KOMPETENZZENTRUM WasserBerlin

Berliner Wasserbetriebe

BERICHT | 19.12.2017

Analyse und Modellierung des Zustands von Abwasserkanälen in Berlin

Abschlussbericht des Forschungsvorhabens SEMA-Berlin

Mathias Riechel (KWB), Nicolas Caradot (KWB), Nic Lengemann (BWB, CO-A)

Inhaltsverzeichnis

1 Vorwort 1

1.1 Historische Entwicklung der Berliner Kanalisation 1
1.2 Berliner Kanalisation als Investitionsgut 2
1.3 Alterungsverhalten und Zustandsprognosen der Kanalisation 4

2 Datengrundlage und -vorbereitung 6

2.1 Übergebene Daten 6
2.2 Datenfilterung und -bereinigung 6
2.3 Beschreibung der Variablen 8
2.3.1 Erklärende Variablen 8
2.3.2 Zielvariable: baulicher Zustand 11
2.4 Ist-Analyse und Datenverteilungen 11
2.4.1 Baujahr, Alter, Material und Abwassertyp 12
2.4.2 Profil, Breite, Höhe und Länge 12
2.4.3 Tiefe, Überdeckung, Gefälle und Straßenklasse 13
2.4.4 Schienenverkehr und Bäume 13
2.4.5 Grundwasserüberdeckung, Bodenart und Rückstau 13
2.4.6 Bezirk und Stadtteil 14
2.4.7 Baulicher Zustand 14

3 Statistische Analyse zum Zustand der Abwasserkanäle 16

3.1 Analyse der Abhängigkeiten der Variablen 16
3.1.1 Methodisches Vorgehen 16
3.1.1.1 Korrelationsuntersuchungen nach Spearman 16
3.1.1.2 Chi-Quadrat-Unabhängigkeitstest und Cramér's V 17
3.1.2 Ergebnisse und Diskussion 20
3.2 Einfluss der Eingangsvariablen auf den Zustand 24
3.2.1 Methodisches Vorgehen 24
3.2.2 Ergebnisse und Diskussion 24
3.3 Analyse der Einzelschäden 30
3.3.1 Methodisches Vorgehen 30
3.3.2 Ergebnisse und Diskussion 31
3.4 Untersuchung der Unsicherheiten bei der Inspektion 39
3.4.1 Datenvorbereitung und erste Analysen 39
3.4.2 Abweichungen der Zustandsklasse zwischen den Doppelbefahrungen 41
3.4.3 Abweichungen der Zustandsklasse mit unterschiedlichen Kameratypen 43
3.4.4 Bewertung der Unsicherheiten der Kanalzustandsbewertung 43
3.4.4.1 Methodik 43
3.4.4.2 Ergebnisse und Diskussion 47

4 Modellierung des Zustands der Abwasserkanäle 49
4.1 Untersuchte Modellansätze 49
4.1.1 GompitZ 49
4.1.2 Random Forest 50
4.1.3 Support Vector Machine 53
4.1.4 Künstliche Neuronale Netze 55
4.2 Methodisches Vorgehen und Bewertungskriterien 58
4.2.1 Datenfilterung 58
4.2.2 Aufteilung der Daten in Trainings- und Testdaten 59
4.2.3 Training der Modelle 60
4.2.4 Test der Modelle 62
4.2.5 Bewertungsindikatoren 63
4.2.6 Simulation der Zustandsentwicklung 65
4.3 Modellbewertung 65
4.3.1 GompitZ 65
4.3.1.1 Bildung der Kohorten und Kalibrierung der Überlebensfunktionen 65
4.3.1.2 Bewertung der Modellgüte 68
4.3.1.3 Einfluss der Datenmenge auf Modellgüte 70
4.3.2 Random Forest 72
4.3.2.1 Einfluss der Modellparameter 72
4.3.2.2 Bewertung der Modellgüte 75
4.3.2.3 Einfluss der Datenmenge auf Modellgüte 77
4.3.3 Support Vector Machine 79
4.3.3.1 Einfluss der Modellparameter 79
4.3.3.2 Bewertung der Modellgüte 82
4.3.4 Künstliche Neuronale Netze 84
4.3.4.1 Einfluss der Modellparameter 84
4.3.4.2 Bewertung der Modellgüte 86
4.3.5 Vergleich der Modelle 88
4.4 Simulation der Zustandsentwicklung 91

