardabw. [m]	Mittelwert [m]
$9-9-1$	0.025	103.8	41.0	3.6
$15-15-1$	0.025	108.8	38.2	0.4

Beide Netze sind in der Lage, die Querabweichung der Trajektorien mit der, aus den tageweise durchgeführten Auswertungen, zu erwartenden Genauigkeit zu prognostizieren. Das Netz mit 15 verdeckten Neuronen je Schicht kann die Querabweichungen noch etwas besser prognostizieren als das mit kleinerer Neuronenzahl. Auch der Mittelwert aller 50 Testbeispiele, die aus dem gesamten Eingangsdatenbereich heraus gewählt wurden[5], liegt nahe 0 und zeigt somit keine systematischen Abweichungen. Da der Tag, der sich am schlechtesten abbilden ließ, ausgeschlossen wurde, ist die Standardabweichung mit ca. $38\,\mathrm{m}$ (Variante $[15-15-1]$) kleiner als der Mittelwert aller neun Tage. Die maximale Abweichung liegt mit $109\,\mathrm{m}$ jedoch etwas darüber.

In den beiden Abbildungen 4.22 und 4.23 sind die Ergebnisse beider Netzvarianten nochmals grafisch gegenübergestellt. Die größten Fehler ergeben sich wiederum bei den extremen Querabweichungen, wobei hier die fünf sehr kleinen Querabweichungen in den Datensätzen Nr. 27, 31, 37, 40 und 49 auffallen. Beide Netzvarianten zeigen ein sehr ähnliches Verhalten und die Fehler in der prognostizierten Querabweichung verlaufen mit einigen Ausnahmen gleichmäßig. Die Amplituden der $[15-15-1]$ Netzvariante sind jedoch meist etwas kleiner, wie bereits die kleinere Standardabweichung in der Tabelle 4.14 gezeigt hat. Es fällt auf, dass die insgesamt sechs Ausreißer, die am schlechtesten von den Netzen abgebildet wer-

[5] Es wurden die Datensätze Nummer $1-10$, $1001-1010$, $2001-2010$, $3001-3010$ sowie $4001-4010$ gewählt um systematische Einflüsse ausschließen zu können.

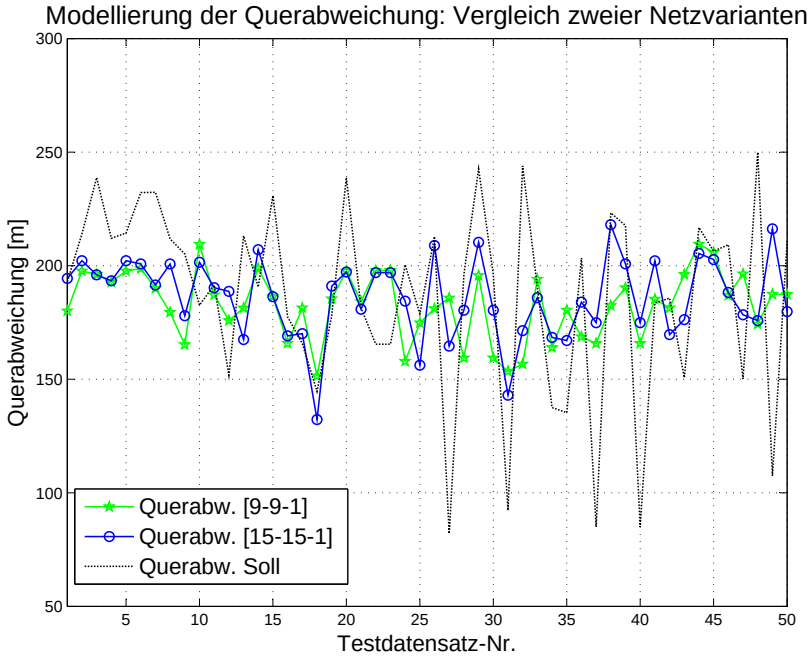


Abbildung 4.22: Modellierung der Querabweichung aus allen acht Tagen: Vergleich der prognostizierten Querabweichungen der zwei Netzvarianten [9-9-1] und [15-15-1]

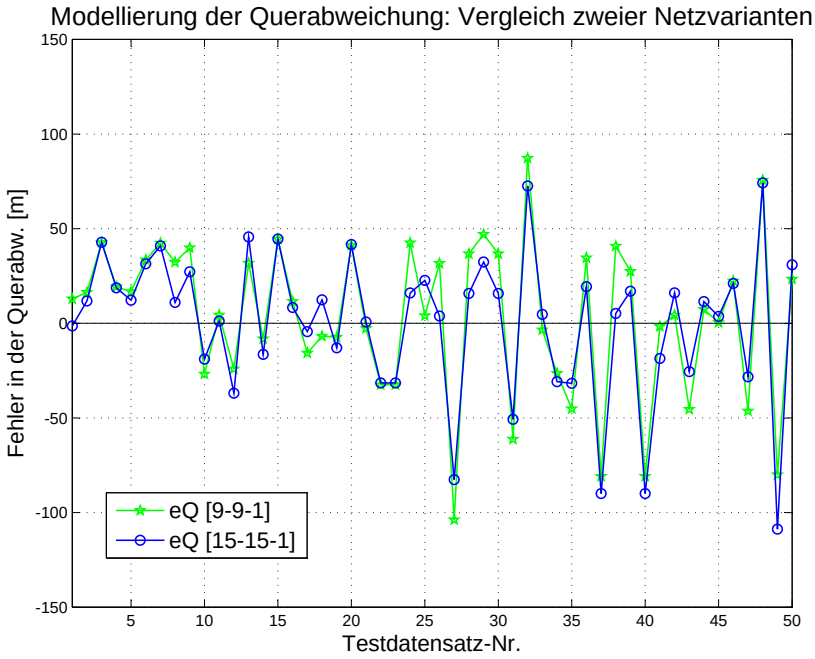


Abbildung 4.23: Modellierung der Querabweichung aus allen acht Tagen: Vergleich der Fehler in der prognostizierten Querabweichungen für die zwei Netzvarianten [9-9-1] und [15-15-1]

den können, alle in der zweiten Hälfte der rechten Grafik auftreten. Aufgrund der zufälligen Wahl der Testbeispiele handelt es sich hier jedoch um eine zufällige Häufung. Außerdem haben weder die Grafiken einzelner Tage für sich betrachtet, noch der Vergleich der einzelnen Tage insgesamt eine derartige

Tendenz gezeigt. Die Fehler der Querabweichungen zeigen keine Vergrößerung in der zweiten Hälfte der Lernbeispiele einzelner Tage und auch keine derartige Tendenz während der acht Evaluierungstage.

Es zeigt sich anhand der sehr ähnlichen Ergebnisse, dass die Anzahl der Lernbeispiele, die bei der tageweisen Auswertung zur Verfügung stand, ausreichend war, um die Netze hinreichend zu trainieren. Die Verwendung der 8-fachen Menge an Lernbeispielen, wie sie bei der zusammengefassten Auswertung zur Verfügung standen, bringt offensichtlich keine weiteren Vorteile.

Abschließend wurde die beste Netzkonfiguration noch mit zwei anderen Transferfunktionen in der Ausgabeschicht untersucht, um die Frage zu klären, wie weit die bisher erzielten Ergebnisse durch Netzvariationen noch optimiert werden können. Es wurde dabei wieder das Netz [15-15-1] zugrundegelegt, mit dem im Falle der gemeinsamen Auswertung aller acht Tage bislang die besten Ergebnisse erzielt wurden. Die Verwendung der Tangens-Hyperbolicus-Funktion oder der Logarithmischen Sigmoidfunktion (vgl. Tabelle 3.5) in der Ausgabeschicht kann sinnvoll sein, da der Wertebereich der Ausgabeparameter auf den Bereich $[-1, 1]$ normiert wird, im Falle der Querabweichung sogar auf $[0, 1]$ begrenzt ist. Damit wird der gesamte Wertebereich der Funktion ausgenutzt, was eine noch bessere Adaption der Aufgabe bewirken kann. Die Ergebnisse sind in Tabelle 4.15 und den Abbildungen 4.24 und 4.25 dargestellt.

Tabelle 4.15: Variation der Transferfunktion in der Ausgabeschicht für das KNN [15-15-1] zur Modellierung der Querabweichung (gemeinsame Auswertung aller acht Tage)

Transferfunktion der Ausgabeschicht	MSE	Maximale Abw. [m]	Standardabw. [m]	Mittelwert [m]
Linearfunktion	0.025	108.8	38.2	0.4
Tangens Hyperbolicus	0.024	80.3	35.7	3.6
Log. Sigmoidfunktion	0.025	90.0	39.1	1.9
zum Vergleich: Unter Ausschluss der 6 extremalen Querabweichungen				
Tangens Hyperbolicus	0.024	80.1	27.9	10.8
Log. Sigmoidfunktion	0.025	84.5	29.8	9.6

Wie in Tabelle 4.15 ersichtlich ist, liefert die Verwendung der Tangens Hyperbolicus-Funktion als Transferfunktion in der Ausgabeschicht mit einer Standardabweichung von ca. $36\,\mathrm{m}$ noch etwas bessere Ergebnisse als die Linearfunktion, die bislang in den Untersuchungen verwendet wurde. Die Verwendung der logarithmischen Sigmoidfunktion hingegen bringt keine Verbesserung mit sich, die Standardabweichung verschlechtert sich sogar geringfügig auf $39\,\mathrm{m}$. Die Abbildungen 4.22 und 4.23 sowie 4.24 und 4.25 zeigen insgesamt sehr ähnliche Verläufe, wobei einzelne Ausreißer mal mehr und mal weniger stark ausgeprägt sind. Die problematischen, schlecht abbildbaren Datensätze bleiben jedoch in allen Fällen die gleichen. Werden die sechs größten Ausreißer -diese stellen gleichzeitig die sechs extremalen Querabweichungen dar[6]- ausgeschlossen , so sinkt die Standardabweichung um etwa $8 - 9\,\mathrm{m}$ auf ca. $28\,\mathrm{m}$ bzw. $30\,\mathrm{m}$. Die größte Abweichung liegt immer noch bei etwa $80 - 85\,\mathrm{m}$. Dabei handelt es sich jedoch um einen einzelnen Ausreißer, der von beiden Netzen nur schlecht abgebildet werden kann. Die zweitgrößte Abweichung liegt jeweils bei nur noch $49\,\mathrm{m}$. Insgesamt verschiebt sich der Mittelwert der Abweichungen nach oben auf ca. $10\,\mathrm{m}$. Dies liegt an den verbleibenden Extremalwerten in den Testbeispielen, die nun tendenziell nach oben (hin zu großen Querabweichungen) ausreißen (vgl. Abbildung 4.24).

[6]Wie in der Abbildung 4.24 ersichtlich, handelt es sich dabei um die Beispieldaten mit den Nummern 27, 31, 37, 40, 48, 49.

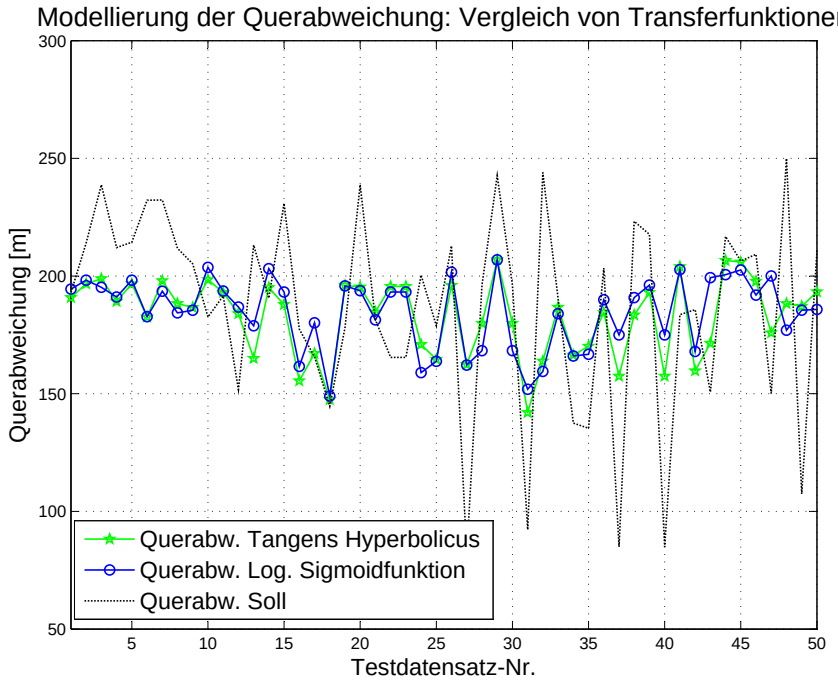


Abbildung 4.24: Modellierung der Querabweichung aus allen acht Tagen: Vergleich der ermittelten Querabweichungen zweier verschiedener Transferfunktionen in der Ausgabeschicht des Netzes [15-15-1]

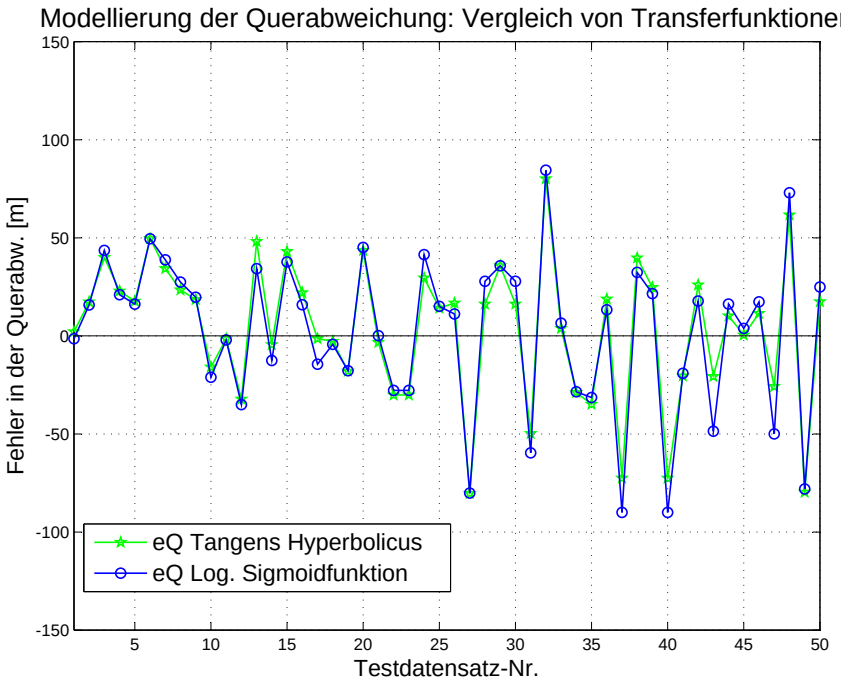


Abbildung 4.25: Modellierung der Querabweichung aus allen acht Tagen: Vergleich der Fehler in den ermittelten Querabweichungen zweier verschiedener Transferfunktionen in der Ausgabeschicht des Netzes [15-15-1]

Damit wurde an einzelnen Beispielen gezeigt, dass das Potenzial der KNN auf Grundlage der Daten aus Do-iT weitgehend ausgereizt ist. Durch Änderung einzelner Parameter in der Netzstruktur können unter Umständen weitere, kleinere Verbesserungen erzielt werden, allerdings ändern diese an der Größenordnung der erzielbaren Abbildungsgenauigkeit kaum etwas. Die KNN wurden ausreichend an die gestellte Aufgabe adaptiert und sind in der Lage, die Querabweichung sämtlicher Trajektorien mit einer Standardabweichung in der Größenordnung von $35 - 45\,\mathrm{m}$ vorauszusagen. Werden Extremwerte von der Prognose ausgeschlossen, so sinkt die Standardabweichung nochmals (in den zufällig gewählten Testdaten bei Ausschluss der sechs Extrema auf unter $30\,\mathrm{m}$).

Im nachfolgenden Abschnitt werden die mit den KNN erzielten Ergebnisse vor dem Hintergrund der Evaluierungsergebnisse aus dem Projekt Do-iT beurteilt. Es werden dazu insbesondere die wenigen ausgewerteten Testfahrten mit GPS mit einbezogen, um die Aussagekraft der Querabweichung anhand der Korrektheit der berechenbaren FPD-Trajektorien besser einschätzen zu können.

4.2.3 Beurteilung der Ergebnisse

Wie im vorigen Kapitel gezeigt wurde, kann die Querabweichung der Trajektorien mit Hilfe der KNN aus den Einflussgrößen Dauer und Länge der Trajektorie, der Anzahl RE, aus denen sich die Trajektorie zusammensetzt sowie der Antennen- und Straßendichte vorhergesagt werden. Dies gelingt im günstigsten Fall mit einer Standardabweichung von etwa $35\,\mathrm{m}$ für Trajektorien, die eine berechnete Querabweichung von weniger als $250\,\mathrm{m}$ aufweisen. Werden alle Trajektorien berücksichtigt, die in Do-iT berechnet wurden, so steigt die Standardabweichung auf mehr als das Doppelte. Die Berechnung der Dichtedifferenzen und -mittelwerte aus den Extrema, anstelle der Dichten am Anfang und Ende der Trajektorien, führt zu etwas schlechteren Ergebnissen (die Standardabweichung steigt um etwa $15 - 20\,\%$). Die Verwendung beider Arten der Berechnung von Dichtedifferenz und Mittelwert für die Netz- und Straßeninfrastruktur kann in weitergehenden Untersuchungen sinnvoll sein, um die Dichteänderungen entlang der Trajektorien besser berücksichtigen zu können. Allerdings wird dadurch auch die Dimension der Eingangsvektoren von sieben auf elf Einflussgrößen erhöht. Damit sind komplexere KNN und eine größere Anzahl an Beispielen notwendig. Eine derartige Untersuchung bleibt jedoch zukünftigen Arbeiten vorbehalten.

Die Adaption der Gewichte der KNN wird in allen untersuchten Fällen bereits früh abgebrochen. Dies deutet auf relativ schlecht zusammenpassende Ein- und Ausgangswerte hin. Um die Ergebnisse besser beurteilen zu können, muss daher zunächst geklärt werden, welche Aussagekraft die als Sollwerte angenommenen Querabweichungen haben und mit welcher Zuverlässigkeit sie bestimmt wurden. Dies kann nur mit Hilfe geeigneter Referenzdaten beurteilt werden.

Die Testfahrten zur Evaluierung der Trajektoriengenerierung aus Abis-Daten im Rahmen des Projektes, wurden im Großraum Karlsruhe durchgeführt. Allerdings beinhalteten diese Fahrten eine Vielzahl von Testszenarien, die der Untersuchung von Identifikationsalgorithmen dienten. Da die meisten dieser Tests statisch oder mit nur kurzen Fahrstrecken durchgeführt wurden, kamen sie bei der Beurteilung der Abis-Daten-Auswertung nicht in Frage. Für eine erfolgversprechende Auswertung mussten die Telefongespräche bzw. die dazu ermittelten Punktfolgen unter anderem eine Mindestlänge von $2\,\mathrm{km}$ aufweisen (vgl. Auflistung zu Beginn des Kapitels 4.2.2). Aus diesen Gründen blieben für die Evaluierung der Trajektoriengenerierung aus Abis-Daten in Do-iT nur insgesamt 11 Fahrten übrig. Die Lage der Trajektorien ist in der Abbildung 4.26 dargestellt, wobei die Testfahrten auf der A5 in orange, die auf der B3 in grün sowie eine Testfahrt auf der B10 in blau hervorgehoben sind. Die Fahrten auf der A5 und B3 sind jeweils doppelt vorhanden, da zwei aktive Mobiltelefone im Fahrzeug waren, und beide Straßen in Nord-Süd- und Süd-Nord-Richtung befahren wurden (die B3 wurde dabei zweimal in Richtung N-S befahren). Bei

der Fahrt auf der B10 konnte nur eines der beiden Telefone bei der Auswertung identifiziert werden, so dass hier nur eine weitere Trajektorie zur Evaluierung hinzu kam.

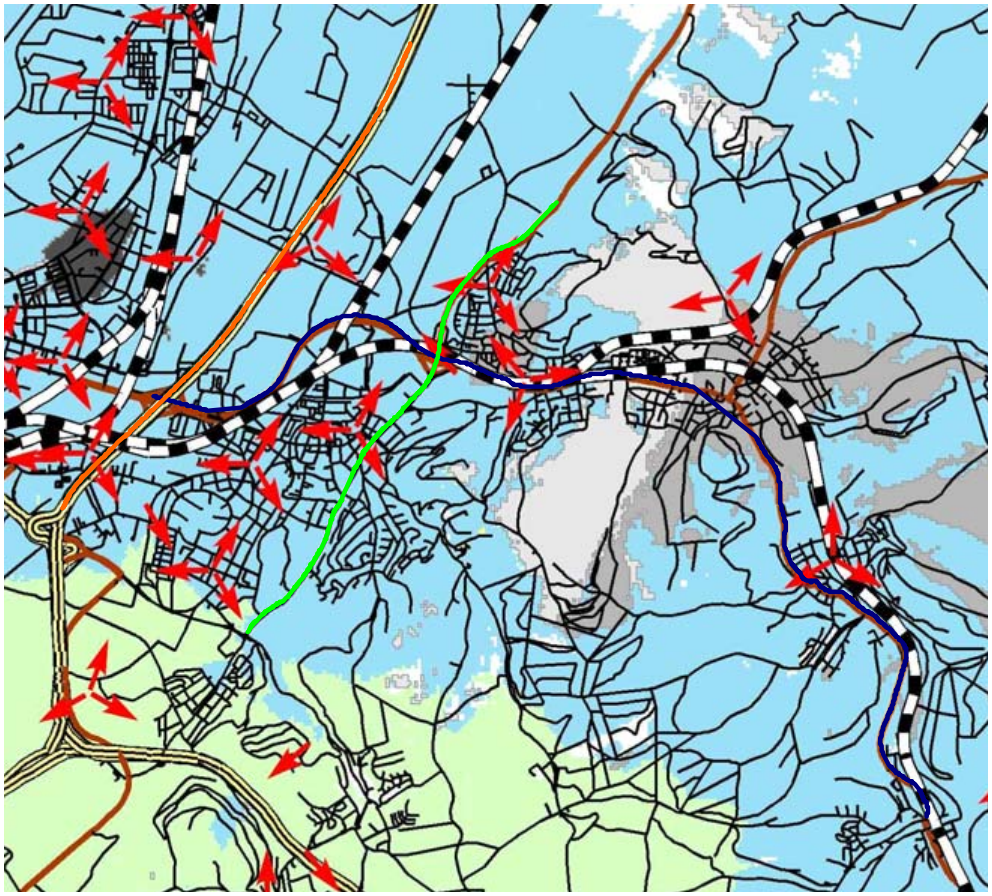


Abbildung 4.26: Lage der Testfahrten mit GPS, die zur Evaluierung des entwickelten Algorithmus im Projekt Do-iT zur Verfügung standen (Do-iT [2008b])

Die Trajektorien wurden aus Positionsfolgen geschätzt, die ihrerseits mit Hilfe der Methode des Signalstärke-Matchings ermittelt wurden. D. h. die Positionen wurden über die räumliche Zuordnungen der gemessenen Signalstärken zu bis zu sechs umliegenden Zellen in den entsprechenden Signalstärkekarten bestimmt. Unter Verwendung eines geeigneten Kalman-Filter können damit im Mittel Standardabweichungen für Einzelpositionen von $560\,\mathrm{m}$ erzielt werden (Do-iT [2008b]). Betrachtet man mit dieser Kenntnis nochmals das Lokalisierungsnetz in Karlsruhe in Abbildung 4.14, so deuten die Maschen des quadratischen Gitters in etwa die Größe der Aufenthaltsbereiche einer mittels RX-Matching[7] ermittelten Fahrzeugposition für eine Sicherheitswahrscheinlichkeit von $95\,\%$ an. Damit wird klar, wie schwer die Zuordnung der korrekten Straßen zu der Punktfolge eines Teilnehmers in Bereichen mit hoher Straßeninfrastruktur ist.

Die Ermittlung der wahrscheinlichsten Kantenfolge des Teilnehmers erfolgte dann durch die Erzeugung eines Korridors, bestehend aus Quadraten um die ermittelten Einzelpositionen. Die Seitenlängen der Quadrate wurde dabei dynamisch an die Einzelpositionsgenauigkeit angepasst, wobei der Maximalwert $800\,\mathrm{m}$ betrug. Nach Festlegung der möglichen Start- und Zielknoten wurden alle theoretisch möglichen

[7] Der Begriff RX-Matching entstand in dem Projekt und bezeichnet die Methode der Positionsbestimmung mit Hilfe der gemessenen Signalstärken und der zugehörigen Signalstärkekarten des Netzbetreibers.

Routen innerhalb des Korridors ermittelt. Unter anderem wurde anhand der Querabweichung der Routen dann die wahrscheinlichste Route ermittelt.

Der Vergleich der berechneten Trajektorie mit der tatsächlichen wurde nun mit Hilfe der beiden Parameter: Zuordnungskorrektheit Typ A und Typ B (vgl. Tabelle 5.1) durchgeführt. Typ A gibt dabei den Anteil der Referenztrajektorie an, die mit der Mobilfunkortung gefunden wurde, Typ B beschreibt den korrekten Anteil der gesamten berechneten Trajektorie.

Eine direkte Korrelation zwischen der Zuordnungskorrektheit und der Querabweichung konnte im Rahmen der Projektevaluierung in den untersuchten Trajektorien nicht klar identifiziert werden. Es konnte jedoch der empirische Zusammenhang formuliert werden, dass eine Querabweichung $< 250\,\mathrm{m}$ in Verbindung mit einer Trajektorienvollständigkeit von $> 97\,\%$ (d. h. mehr als $97\,\%$ der berechneten Trajektorien eines Teilnehmers konnten zur Ermittlung der Trajektorie verwendet werden) eine Zuordnungskorrektheit Typ B von $> 70\,\%$ nach sich zieht (Do-iT [2008b]). Betrachtet man allerdings die Daten der 9 Evaluierungstage, so erfüllen lediglich $4\,\%$ aller Trajektorien diese beiden Anforderungen. Insgesamt ergibt sich daher eine relativ hoher Anteil an weniger als $70\,\%$ korrekten Trajektorien im Datenmaterial. Die nicht korrekt zugeordneten Teilstücke der Trajektorien -in der Regel handelt es sich dabei um das Anfangs- und Endstück- verschlechtern die Querabweichung erheblich.

Damit wird klar, dass der Parameter der Querabweichung keine scharfen quantitativen Aussagen hinsichtlich der Korrektheit der Trajektorie zulässt. Er dient vielmehr als ein Indikator für die Beurteilung der Trajektorien. Eine exakte Prognose dieses Parameters mit Hilfe von KNN ist daher nicht möglich. Unter diesem Gesichtspunkt betrachtet ist eine Prognose dieses Parameters mit einer Standardabweichung von ca. $40\,\mathrm{m}$ nach Meinung des Autors als zufriedenstellend einzuschätzen. Die Eignung von KNN zur Prognose von Querabweichungen von FPD-Trajektorien aus Abis-Daten wird damit in den festgestellten Grenzen bestätigt.

Die Berücksichtigung weiterer Eingangsparameter, die bei der Berechnung der Trajektorien verwendet wurden, kann die Ergebnisse unter Umständen noch etwas verbessern. Dies können zum Beispiel die gemessenen Signalstärken und deren Differenzen zu den auf Ausbreitungsmodellen basierenden Signalstärkenkarten oder die aus den Extrema entlang der Trajektorien berechneten Dichtedifferenzen und -mittelwerte der Infrastruktur (Mobilfunk- und Straßennetz) sein. Ob damit die Prognose der Querabweichungen tatsächlich noch wesentlich verbessert werden kann, muss in weiteren Arbeiten untersucht werden.

4.3 Beurteilung der Eignung von KNN

Die Eignung der künstlichen neuronalen Netze zur Modellierung von Datenqualität in Prozessen wurde anhand zweier praktischer Beispiele detailliert untersucht. Zunächst diente die geodätische Grundaufgabe, das polare Anhängen von Punkten an einen bekannten Standpunkt, als Beispiel. Es wurden mit Matlab eine große Anzahl Testdaten simuliert, um die Abbildung von verschiedenen Qualitätsparametern zu untersuchen. Die grundsätzliche Eignung der KNN zur Abbildung von Qualitätsübergängen in Prozessen konnte damit gezeigt werden. Es gelang, die beiden Genauigkeitsparameter mit Submillimeter-Genauigkeit abzubilden. Ebenso konnten simulierte Mängel in der Datenverfügbarkeit, Vollständigkeit und Konsistenz der Eingangsdaten zufriedenstellend abgebildet werden. Es wurden außerdem erste Erfahrungen in der Architektur und der Adaption der KNN an die zu lösende Aufgabe gesammelt. So ist eine Normierung der Daten auf den Wertebereich $[0, 1]$ ebenso wie eine ausreichende, jedoch nicht zu großzügige Dimensionierung der Netze dringend zu empfehlen.

Im Anschluss an das simulierte Beispiel wurde ein reales Beispiel aus der Praxis herangezogen. Es handelte sich dabei um die Generierung von Trajektorien von Straßenverkehrsteilnehmern aus Mobilfunkdaten, welche im Rahmen des Projektes Do-iT untersucht wurde. Es wurde ein Algorithmus entwickelt und implementiert, mit dem aus den Abis-Daten Trajektorien von Straßenverkehrsteilnehmern generiert werden können. Dieser Prozess beinhaltet zahlreiche Zwischenschritte und Einzelentscheidungen, so dass eine formelmäßige Beschreibung nur schwer möglich ist. Entsprechend komplex sind daher die Zusammenhänge zwischen der Eingangs- und Ausgangsdatenqualität. Auf Grundlage der archivierten Daten aus der Evaluierungsphase des Projektes konnte die Querabweichung der Trajektorien mit Hilfe von KNN modelliert werden. Insgesamt sieben Einflussgrößen wurden dabei als Eingänge berücksichtigt. Die Querabweichung der Trajektorien kann in der Regel mit einer Standardabweichung besser als $40\,\mathrm{m}$ vorhergesagt werden, was in Anbetracht der relativ großen Unsicherheiten, die die Mobilfunkortung in Ballungsräumen mit dichter Straßeninfrastruktur mit sich bringt, ein sehr gutes Ergebnis ist. Dies konnte mit Hilfe der Evaluierungsergebnisse aus dem Projektbericht (Do-iT [2008b]) bestätigt werden.

Sinnvoll wäre im Weiteren noch die Untersuchung der Modellierbarkeit binär beschreibbarer Parameter der Vollständigkeit und der Verfügbarkeit sowie von lückenhaften Datenquellen, d. h. nicht jederzeit verfügbaren Daten. Dies konnte im Rahmen dieser Arbeit nur mit künstlich generierten Testdaten ansatzweise gezeigt werden, da einerseits keine realen Daten in ausreichender Menge vorhanden waren und andererseits im Rahmen dieser Arbeit nur die grundlegende Eignung der gewählten Methode gezeigt werden sollte. Weitere Untersuchungen an anderen geeigneten Beispielen bleiben daher nachfolgenden Arbeiten vorbehalten.

Die beiden untersuchten Beispiele aus sehr unterschiedlichen technischen Bereichen deuten jedoch bereits das Potenzial an, welches die künstlichen Neuronalen Netze zur Modellierung von Datenqualität bieten. Dies gilt bislang nur für Qualitätsparameter, die als statisch betrachtet werden können. Sich zeitlich schnell verändernde Parameter wurden hier nicht betrachtet. Für eine abschließende und vor allem allgemeingültige Beurteilung der Eignung von KNN zur Modellierung statischer Qualitätsparameter müssten jedoch wesentlich mehr Testbeispiele mit den unterschiedlichsten Daten in ausreichender Menge untersucht werden. D. h. derzeit muss im Einzelfall untersucht werden, ob ein KNN die Datenqualität in ausreichender Güte abbilden kann. Dazu ist Erfahrung erforderlich und es müssen jeweils eine Reihe von Netzkonfigurationen getestet werden, um eine der optimalen Lösungen für die gestellte Aufgabe zu finden.

Künstliche neuronale Netze können die Entwicklung von Prozessen simulatorisch unterstützen oder die Abläufe in bestehenden Prozessen nachbilden. Damit ist unter anderem die Abschätzung geplanter Qualitätsverbesserungsmaßnahmen und deren Auswirkungen vor der Einführung möglich. Wichtig ist dabei jedoch, dass die dem Netz zugeführten Eingangsdaten ausschließlich aus dem Eingangsdatenraum stammen, da KNN grundsätzlich nur schlecht oder gar nicht zur Extrapolation geeignet sind.

5 Ein Qualitätsmanagementkonzept für Daten

Ein Qualitätsmanagement (QM) für Daten umfasst die organisierten Maßnahmen zur Sicherung und Verbesserung der Datenqualität. Es stellt den Rahmen für eine einheitliche Beschreibung und Ermittlung von Datenqualität sowie zu deren systematischen Beurteilung, Sicherung und Verbesserung bereit. Damit bildet das Management der Datenqualität einen wichtigen Bestandteil eines modernen und ganzheitlichen unternehmerischen Qualitätsmanagements, in dem alle Bereiche einer Organisation involviert sind. In Kapitel 2.4 wurden bereits die wichtigsten Qualitätsmanagementsysteme mit ihren wesentlichen Eigenschaften vorgestellt.

In diesem Abschnitt wird ein Qualitätsmanagementkonzept für Daten vorgestellt, welches ursprünglich im Rahmen des Projektes Do-iT entwickelt und teilweise umgesetzt wurde (Do-iT [2007] und Laufer [2008]). Das Konzept besteht zum einen aus dem bereits im Kapitel 2.2 vorgestellten Qualitätsmodell, mit dem eine umfassende Beschreibung der systemrelevanten Daten ermöglicht wird und den geeigneten Messmethoden zur Ermittlung der Parameterwerte. Zum anderen stellt das Konzept geeignete Analysemethoden zur Verfügung, die die Entwicklung eines besseren Verständnisses der Beteiligten für die Prozesse fördern. Dies ist die Grundlage zur Identifikation von Verbesserungspotenzial sowie zur Definition, Umsetzung und Überprüfung von Maßnahmen zur Sicherung und Verbesserung der Datenqualität. Die Methode zur Modellierung von Qualität in Prozessen mit Hilfe künstlicher neuronaler Netze wird in den Gesamtzusammenhang des Qualitätsmanagementkonzepts gestellt und die Schnittstellen zu den anderen Komponenten werden erläutert. Damit wird das Konzept um eine praktikable, quantitative Analysemethode erweitert, die das bestehende Verfahren auf Basis Boolescher Algebra ergänzen oder ersetzen kann. Wie bereits im Kapitel 3 ausführlich erläutert, ist das bestehende Rechenverfahren, welches von Wiltschko [2004] vorgeschlagen wurde, sehr konzeptionell und nur in wenigen Fällen praktisch einsetzbar, da die Modellierung der Datenqualität meist auf Merkmalsebene verbleibt (vgl. 3.1.4). Eine Modellierung der Verfügbarkeit ist mit dem Verfahren in der Regel möglich, sofern Kenntnisse über die Verfügbarkeit einzelner Komponenten vorliegen. Hier ist eine Vereinfachung der Abläufe durch Boolesche Operatoren sinnvoll.

Die wesentlichen Eigenschaften des QM-Konzepts sowie die Möglichkeiten, die dessen Umsetzung bietet, sind im Folgenden nochmals kurz aufgeführt:

- Datenqualität ist in einem einheitlichen Modell darstellbar,
- Vergleich von Daten aus unterschiedlichen Quellen ist möglich,
- Relevante DV-Prozesse können analysiert werden,
- Systematische Identifizierbarkeit von Verbesserungspotenzial,
- Nachweis der Wirksamkeit von Maßnahmen zur Sicherung und Verbesserung der Qualität,
- Evaluierung der neu entwickelten oder modifizierten DV-Prozesse,
- Modellierung und Simulation der Datenqualität in Prozessen auf Parameterebene.

Neben der theoretischen Erläuterung des Konzepts und seiner einzelnen Bestandteile in den folgenden Abschnitten, werden zur Veranschaulichung auch Beispiele aus den Arbeiten in dem bereits abgeschlossenen Projekt Do-iT herangezogen.

5.1 Übersicht über das Konzept

Das Konzept zum Management von Datenqualität ist in der Abbildung 5.1 mit seinen wesentlichen Komponenten schematisch dargestellt. Neu an dieser Darstellung ist die übersichtliche Zusammenstellung aller wesentlichen Bestandteile eines Qualitätsmanagementkonzepts für Daten. Bei den einzelnen Bestandteilen und Methoden des Konzepts handelt es sich weitgehend um bekannte Verfahren bzw. bestehendes Wissen. Eine Ausnahme stellt hier die bereits im Kapitel 4 zur Fortpflanzung von Datenqualität untersuchten künstlichen neuronalen Netze dar.

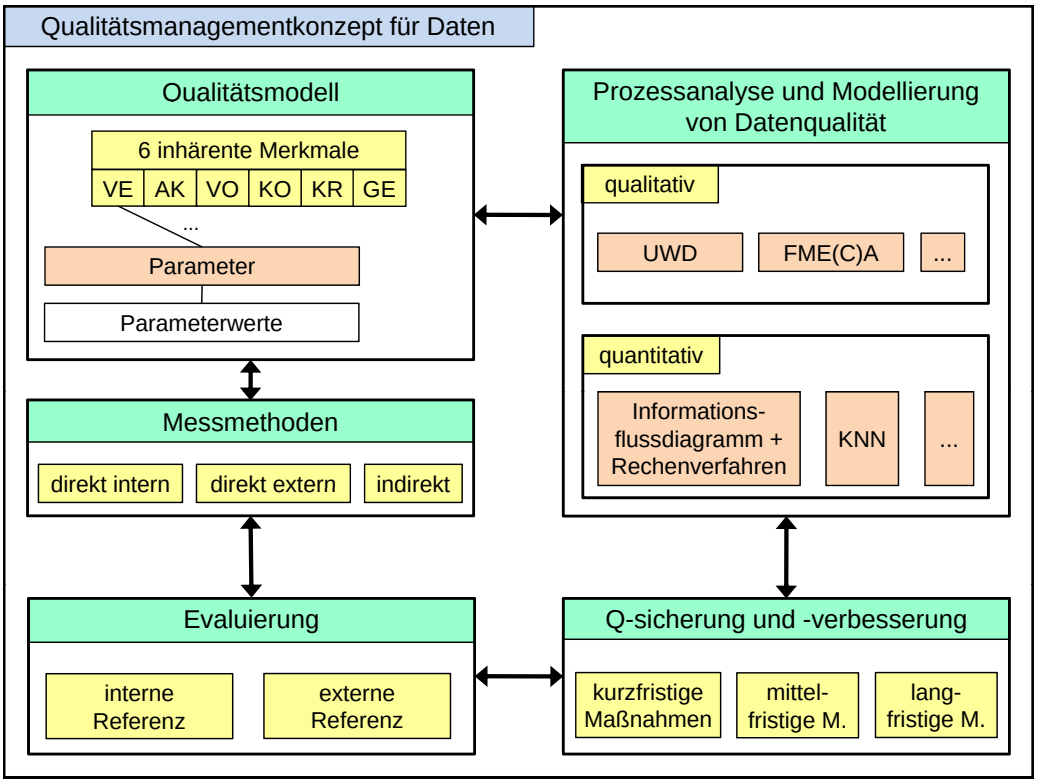


Abbildung 5.1: Übersicht der wesentlichen Bestandteile des Qualitätsmanagementkonzeptes

Die Abbildung ist nicht als Ablaufschema zu verstehen, vielmehr sollen die Doppelpfeile lediglich die vorhandenen, engen Verknüpfungen zwischen den Hauptbestandteilen des Konzeptes andeuten. Die fünf Hauptbestandteile *Qualitätsmodell*, *Prozessanalyse & Qualitätsmodellierung*, *Messmethoden*, sowie die *Qualitätssicherung und -verbesserung* und die *Evaluierung* geben einen ersten groben Überblick über die wesentlichen Tätigkeiten, die für ein erfolgreiches Management von Datenqualität eine wichtige Rolle spielen. Der Ablauf gemäß dem Konzept kann dabei wie folgt aussehen:

- Eine erste Analyse der vorliegenden oder geplanten Prozesse sowie die Diskussion mit allen Beteiligten gibt Aufschluss über die auftretenden Datenarten und
- ermöglicht die Konkretisierung der Qualitätsmodelle mit Hilfe geeigneter Parameter sowie
- die Erarbeitung der zugehörigen Messmethoden zur Quantifizierung der Parameter.
- Damit ist die Qualität der Daten beschreibbar und kann erstmals ermittelt werden.
- Eine detaillierte Prozessanalyse ist Voraussetzung zur Beschreibung und Modellierung von Zusammenhängen zwischen den Qualitätsparametern innerhalb der Prozesse, sowie

- für die Erarbeitung und Umsetzung von Qualitätssicherungs- und -verbesserungsmaßnahmen.
- Die Wirksamkeit einzelner Maßnahmen kann ggf. zunächst simuliert werden.
- Zur Evaluierung der Maßnahmen und der aktuell vorliegenden Datenqualität sind wiederum die definierten Qualitätsparameter mit ihren Messmethoden und/oder geeignete Analysemethoden notwendig.