5 Fazit 94
5.1 Zusammenfassung 94
5.2 Offene Fragen 96
5.3 Ausblick 96

Anhang A: Zustandsverteilung und Einflussfaktoren 97
Anhang B: Abhängigkeiten der wichtigsten Variablen 98
Anhang C: Schadenstypen und ihre Einflussfaktoren 104
Anhang D: GompitZ: Kohorten und Überlebenskurven 107
Anhang E: Einfluss der Parameter auf Modellgüte 111

Anhang F: Zusammengefasste Modellergebnisse 120
Tabellenverzeichnis 121
Abbildungsverzeichnis 123
Literaturverzeichnis 128

1 Vorwort

1.1 Historische Entwicklung der Berliner Kanalisation

Die Bevölkerung in Deutschland mit Trinkwasser zur versorgen und deren Abwasser ordnungsgemäß zu entsorgen, ist eine Aufgabe die mit höchsten Ansprüchen ausgeführt werden muss. Diese Aufgabe wird in unterschiedlichem Maße durch eine Infrastruktur von Pump-, Wasser- und Klärwerken realisiert. Über Trinkwasserleitungen, Kanäle und Abwasserdruckrohrleitungen erreicht das Wasser bzw. Abwasser seinen jeweiligen Bestimmungsort.

Die Ausprägung der modernen Berliner Infrastruktur ist das Ergebnis einer langfristigen technologischen Entwicklung die ihren Ursprung im 19. Jahrhundert hat. Die zentrale Wasserversorgung Berlins war im Jahr 1856 die erste größere Netzinfrastruktur in der Stadt und löste die weit verbreitete dezentrale Brunnenversorgung ab.

Die ersten Fachplanungen für eine unterirdische Ableitung von Abwasser und Regenwasser fanden im Jahr 1861 durch Friedrich Eduard Salomon Wiebe und anschließend durch James Hobrecht 1871 erfolgreich statt.

Abbildung 1: Der Hobrechtplan zur Kanalisation von Berlin, 1871 (aus Bärthel 2003)

Der durch Hobrecht erarbeitete Generalbericht über die „Canalisation von Berlin“ sieht eine Unterteilung des Kanalsystems in zwölf Entwässerungssysteme, auch Radialsysteme genannt, vor. Diese grenzen sich durch die geologischen Berliner Gegebenheiten z.B. durch Flüsse oder Höhenrücken ab. In jedem dieser Teilsysteme enden die Kanäle an einem geografischen Tiefpunkt möglichst nahe an einem Wasserlauf. An dieser Stelle wurde ein

Pumpwerk errichtet, das über Druckleitungen das Abwasser auf Rieselfelder vor den Toren der Stadt pumpt. Die Zusammenführung von Schmutz- und Regenwasser zu einer Mischkanalisation ist heute noch im Innenstadtbereich vorherrschend.

Der Spatenstich für die Berliner Kanalisation erfolgte am 14. August 1873; die bauliche Realisierung für die Kernstadt dauerte bis 1893. Bereits im Jahr 1897 betrug die Länge des gebauten Kanalnetzes 1.167 km (Bärthel 2003). Die Überlegungen zu einer getrennten Ableitung von Schmutz- und Regenwasser, genannt Trennsystem, kamen erstmals im Jahr 1895 im Zuge der Erschließung von Steglitz auf.

1.2 Berliner Kanalisation als Investitionsgut

Heutzutage ist die Kanalisation ein deutschlandweit weitestgehend unbeachtetes Investitionsgut der Kommunen, Länder oder des Bundes. Dabei erstrecken sich die öffentlichen Abwasserkanäle deutschlandweit über eine Länge von ca. 500.000 km mit einem Anlagevermögen von schätzungsweise 700 Mrd. EUR (Biegel 2016).

Die Berliner Wasserbetriebe, als kommunaler Betreiber der Berliner Wasserver- und Abwasserentsorgung, haben mit über 9.700 km Kanalnetz mit einem Anlagenbuchwert von rd. 2,7 Mrd. EUR (12-2016) einen ihrer größten Vermögenswerte unter der Erde. Das Kanalnetz unterteilt sich in rd. 2.000 km Mischwasserkanal, 3.300 km Regenwasserkanal und 4.400 km Schmutzwasserkanal.