Hinsichtlich des Ziels der ständigen Verbesserung der Datenqualität als Bestandteil eines ganzheitlichen, unternehmerischen Qualitätsmanagements nach dem Vorbild des Kaizen (*jap.*: Veränderung zum Besseren) bzw. dem KVP-Prinzips (KVP steht für **K**ontinuierlicher **V**erbesserungs**p**rozess), wiederholen sich die Tätigkeiten in regelmäßigen Abständen und die Festlegungen werden kritisch geprüft und ggf. überarbeitet (Kamiske und Brauer [2008]). Damit soll der ständigen, zeitlichen Veränderungen Rechnung getragen und das Verbesserungspotenzial kontinuierlich ausgeschöpft werden. Die Grundidee wird Deming [1982] zugeschrieben, der den PDCA-Zyklus (oder auch Deming-Z.) bestehend aus den sich immer wiederholenden Tätigkeiten Plan-Do-Check-Act entwickelt hat. Dieser Kreislauf spielt auch in der DIN EN ISO 9000 [2005] eine wesentliche Rolle (vgl. Kapitel 2.4.1). Eine Interpretation der Bestandteile des Zyklus zur Erläuterung der Abläufe zur ständigen Verbesserung der Datenqualität ist im Abschnitt 5.2.6 in Abbildung 5.5 dargestellt.

Dabei spielt in der Praxis neben der Qualitätsverbesserung insbesondere auch die Sicherung des erreichten Qualitätsniveaus eine entscheidende Rolle. In der Übersicht in Abbildung 5.1 werden beide Begriffe wegen ihrer engen Zusammengehörigkeit in einem Kasten zusammengefasst.

Der Vorteil des Konzepts liegt in der einfachen und übersichtlichen Zusammenstellung der Verfahren in einer Grafik. Erstmals werden hier alle für die Sicherung und Verbesserung von Datenqualität erforderlichen Schritte gemeinsam dargestellt. Es werden Verfahren vorgeschlagen, mit denen das Management von Datenqualität -von der Beschreibung der Datenqualität bis zu deren Fortpflanzung- erfolgen kann. Das Konzept bleibt jedoch offen für weitere Verfahren zur Prozessanalyse und Modellierung von Datenqualität, was in der Übersicht durch die leeren Kästen angedeutet wird.

5.2 Bestandteile des Konzepts

In den folgenden Abschnitten werden die einzelnen Bestandteile des Konzepts zum Qualitätsmanagement von Daten einzeln dargestellt und kurz erläutert. Dabei dienen einige praktische Beispiele aus dem Projekt Do-iT -soweit vorhanden- der besseren Veranschaulichung des Konzepts. Der Vollständigkeit halber werden einzelne Bestandteile wie z. B. das Qualitätsmodell, trotz der Darstellung an anderer Stelle, in der Arbeit hier nochmals kurz aufgeführt. Die Beschreibung beschränkt sich in diesen Fällen jedoch auf neue Aspekte im Zusammenhang mit dem Qualitätsmanagementkonzept sowie auf die Darstellung von konkreten Beispielen zur Erläuterung.

5.2.1 Qualitätsmodell

Das Qualitätsmodell mit seinen Qualitätsmerkmalen und den datenindividuell definierten Parametern wurde bereits im Kapitel 2.2 erläutert. Das Modell dient der umfassenden Beschreibung der relevanten Qualitätsaspekte von Daten mit Hilfe geeigneter Qualitätsparameter, die den sechs festgelegten Qualitätsmerkmalen zugeordnet werden können. Eine umfassende Beschreibung der Datenqualität ist insbesondere zur Beurteilung der Datenqualität, zum Vergleich verschiedener Datenquellen, zur Quantifizierung von Kundenanforderungen sowie zur Archivierung der Daten erforderlich. Werden die Daten zu einem

späteren Zeitpunkt in einer neuen Anwendung verwendet, so ist es notwendig, zunächst die Qualität der gespeicherten Daten mit Hinblick auf die neue Anwendung beurteilen zu können.

Die Tabelle 5.1 zeigt beispielhaft einen Auszug aus dem Qualitätsmodell für die aus Mobilfunkdaten generierten Fahrzeugtrajektorien (FPD-Trajektorien). Dabei wurden sowohl Parameter aufgenommen, die nur intern während der Entwicklungsphase der Algorithmen von Bedeutung waren, als auch solche, die zur Evaluierung der Daten definiert wurden. Das Modell beinhaltet neben der Zuordnung der Parameter zu einem der sechs Qualitätsmerkmale auch deren Einheit und eine kurze Beschreibung. Eine Nummer, zusammengesetzt aus der Nummer des Merkmals und einer laufenden Nummer erleichtert die Übersicht.

Tabelle 5.1: Auszug aus dem Qualitätsmodell für FPD-Trajektorien (Do-iT [2009a])

Nr.	Merkmal	Parameter	Kürzel	Definition
1.x	VE	...	...	...
2.x	AK	...	...	...
3.1	VO	Trajektorien-vollständigkeit	$\mathrm{VO}_{\mathrm{FPD}}$ $[\%]$	Anteil der verwendeten Positionen einer Positionsfolge eines Teilnehmers, der für die Trajektorie verwendet wird
3.2		Durchdringung mit FPD	d $[\%]$	Anteil des Verkehrs, der mit FPD erfasst werden kann; bezogen auf den gesamten Verkehr
4.x	KO	Die Konsistenz wurde durch Einhaltung des Datenmodells gewährleistet		
5.1	KR	Zuordnungs-korrektheit Typ A	$\mathrm{KR}_{\mathrm{ZuA}}$ $[\%]$	Korrekter Streckenanteil der FPD-Route, der sich mit der GPS-Route deckt, bezogen auf die Länge der GPS-Route
5.2		Zuordnungs-korrektheit Typ B	$\mathrm{KR}_{\mathrm{ZuB}}$ $[\%]$	Korrekter Streckenanteil der FPD-Route, der sich mit der GPS-Route deckt, bezogen auf die Länge der ermittelten FPD-Route
5.3		Trajektorienlänge	$L_{\mathrm{FPD}}[m]$	Länge der FPD-Trajektorie
6.1	GE	Mittlere Querabweichung	$\mathrm{QA}_{\mathrm{FPD}}$ $[\mathrm{m}]$	Mittel der orthogonalen Abweichungen der verwendeten Positionen von der wahrscheinlichsten Route
6.2		Standardabw. der Geschwindigkeit	s_v $[\mathrm{km/h}]$	Genauigkeit der aus den Trajektorien ermittelbaren Geschwindigkeit

Es wurden hier aus Platzgründen nur einige wichtige Parameter herausgegriffen, die die Vollständigkeit, Korrektheit und Genauigkeit der Daten beschreiben. Die weiteren Parameter zur Beschreibung der Merkmale Verfügbarkeit, Aktualität und Konsistenz der Daten sind in der Tabelle nur als Platzhalter dargestellt, um die durchgehende Nummerierung zu verdeutlichen. Insgesamt wurden 15 Parameter (ohne räumlich oder zeitlich aggregierte Parameter) definiert, um alle wesentlichen Aspekte der FPD-Trajektorien hinreichend beschreiben zu können. Das vollständige Qualitätsmodell ist im Anhang B dargestellt.

Das Qualitätsmodell stellt eine umfassende Sammlung aller relevanten Qualitätsparameter für eine am Prozess beteiligte Datenart dar. Gegebenenfalls kann und muss eine geeignete Auswahl getroffen werden, wenn die Daten an Kunden abgegeben werden, da in der Regel nicht alle Parameter für den Kunden von

Interesse sind. Die Weitergabe aller Parameter ist nur in wenigen Fällen erforderlich, da viele Parameter vorwiegend zur Prozessoptimierung intern Verwendung finden.

5.2.2 Messmethoden

Qualitätsparameter stellen immer eine quantitative Beschreibung der Daten dar. Daher sind geeignete Messmethoden erforderlich, die der quantitativen Bestimmung von Qualitätsparameterwerten dienen. Es können grundsätzlich drei Methoden zur Bestimmung der Qualitätsparameterwerte unterschieden werden, die in der Tabelle 5.2 dargestellt sind.

Tabelle 5.2: Messmethoden zur Bestimmung von Qualitätsparameterwerten

Messmethode	**Beschreibung**	**Beispiele**
direkt extern	Vergleich mit Referenzdaten	Standardabweichung in Längs- und Querrichtung eines Polarpunktes (vgl. Abschnitt 4.1) Do-iT: Trajektorienvollständigkeit
direkt intern	Parameterwert aus DV-Prozess heraus bestimmt	Berechnungszeit Do-iT: Mittlere Querabweichung (vgl. Kap. 4.2.2)
indirekt	Parameterwert aus Expertenwissen	Instrumentengenauigkeit Do-iT: Zählgenauigkeit einer Induktionsschleife

Die Aussagekraft der Parameter hängt dabei von einer Vielzahl von Faktoren ab. So sollte die Qualität der Referenz unbedingt überprüft und beurteilt werden, bevor mit deren Hilfe aussagekräftige Parameterwerte bestimmt werden können. Wird beispielsweise die Standardabweichung einer Messung als Parameter gewählt, so hängt die Zuverlässigkeit der Aussage entscheidend vom Stichprobenumfang ab, aus der die Genauigkeit empirisch bestimmt wird.

Grundsätzlich sind die ersten beiden Methoden zu bevorzugen, da die Aussagekraft der Parameter in der Regel gut beurteilt werden kann. Kann eine Bestimmung der Parameterwerte nur auf Grundlage von Expertenwissen oder Angaben Dritter, wie z. B. aus einem Herstellerhandbuch erfolgen, so ist die Qualität der Quelle nur schwer beurteilbar.

Zur Verdeutlichung der unterschiedlichen Messmethoden werden im Folgenden aus dem Qualitätsmodell für FPD-Trajektorien (Tabelle 5.1) zwei Beispiele angeführt:

- **Direkt externe Messmethode:** Der Wert kann nur mit Hilfe von Referenzdaten höherer Genauigkeit ermittelt werden. Beispiel: Rangkorrelation der Verkehrsstärke aus FPD mit der Referenzverkehrsstärke aus SES über 24 Stunden (vgl. Parameter 5.3 in Tabelle 5.1):

$$r = \frac{1 - 6 \cdot \sum D^2}{n(n^2 - 1)}. \tag{5.1}$$

  mit den Differenzen D der n Rangpaare der absteigend sortierten Messreihen. In der Abbildung 5.2 sind als Beispiel die aus Mobilfunkdaten ermittelten Verkehrsstärken denen aus einer Induktionsschleife für einen Tag und an einem Schleifensensor gegenübergestellt. In diesem Beispiel ergab sich eine Rangkorrelation von 0.79.

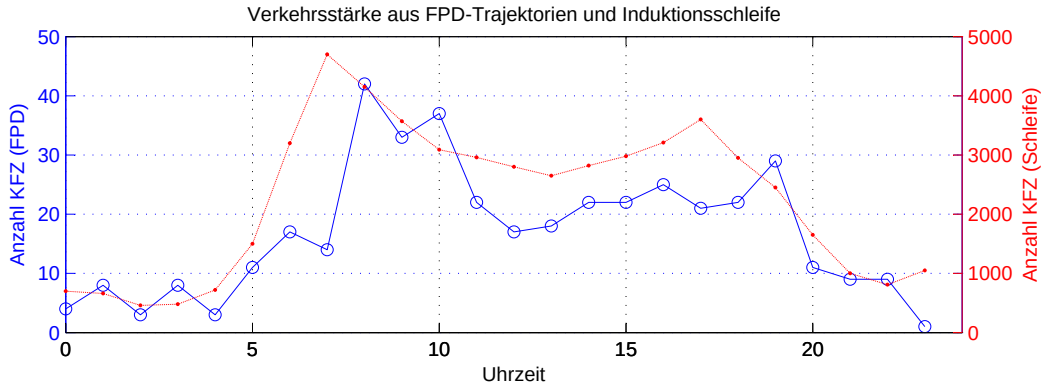


Abbildung 5.2: Verkehrsstärken eines Tages aus Mobilfunkdaten und einer Induktionsschleife gegenübergestellt (Quelle: Do-iT [2009b])

- **Direkt interne Messmethode:** Der Wert kann aus dem Prozess heraus ohne Referenzdaten ermittelt werden. Beispiel: Trajektorienvollständigkeit (vgl. Parameter 3.1 in Tabelle 5.1):

$$\mathrm{VO_{FPD}} = \frac{\text{Anz. verwendeter Positionen}}{\text{alle Positionen der Folge}} \cdot 100\,[\%] \tag{5.2}$$

  Die Trajektorienvollständigkeit wurde intern bestimmt und als Parameter zur Beurteilung der Qualität der berechneten Trajektorien eingesetzt. Dieser Parameter gibt den Anteil der aus Mobilfunkdaten ermittelten Einzelpositionen eines Mobilfunkteilnehmers an, der tatsächlich für die Berechnung der Trajektorie im Straßennetz verwendet werden konnte.

- **Indirekte Methode:** Der Wert kann nicht direkt berechnet werden, sondern muss aus anderer Quelle abgeschätzt werden. In Do-iT konnte bei den berechneten Datenarten auf nur indirekt bestimmbare Parameter verzichtet werden. Allerdings musste bei der Abschätzung der Genauigkeit der mit SES ermittelten Verkehrsstärken, die als Referenz herangezogen wurden, auf Expertenwissen zurückgegriffen werden.

5.2.3 Qualitative Analyseverfahren

Grundsätzlich können qualitative und quantitative Analyseverfahren, je nach der Art der gewonnenen Erkenntnisse, unterschieden werden. Zunächst liegt der Fokus auf den qualitativen Analysen, die ein besseres Verständnis der vorliegenden Prozesse ermöglichen sollen. Aufbauend auf diesen Ergebnissen kann dann eine quantitative Analyse der Prozesse oder Teilprozesse und damit beispielsweise eine Modellierung der Datenqualität mit KNN erfolgen (vgl. Abschnitt 5.2.4).

Eine sehr einfache Darstellung vorliegender oder geplanter Prozesse in Form eines Funktionsschemas unterstützt zunächst die Identifikation aller beteiligten Datenarten. Im Funktionsschema werden alle wesentlichen Elemente und Prozesse eines Systems sowie deren Verknüpfungen grafisch dargestellt, vergleichbar mit einem Blockdiagramm (Datacom [2010]). Dazu sollten bereits die entsprechenden Fachleute an einen Tisch gebracht werden, um das Expertenwissen zu bündeln. Für die nachfolgenden Detailanalysen ist dies jedoch eine notwendige Voraussetzung, da die Qualität der Untersuchungen unmittelbar von dem, in der Regel verteilt vorliegenden, Wissen der Beteiligten aus unterschiedlichen Fachbereichen abhängt.

Nachdem die einzelnen Datenarten identifiziert wurden, können geeignete Qualitätsparameter und zugehörige Messmethoden zu deren quantitativen Bestimmung definiert werden. Damit ist bereits eine Bestandsaufnahme des aktuellen Qualitätsniveaus möglich. In der Entwicklungsphase von Prozessen kann der Fortschritt so überwacht und Änderungen schnell erkannt und ggf. näher identifiziert werden. Liegen bereits bestehende Prozesse vor, so ist eine regelmäßige Überprüfung der Datenqualität möglich. Damit kann auch die Einhaltung des vom Kunden oder der Anwendung geforderten Qualitätsniveaus kontrolliert werden und Schwankungen der Qualität lassen sich besser und schneller feststellen.

Um nach dem Prinzip der ständigen Verbesserung Optimierungspotenzial zu identifizieren und die Datenqualität regelmäßig zu steigern, mindestens jedoch zu sichern, ist eine detaillierte Analyse der Prozesse notwendig. Dazu stehen eine Reihe von Standardmethoden zur Verfügung, die ein strukturiertes Vorgehen sicherstellen und helfen, das vorhandene Expertenwissen aufzuspüren, zu bündeln und festzuhalten sowie Optimierungspotenzial zu identifizieren. Aus den dokumentierten und priorisierten Optimierungspotenzialen können schließlich Maßnahmen zur Qualitätssicherung und/oder -verbesserung abgeleitet und umgesetzt werden. Deren Erfolg kann anhand geeigneter Kennzahlen bzw. Qualitätsparameter überprüft werden.

Als rein qualitative Analysemethoden wurden im Projekt Do-iT das Ursachen-Wirkungs-Diagramm (UWD) sowie die Fehlermöglichkeits- und -einflussanalyse (FMEA) eingesetzt, die bereits im Abschnitt 3.2.1 erläutert wurde. Beim UWD und der FMEA handelt es sich um teamorientierte und textbasierte Verfahren, die ganz allgemein der systematischen Ermittlung der Zusammenhänge zwischen Ursache (Fehler) und Wirkung in einem Produkt, Bauteil oder informationsverarbeitenden System dienen. In diesem Zusammenhang wurden die Verfahren etwas abgewandelt und insbesondere die Wirkung auf die Datenqualität berücksichtigt. Die beiden Verfahren unterscheiden sich in der Herangehensweise an die zu erörternde Problematik. Das UWD wurde bereits 1943 von dem Japaner Kaoro Ishikawa entwickelt und wird daher auch als Ishikawa-Diagramm bezeichnet. Bei der Aufstellung eines UWD werden mögliche Ursachen und Nebenursachen für ein zuvor festgelegtes, nicht gewünschtes Ereignis gesucht, daher auch die Bezeichnung „top-down-Ansatz“(Zurhausen [2002]). Bei der FMEA-Analyse hingegen handelt es sich um einen „bottom-up“Ansatz. Das bedeutet, es werden Fehlermöglichkeiten in einzelnen Bauteilen ermittelt und anschließend deren Auswirkungen auf das gesamte System näher untersucht. Aufgrund der Charakteristik wird letzteres auch als induktives Verfahren und das UWD als deduktives Verfahren bezeichnet.

Wie sich gezeigt hat, kann die Aufstellung eines UWD auch als Vorbereitung auf eine FMEA zum Einsatz kommen (Do-iT [2008a]). Diese universellen Methoden zur Suche von Fehlerursachen und -möglichkeiten werden hier gezielt mit Hinblick auf Auswirkungen auf die Qualitätsmerkmale und -parameter hin verwendet. Mängel hinsichtlich einzelner Qualitätsmerkmale oder auch -parameter werden im UWD als unerwünschte Wirkung vorgegeben. Bei der FMEA hat sich gezeigt, dass die tabellarische Dokumentation der Erkenntnisse um die Darstellung der möglichen Auswirkungen auf die Qualitätsmerkmale erweitert werden muss (Do-iT [2008a]). Zusätzlich wurden Risikoprioritätszahlen ermittelt, um die Dringlichkeit der Maßnahmen quantifizieren zu können. Dabei ist es ratsam, die Beurteilungstabellen für Auftreten, Erkennbarkeit und Behebbarkeit dem zu beurteilenden System anzupassen, um den unterschiedlichen Systemen gerecht zu werden. Die Risiken einer Klimamessstation müssen mit anderen Maßstäben beurteilt werden, wie beispielsweise der Ausfall eines Inertialmesssystems im Flugzeug.

Zur Verdeutlichung des Vorgehens werden zwei Analysebeispiele aus dem Projekt Do-iT dargestellt (Abbildung 5.3 und Tabelle 5.3). Im Rahmen des Projektes war es möglich, die von den Partnern zur Verfügung gestellten Referenzdatenquellen mit den oben beschriebenen Methoden im Detail zu analysieren. Es handelt sich dabei um stationäre Verkehrsdatenerfassungssysteme (SES), die im Projektnetz von Do-iT installiert sind und von den Projektpartnern betrieben werden. Die SES erfassen in erster Li-

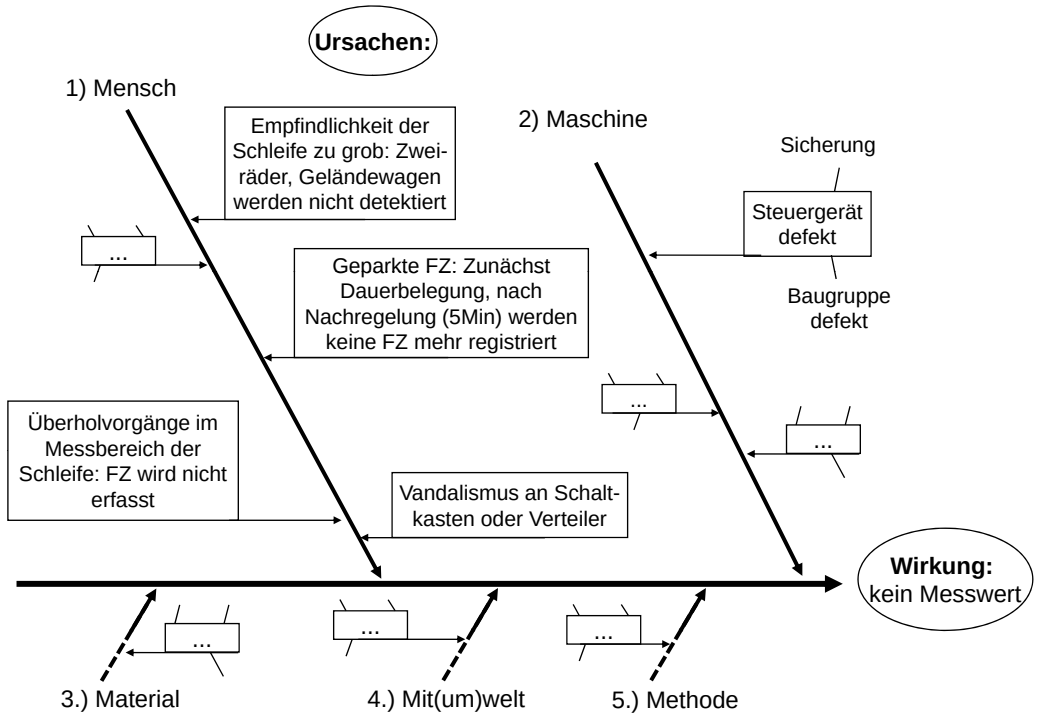


Abbildung 5.3: Auszug aus dem UWD zur Analyse der Verfügbarkeit von Daten der Schleifensensoren in Karlsruhe

nie Verkehrsstärken und die Geschwindigkeit des Verkehrsflusses. Im Rahmen des Projekts standen zum einen Induktionsschleifen zur Verfügung und andererseits sogenannte „Traffic-Eyes“, bei denen es sich um passive Infrarotsensoren handelt, die über der Fahrbahn installiert sind. Die Abbildung 5.4 zeigt zwei Beispiele für die beiden Arten von stationären Verkehrsdatenerfassungssysteme: Auf der linken Seite ist ein Traffic-Eye abgebildet und die rechte Seite zeigt zwei in den Fahrspuren verlegte Induktionsschleifen. Die in diesem Abschnitt beispielhaft aufgeführten Analyseergebnisse stammen aus der Analyse der Schleifensensoren der Stadt Karlsruhe.

Die Abbildung 5.3 stellt in Auszügen das Ergebnis der Analyse der Schleifensensoren in Karlsruhe bezüglich der Verletzung der Verfügbarkeit dar. Mögliche Ursachen für die unerwünschte Wirkung „kein Messwert“und damit eine Verletzung der Verfügbarkeit werden im Allgemeinen den sogenannten 5Ms, den Einflussfaktoren Mensch, Maschine, Material, Mit(um)welt und Methode zugeordnet. Die Darstellung erfolgt, wie in der Abbildung angedeutet, in Form von Wirkungspfeilen und ähnelt daher in der Anordnung einem Fischgerippe.

Wie an diesem Beispiel erkennbar ist, erscheinen die aufgedeckten Ursachen in vielen Fällen trivial, allerdings gibt es selten umfassende und systematische Dokumentationen der oft großen Anzahl möglicher Fehlerursachen. Die Stärke dieser Methodik liegt in der Bündelung des Expertenwissen. Dabei ist allerdings die richtige Zusammensetzung des Teams von großer Bedeutung.

Die verschiedenen Ursache-Wirkungs-Diagramme aller Bauteile des Systems, die je nach Aufwand und Zusammensetzung des Teams eine mehr oder weniger vollständige Sammlung an Ursachen für unterschiedliche unerwünschte Wirkungen darstellen, dienten im Anschluss als Grundlage zur Durchführung der FMEA bzw. der FMECA. Die Bestimmung der RPZ wurde hier ebenfalls durchgeführt. Die aufgestellten Diagramme dienten dabei als Gedächtnisstütze bei der Suche nach den Fehlermög-

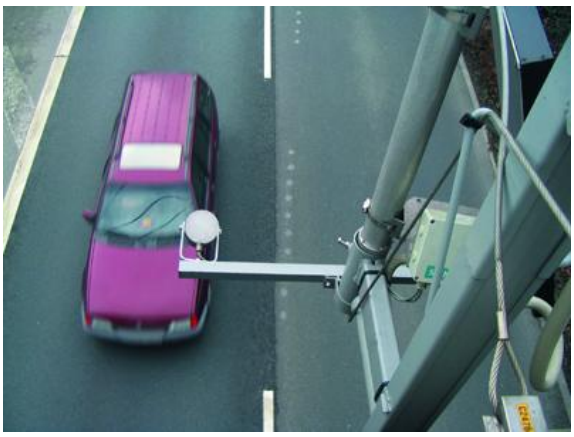
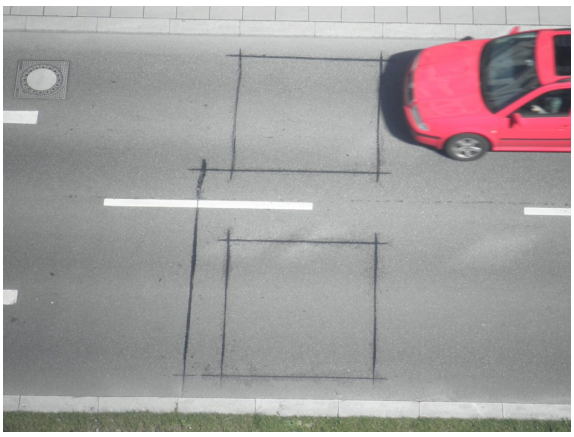



Abbildung 5.4: Stationäre Erfassungssysteme: Traffic-Eye (links, Quelle: http://www.siemens.com) und Induktionsschleife (rechts)

lichkeiten einzelner Bauteile. In der Tabelle 5.3 sind exemplarisch einige Ergebnisse der Untersuchung des Bauteils Induktionsschleife dargestellt. Die Beurteilung der Auswirkungen hat sich hierbei auf die Qualitätsmerkmale beschränkt, zukünftig sollen jedoch zusätzlich die Auswirkungen auf die im Qualitätsmodell definierten Parameter festgehalten werden.

Tabelle 5.3: Auszug aus der FMECA der Schleifensensoren in Karlsruhe (Quelle: Do-iT [2008a])

lfd. Nr.	Bauteil	Funktion	potenzielle Ausfallart	mögliche Ursachen	lokale Auswirkung	Auswirkung auf System u. Qualität	Erkennung	vorsorgliche Gegenmaßnahmen	A/B/E RPZ
1.1	Leiter-Schleife	Induktion	Bruch/Abriß	mech. Belastung; Baumaßnahmen; Alterung; Witterung	keine Induktion	keine Messdaten VE verletzt	Meldung an VR Status: braun	gute Vorplanung von Baumaßnahmen; exaktes Leitungskataster	4/7/2 56
1.2			Kurzschluss					keine	2/7/2 28
1.3			Wackel-kontakt		zeitweise keine Ind.	sporadisch keine Daten VE,KR verletzt	Statuswechsel grün/braun	keine	2/8/5 80
2.1	Kabel-muffe	Verbindung Schleife mit Zuleitung	...	...	...	...	...	...	...
...									

Der Auszug aus den FMECA-Ergebnissen zeigt die Probleme, die in dem Bauteil *Leiterschleife* auftreten können sowie deren Bewertung mit Hilfe der RPZ. Wie der Tabelle zu entnehmen ist, stellen auftretende Wackelkontakte in der Schleife die größten Probleme dar, insbesondere aufgrund der schlechten Erkennbarkeit. Im Hinblick auf die Entwicklung von Qualitätsverbesserungsmaßnahmen ist die gleichzeitige stichwortartige Dokumentation von möglichen Gegenmaßnahmen wichtig. Diese können dann nach und nach mit Rücksicht auf die Ressourcen im Detail ausgearbeitet und implementiert werden, wobei sich die Priorisierung an den RPZ orientieren sollte. Das vollständige FMECA-Formblatt als Ergebnis der Untersuchung des Subsystems „Induktionsschleife“ist im Anhang D dargestellt.

5.2.4 Quantitative Analyseverfahren

Unter *quantitativer Analyse* von Prozessen der Datenverarbeitung ist hier das Analysieren und Modellieren von Zusammenhängen der Datenqualität von Eingangs- und Ausgangsdaten zu verstehen. Ganz

konkret konzentriert sich dabei die Modellierung auf die Qualitätsparameter, die zur quantitativen Beschreibung der Datenqualität definiert wurden.

Der Schwerpunkt der vorliegenden Arbeit war die Suche nach einem quantitativen Analyseverfahren, welches zur Fortpflanzung von Datenqualität geeignet ist. Wie im vorigen Kapitel an praktischen Beispielen gezeigt wurde, verfügen künstliche neuronale Netze über das erforderliche Potenzial zur Modellierung von Datenqualität auf Parameterebene. Aufwendig gestaltet sich jedoch die Beschaffung einer ausreichenden Zahl geeigneter Lernbeispiele, insbesondere bei komplexeren Netzen mit zahlreichen Ein- und/oder Ausgangsgrößen. Daher ist eine Beschränkung auf die für die Qualitätsmodellierung relevanten Parameter anzustreben. Dabei muss jedoch auf mögliche Abhängigkeiten verschiedener Qualitätsparameter geachtet werden und es sind ggf. weitere Parameter mit in die Modellierung der Datenqualität aufzunehmen, die nur indirekt von Interesse sind.

Das Rechenverfahren von Wiltschko, wie es im Abschnitt 3.1.2 vorgestellt und an einem Beispiel erläutert wurde, ist auf konzeptioneller Ebene einsetzbar, sofern die Prozesse sich mit den Mitteln der Booleschen Algebra im Detail beschreiben lassen. Daher kann es in geeigneten Fällen weiterhin zur Modellierung der Datenqualität -in der Regel auf Merkmalsebene- zum Einsatz kommen. Insbesondere die Verfügbarkeit von Daten kann ggf. mit dem Verfahren dargestellt werden, da es sich hier meist um einen Erfüllungsgrad in Prozent handelt, der sich nach den Booleschen Rechenregeln fortpflanzen lässt. Im Gegensatz zu den anderen Qualitätsmerkmalen wird die Verfügbarkeit meist nur durch einen Parameter repräsentiert, was den trivialen Übergang von der Merkmals- auf die Parameterebene ermöglicht. Des Weiteren sind für die Verfügbarkeit einzelner Bauteile meist empirische Werte (z. B. als Erfahrungswerte oder Angaben des Herstellers) vorhanden, die zu deren Fortpflanzung eingesetzt werden können.

Beide Verfahren werden als Werkzeuge in das Qualitätsmanagementkonzept von Daten aufgenommen und können sich in manchen Fällen sinnvoll ergänzen. Das auf Boolescher Algebra basierende Verfahren ist -mit den genannten Einschränkungen- in der Regel nur in der Planungsphase eines DV-Prozesses sinnvoll einsetzbar. Die KNN können hingegen beliebig komplexe DV-Prozesse abbilden, sofern sie ausreichend dimensioniert sind und genügend Lernbeispiele zur Verfügung stehen oder generiert werden können. Eine Einschränkung auf die relevanten Parameter und ggf. die getrennte Betrachtung einzelner Teilprozesse ist daher sinnvoll. Die KNN bieten darüber hinaus die Möglichkeit nach dem erfolgreichen Training auch die Änderung einzelner Eingangsparameter zu simulieren. Damit ist die Überprüfung von einzelnen Qualitätssicherungsmaßnahmen bereits vor deren tatsächlichen Umsetzung möglich. Schließlich bietet die Echtzeitfähigkeit der KNN weitere Einsatzmöglichkeiten, beispielsweise zur Überwachung realer Prozesse.

Neben den bereits genannten Verfahren wurde für die Abbildung der Genauigkeit auch die Kovarianzfortpflanzung im Kapitel 3.1.3 erläutert. Zur Anwendung dieses Verfahrens ist jedoch ebenfalls die Kenntnis des funktionalen Zusammenhangs erforderlich. Ist dies nicht der Fall, so kann die Genauigkeit der Daten mit Hilfe der KNN fortgepflanzt werden, wie bereits im Kapitel 4.1 an einem einfachen Beispiel aus der Geodäsie gezeigt wurde. In geeigneten Fällen kann das Kovarianzfortpflanzungsgesetz eine gute Alternative sein, da damit auch die Bestimmung von Kovarianzen zwischen einzelnen Größen problemlos möglich ist. Daher ist diese Methodik weiterhin Bestandteil des Werkzeugkastens des Qualitätsmanagementkonzeptes und steht als quantitative Methode zur Prozessanalyse und Modellierung der Datenqualität zur Verfügung.

Weitere Methoden sind hier auch denkbar. So kann die Kovarianzfortpflanzung und damit die Modellierung der Genauigkeit bei unbekanntem formelmäßigem Zusammenhang auch durch eine geeignete Monte-Carlo-Simulation (Kapitel 3.2.3) ermittelt werden. Das Konzept bietet daher nur eine Auswahl geeigneter Rechen- bzw. Fortpflanzungsverfahren an und bleibt offen für weitere Verfahren.

5.2.5 Evaluierung

Eine regelmäßige Bestimmung und Beurteilung der aktuellen Datenqualität ist ein wesentlicher Bestandteil eines Qualitätsmanagementkonzeptes für Daten. Nur so kann ein belastbarer Nachweis für die Einhaltung der versprochenen Datenqualität erfolgen. Die Evaluierung dient damit auch der Bewertung von durchgeführten Maßnahmen zur Qualitätssicherung und -verbesserung.

Die Evaluierung der Qualität von Daten erfordert in der Regel Referenzinformationen, die einen Soll-Ist-Vergleich ermöglichen. Relative Änderungen der Qualität können mithilfe einer internen Referenz erfasst werden. Dabei werden die aktuelle ermittelten Qualitätsparameterwerte mit aus der Historie zu erwartenden Werten verglichen. Unerwartete Änderungen der Datenqualität können somit schnell erkannt werden. Deren mögliche Ursachen sind anschließend zu identifizieren und näher zu untersuchen bzw. zu bekämpfen (z. B. mithilfe einer FMEA).

Die Qualitätsparameter, die der Evaluierung der Daten dienen, sind oft erst zu diesem Zweck definiert worden und richten sich nicht zuletzt nach der Art der Referenzdaten, die zur Verfügung stehen oder mit vertretbarem Aufwand ermittelbar sind. In dem bereits mehrfach angeführten Projektbeispiel Do-iT wurde die Qualität von FPD-Trajektorien evaluiert. Dazu standen einige wenige GPS-gestützte Testfahrten zur Verfügung sowie die Zählwerte zahlreicher stationärer Verkehrsdatenerfassungssysteme im Projektstraßennetz. Die Eignung der Daten als Referenz wurde im Falle der GPS-Messungen durch langjährige Erfahrung bestätigt und konnte im Falle der SES im Rahmen einer Studienarbeit exemplarisch durch manuelles Nachzählen belegt werden (Karrer [2008]). Schließlich wurden die beiden Parameter *Durchdringung mit FPD* und *Rangkorrelation mit der Verkehrsstärke* definiert, wie sie in der Tabelle 5.1 definiert sind.

In Abbildung 5.2 ist beispielhaft eine Gegenüberstellung der Verkehrsstärken für einen Autobahnabschnitt im Untersuchungsgebiet nahe Karlsruhe dargestellt. Es handelt sich dabei um die in rot aufgetragene stundenweise aggregierte Referenzverkehrsstärke, die von einem stationären Schleifensensor über den Tag gezählt wurde und um die Anzahl FPD-Trajektorien, die im selben Zeitraum auf dem entsprechenden Straßenabschnitt berechnet werden konnten (in blau mit Kreisen dargestellt). Für die beiden dargestellten Kurven ergab sich eine Rangkorrelation von etwa 0.79. Die Durchdringung berechnet sich als Quotient aus der, an einer bestimmten Stelle, aus Mobilfunkdaten detektierten Anzahl Trajektorien und der Referenzverkehrsstärke. An dem in der Abbildung dargestellten Schleifensensor lag die Durchdringung während des Evaluierungszeitraumes im Durchschnitt bei $1.1\,\%$. Beide Parameter zusammen waren gut zur Evaluierung der Qualität der FPD-Trajektorien und damit zur Evaluierung wesentlicher Projektergebnisse geeignet. Das Potenzial der Mobilfunkortung konnte damit besser beschrieben und es konnten Schwachpunkte der Verfahren detektiert werden.

5.2.6 Qualitätssicherung und -verbesserung

Die Sicherung und Verbesserung der Datenqualität sind die zentralen Aufgaben in einem Qualitätsmanagementkonzept für Daten. Die Potenziale zur stetigen Verbesserung der Qualität sollten in regelmäßigen Abständen neu analysiert und ausgeschöpft werden. Dabei können beispielsweise die im Rahmen einer FMECA ermittelten RPZ als Orientierung dienen. In vielen Fällen ergeben sich mögliche Maßnahmen bereits unmittelbar bei der gemeinsamen Suche und Erörterung von Problemen und Fehlerfolgen, die in einzelnen Systemkomponenten oder Prozessschritten auftreten können. Das Wissen und Vorstellungsvermögen aller Mitglieder des Teams wird bei der Durchführung einer FMECA gebündelt und die Ideen

und Ansätze können direkt in der Runde diskutiert werden. In einem zweiten Schritt kann dann die konkrete Ausformulierung und Planung der Umsetzung der festgehaltenen Ideen zur Qualitätssicherung und -verbesserung erfolgen. Dabei können die Maßnahmen mit Blick auf den Zeithorizont in

- kurzfristige (direkt umsetzbare),
- mittelfristige (benötigen einigen Vorlauf) und
- langfristige (Umsetzung erst in einer der nächsten Systemgenerationen)

Maßnahmen eingeteilt werden. An die Umsetzung einzelner Maßnahmen schließt sich der Nachweis der Wirksamkeit an. Der Nachweis dient insbesondere der Rechtfertigung zusätzlicher Mittel für Material oder Personal, die in der Regel mit der Einführung von Qualitätssicherungs- und -verbesserungsmaßnahmen einher gehen.

Die Wirksamkeit der eingeführten Maßnahmen kann auf verschiedene Arten nachgewiesen werden. Das QM-Konzept bietet dazu zum einen die Möglichkeit, Qualitätsparameter vor und nach der Einführung einzelner Maßnahmen zu bestimmen und aus deren Änderung die Wirksamkeit der Maßnahmen zu quantifizieren. Zum anderen kann mithilfe einer erneuten Durchführung der FMECA eine Wirksamkeit anhand der Verkleinerung einzelner oder mehrerer RPZ nachgewiesen werden. Dabei handelt es sich allerdings nur bedingt um eine quantitativ beurteilbare Aussage, vielmehr kann die Wirksamkeit der Maßnahme auf die Fehlerfolge im Vergleich zu den RPZ anderer Fehlerfolgen erneut beurteilt und die Priorisierung aktualisiert werden.

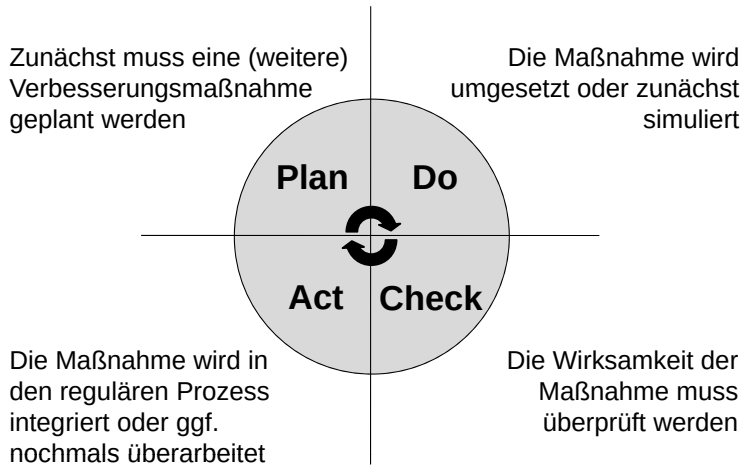


Abbildung 5.5: Mögliche Interpretation des PDCA-Zyklus für Prozesse der Datenverarbeitung

Die einzelnen Phasen des Zyklus der ständigen Qualitätsverbesserung sind in Abbildung 5.5 dargestellt und können hier wie folgt interpretiert werden:

Im ersten Schritt wird eine geeignete Maßnahme zur Verbesserung der Datenqualität geplant (*Plan*-Phase). Diese wird im zweiten Schritt umgesetzt oder im Falle sehr umfangreicher Maßnahmen zunächst

simuliert (*Do*-Phase) und im Anschluss auf ihre Wirksamkeit hin überprüft (*Check*-Phase). Ist die Maßnahme erfolgreich, so wird sie in den regulären Prozess integriert. Ist deren Wirksamkeit nicht wie gewünscht, so muss die Maßnahme gegebenenfalls überarbeitet oder ganz verworfen werden (*Act*-Phase). Nach dem erstmaligen Durchlauf beginnt der Zyklus wieder mit der Planungsphase und es werden weitere Ansatzpunkte verfolgt und Maßnahmen entwickelt oder überarbeitet. Dieser Kreislauf kann nahezu beliebig fortgeführt werden, in aller Regel wird jedoch eine Verlangsamung eintreten, da das Potenzial für Verbesserungen mit der Zeit abnimmt. Aufgrund sich naturgemäß ändernder Rahmenbedingungen, kommt der kontinuierliche Verbesserungsprozess jedoch selten und wenn, dann nur temporär zum Erliegen. Die Wiederholung der Analysen in regelmäßigen Abständen ist daher sehr wichtig, um den aktuellen Entwicklungen Rechnung zu tragen.