Ein nachhaltiger Substanzerhalt durch künftige Investitionen in das Kanalnetz ist im Gegensatz zu überirdischen baulichen Anlagen ein erschwerter Prozess, da eine Zustandsbewertung nur mit aufwendigen technologischen Hilfsmitteln durchführbar ist. Für die Zustandsbewertung werden in der Regel Kamerainspektionen eingesetzt. Die Inspektion dient der Untersuchung und Dokumentation des Kanalzustands in Bezug auf seine Standsicherheit (S), Betriebssicherheit (B) und Dichtheit (D).

Die Inspektions-Kamera ist ein Roboter, der durch den Kanal gesteuert wird und dabei eine Video-Aufnahme (Dreh-Schwenkkopf-Kamera) oder eine aneinander gereihte 360-Grad-Bilder-Serie (Kugelbildscanner) macht. Auf dieser Grundlage ist es durch den Inspekteur möglich, die erkannten Schäden zu verorten und ihr Ausmaß zu beschreiben. In Berlin werden aktuell die Schäden nach dem Berliner Schadenskatalog (V11) kodiert und nach den Schutzzielen Standsicherheit, Betriebssicherheit und Dichtheit in die nachstehenden Schadensklassen übersetzt (Abbildung 2).

Dabei wird unterschieden zwischen *dringendem* Sanierungsbedarf (SK 1A), *kurz- bis mittelfristigem* Sanierungsbedarf (SK 1B-2A), *mittel- bis langfristigem* Sanierungsbedarf (SK 2B-4) und keinem Handlungsbedarf (SK 5). Haltungen in einem Wasserschutzgebiet erfahren einen strengeren Sanierungsbedarf als außerhalb eines Wasserschutzgebietes, da hier zusätzlich noch die dichtheitsrelevanten Schäden der Schadensklasse 2b und 3 als kurz- bis mittelfristig sanierungsbedürftig eingestuft.

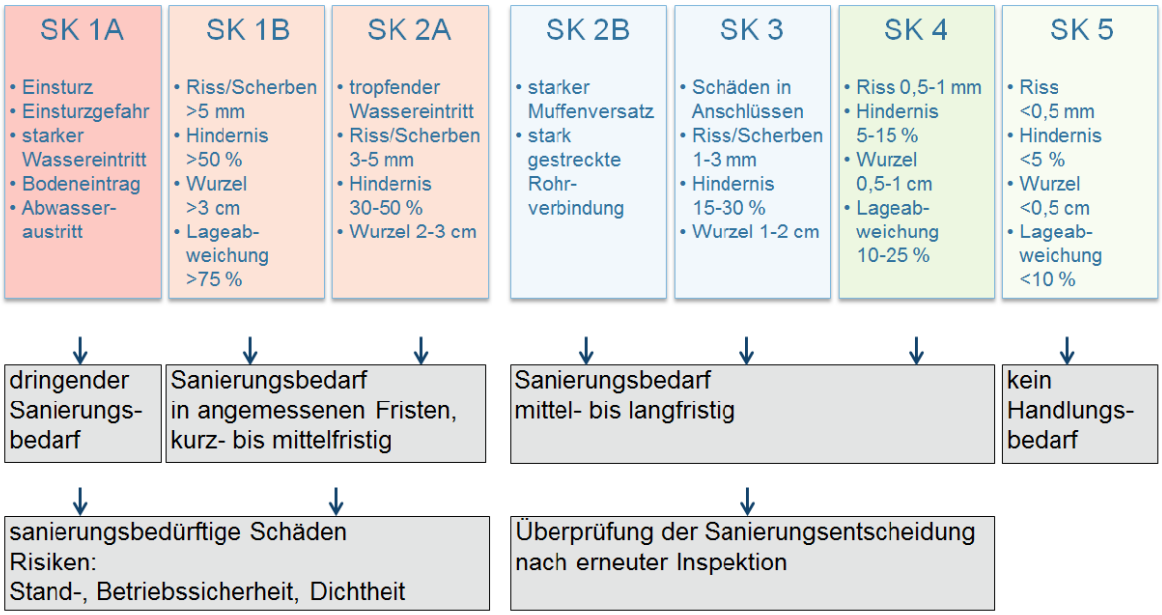


Abbildung 2: Definition des Sanierungsbedarfs außerhalb der Wasserschutzgebiete nach Schadensklassen

Umfragen der DWA unter Abwasserentsorgern in Deutschland zeigen, dass der Anteil an Haltungen mit schweren Schäden (Klasse 1A-2A) zwischen 2009 und 2015 deutschlandweit von 17% (Berger und Falk 2011) auf 24% (Berger et al. 2015) angestiegen ist. Abbildung 3 zeigt ausgewählte Schadensbilder aus der Inspektion.

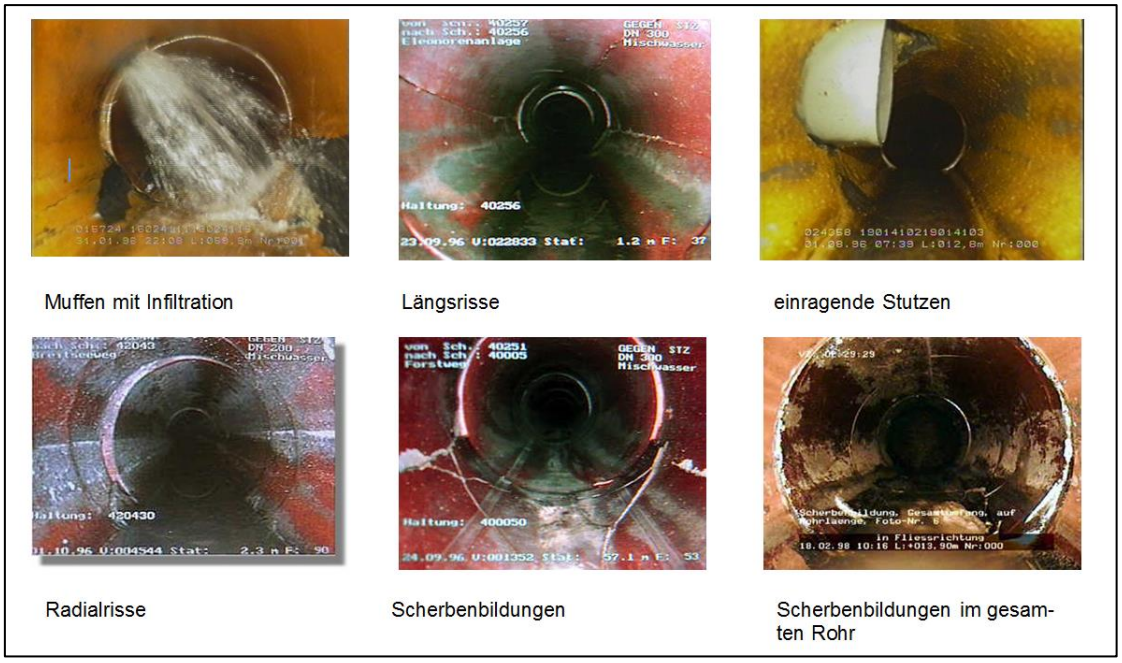


Abbildung 3: Mögliche Schadensbilder in einem Kanal (nicht klassifiziert)

Die Klassifizierung der Einzelschäden dient als Grundlage für die priorisierte Kanalsanierung nach Schwere und Gefährdungspotential im Rahmen der Kanalsanierungsstrategie 2011 (KSS 2011).

Mit Erfassung der Einzelschäden erfolgt auch eine Berechnung einer Zustandspunktzahl für die gesamte Haltung[1]. Diese wird im Wesentlichen durch die Schadensklasse des schwersten Einzelschadens bestimmt. Im Detail wird zuerst die Schadenspunktzahl aus der Bewertung der Schadensintensität (Schadensklasse Einzelschaden), der Bewertung des Gefährdungspotentials (Umweltgefährdung, Betriebsgefährdung) und einem Streckenzuschlag aus dem Quotient von Schadenslänge zu Haltungslänge ermittelt. Anschließend ergibt sich über eine Gewichtung der Schadenspunktzahl eine Zustandspunktzahl, anhand der eine Zustandsklasse abgeleitet wird. Es ergibt sich eine gesamte Zustandsklasse von „1“-umgehende Schadensbehebung bis „6“-ohne Schaden. Für einige Analysen, insbesondere die Modellierung, wurden die Zustandsklassen in Absprache mit der Abteilung Abwasserentsorgung (BWB-AE) in drei Zustandsbereiche eingestuft. Die Einstufung ist in Kap. 2.3.2 erläutert.