Die Simulation einzelner Maßnahmen zur Verbesserung der Datenqualität und damit die Abschätzung von Kosten und Nutzen bereits vor deren Umsetzung kann Ressourcen sparen. Die Modellierung der Datenqualität in einzelnen Prozessen mit KNN kann eine derartige Abschätzung ermöglichen. Dazu ist eine Abschätzung der durch die Maßnahme hervorgerufenen Änderungen in den Eingangswerten des Prozesses erforderlich. Das gut trainierte KNN ist anschließend in der Lage, die zu erwartenden Ausgangswerte zu simulieren und ermöglicht somit die Beurteilung der Maßnahme vor der tatsächlichen Umsetzung der Maßnahme. Mit KNN kann des Weiteren mit Hilfe geeigneter Qualitätsparameter eine Prozessüberwachung und damit auch eine Qualitätssicherung in Echtzeit erfolgen.

5.3 Zusammenfassung

Das vorgestellte Konzept bietet die Möglichkeit, Datenqualität ganzheitlich zu behandeln. Es werden alle wesentlichen Aspekte zum praktischen Qualitätsmanagement von Daten in einem übersichtlichen Schaubild zusammengefasst. Eine Zusammenstellung in ähnlicher Form ist derzeit in der Literatur nicht zu finden, trotz der besonderen Bedeutung des Datenqualitätsmanagements als Bestandteil eines umfassenden, unternehmerischen Qualitätsmanagement. Einerseits ist die vorliegende und die über längere Zeiträume garantierbare Qualität von Daten der wesentliche Verkaufsfaktor auf dem Datenmarkt und bietet Wettbewerbsvorteile. Andererseits kann eine unbemerkte Verschlechterung der Qualität unter Umständen fatale Folgen für die Datenanwendungen und damit für die daraus resultierenden Entscheidungen haben (Planungsfehler, Navigationsfehler, etc.).

Das Konzept bietet zunächst einmal den Rahmen für eine einheitliche und umfassende Beschreibung der Qualität von gemessenen oder weiterverarbeiteten Daten und ermöglicht damit überhaupt erst das Messen und Beurteilen der Datenqualität. Es werden qualitative Analyseverfahren vorgeschlagen und erläutert, die der Schaffung einer Wissensbasis dienen und damit die Planung und Entwicklung von Qualitätssicherungs- und -verbesserungsmaßnahmen ermöglichen. Insbesondere die FMECA hat sich bereits als Analysewerkzeug bewährt. Ebenso kann sie der Beurteilung der Wirksamkeit geplanter oder bereits umgesetzter Maßnahmen dienen.

Es wird neben einem bereits bestehenden, in der Literatur beschriebenen Verfahren zur quantitativen Analyse von Datenqualität, mit den künstlichen neuronalen Netzen ein neues Verfahren vorgeschlagen. Dieses Verfahren ermöglicht die Modellierung von Qualitätsübergängen in Prozessen auf Parameterebene und damit beispielsweise auch die Simulation von Qualitätsverbesserungsmaßnahmen. Deren Wirksamkeit kann somit vor der Umsetzung zunächst abgeschätzt werden. Die grundsätzliche Tauglichkeit des Verfahrens wurde im Kapitel 4 an zwei Beispielen gezeigt und kritisch beurteilt. Die entscheidenden Vorteile liegen in der Möglichkeit, Qualität auf Parameterebene zu modellieren sowie in der Echtzeitfähigkeit der KNN.

Anhand der definierten Qualitätsparameter und den entwickelten Messmethoden zu deren quantitativen Bestimmung kann jederzeit eine Evaluierung der Qualität des Datenbestandes erfolgen. Damit ist durch eine regelmäßige Evaluierung auch langfristig eine entsprechende Datenqualität zu gewährleisten. Probleme werden so schnell erkannt und können oft bereits in den bestehenden FMEA-Formblättern, in denen bereits viele Fehlermöglichkeiten festgehalten wurden, identifiziert werden. Falls dies nicht der Fall ist, muss gegebenenfalls mit einer Ursachen-Wirkungs-Analyse eine weitere Untersuchung angeregt werden, die in der Regel zur schnellen Identifizierung des Problems führt. Daraus lassen sich oft bereits Gegenmaßnahmen ableiten.

Das Konzept ist sehr flexibel anwendbar, jederzeit erweiterbar und lässt Freiräume für die Einbindung weiterer Methoden und Verfahren. Damit kann es einen wichtigen Teil eines unternehmerisches Qualitätsmanagements -drei weit verbreitete Qualitätsmanagementsysteme wurden bereits im Abschnitt 2.4 vorgestellt- darstellen.

6 Zusammenfassung und Ausblick

In der vorliegenden Arbeit ist es gelungen, künstliche neuronale Netze erfolgreich für die Beschreibung von Datenqualität in Prozessen einzusetzen. Damit wurde mit diesem universalen Verfahren eine neue Methode zur Modellierung von Abhängigkeiten zwischen der Qualität der Eingangs- und Ausgangsdaten gefunden, die insbesondere eine Behandlung der Qualität auf Parameterebene ermöglicht. Das bisher bekannte Verfahren von Wiltschko war dazu nicht in der Lage. Die Modellierung von Qualität war bisher auf Merkmalserfüllungsgrade und damit auf die abstrakte Ebene der Qualitätsmerkmale beschränkt. Einzig die Genauigkeit hat sich als isoliertes Merkmal mit Hilfe des seit langem bekannten Verfahren der Kovarianzfortpflanzung bereits auf Parameterebene behandeln lassen.

Hier ist ein entscheidender Durchbruch gelungen, der zukünftig neue Möglichkeiten im Hinblick auf die Planung von Prozessen oder die Simulation von Änderungen eröffnet. Die KNN sind dabei als wichtige Ergänzung der bestehenden Verfahren zu sehen und wurden in die Sammlung der zur Verfügung stehenden Werkzeuge mit aufgenommen. Mit dem untersuchten Verfahren auf Basis der KNN ist nun auch die Überwachung und Fortpflanzung der Datenqualität in realen Prozessen der Datenverarbeitung in Echtzeit möglich.

Mit Hilfe normierter Testbeispiele konnten KNN für die Fortpflanzung von Genauigkeitsparametern mit hinreichender Güte trainiert werden. Die Hinzunahme und Identifikation echter Nullen in den Eingangs- und Ausgangswerten ermöglichte zusätzlich die Darstellung von nicht verfügbaren Daten. Durch die Definition von Datensätzen zur Unterscheidung von Mängeln in der Verfügbarkeit und der Vollständigkeit von Daten konnte auch die Fortpflanzung der Datenvollständigkeit exemplarisch gezeigt werden. Diese wurde durch das Anfügen einer weiteren Stelle im Eingangs- und Ausgangsdatenvektor ermöglicht, der mit dem Wert 1 (Datensatz ist verfügbar) oder dem Wert 0 (Datensatz ist nicht verfügbar) belegt sein kann. Es können für jeden weiteren, definierten Datensatz in den zu modellierenden Daten, auch weitere Stellen zur Abbildung angefügt werden. Die Größe der Ein- und Ausgangsvektoren ist grundsätzlich nur durch die steigende erforderliche Komplexität des Netzes und der damit verbundenen größeren Anzahl an Lernbeispielen sowie der höheren Rechnerkapazität beschränkt. Liegt ein geeignetes Datenmodell vor, so kann durch Hinzunahme weiterer 0/1-Stellen auch die Konsistenz der Daten beschrieben werden. Dabei erfolgt die Unterscheidung der Parameter der Vollständigkeit und der Parameter der Konsistenz durch die Lage im Eingangs- bzw. Ausgangsvektor.

Zur Beurteilung der Korrektheit sind stets Referenzdaten erforderlich, die in aller Regel nicht kontinuierlich in Echtzeit zur Verfügung stehen. Da der größte Vorteil der KNN-Methode jedoch in der Echtzeitfähigkeit liegt, erschien die Untersuchung der Modellierbarkeit von Korrektheitsparametern hier nicht sinnvoll. Des Weiteren besteht funktional eine große Ähnlichkeit zur Modellierung der Konsistenz, so dass grundsätzlich eine Modellierung in gleicher Art und Weise durch Erweiterung der Datenvektoren möglich ist.

Das Merkmal Aktualität wurde bei den Untersuchungen nicht betrachtet, da sich diese Arbeit auf die Untersuchung statischer KNN beschränkt hat. Für die Modellierung von sich zeitlich schnell ändernder Größen muss die zeitliche Reihenfolge der Lernbeispiele beim Trainieren der KNN berücksichtigt werden, daher sind dazu dynamische Netze erforderlich.

Die Fortpflanzung von Datenqualität in sehr komplexen Prozessen wurde am Beispiel der Generierung von Fahrzeugtrajektorien aus Mobilfunkdaten gezeigt. Zunächst ist es erforderlich, die wichtigsten Qualitätsparameter zu identifizieren, die einen Einfluss auf die Qualität der Ausgangsgröße haben. Im Beispiel wurden insgesamt sieben Parameter ausgewählt, mit denen die Prognose der Querabweichung in zufriedenstellender Genauigkeit gelang. Die Eignung der KNN konnte damit auch exemplarisch an einem realen Prozess gezeigt werden. Nach dem Training waren die KNN in der Lage, die Qualität von Trajektorien aus den Eingangsgrößen in Bruchteilen von Sekunden vorherzusagen. Damit ist auch der Einsatz in Echtzeit, d. h. in diesem Beispiel unmittelbar der eigentlichen Trajektorienberechnung nachgeschaltet, problemlos möglich. In dieser Arbeit wurden aus datentechnischen Gründen keine gemessenen Signalstärken als Eingangsparameter für die Qualitätsfortpflanzung verwendet. Dies erscheint jedoch sehr vielversprechend und sollte daher in weiteren Arbeiten noch untersucht werden.

Trotz der beiden ausführlich behandelten Beispiele bleiben noch einige Fragen im Bezug auf die Anwendbarkeit von KNN in diesem Zusammenhang offen. Bislang noch nicht untersucht wurde unter anderem

- die Abbildbarkeit zeitlich veränderlicher Qualitätsparameter, die dem Merkmal Aktualität zugeordnet werden können mit Hilfe dynamischer KNN,
- das Mischen unterschiedlicher Transferfunktionen in der selben Neuronenschicht, um den teils sehr inhomogenen Eingangs- und Ausgangsdaten Rechnung zu tragen,
- das Potenzial weiterer, frei definierter Transferfunktionen in den Netzen,
- die Verwendung rückgekoppelter Netze oder
- die Kombination der KNN mit weiteren Verfahren und Methoden (z. B. Fuzzy-Verfahren) sowie
- die Kombination von Parametern verschiedener Merkmale in den Ein- und Ausgangsdaten in realen Beispielen.

Diese Ansatzpunkte können helfen, die Anwendung künstlicher neuronaler Netze im Bereich der Datenqualität weiter zu vereinfachen und noch bessere Ergebnisse zu erzielen. Grundsätzlich muss jedoch klar sein, dass die Modellierung von Datenqualität nicht mit beliebiger Güte möglich ist. Viele Qualitätsparameterwerte stellen nur Größenordnungen dar, deren Aussagekraft nicht überschätzt werden darf. Standardabweichungen beschränken sich immer auf die Angabe eines Bereichs, in dem die beschriebene Größe mit einer gewissen Wahrscheinlichkeit liegt (z. B. $68\,\%$ oder $95\,\%$) und die Korrektheit wird immer trotz des binären Charakters nur mit einer gewissen Wahrscheinlichkeit den wahren Wert angeben. Daher ist die Präzision, mit der das KNN die gesuchten Ergebnisse angeben soll, kritisch einzuschätzen. Aus vagen Angaben der Qualität von Eingangsgrößen können auch nur vage Ausgangsgrößen ermittelt werden, unabhängig von der Komplexität des KNN, der Anzahl Iterationen oder der Menge an Lernbeispielen, die zur Verfügung steht.

Das bestehende Qualitätsmodell hat sich bewährt und wurde daher in dieser Arbeit nicht in Frage gestellt. Durch dessen Flexibilität kann insbesondere auch auf die durch den Kunden gewünschten Parameter der Datenqualität eingegangen werden. Die Definition und Auswahl der Parameter kann sich an den Anforderungen orientieren, eine einfache Zuordnung zu den vorgegebenen Qualitätsmerkmalen ist erfahrungsgemäß fast immer möglich.

Mit dem eingeführten Qualitätsmanagementkonzept konnten alle relevanten Bestandteile der datenqualitätsbezogenen Tätigkeiten in einen Zusammenhang gestellt werden. Die neue Methode zur Fortpflanzung von Datenqualität auf Basis von KNN konnte dabei in diesen Rahmen des Datenqualitätsmanagements eingegegliedert werden. Das Konzept kann unter anderem der Erläuterung der gesamten

Problematik dienen und das Bewusstsein für die Notwendigkeit der einzelnen Arbeitsschritte im Qualitätsmanagement schaffen. Mit dem Konzept wird nicht zuletzt die Kundenanforderung in den Mittelpunkt der qualitätsbezogenen Arbeiten gerückt. Eine übersichtliche und umfassende Darstellung der Tätigkeiten mit dem Fokus auf die Datenqualität schafft Transparenz und ermöglicht die Identifikation der Schnittstellen mit dem Auftraggeber.

Einzelne Bestandteile des Konzeptes wurden dargestellt und deren Funktionsweise an realen Projektbeispielen gezeigt. So wurden für stationäre Induktionsschleifen zur Verkehrsdatenerfassung Ergebnisse aus den Ursachen-Wirkungs-Diagramm gezeigt sowie die Vorgehensweise bei der FMECA anhand des Formblatts für das Bauteil Induktionsschleife erläutert. Diese Methoden spielen eine entscheidende Rolle bei der Bündelung von Fachwissen und der Suche nach weiterem Potenzial zur Verbesserung der Datenqualität.

Wünschenswert ist hier jedoch noch eine klarere Definition der Schnittstellen zwischen dem Qualitätsmanagementkonzept zur Behandlung von Datenqualität und dem übergeordneten unternehmerischen Qualitätsmanagement. Insbesondere die monetären und personellen Aspekte wurden bei den bisherigen Ausführungen vollständig ausgeblendet. Vor der Umsetzung des Konzeptes müssen diese betriebswirtschaftlichen Aspekte zunächst ausreichend untersucht werden, nicht zuletzt um die Wirtschaftlichkeit des derartigen Vorgehens nachweisen zu können.

Literaturverzeichnis

[Ambrosius 2008] AMBROSIUS, J.-P.: *Monte Carlo Simulation*. Vortrag im Rahmen des Stochastik Seminars von PD Dr. Flavius Guias im Sommersemester 2008 an der TU Dortmund, 2008. – URL `http://www.mathematik.uni-dortmund.de/lsi/fguias/monte_carlo.pdf`. – letzter Zugriff: 26.04.2010

[Benning 2010] BENNING, W.: *Statistik in Geodäsie, Geoinformation und Bauwesen*. Bd. 3. überarbeitete und erweiterte Auflage. Heidelberg : Wichmann, 2010

[Berkhahn u. a. 2010] BERKHAHN, V. ; BERNER, F. ; HIRSCHNER, J. ; KUTTERER, H. ; REHR, I. ; RINKE, N. ; SCHWEITZER, J. ; SCHWIEGER, V.: Effizienzoptimierung und Qualitätssicherung ingenieurgeodätischer Prozesse im Hochbau. In: *Der Bauingenieur* Nr. 11 (2010), S. 491–501

[Bill 1999] BILL, R.: *Grundlagen der GEO-Informationssysteme - Band 1*. 4. völlig neubearb. und erw. Aufl. Heidelberg : Wichmann, 1999

[Bill und Zehner 2001] BILL, R. ; ZEHNER, M. L.: *Lexikon der Geoinformatik*. Heidelberg : Wichmann, 2001

[Capurro 1978] CAPURRO, R.: *Information: Ein Beitrag zur etymologischen und ideengeschichtlichen Begründung des Informationsbegriffs*, Universität Düsseldorf, Dissertation, 1978

[Cardoso u. a. 1998] CARDOSO, J. u. a. ; CARDOSO, Janette (Hrsg.) ; CAMARGO, Heloisa (Hrsg.): *Fuzziness in Petri Nets*. Heidelberg : Physica Verlag, 1998 (Studies in fuzziness and soft computing)

[Claus und Schwill 2006] CLAUS, V. ; SCHWILL, A.: *Duden Informatik A-Z*. Mannheim, Leipzig, Wien, Zürich : Dudenverlag, 2006

[Czommer 2000] CZOMMER, R.: *Leistungsfähigkeit fahrzeugautonomer Ortungsverfahren auf der Basis von Map-Matching-Techniken*, Universität Stuttgart, Institut für Anwendungen der Geodäsie im Bauwesen (IAGB), Dissertation, 2000

[Datacom 2010] DATACOM BUCHVERLAG GMBH: *Online-Lexikon für Informationstechnologie*. Peterskirchen : Datacom Buchverlag GmbH, 2010. – URL `http://www.itwissen.info`. – letzter Zugriff: 10.03.2010

[Deming 1982] DEMING, W. E.: *Out of the crisis*. Cambridge, MA : MIT Press, 1982

[DIN 25419 1985] DIN DEUTSCHES INSTITUT FÜR NORMUNG: *Ereignisablaufanalyse*. Berlin : Beuth, 1985

[DIN 25424 Teil 1 1981] DIN DEUTSCHES INSTITUT FÜR NORMUNG: *Fehlerbaumanalyse - Methode und Bildzeichen*. Berlin : Beuth, 1981

[DIN 25424 Teil 2 1990] DIN DEUTSCHES INSTITUT FÜR NORMUNG: *Fehlerbaumanalyse - Handrechenverfahren zur Auswertung eines Fehlerbaums*. Berlin : Beuth, 1990

[DIN 25448 1990] DIN DEUTSCHES INSTITUT FÜR NORMUNG: *Ausfalleffektanalye (FMEA)*. Berlin : Beuth, 1990. – zurückgezogen und ersetzt durch DIN EN 60812

[DIN EN 60812 2006] DIN DEUTSCHES INSTITUT FÜR NORMUNG: *Analysetechniken für die Funktionsfähigkeit von Systemen - Verfahren für die Fehlerzustandsart- und -auswirkungsanalyse (FMEA)*. Berlin : Beuth, 2006

[DIN EN ISO 19113 2005] DIN DEUTSCHES INSTITUT FÜR NORMUNG: *Geoinformation - Qualitätsgrundsätze*. Berlin : Beuth, 2005

[DIN EN ISO 9000 2005] DIN DEUTSCHES INSTITUT FÜR NORMUNG: *Qualitätsmanagementsysteme - Grundlagen und Begriffe*. Berlin : Beuth, 2005

[DIN EN ISO 9001 2008] DIN DEUTSCHES INSTITUT FÜR NORMUNG: *Qualitätsmanagementsysteme - Anforderungen*. Berlin : Beuth, 2008

[Do-iT 2006] DO-IT: *Nutzerspezifische Qualitätsangaben*. Interner Do-iT Projektbericht, 2006

[Do-iT 2007] DO-IT: *Analyseverfahren zur Bewertung der Datenqualität*. Interner Do-iT Projektbericht, 2007

[Do-iT 2008a] DO-IT: *Analyseergebnisse der vorliegenden Informationsstrukturen und -Ketten*. Interner Do-iT Projektbericht, 2008

[Do-iT 2008b] DO-IT: *Qualitätsbewertung Map-Matching gestützter Positionsbestimmungsverfahren*. Interner Do-iT Projektbericht, 2008

[Do-iT 2009a] DO-IT: *Abschlussbericht zur Evaluierung des Projektes Do-iT*. Interner Do-iT Projektbericht, 2009