1.3 Alterungsverhalten und Zustandsprognosen der Kanalisation

Die Inspektion des Berliner Kanalnetzes mit einer Länge von 9.700 km ist ein fortlaufender Vorgang. Pro Jahr werden circa 750 km inspiziert. Daraus ergibt sich rechnerisch für die einmalige Inspektion des gesamten Kanalnetzes eine Gesamtdauer von ca. 13 Jahren. In der Realität sind turnusmäßige Inspektionsintervalle von ca. 20 Jahren (mit Ausnahme der Wasserschutzzonen II, III und IIIa) vorgesehen.

Mit fortschreitender Zeit kann sich der Zustand des Netzes infolge Nutzung und Ereignissen verändern, so dass die Aktualität einer Inspektion abnimmt. Sie ist nur eine Momentaufnahme des Kanalzustands zum Zeitpunkt der Inspektion, die das Gesamt-System "Umgebung - Kanal" nicht erfassen kann.

Um längerfristige Strategien entwickeln zu können und die weitere Entwicklung des Kanalnetzes auf der Grundlage des derzeitigen bzw. eines früheren inspizierten Zustands zu prognostizieren, sind Alterungsmodelle entwickelt worden. Alterungsmodelle können verwendet werden, um (i) die Zustandsklasse von nicht inspizierten Haltungen zu simulieren und (ii) die Entwicklung des Zustands des Netzes zu prognostizieren.

Mehrere Modellansätze/ -theorien sind inzwischen verfügbar, aber werden nur in geringem Umfang von Kanalisationsbetreibern und Kommunen verwendet, um Strategien zu unterstützen. Die angewandten Modelle haben bisher oft nicht den Nachweis erbracht, dass sie den zukünftigen Zustand angemessen prognostizieren können.

Die Validierung von Alterungsmodellen ist damit eine primäre Aufgabe, die erfüllt werden muss, um (i) den Nutzen von Modellierungsansätzen im Hinblick auf die Festlegung von Inspektions- und Sanierungsstrategien nachzuweisen und (ii) Vertrauen auf der Ebene der Anwender (Stadtwerke, Kommunen, Verbände) in Bezug auf die Verwendung von Modellen aufzubauen.

Das Projekt „SEMA-Berlin“ soll die Anwendungsmöglichkeit von Kanalalterungsmodellen bei den Berliner Wasserbetrieben unter Berücksichtigung der Prognosequalität überprüfen. Neben den Kanaleigenschaften soll die Berücksichtigung von berlinspezifischen Umgebungsparametern (z.B. Grundwasser, Baumbestand, etc.) dabei helfen. Die Hauptfragestellungen sind:

[1] Kanalabschnitt zwischen zwei Einstiegsschächten

- Welche sind die wichtigsten Einflussfaktoren für den Zustand der Berliner Kanalisation?
- Welches sind die häufigsten Schäden der Kanäle und wovon hängen diese ab?
- Wie hoch sind die Unsicherheiten in der Zustandsbewertung basierend auf Inspektionen?
- Mit welcher Genauigkeit lässt sich der Zustand der Kanalisation vorhersagen?
- In wie weit können die Modelle die strategische Investitionssteuerung oder die Identifizierung prioritärer Haltungen für Inspektionen oder Sanierungen unterstützen?

Das Projekt wurde in mehrere Arbeitspakete unterteilt:

I. Datenvorbereitung,
II. Statistische Analysen zum Zustand der Kanäle, Art und Häufigkeit von Einzelschäden und Einflussfaktoren.
III. Aufbau und Test verschiedener Modellansätze.

Das Projekt „SEMA-Berlin“ ist ein Forschungsprojekt mit Beteiligung des Kompetenzzentrums Wasser Berlin und den Berliner Wasserbetrieben. Im Fokus des Projekts sind zwei Anwendungsstränge: eine operative Anwendung zur Unterstützung einer zielgerichteten Kanalbefahrung und eine strategische Ausprägung für die Simulation des zukünftigen Netzzustandes unter variablen Sanierungsverhalten.

Im vorliegenden Bericht werden zunächst die verwendeten Daten (Variablen, Anzahl Datensätze, Häufigkeitsverteilung, etc.) sowie das Vorgehens bei der Datenvorbereitung beschrieben (Kapitel 2). In Kapitel 3 werden Methodik und Ergebnisse der statistischen Analyse zur Zustandsverteilung, Schadensauftreten, den wichtigsten Einflussfaktoren sowie den Inspektionsunsicherheiten vorgestellt. In Kapitel 4 werden die getesteten Modelle beschrieben, die allgemeine Methodik beim Modellaufbau beschrieben und die Ergebnisse der Modellierung diskutiert. Kapitel 5 enthält Empfehlungen und Schlussfolgerungen, die sich aus den Arbeiten ergeben haben.

2 Datengrundlage und -vorbereitung

2.1 Übergebene Daten

Die für die statistische Analyse und die Modellierung verwendeten Daten wurden in Form einer MS-Access-Datenbank von den Berliner Wasserbetrieben übergeben (Stand Januar 2017). Die Datenbank besteht aus mehreren Tabellen, die sich in i) Daten zum Zustand des Kanalnetzes und ii) Informationen zu den Eigenschaften und Umweltfaktoren der einzelnen Haltungen unterscheiden lassen. Es wurden nur die nach BWB-Schadenskatalog 11 (BWB 2001) erfassten Inspektions- und Schadensdaten verwendet (Zeitraum 2001 bis 2016).

Die Daten zur Zustandsbeschreibung beinhalten die haltungsscharfen Bewertungsergebnisse aus den Kamerabefahrungen (140.690 Datensätze, 4.825 km inspizierte Kanallänge, siehe Kap. 2.3.2). Sie werden ergänzt durch Informationen zum verwendeten Kameratyp (z.B. Dreh-Schwenkkopf-Kamera oder Kugelbildscanner) und einer Tabelle mit allen kodierten Einzelschäden mit der jeweiligen Schadensklasse (1998018 Datensätze). Die Verknüpfung der Daten erfolgt über eine Inspektions-ID.

Die Daten zu den Kanaleigenschaften und Umweltfaktoren beinhalten die sogenannten Stammdaten der Berliner Wasserbetriebe mit Informationen zu baulichen und betrieblichen Eigenschaften aller Haltungen sowie ausgewählten äußeren Einflussfaktoren (235.988 Datensätze, 9.661 km Gesamtlänge). Sie werden ergänzt um Informationen zur geografischen Lage (Bezirk, Ortsteil), zur Nähe zu Bäumen, zur Grundwasser-Überdeckung, zur Bodenart und zum U-Bahn-Verkehr, die über die Senatsverwaltung für Umwelt, Verkehr und Klimaschutz bezogen wurden. Nähere Erläuterungen dazu sind in Kapitel 2.3.1 zu finden. Die Verknüpfung der Daten erfolgt über die Haltungs-ID.

Die Zustandsdaten und die Informationen zu den Eigenschaften und Umweltfaktoren der Haltungen wurden über eine gemeinsame Haltungs- bzw. Objekt-ID miteinander verknüpft. Die 140.690 übergebenen Inspektionsdatensätze beziehen sich auf 118.474 unabhängige inspizierte Haltungen. Das heißt, für 50,2% aller Haltungen der Berliner Abwasserkanalisation liegt aus dem Zeitraum 2001 bis 2016 mindestens eine Zustandsbewertung vor (nicht eingenommen der Nachträge nach Januar 2017). 22.216 Inspektionsdatensätze (15,8%) beziehen sich auf wiederholte Befahrungen. Die Zahlen beziehen sich auf die unbereinigten und ungefilterten Rohdaten.