[Do-iT 2009b] DO-IT: *Qualitätsbewertung Map-Matching gestützter Positionsbestimmungsverfahren auf Abis und A-Ebene sowie Evaluierung der Positionsbestimmung durch höherwertig Ortungsverfahren*. Interner Do-iT Projektbericht, 2009

[Evans 1975] EVANS, D. H.: Statistical Tolerancing: The State of the Art, Part III: Shifts and Drifts. In: *Journal of Quality and Technology* (1975)

[Fank 2001] FANK, M.: *Einführung in das Informationsmanagement*. 2. Auflage. München : Oldenbourg, 2001

[Fischer und Hofer 2008] FISCHER, P. ; HOFER, P.: *Lexikon der Informatik*. 14. überarbeitete Auflage. Berlin, Heidelberg : Springer, 2008

[Frank 2010] FRANK, J.: *Untersuchung der Anwendbarkeit künstlicher neuronaler Netze bei der Beschreibung von ingenieurgeodätischen Prozessen*, Universität Stuttgart, Institut für Anwendungen der Geodäsie im Bauwesen (IAGB), Diplomarbeit, 2010. – unveröffentlicht

[Frankfurter Rundschau 2008] FRANKFURTER RUNDSCHAU: *701.000 Flugbewegungen und 260 Klagen*. Frankfurter Rundschau, 2008. – URL `http://www.fr-online.de`. – letzter Zugriff: 24.03.2010

[Gabler Verlag 2010] GABLER VERLAG: *Gabler online Wirtschaftslexikon - Die ganze Welt der Wirtschaft*. Wiesbaden : Gabler Verlag; Springer Fachmedien GmbH, 2010

[Hagan u. a. 1996] HAGAN, M. T. ; DEMUTH, H. B. ; BEALE, M.: *Neural Network Design*. USA : PWS Publishing Company, division of Thomson Learning, 1996

[Hake u. a. 2002] HAKE, G. ; GRÜNREICH, D. ; MENG, L.: *Kartographie*. 8. völlig neubearb. und erw. Auflage. Berlin : de Gruyter, 2002

[Heine 1999] HEINE, K.: *Beschreibung von Deformationsprozessen durch Volterra- und Fuzzy-Modelle sowie neuronale Netze*, Technische Universität Braunschweig, Fachbereich Bauingenieurwesen, Dissertation, 1999

[Hengartner und Theodorescu 1978] HENGARTNER, W. ; THEODORESCU, R.: *Einführung in die Monte-Carlo-Methode*. München, Wien : Carl Hanser, 1978

[Hornik u. a. 1989] HORNIK, K. M. ; STINCHCOMBE, M. ; WHITE, H.: Multilayer feedforward networks are universal approximators. In: *Neural Networks* Vol. 2 (1989), S. 359–366

[Horton u. a. 1998] HORTON, G. ; KULKARNI, V. G. ; NICOL, D. M. ; TRIVEDI, K. S.: *Fluid stochastic Petri Nets: Theory, applications and solution techniques*. European Journal of Operational Research, S. 184-201, Elsevier Science B.V., 1998

[Höpcke 1980] HÖPCKE, W.: *Fehlerlehre und Ausgleichungsrechnung*. Berlin, New York : de Gruyter, 1980

[ISO TS 19138 2006] ISO INTERNATIONAL ORGANIZATION FOR STANDARDIZATION: *Geographic information - Data quality measures*. Geneva : ISO copyright office, 2006

[Kamiske und Brauer 2008] KAMISKE, G. F. ; BRAUER, J.-P.: *Qualitätsmanagement von A-Z*. Bd. 6. Auflage. München : Carl Hanser, 2008

[Karabork u. a. 2008] KARABORK, H. ; BAYKAN, O. K. ; ALTUNTAS, C. ; YILDZ, F.: *Estimation of unknown height with artificial neural*

network on digital terrain model. Proceedings on isprs International Society for Photogrammetry and Remote Sensing 2008, Bejing, China, 2008

[Karrer 2008] KARRER, K.: *Untersuchungen der Qualität stationärer Erfassungssysteme auf der A81*. Studienarbeit am IAGB, Universität Stuttgart, 2008. – nicht veröffentlicht

[Kinnebrock 1992] KINNEBROCK, W.: *Neuronale Netze - Grundlagen, Anwendungen, Beispiele*. München, Wien : Oldenbourg, 1992

[Laufer 2008] LAUFER, R.: *Usage of a quality management concept for data exemplified with stationary traffic data acquisition systems*. Proceedings on 4th International Symposium Networks for Mobility, Stuttgart, 2008

[Leandro u. a. 2007] LEANDRO, R. F. ; SILVA, C. a. U. da ; SEGANTINE, P. C. L. ; SANTOS, M. C.: Processing GPS Data with Neural Networks. In: *GPS World* 18 (2007), Nr. 9, S. 60–65

[Leica 2009] LEICA: *Leica TPS1200+ Serie - Technische Daten*. Leica-Geosystems AG, 2009. – URL `http://www.leica-geosystems.de`. – letzter Zugriff: 24.05.2010

[Lexis 2010] LEXISNEXIS DEUTSCHLAND GMBH: *Lexikon der Unternehmensführung - Organisation*. LexisNexis Deutschland GmbH, 2010. – URL `http://www.steuerlinks.de`. – letzter Zugriff: 08.03.2010

[Matlab® 2010] HAGAN, M. T. ; DEMUTH, H. B. ; BEALE, M.: *Neural Network Toolbox™ 6 User´s Guide*. The MathWorks^{TM}, 2010. – URL `http://www.mathworks.com/access/helpdesk/help/pdf_doc/nnet/nnet.pdf`. – letzter Zugriff: 15.04.2010

[McCulloch und Pitts 1943] MCCULLOCH, W. ; PITTS, W.: *A logical calculus of the ideas immanent in nervous activity*. Bulletin of Mathematical Biophysics, Vol. 5, S. 115-133, 1943

[Metropolis und Ulam 1949] METROPOLIS, N. ; ULAM, S.: The Monte Carlo Method. In: *Journal of the American Statistical Association* 44 (1949), Nr. 247, S. 335–341

[Meyna 1982] MEYNA, A.: *Einführung in die Sicherheitstheorie*. München, Wien : Hanser, 1982

[Minsky und Papert 1969] MINSKY, S. ; PAPERT, M.: *Perceptrons*. Cambridge, MA : MIT Press, 1969

[Müller und Tietjen 2000] MÜLLER, D. H. ; TIETJEN, T.: *FMEA-Praxis*. München, Wien : Hanser, 2000

[Neuner und Kutterer 2010] NEUNER, H. ; KUTTERER, H.: *Ingenieurvermessung10 - Beiträge zum 16. internationalen Ingenieurvermessungskurs, München*. Kap. Modellselektion in der ingenieurgeodätischen Deformationsanalyse, Berlin : Wichmann, 2010

[Otto 1995] OTTO, P.: Identifikation nichtlinearer Systeme mit Künstlichen Neuronalen Netzen. In: *at-Automatisierungstechnik* 43 (1995), Nr. 2, S. 62–68

[PAS 1071 2007] DIN DEUTSCHES INSTITUT FÜR NORMUNG: *Qualitätsmodell für die Beschreibung von Geodaten*. Berlin : Beuth, 2007

[Peters 2008] PETERS, R.: *Künstliche neuronale Netze zur Beschreibung der Hydrodynamischen Prozesse für den Hochwasserfall unter Berücksichtigung der Niederschlags-Abfluss-Prozesse im Zwischeneinzugsgebiet*, Technische Universität Dresden, Institut für Hydrologie und Meteorologie, Dissertation, 2008

[Pieruschka 1963] PIERUSCHKA, E.: *Principles of Reliability*. Englewood Cliffs, N. J. : Prentice-Hall, 1963

[Poddig und Sidorovitch 2001] PODDIG, T. ; SIDOROVITCH, I.: *Handbuch Data Mining im Marketing*. Kap. Künstliche neuronale Netze: Überblick, Einsatzmöglichkeiten und Anwendungsprobleme, S. 363–402, Braunschweig, Wiesbaden : Vieweg, 2001

[Ramm und Schwieger 2008] RAMM, K. ; SCHWIEGER, V.: Mobile Positioning for Traffic State Acquisition. In: *Journal of Location Based Services* (2008)

[Rosenblatt 1961] ROSENBLATT, F.: *Principles of Neurodynamics*. Washington D.C. : Spartan Press, 1961

[Rothlauf 2001] ROTHLAUF, J.: *Total Quality Management*. München : Oldenbourg, 2001

[Rummelhard und McClelland 1986] RUMMELHARD, D. E. ; MCCLELLAND, J. L.: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Vol. 1, Cambridge, MA : MIT Press, 1986

[Sachs und Hedderich 2006] SACHS, L. ; HEDDERICH, J.: *Angewandte Statistik*. 12. Auflage. Berlin, Heidelberg, New York : Springer, 2006

[Schweitzer und Schwieger 2011] SCHWEITZER, J. ; SCHWIEGER, V.: Modeling of Quality for Engineering Geodesy Processes in Building Constructions. In: *Journal of Applied Geodesy* Berlin, New York : de Gruyter (2011). – eingereicht

[Sobol 1991] SOBOL, I. M.: *Die Monte-Carlo-Methode*. 4. Auflage. Berlin : Deutscher Verlag der Wissenschaften, 1991

[Stopp 1973] STOPP, F.: *Lehr- und Übungsbuch Mathematik: Band IV*. Kap. Monte-Carlo-Methoden, S. 409ff, Thun : Harry Deutsch, 1973

[Tennant 2001] TENNANT, G.: *Six Sigma: SPC and TQM in Manufacturing and Services*. Hampshire : Gower Publishing Ltd., 2001

[Toutenburg und Knöfel 2008] TOUTENBURG, H. ; KNÖFEL, P.: *Six Sigma*. Berlin, Heidelberg : Springer, 2008

[Trimble 2010] TRIMBLE: *Datenblatt Trimble S8 Total Station*. Trimble Navigation Limited, 2010. – URL `http://www.trimble.com`. – letzter Zugriff: 24.05.2010

[VDI 4001 Blatt 2 2006] VDI VEREIN DEUTSCHER INGENIEURE: *Terminologie der Zuverlässigkeit*. Berlin : Beuth, 2006

[VDI 4003 2007] VDI VEREIN DEUTSCHER INGENIEURE: *Zuverlässigkeitsmanagement*. Berlin : Beuth, 2007

[VDI 4008 Blatt 2 1998] VDI VEREIN DEUTSCHER INGENIEURE: *Boolesches Modell*. Berlin : Beuth, 1998

[VDI 4008 Blatt 4 2008] VDI VEREIN DEUTSCHER INGENIEURE: *Methoden der Zuverlässigkeit - Petri-Netze*. Berlin : Beuth, 2008

[VDI 4008 Blatt 6 1999] VDI VEREIN DEUTSCHER INGENIEURE: *Monte-Carlo-Simulation*. Berlin : Beuth, 1999

[Völz 1999] VÖLZ, H.: *Das Mensch-Technik-System: physiologische, physikalische und technische Grundlagen - Software und Hardware*. Renningen-Malmsheim : expert-Verlag; Wien : Linde, 1999

[Wahrig 2002] WISSEN MEDIA VERLAG GMBH: *Wahrig Deutsches Wörterbuch*. Bd. 7. vollst. neu bearbeitete und akt. Auflage. Gütersloh, München : Wissen Media Verlag GmbH, 2002

[Wiltschko 2004] WILTSCHKO, T.: *Sichere Information durch infrastrukturgestützte Fahrerassistenzsysteme zur Steigerung der Verkehrssicherheit an Straßenknotenpunkten*, Universität Stuttgart, Institut für Anwendungen der Geodäsie im Bauwesen (IAGB), Dissertation, 2004

[Wiltschko und Kaufmann 2005] WILTSCHKO, T. ; KAUFMANN, T.: *Modellierung und Bewertung von Prozessen in der Geodatenverarbeitung*. AGIT Symposium und Fachmesse angewandte Geoinformatik, Universität Salzburg, 2005

[Wiltschko und Möhlenbrink 2005] WILTSCHKO, T. ; MÖHLENBRINK, W.: *Save Information for Telematic Applications*. HITS 5^{th} European Congress and Exhibition on Intelligent Transport Systems and Services, Hannover, 2005

[Wiltschko u. a. 2007] WILTSCHKO, T. ; SCHWIEGER, V. ; MÖHLENBRINK, W.: *Acquisition of traffic state information by mobile phone positioning*. Proceedings on 6^{th} European Congress on ITS, Aalborg, Denmark, 2007

[Winkler 2006] WINKLER, P.: *Computer Lexikon*. München : Markt und Technik Verlag, 2006

[Zell 1997] ZELL, A.: *Simulation Neuronaler Netze*. Bd. 2. Auflage, unveränderter Nachdruck. München, Wien : Oldenbourg, 1997

[Zollondz 2002] ZOLLONDZ, H.-D.: *Grundlagen Qualitätsmanagement*. 2. Auflage. München : Oldenbourg, 2002

[Zurhausen 2002] ZURHAUSEN, M. S.: *Organisation*. 3. Auflage. München : Vahlen, 2002

Glossar

Adaline: Spezieller einstufiger vorwärts gerichteter Typ eines künstlichen neuronalen Netzes aus den Anfängen der KNN-Theorie; Ein- und Ausgangswerte folgen der Signumfunktion und sind entweder +1 oder -1.

Aktivierungsfunktion: Die Aktivierungsfunktion berechnet aus der Netzeingabe den Ausgang, der den Eingang der Ausgabefunktion darstellt. In der Regel handelt es sich dabei um den Ausgang des Neuronalen Netzes, da als Ausgabefunktion meist die Identitätsfunktion Verwendung findet. Es gibt drei häufig verwendete Arten von Aktivierungsfunktionen: Schwellwertfunktion, Lineare Funktion, Sigmoide Funktion.

Aktualität (AK): Die AK gibt das Ausmaß der Übereinstimmung der Information mit der sich zeitlich ändernden konzeptionellen Realität an.

Ausgabefunktion: Der Aktivierungsfunktion nachgeschaltete Funktion, die eine weitere Modellierbarkeit des Netzes ermöglicht. Meist wird hier jedoch die Identitätsfunktion $f(x) = x$ verwendet.

Backpropagation-Algorithmus: Der B-Algorithmus ist die am weitesten verbreitete und effektivste Methode, KNN zu trainieren. Es handelt sich um ein Gradientenabstiegsverfahren zur Suche nach dem globalen Minimum in der multidimensionalen Fehlerfläche.

Base Station Controller (BSC): Organisationseinheit der GSM-Netzarchitektur an die mehrere (BTS) Base Tranceiver Stations angeschlossen sind.

Base Tranceiver Station (BTS): Kleinste übergeordnete Organisationseinheit der GSM-Netzstruktur, die ein bis maximal sechs Funkzellen bzw. Einzelantennen versorgt.

Datenverarbeitung (DV): Die DV umfasst laut Wahrig [2002] das „Sammeln, Sichten, Speichern, Bearbeiten u. Auswerten von Informationen, die als Größen und Werte miteinander in Beziehung gesetzt werden können“.

Eingangsfunktion: Der Aktivierungsfunktion vorgeschaltete Funktion. Dabei handelt es sich um die gewichtete Summe der Eingangswerte und des Schwellwerts.

Feed-Forward-Netz (FFN): Eine häufig zum Einsatz kommende Form der KNN, bei denen der Informationsfluss gerichtet ist und keine Rückkopplungen erlaubt sind.

Fehlermöglichkeits- und -einflussanalyse (FMEA): Die FMEA ist eine Teammethode zur detaillierten Analyse von Systemen. Ziel ist es, Fehlermöglichkeiten in Bauteilen zu erkennen und zu charakterisieren sowie Gegenmaßnahmen zu entwickeln.

Floating Phone Data (FPD): Mit FPD werden die aus Mobilfunkdaten sich bewegender Teilnehmer generierten Verkehrsinformationen bezeichnet. In der Regel handelt es sich hierbei um Bewegungsprofile, die als FPD-Trajektorien bezeichnet werden.

FMECA: Fehlermöglichkeits- und -einflussanalyse mit zusätzlicher Quantifizierung der Kritizität (*engl.*: Criticality) in Form der RPZ; dies ermöglicht die Erstellung einer Priorisierungsliste.

Genauigkeit (GE): Die GE gibt den Zusammenhang zwischen dem ermittelten (meist gemessenen) und dem wahren bzw.plausibelsten Wert an.

Graphical User Interface (GUI): Grafische Benutzeroberfläche eines Software-Programms zur Vereinfachung der Bedienung.

Konsistenz (KO): Die KO gibt das Ausmaß der Übereinstimmung der Information mit dem Informationsmodell an.

Korrektheit (KR): Die KR gibt das Ausmaß der Übereinstimmung der Information mit der konzeptionellen Realität bei vorausgesetzter Aktualität an.

Künstliches Neuronales Netz (KNN): en.: *Artificial Neural Network (ANN)* Der Gehirnstruktur nachempfundenes Netz von Neuronen, die durch gewichtete Verbindungen verknüpft sind. KNN sind lernfähig und dienen der Informationsverarbeitung.

Lernrate/Lernfaktor η: Es handelt sich dabei um den variablen Parameter der Schrittweite, mit dem die Anpassung der Gewichte beim Gradientenabstiegsverfahren (Backpropagation-Algorithmus) erfolgt.

Levenberg-Marquardt-Algorithmus: Bei dem von Levenberg und Marquardt entwickelten Algorithmus handelt es sich um eine zeitoptimierte Weiterentwicklung des Back-Propagation-Algorithmus zum Training künstlicher neuronaler Netze.

Location Area Code (LAC): Bezeichnung für eine administrative Einheit in der GSM-Architektur. Diese beinhaltet in der Regel mehrere BTS (Base Tranceiver Station) und BSC (Base Station Controller).

Mobile-services Switching Center (MSC): Zentraler Bestandteil der GSM-Netzstruktur. Die MSC ist eine digitale Vermittlungsstelle zur Steuerung des Mobilfunkverkehrs und verwaltet mehrere BSC (Base Station Controller).

Mobilstation (MS): Allgemeine Bezeichnung für ein mobiles Gerät im Mobilfunknetz, welches mit einer SIM-Karte betrieben wird. Meist handelt es sich hierbei um ein Mobiltelefon oder Datenmodem.

Momentum-Term: Modifikation des Backpropagation-Lernalgorithmus zur Überwindung flacher Plateaus bei der Suche globaler Minima mit dem Gradientenverfahren. Durch Berücksichtigung der jeweils vorhergehenden Gewichtsänderung (Momentum-Term) in der Berechnung der neuen, wird die Suche beschleunigt.

Multi-Layer Feed-Forward-Netz (MLFFN): Dabei handelt es sich um eine Klasse von KNN mit vorgegebener Fließrichtung der Informationen und grundsätzlich beliebiger Anzahl verdeckter Schichten, die durch überwachtes Lernen trainiert werden können.

Perzeptron: Sehr frühe Netzsorte, die von dem amerikanischen Psychologe und Informatiker Frank Rosenblatt (vgl. Rosenblatt [1961] bereits Ende der 1950er Jahre entwickelt wurde. In der ursprünglichen Version als Feed-Forward-Netz bestehend aus einem Neuron entstanden bald komplexere Netze aus mehreren Neuronen, die in mehreren Schichten angeordnet werden (Multi-Layer-Perzeptron).

Prozess: In der Datenverarbeitung: Eine Abfolge von Aktivitäten (sequentiell, parallel, alternativ), die Eingangsdaten in Ausgangsdaten umwandeln.

Qualität: Grad, in dem ein Satz inhärenter Merkmale Anforderungen erfüllt DIN EN ISO 9000 [2005].

Qualitätsparameter (QP): Abkürzung QP; konkretisiertes Qualitätsmerkmal zur quantitativen Beschreibung eines Qualitätsaspektes.

Qualitätsmanagement (QM): Aufeinander abgestimmte Tätigkeiten zum Leiten und Lenken einer Organisation bezüglich Qualität (DIN EN ISO 9000 [2005]).

Qualitätsmanagementkonzept (QMK): Konzept zur konkreten praktischen Umsetzung der abgestimmten Tätigkeiten des Qualitätsmanagements.

Qualitätsmanagementsystem: Managementsystem zum Leiten und Lenken einer Organisation bezüglich der Qualität (DIN EN ISO 9000 [2005]).

Road Element (RE): Die Abkürzung RE steht für **R**oad **E**lement und bezeichnet ein Straßenelement zwischen zwei Knotenpunkten in der digitalen Straßenkarte. Bei baulich getrennten Fahrbahnen werden für ein Straßenabschnitt zur richtungsscharfen Abbildung zwei RE definiert.

Risikoprioritätszahl (RPZ): Die Risikoprioritätszahl dient der Quantifizierung der Kritizität von Fehlermöglichkeiten bei der FMEA. Die RPZ ist das Produkt aus den Parametern Auftreten, Bedeutung und Erkennbarkeit bzw. Behebbarkeit des Fehlers jeweils auf einer Skala von $1-10$ und kann daher Werte zwischen $1-1000$ annehmen.

RX-Matching: Der Begriff *RX-Matching* entstand in dem Projekt Do-iT und bezeichnet die Methode der Positionsbestimmung von Mobiltelefonen aus Abis-Daten mit Hilfe der gemessenen Signalstärken und der zugehörigen Signalstärkekarten des GSM-Netzbetreibers.

Stationäre Erfassungssysteme (SES): Dabei handelt es sich um stationäre Sensoren zur Erfassung von Verkehrsparametern, insbesondere der Verkehrsstärke, im Straßennetz. Es gibt eine Reihe unterschiedlicher Sensoren, neben den am häufigsten auftretenden Induktionsschleifen, kann die Erfassung beispielsweise auch mit passiven Infrarotsensoren oder durch Ultraschallsensoren erfolgen.

Timing Advance (TA): Der TA-Wert dient der zeitlichen Synchronisation der im Netz zu übertragenden Daten eines Mobilfunkteilnehmers. Der Wert wird aus der Signallaufzeit abgeleitet und stellt daher eine grobe Entfernungsangabe zwischen Antenne und MS dar.

Ursachen-Wirkungs-Diagramm (UWD): Das Ursachen-Wirkungs-Diagramm dient der Recherche und grafischen Darstellung von Ursachen für unerwünschte Wirkungen in einem System. Dabei werden Haupt- und Nebenursachen unterschieden und die Wirkungen jeweils einzeln betrachtet.

Verfügbarkeit (VE): Die VE gibt das Ausmaß des Vorhandenseins der Information zu einem definierten Zeitpunkt an einem bestimmten Ort an.

Vollständigkeit (VO): Die VO gibt das Ausmaß des Vorhandenseins sämtlicher zur Beschreibung der konzeptionellen Realität erforderlichen Informationen an.

Formelzeichen

$\mathbf{a}$	Matrix der Ausgangswerte eines KNN
A	Parameter zur Quantifizierung des Auftretens eines Systemfehlers (FMEA)
A^{QM}	Ausgangswert eines Prozesses als Merkmalserfüllungsgrad zwischen 0 und 1 für das Qualitätsmerkmal QM
b	Bias oder Schwellwert eines künstlichen Neurons
B	Parameter zur Quantifizierung der Bedeutung eines Systemfehlers (FMEA)
dT_{FPD}	Zeitliche Länge einer FPD-Trajektorie
E	Parameter zur Quantifizierung der Erkennbarkeit und Behebbarkeit eines Systemfehlers (FMEA)
E^{QM}	Eingangswert eines Prozesses als Merkmalserfüllungsgrad zwischen 0 und 1 für das Qualitätsmerkmal QM
f	Aktivierungsfunktion, mit der der Ausgang eines Neurons berechnet wird
$\mathbf{F}$	Jacobi-Matrix mit den partiellen Ableitungen
h	Anzahl der Neuronen in der verdeckten Schicht eines Feed-Forward-Netzes mit nur einer verdeckten Schicht oder mehrerer verdeckter Schichten mit identischer Neuronenzahl
l	Längsfehler des Neupunktes beim polaren Anhängen
$\mathrm{L}_{\mathrm{FPD}}$	Metrische Länge einer FPD-Trajektorie
m	Anzahl der Ausgangswerte bzw. Ausgangsneuronen
n	Anzahl der Eingangswerte bzw. Eingangsneuronen
MSE	Mittlerer quadratischer Fehler (en.: Mean Square Error)
$\mathbf{p}$	Eingangsvektor der Trainingsbeispiele
ppm	Atmosphärischer Einfluss auf die elektrooptische Streckenmessung in parts per Million
q	Querfehler des Neupunktes beim polaren Anhängen
Q	Anzahl der Lernbeispiele für das Training eines KNN
r	Horizontale Richtung
s	Horizontale Strecke
S	Standardabweichung einer fehlerbehafteten Größe
$\sum$	Übertragungsfunktion, mit der die Eingänge eines Neurons zusammengeführt werden
$\sum_{ll}$	Varianz-Kovarianzmatrix der fehlerbehafteten Größen

$\sum_{xx}$	Varianz-Kovarianzmatrix der gesuchten Größen
$\mathbf{Q}_F$	Qualitätstupel zur zusammenfassenden Darstellung aller Qualitätsmerkmale in einem Ausdruck
$\mathbf{t}$	Target-Vektor: Ausgangsvektor der Trainingsbeispiele
v	Verbesserung als Differenz aus Sollwert-Istwert
$w_{i,j}$	Gewicht zwischen Neuron i und Eingangswert j im KNN
$\mathbf{W}$	Gewichtsmatrix mit den Gewichten $w_{i,j}$ eines Multilayer Feed-Forward-Netz

Anhang A: Das GUI nntool von Matlab

Das in der KNN-Toolbox zur Verfügung stehende GUI zur Definition und Simulation von KNN erleichtert den Einsatz der zahlreichen zur Verfügung stehenden Funktionen. Für die in Kapitel 4 berechneten Beispiele wurden die folgenden Programmversionen verwendet:

- Matlab-Software: V 7.8.0.347 (R2009a)
- Neural Network Toolbox-Software: V 6.0.2

Grundsätzlich stehen während der Verwendung des nntools alle Funktionen von Matlab parallel zur Verfügung. Der Zugriff auf alle Variablen in der Workspace eines Matlab-Projekts ist jederzeit möglich. Somit erfolgt im Fall der simulierten Beispiele in Kapitel 4.1 die Generierung der Trainingsdaten sowie der unabhängigen Testdatensätze in Matlab. Die erzeugten Vektoren und Matrizen können dann im Anschluss einfach in das nntool eingelesen werden, solange die Variablen im temporären Speicher vorhanden sind. Für eine spätere Verwendung der erzeugten Datensätze oder zum erneuten Nachvollziehen der Ergebnisse, werden diese jedoch als ASCII-Dateien exportiert. Das nntool bietet auch die Möglichkeit, die Datensätze als Dateien einzulesen, wodurch der Umgang mit der GUI noch flexibler wird.

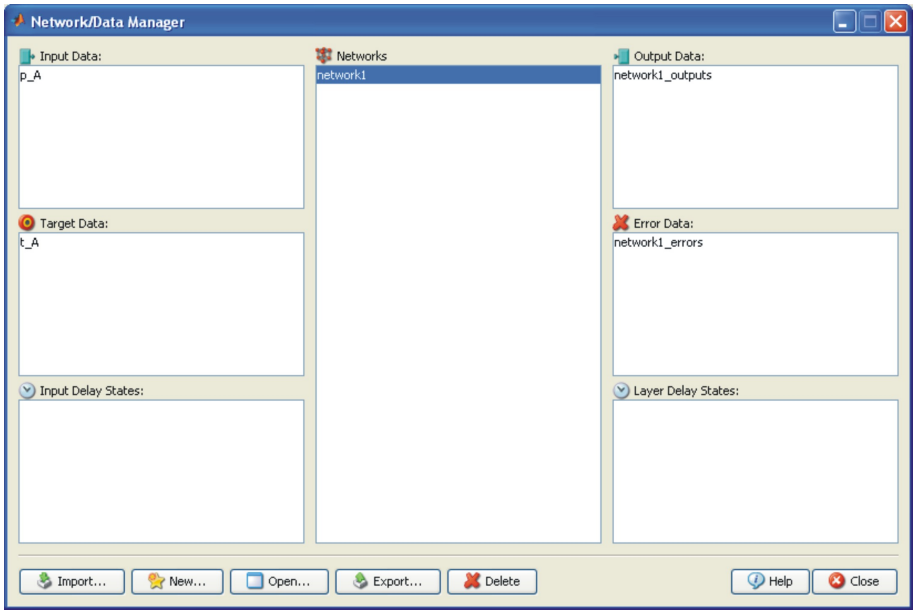


Abbildung 6.1: Startfenster der grafischen Benutzeroberfläche des nntool

Mit dem Befehl nntool kann das GUI aus der Matlab Benutzeroberfläche gestartet werden. Es erscheint das Startfenster des nntools, wie es in der Abbildung 6.1 abgebildet ist. Das Startfenster gibt einen Überblick über alle eingelesenen oder erzeugten Variablen. Dies sind auf der linken Seite in den beiden oberen Fenstern die originären Eingangs- und Ausgangsdaten (*Input Data* und *Target Data*), die zum Trainieren des Netzes zur Verfügung stehen. In der Mitte erscheinen die bereits definierten Netze und in den beiden oberen rechten Fenstern stehen, sofern bereits ein Netz trainiert und eingesetzt wurde, die berechneten

Ausgangsdaten (*Output Data*) und die Restfehler (*Error Data*[1]) nach der letzten Iteration zur Verfügung. Die beiden unteren Fenster sind nur in Verbindung mit dynamischen Netzen erforderlich. Diese werden im Rahmen dieser Arbeit jedoch nicht weiter betrachtet. In der unteren Funktionsleiste stehen die Möglichkeiten neue Daten zu importieren, ein neues Netz zu erzeugen, zu öffnen oder zu löschen sowie der Datenexport in eine Datei zur Verfügung.

Die Definition eines neuen Netzes erfolgt in einem weiteren interaktiven Fenster, welches Abbildung 6.2 zeigt. Nach der Vergabe eines eindeutigen Namens für das Netz muss der Netztyp mit Hilfe eines drop-down Menüs gewählt werden. Abhängig von diesem, müssen nachfolgend die für das Training zur Verfügung stehenden Lernbeispiele gewählt werden. Für die Modellierung der Datenqualität werden hier -wie bereits erwähnt- stets Feed-Forward-Netze gewählt. Dieser Netztyp bedingt die Verwendung eines Backpropagation-Algorithmus für das Training des Netzes, daher sind im Weiteren die zur Verfügung stehenden Eingangs- und Ausgangsdaten zu wählen. Über die drop-down Menüs kann einfach aus den bereits vorab über das Startfenster (vgl. Abbildung 6.1) eingelesenen oder erzeugten Lerndaten gewählt werden.

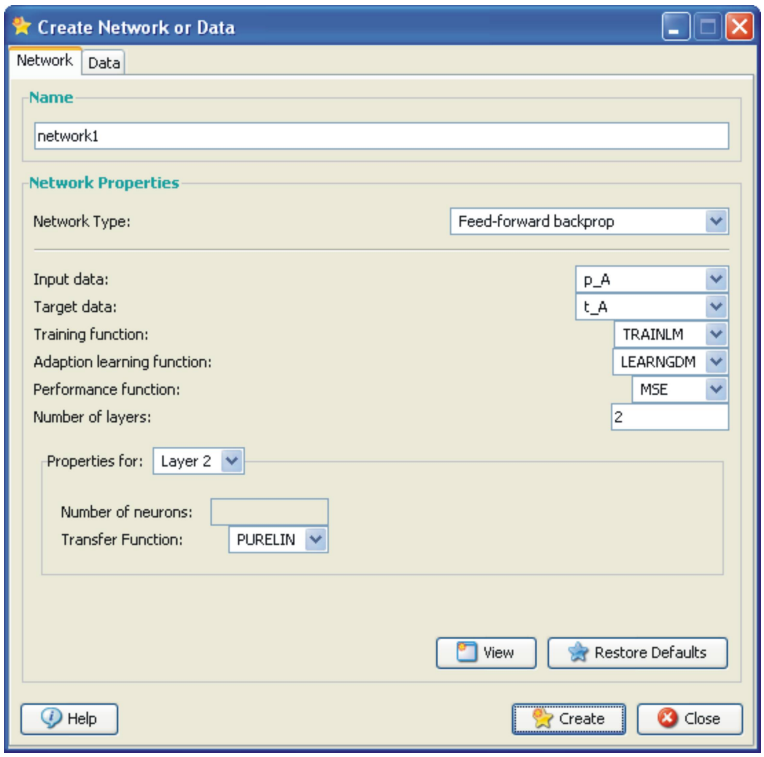


Abbildung 6.2: GUI nntool: Interaktives Fenster zur Definition eines neuen KNN

Die Führung des Nutzers von oben nach unten durch die Einstellmöglichkeiten hindurch ist intelligent gelöst. Es sind immer nur die nach der vorhergehenden Auswahl sinnvollen Auswahlmöglichkeiten vorhanden. So kann nach der Entscheidung für die Feed-Forward-Netze auch nur noch zwischen einigen Lernalgorithmen, die alle auf der Backpropagation-Methode basieren, gewählt werden. Aufgrund der Leistungsfähigkeit wird für alle weiteren Auswertungen der von Levenberg und Marquardt entwickelte Algorithmus *TrainLM* eingesetzt. Grundsätzlich können hier auch andere Lernalgorithmen gewählt

[1] In Matlab werden die Differenzen aus Sollwert-Istwert als *Error* bezeichnet. Aus geodätischer Sicht handelt es sich dabei jedoch um Verbesserungen, daher wird in diesem Zusammenhang der englische Begriff *Error* mit *Verbesserung* gleichgesetzt.

werden, wenn beispielsweise nicht genügend virtueller Speicher für den speicherintensiven LM-Algorithmus zur Verfügung steht. Zusätzlich kann in der Software auch der Momentum-Term berücksichtigt werden, um flache Plateaus in der Fehlerfläche besser überwinden zu können und das Oszillieren in tiefen Schluchten zu verhindern. Die Gewichtung des Terms ist zwischen 0 und 1 frei wählbar.

Der Nutzer kann als weitere Grundeinstellung vor der Definition der Netzwerkarchitektur schließlich noch die gewünschte Performance- bzw. Leistungsfunktion wählen. Neben dem im Folgenden verwendeten mittleren quadratischen Fehler (*MSE*), stehen die Summe der Fehlerquadrate (*SSE*) und eine erweiterte Form des MSE zur Verfügung, bei der das mittlere quadratische Gewicht mit in die Performance-Funktion einfließt. Dies soll insgesamt zu kleineren Gewichten und Schwellwerten führen und damit dem Overfitting vorbeugen (Matlab® [2010]).

Im Unteren Bereich des Eingabefensters in Abbildung 6.2 erfolgt die Definition der Netzwerkarchitektur. Nach der Definition der Schichtenanzahl (*Number of layers*), wobei die Summe aus verdeckten Schichten und der Ausgabeschicht gemeint ist, kann für jede Schicht getrennt die Anzahl der Neuronen und die Transferfunktion festgelegt werden. Für die Ausgabeschicht hat Matlab bereits die Dimension des zuvor gewählten Ausgabevektors t für die Anzahl der Neuronen eingesetzt. Die Wahl der Transferfunktion für die Ausgabeschicht hängt von dem gewünschten Wertebereich der Ausgangsdaten ab, aufgrund des Lernalgorithmus muss diese jedoch stetig differenzierbar sein. Grundsätzlich kommen daher aus der Tabelle 3.5 nur die häufig verwendete Linearfunktion, bei der die Ausgabe nicht auf einen Wertebereich festgelegt wird sowie die logarithmische Sigmoidfunktion und der Tangens Hyperbolicus in Frage.

Es besteht an dieser Stelle bereits die Möglichkeit, mit *View* eine visuelle Darstellung des definierten Netzes aufzurufen. Diese Darstellung erhält man jedoch ebenfalls nach Erzeugung des definierten Netzes mit *Create* und anschließendem Aufruf des neuen Netzes durch Auswahl und Bestätigung mit *Open...* im Startfenster (vgl. Abbildung 6.1). Wie die Abbildung 6.3 exemplarisch zeigt, können in der grafischen Darstellung nochmals die Anzahl der Schichten sowie die gewählten Transferfunktionen überprüft werden (diese werden symbolisch analog zu Tabelle 3.5 dargestellt). Derzeit wird die definierte Anzahl der Neuronen in den Schichten leider in der aktuellen Version nicht angezeigt.

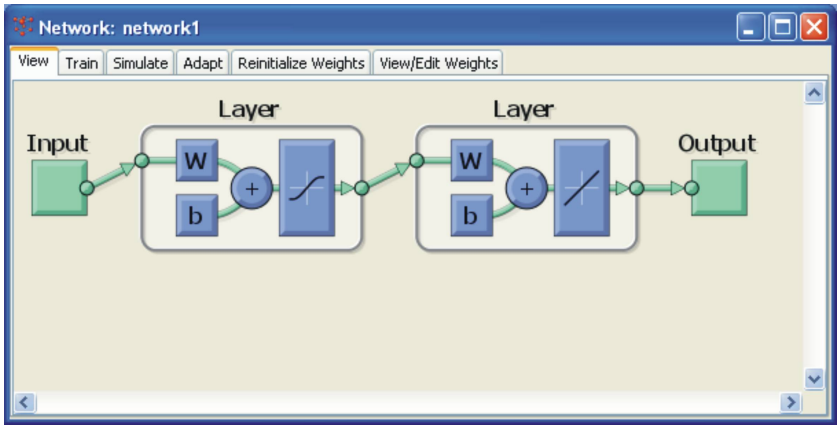


Abbildung 6.3: GUI nntool: Interaktives Fenster zur Modifikation eines erzeugten KNN

Unter den verschiedenen Reitern der grafischen Benutzeroberfläche in Abbildung 6.3 können zunächst die default-Einstellungen für das Training betrachtet und ggf. editiert werden. Das Netz kann entweder unter dem Reiter *Train* trainiert werden, wobei die Reihenfolge der Lernbeispiele zufällig ist, oder es kann ein sogenanntes adaptives Training erfolgen (Reiter: *Adapt*). Beim adaptiven Training folgt die

Reihenfolge der Lernbeispiele der Reihenfolge der Beispiele in der Matrix der Eingangsdaten. Diese Methode wird für dynamische Netze eingesetzt, in denen die Ein- und Ausgangsvariablen von der Zeit abhängen.

Neben der Überprüfung und ggf. manuellen Anpassung der Startgewichte und -schwellwerte (Reiter: *View/Edit Weights*), können verschiedene Parameter und Abbruchkriterien geändert oder deaktiviert werden. Unter anderem sind unter dem Reiter *Train→Training Parameters* die folgenden Parameter frei wählbar:

epochs	Maximale Anzahl der Iterationen/Epochen bis zum Abbruch
time	Maximale Dauer des Iterationsprozesses bis zum Abbruch
goal	Zielwert für die Performance-Funktion (hier: MSE)
max_fail	Maximaler Validierungsfehler
min_grad	Minimaler Performance Gradient
mu	Lernparameter aus dem LM-Algorithmus

Der maximale Validierungsfehler *max_fail* wird von Matlab für das *Early Stopping* benötigt. Der Parameter ist frei wählbar und gibt die maximale Anzahl an kontinuierlichen Verschlechterungen des Approximationsfehlers an, nach der das Training abgebrochen wird. Diese Methode soll ein Überlernen des Netzes verhindern, in dem zunächst eine Dreiteilung der Lerndatenmenge in Lern-, Validierungs- und Testdaten erfolgt. Die zufällige Aufteilung erfolgt in 60% Lerndaten und je 20% Validierungs- und Testdaten. Es werden in jeder Epoche nur die Lerndaten im Backpropagation-Algorithmus verwendet, um die Netzparameter zu verbessern. Nach jeder Epoche wird das Netz mit dem Validierungsdatensatz getestet. Solange der Restfehler sowohl bei den Lern- als auch bei den Validierungsdaten sinkt, wird weiter iteriert. Sobald jedoch der Fehler im Validierungsdatensatz mehr als *max_fail*-mal in Folge steigt, wird das Training beendet. Erst im Anschluss dient der Testdatensatz der Kontrolle des trainierten Netzes.

Der skalare Parameter μ (im nntool mit *mu* bezeichnet) spielt eine wichtige Rolle in dem Trainingsalgorithmus von Levenberg und Marquardt (vgl. Formel 3.34). Laut Hagan u. a. [1996] ist das Newton-Verfahren in der Nähe der Minima in der Fehlerfigur schneller und genauer als das Verfahren von Levenberg und Marquardt. Daher wird μ nach jeder Iteration durch Multiplikation mit dem Abnahmefaktor *mu_dec* verkleinert, es sei denn, die Performance-Funktion steigt an, was eine Vergrößerung von μ um den Faktor *mu_inc* zur Folge hat. Damit wird ein möglichst schnelles Finden der Minima in der Fehlerfläche und damit der Übergang zum Newton-Verfahren angestrebt. Die Parameter mu, mu_desc und mu_inc sind in dem GUI frei wählbar. In der Regel können jedoch die bereits vorgeschlagenen Parametereinstellungen übernommen werden. Die Vergrößerung von μ nach oben kann zusätzlich mit Hilfe des Parameters *mu_max* begrenzt werden. Das Überschreiten von mu_max führt zum Abbruch des Iterationsverfahrens. Standardmäßig ist dieser Parameter auf 10^{-9} gesetzt und damit als Abbruchkriterium faktisch deaktiviert.