2.2 Datenfilterung und -bereinigung

Bevor die Daten für statistische Analyse und die Modellierung verwendet werden konnten, mussten einige Datensätze gefiltert werden. Dabei wurden beispielsweise Datensätze entfernt, die sich auf renovierte, reparierte und teilerneuerte Kanäle beziehen. Außerdem wurden Datensätze von erneuerten Kanälen gefiltert, die sich auf den Zustand vor der Erneuerung beziehen, bei denen die entsprechenden Stammdaten (Alter, Material, etc.) aber überschrieben wurden. Datensätze, die doppelt vertreten waren (Duplikate) und Inspektionen, die nicht dem Zweck der Zustandsbewertung dienten, wurden ebenfalls entfernt. Tabelle 1 zeigt die berücksichtigten Filterschritte. Für die Filterung wurden zusätzliche Tabellen mit Informationen zu renovierten (Liner) und teilerneuerten Kanälen (Sanierungen) übergeben.

Tabelle 1: Filterschritte zur Entfernung nicht zu bewertender Datensätze

Filterschritte	Filterkriterien	Anzahl Datensätze [5]
Ausschluss aller Drainagekanäle	Entwässerungstyp „SD“, „RD“ oder „MD“	140
Ausschluss aller Kompletterneuerungen (vor 2012)	Inspektionsdatum < Baudatum [1]	3.359
Ausschluss der reparierten Kanäle	Codierung „K“ in Schaden-Tabelle	12.912
Ausschluss der renovierten Kanäle (Liner)	Kommen in Liner-Tabelle vor; Linerdatum < Inspektionsdatum	2.227
Ausschluss der teilerneuerten Kanäle	Kommen in Sanierungs-Tabelle vor; Sanierungsdatum < Inspektionsdatum	6.502
Ausschluss der Doppelinspektionen am selben Tag	Wenn Haltung zweimal am selben Tag inspiziert wurde, behalte nur eine [2]	265
Ausschluss nicht untersuchter Kanäle	Untersuchungstyp „A“, „B“, „K“, „N“ oder Unbekannt [3]	5.769
Ausschluss nicht bewerteter Kanäle	Zustandsklasse 7 oder 0 [4]	2.352

[1] *Bis 2012 wurden Stammdaten bei Erneuerung überschrieben, so dass Inspektionen der ursprünglichen Haltung (vor Erneuerung) nicht ausgewertet werden können.*
[2] *In den meisten Fällen lieferten beide Inspektionen dasselbe Ergebnis; falls nicht, wird das schlechtere behalten.*
[3] *Untersuchungstypen: „A“ – Anschlussnummerierung, „B“ – Begehung, „K“ – Druckprüfung, „N“ – nicht untersucht. Untersuchungen dienten nicht der Zustandserfassung.*
[4] *Befahrungslänge < 30% der Haltungslänge → Kanäle erhalten Zustandsklasse 7 oder 0.*
[5] *Einige Datensätze erfüllen mehrere Filterkriterien. Daher ist die Anzahl der insgesamt gefilterten Datensätze kleiner als die Summe der Spalte „Anzahl Datensätze (je Kriterium)“.*

Insgesamt wurden 25.432 Datensätze (18%) entfernt. Nach Filterung der Daten bleiben noch 115.258 von ursprünglich 140.690 Datensätzen erhalten (82%).

Über die Filterung hinaus, wurden unplausible Werte aus den Daten entfernt. Dabei wurde nur der betroffene Eintrag entfernt (z.B. Breite = 0 mm), die anderen Informationen zur entsprechenden Haltung (z.B. Alter, Material, etc.) bleiben erhalten. Das heißt, der Datensatz bleibt bestehen, weist aber im entsprechenden Feld eine Lücke („NA“ - nicht angegeben) auf. Folgende Datenbankeinträge wurden bereinigt (Tabelle 2):

Tabelle 2: Auflistung der entfernten Datenbankeinträge

Variable	Wert(ebereich)	Anzahl Einträge [1]
Material	„Ohne“	92.170
Höhe	0 m	2
Breite	0 mm	2
Länge	0 m	4
Baujahr	≥ 2017	30.321
Geländeoberkante (am Anfangsschacht)	≥ 99 m; = 0 m	101; 97
Geländeoberkante (am Endschacht)	≥ 99 m; = 0 m	139; 76
Höhe Kanalsohle (am Anfangsschacht)	≥ 99 m	2
Höhe Kanalsohle (am Anfangsschacht)	≥ 99 m	2
Tiefe	< 0 m	15
Überdeckung	< 0 m	7

[1] *Die Anzahl bezieht sich auf alle 235.988 inspizierten und nicht inspizierten Haltungen.*

2.3 Beschreibung der Variablen

Hauptziel der statistischen Analyse und der Modellierung ist es, einen Zusammenhang zwischen dem baulichen Zustand (Zielvariable) und verschiedenen Kanaleigenschaften sowie äußeren Einflussfaktoren (erklärende Variablen) herzustellen und modellhaft nachzubilden.