Schließlich bietet die Definition des minimalen Performance-Gradienten min_grad dem Nutzer die Möglichkeit, das Lernziel festzulegen. Die finalen Netzwerkparameter wurden gefunden, wenn der Gradient bei weiteren Iterationen den vorgegebenen Grenzwert nicht mehr übersteigt, d. h. keine signifikante Verbesserung der Parameter mehr zu erwarten ist.

Unter dem Reiter *View/Edit Weights* auf der grafischen Benutzeroberfläche (vgl. Abbildung 6.3) können die Initialisierungsgewichte und Schwellwerte betrachtet und ggf. geändert werden. In den Beispielen in Kapitel 4 werden stets die vorgeschlagenen Gewichte verwendet, da keine besseren Startwerte

bekannt sind. Nach einem abgeschlossenen Training besteht die Möglichkeit, diese Gewichte vor einem erneuten Training zunächst zu reinitialisieren, um beispielsweise das erste Iterationsergebnis mit neuen Startwerten nochmals zu überprüfen (Reiter: *Reinitialize Weights*). Im Anschluss kann das trainierte Netz zur Prozessierung neuer Daten eingesetzt werden (Reiter: *Simulate*). Dabei besteht auch die Möglichkeit, das Netz mit neuen, unabhängigen Daten zu evaluieren. Nach Angabe der Solldaten erfolgt automatisch ein Vergleich mit dem produzierten Netzausgang.

Sämtliche definierbaren Abbruchkriterien werden im Sinne einer ODER-Verknüpfung behandelt. Das Erfüllen eines einzelnen Abbruchkriteriums beendet damit das Iterationsverfahren. Der Verlauf in Echtzeit sowie die Ergebnisse des Verfahrens werden anschließend in einer Grafik übersichtlich dargestellt (vgl. Abbildung 6.4).

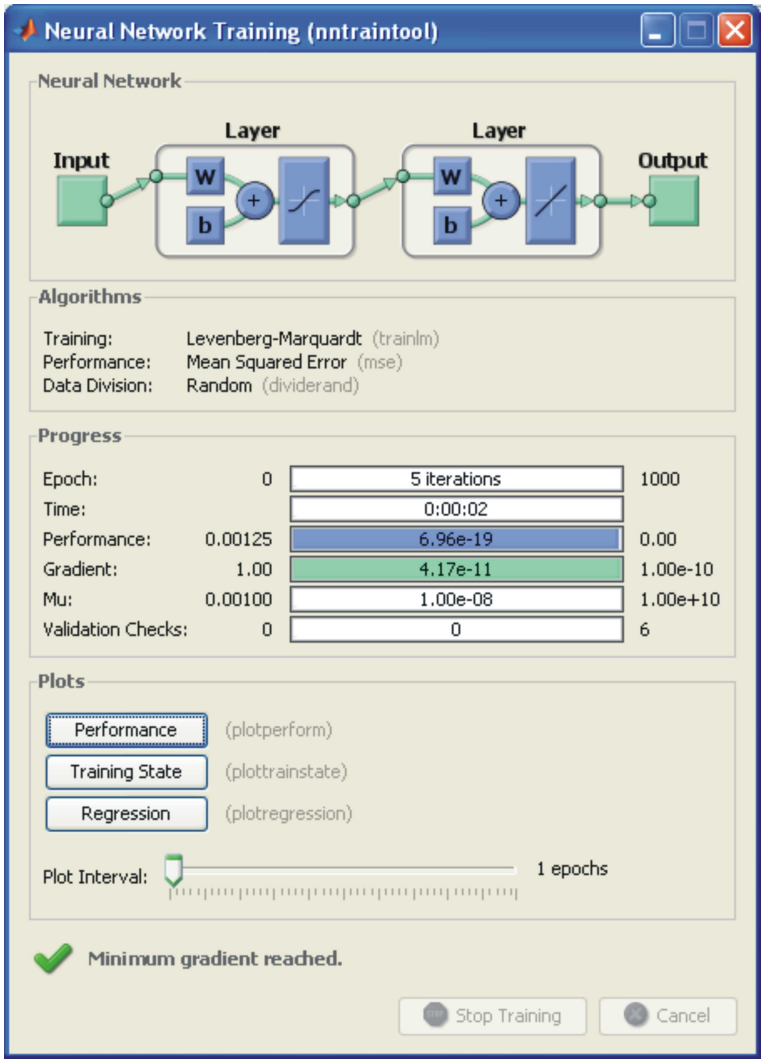


Abbildung 6.4: GUI nntool: Darstellung der Trainingsergebnisse

Die grafische Darstellung der Trainingsergebnisse (vgl. Abbildung 6.4) gibt zunächst nochmals Aufschluss über die Architektur des Netzes und die Wahl des Lernalgorithmus. Im Weiteren wird das für den Abbruch verantwortliche Kriterium (im Beispiel wurde der Minimalwert für den Gradienten unterschritten) sowie die weiteren Parameterwerte nach Abbruch des Trainings angegeben. Dabei werden die einzelnen Parameter sowohl in Zahlen als auch als farbige horizontale Balkendiagramme dargestellt.

Links der Balken sind die Startwerte, rechts der Balken die Grenzwerte angegeben. Greift keines der Abbruchkriterien, so kann das Training ggf. auch per Hand gestoppt werden.

Dem Nutzer werden unter der Rubrik *Plots* diverse Grafiken zur Interpretation der Lernphase zur Verfügung gestellt. Diese Grafiken bilden die Grundlage zur Beurteilung des Trainingserfolgs und damit der Eignung des trainierten Netzes für die jeweilige Aufgabe. Der Performance-Plot zeigt die Änderung der Performance-Funktion für die Trainings-, die Validierungs- sowie die Testdaten über alle Epochen hinweg. Die Veränderung der Parameter Abstiegsgradient, Lernparameter μ sowie der Validierungsfehler können unter *Training State* in einem Plot angezeigt und damit das Lernverhalten interpretiert werden. Schließlich werden unter *Regression* die Regressionen für die Trainings-, Validierungs- und Testdaten sowie für die gesamten Lernbeispiele abgerufen. Allerdings besteht hier nur die Möglichkeit, jeweils den ersten Ausgangsparameter mit dem entsprechenden Sollwert zu vergleichen. Dies gibt jedoch einen ersten Hinweis auf die Qualität der erzielten Ergebnisse.

Zur Beurteilung der Eignung von KNN für die Modellierung und Fortpflanzung von Datenqualität in Prozessen dienen diese, von der Toolbox zur Verfügung gestellten Plots, nur der ersten groben Beurteilung der Resultate. Die verschiedenen Versuche in Kapitel 4 werden zusätzlich mit Hilfe weiterer Testdatensätze beurteilt. Diese Testdatensätze werden erst nach erfolgreichem Training mit dem Netz berechnet, um eine unabhängige Kontrolle sicherzustellen. Im Anschluss erfolgt jeweils die grafische Darstellung der Restfehler aus dem Soll-Ist-Vergleich der bekannten Ergebnisse mit den durch das KNN berechneten.

Aus Platzgründen werden hier nur die wichtigsten Einstellungen des GUI näher erläutert. Für weitere Detailinformationen wird dem interessierten Leser die sehr hilfreiche Programmdokumentation von Matlab® [2010] und das von den selben Autoren verfasste Standardwerk Hagan u. a. [1996] empfohlen.

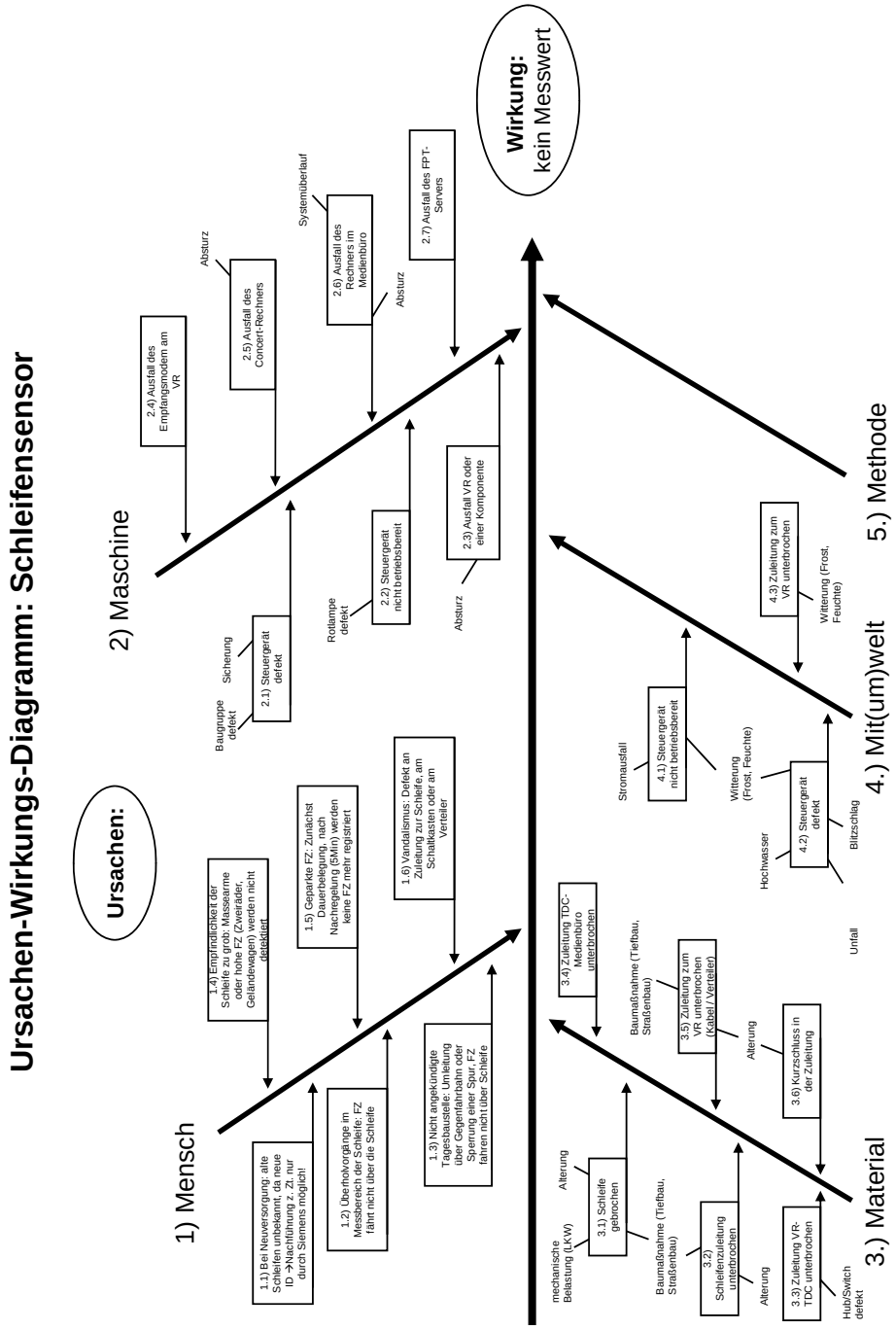
Anhang B: Das Qualitätsmodell für FPD-Trajektorien

Tabelle 6.1: Qualitätsmodell für FPD-Trajektorien aus Abis-Daten (Quelle: Do-iT [2009a])

Nr.	Merkmal	Parameter	Kürzel	Definition
1.1	VE	Globale Ausfallrate	$\mathrm{AUR}_{\mathrm{global}}$ [%]	Beschreibt die Verfügbarkeit des FPD-Servers in einem bestimmten Zeitraum
1.2		Trajektoriendichte	ρ_{Tra} $[1/\mathrm{s}]$	Summe(Routenlängen)/L/T mit L: Länge des betrachteten Streckenabschnitts; T: Zeitraum
2.1	AK	Berechnungszeit	$\Delta t_{\mathrm{ber}}[\mathrm{s}]$	Berechnungszeit für eine FPD-Trajektorie
2.2		Trajektorienalter	Δt_{FPD} [min]	Alter der Trajektorie als Differenz: Aktuelle Zeit-Zeitstempel der 1. Position der Trajektorie
3.1	VO	Trajektorien-vollständigkeit	$\mathrm{VO}_{\mathrm{FPD}}$ [%]	Anteil der verwendeten Positionen einer Positionsfolge eines Teilnehmers, der für die Trajektorie verwendet wird
3.2		Durchdringung mit FPD	d [%]	Anteil des Verkehrs, der mit FPD erfasst werden kann; bezogen auf den gesamten Verkehr
3.3		Abdeckungsrate	$\mathrm{ABR}_{\mathrm{FPD}}$ [%]	Grad der Abdeckung des betrachteten Straßennetzes mit FPD
4.x	KO	Die Konsistenz wurde durch Einhaltung des Datenmodells gewährleistet		
5.1	KR	Klassifizierungs-korrektheit	$\mathrm{KR}_{\mathrm{Kl}}$ [%]	Wahrscheinlichkeit mit der die richtigen Teilnehmerklassen identifiziert wurden
5.2		Trajektorienlänge	$\mathrm{L}_{\mathrm{FPD}}[\mathrm{m}]$	Länge der FPD-Trajektorie
5.3		Zuordnungs-korrektheit Typ A	$\mathrm{KR}_{\mathrm{ZuA}}$ [%]	Korrekter Streckenanteil der FPD-Route, der sich mit der GPS-Route deckt, bezogen auf die Länge der GPS-Route
5.4		Zuordnungs-korrektheit Typ B	$\mathrm{KR}_{\mathrm{ZuB}}$ [%]	Korrekter Streckenanteil der FPD-Route, der sich mit der GPS-Route deckt, bezogen auf die Länge der ermittelten FPD-Route
5.5		Rangkorrelation nach Spearman	r [-]	Ähnlichkeit der Tagesganglinien der Verkehrsstärke aus FPD und aus SES

Nr.	Merkmal	Parameter	Kürzel	Definition
6.1	GE	Mittlere Querabweichung	$\mathrm{QA}_{\mathrm{FPD}}$ $[\mathrm{m}]$	Mittel der orthogonalen Abweichungen der verwendeten Positionen von der wahrscheinlichsten Route
6.2		Standardabw. der Geschwindigkeit	s_v $[\mathrm{km/h}]$	Genauigkeit der aus den Trajektorien ermittelbaren Geschwindigkeit
6.3		Differenz der Durchschnittsgeschwindigkeiten	$\delta[\mathrm{km/h}]$	Vorzeichenfreie Differenz zwischen Durchschnittsgeschwindigkeit aus GPS und aus Mobilfunkdaten

Anhang C: Beispiel: Ursachen-Wirkungs-Diagramm



Anhang D: Beispiel: FMECA

Fehler-Möglichkeits- und Einflussanalyse: System-FMECA

Bearbeiter/Team: *Hr. Heisler, Hr. Pfleiderer, Hr. Schwartz (TBA KA)*
Hr. Laufer (IAGB)

Datum: *02.08.2007*

System: *Schleifensensor*

Subsystem: *Induktionsschleife*

Betriebsdauer: *sehr unterschiedlich, bis zu 20 Jahren*

Nr.	Bauteil	Funktion	potentielle Ausfallart	mögliche Ursache	Fehlerfolge / lokale Auswirkung	Auswirkung auf das System	Erkennungsmethode	vorsorgliche Gegenmaßnahmen	A B E	RPZ
1.1	Induktionsschleife	Induktion von Spannung	Bruch / Abriß	mechanische Belastung; Baumaßnahmen, Alterung, Witterung	keine Induktion	keine Messdaten	Detektion von der Baugruppe und Meldung an VR (Status braun); Meldung wird dort protokolliert	gute Vorplanung von Baumaßnahmen, exaktes Leitungskataster	4 7 2	56
1.2			Kurzschluss					-	2 7 2	28
1.3			sporadischer Kurzschluss		sporadisch keine Induktion	sporadisch keine Messdaten	s. o. mit Statuswechsel zwischen braun und nicht braun	-	2 8 5	80
2.1	Muffe	Verbindung der Induktionsschleife mit der Zuleitung zur Auswertebaugruppe	Bruch / Abriß	mechanische Belastung; Baumaßnahmen, Alterung, Witterung	keine Signalübertragung	keine Messdaten	Detektion von der Baugruppe und Meldung an VR (Status braun); Meldung wird dort protokolliert	Bauteil mit hoher Qualität; gute Vorplanung von Baumaßnahmen, exaktes Leitungskataster	3 7 2	42
2.2			Kurzschluss					sorgfältiges Arbeiten	1 7 2	14
2.3			sporadischer Kurzschluss		sporadisch keine Signalübertragung	sporadisch keine Messdaten	s. o. mit Statuswechsel zwischen braun und nicht braun		1 8 5	40
3.1	Zuleitung vom Kleinschacht zum Schaltkasten	Verbindung der Induktionsschleife mit der Baugruppe	Bruch / Abriß	mechanische Belastung; Baumaßnahmen, Alterung, Witterung	keine Signalübertragung	keine Messdaten	Detektion von der Baugruppe und Meldung an VR (Status braun); Meldung wird dort protokolliert	gute Vorplanung von Baumaßnahmen, exaktes Leitungskataster	3 7 2	42
3.2			Kurzschluss					Witterungsbedingte Ausfälle ausschließen, durch Verwendung moderner (fettverpresster) Kabel	1 7 2	14
3.3			sporadischer Kurzschluss		sporadisch keine Signalübertragung	sporadisch keine Messdaten	s. o. mit Statuswechsel zwischen braun und nicht braun		2 8 5	80

Bemerkungen zu Nr.: 1,2,3: Leitungsabriß durch mechanische Belastung ist weniger häufig, als Abriß durch Baumaßnahmen, daher die Bewertung A = 3 bzw. 4
1,2,3: Ist im Steuergerät die Brücke gesetzt, wird der Verlauf im VR mitprotokolliert, daher E=5 ohne und E=2 mit gesetzter Brücke

Danksagung

Ich möchte allen danken, die mich bei dieser Arbeit unterstützt haben. Das diese Arbeit in der schwierigen Umbruchphase am Institut, ausgelöst durch den viel zu frühen Tod von Herrn Prof. Dr.-Ing. Wolfgang Möhlenbrink und der damit verbundenen Zeit der Ungewissheit dennoch zu einem guten Ende gebracht werden konnte, ist insbesondere auf die gute Unterstützung durch Prof. Dr.-Ing. habil. Volker Schwieger und Dr.-Ing. Martin Metzner zurückzuführen, denen ich hiermit meinen besonderen Dank ausspreche. Ich möchte auch insbesondere Alexander Beetz danken für die vielen gewinnbringenden Gespräche, Diskussionen und Ermunterungen.

Schließlich möchte ich mich bei meiner Lebensgefährtin Soña, meinen Eltern und meiner Schwester Birgit herzlich für die ständige uneingeschränkte seelische und moralische Unterstützung in dieser Zeit und darüber hinaus bedanken.

Lebenslauf

Zur Person	
Name	Ralf Laufer
Geburtsdatum	12.05.1977
Geburtsort	Villingen-Schwenningen
Berufstätigkeit	
seit 03/2011	Mitarbeiter der Vorarlberger Illwerke AG, Schruns, Österreich
05/2005 - 02/2011	Wissenschaftlicher Mitarbeiter am Institut für Anwendungen der Geodäsie im Bauwesen (IAGB), seit Herbst 2010 umbenannt in Institut für Ingenieurgeodäsie (IIG), Universität Stuttgart
03/2005 - 04/2005	Leica-Geosystems AG, Metrology Division, Unterentfelden, Schweiz Vermessungsingenieur Abteilung Research&Development
Ausbildung	
10/2000 - 04/2005	Universität Stuttgart Studium Studienrichtung Geodäsie und Geoinformatik, Abschluss: Dipl.-Ing. Geodäsie und Geoinformatik
10/2004 - 02/2005	Leica-Geosystems AG, Metrology Division, Unterentfelden, Schweiz Diplomarbeit: Untersuchungen des T-Scan-Systems von Leica-Geosystems hinsichtlich Funktionssicherheit und Genauigkeit Auszeichnung mit dem Preis des Vereins der Freunde des Studiengangs
10/1997 - 09/2000	ÖbVI Grießhaber, Villingen-Schwenningen Ausbildung zum Vermessungstechniker Abschluss als Jahrgangsbester in BW
09/1987 - 09/1996	Gymnasium am Hoptbühl, Villingen-Schwenningen Abschluss: Allgemeine Hochschulreife
09/1983 - 07/1987	Grundschule im Steppach, Villingen-Schwenningen