2.3.1 Erklärende Variablen

Die erklärenden Variablen bestehen aus baulichen und betrieblichen Eigenschaften sowie äußeren Einflussfaktoren und können in numerische und kategorische Variablen unterschieden werden. Während numerische Variablen aus kontinuierlich skalierbaren Zahlenwerten bestehen (z.B. Alter, Breite, etc.), haben kategorische Variablen festdefinierte Ausprägungen, d.h. Zustände oder Kategorien (z.B. Material: Beton, Steinzeug, etc.). In Tabelle 3 sind die verschiedenen Variablen aufgelistet und erläutert.

Für bestimmte Analyseverfahren (z.B. Chi-Quadrat-Unabhängigkeitstest, siehe Kap. 3.1.1.2) müssen die numerischen Variablen in kategorischen Variablen überführt, d.h. klassifiziert, werden. Die Klassifizierung wurde in Absprache mit BWB vorgenommen und ist ebenfalls in Tabelle 3 beschrieben.

Tabelle 3: Eingangsvariablen für die statistische Analyse und die Modellierung

Variable	Erläuterungen	Typ (n / k) [1]	Klassifizierung
Baujahr	Jahr, in dem die Baumaßnahme abgeschlossen wurde	n, k	< 1900; 1900-1949; 1950-1990; >=1991
Alter	Alter zum Zeitpunkt der Inspektion	n, k	0-25 a; 26-50 a; 51-75 a; 76-100 a; >= 100 a
Material	Werkstoff, aus dem die Haltung gefertigt ist	k	Steinzeug; Beton; Beton mit hoher Tragfähigkeit (wandverstärkter Beton und Stahlbeton); Asbest-Zement; Mauerwerk; PVC-U und Andere (Grauguss, Edelstahl, etc.)
Abwassertyp	Medium, das im Kanal transportiert wird	k	Mischwasser; Regenwasser; Schmutzwasser
Profil	Form des Kanalquerschnitts	k	Kreis; Kreis im Vortrieb; Ei; Maul und Andere (Kasten- und andere Sonderprofile)
Breite	Maximale Breite der Haltung	n, k	< 250 mm; 250-399 mm; 400-599 mm; 600-999 mm; >= 1000 mm
Höhe	Maximale Höhe der Haltung	n, k	< 250 mm; 250-399 mm; 400-599 mm; 600-999 mm; >= 1000 mm
Länge	Länge der Haltung	n, k	< 20 m; 20-40 m; 40-60 m; >= 60 m [2]
Tiefe	Differenz zwischen Geländeoberkante und Kanalsohle	n, k	< 1 m; 1-2 m; 2-3 m; 3-4 m; >= 4 m [2]
Überdeckung	Differenz zwischen Geländeoberkante und Kanaldecke	n, k	< 1 m; 1-2 m; 2-3 m; 3-4 m; >= 4 m [2]
Gefälle	Höhendifferenz zwischen Kanalsohle am Anfangs- und am Endschacht geteilt durch Länge der Haltung	n, k	< 0,3%; 0,3-0,6%; 0,6-0,9%; >= 1,2% [2]
Straßenklasse	Befahrungsgrad der anliegenden Straßen; ermittelt über GIS-Verschneidung mit Straßenkarte; 20 m Puffer um jede Straße (je 10 m rechts und links)	k	Straßenklassen 1, 2, 3 und 4 (viel bis wenig Verkehr) und Andere (nicht in Berliner Klassifizierung erfasste Nebenstraßen)
…	…	…	…

Variable	Erläuterungen	Typ (n / k) [1]	Klassifizierung
Schienen-verkehr	Beeinflussung durch Schienenverkehr: liegen im 3-m-Radius um Haltung Gleise oder kreuzen diese?; ermittelt über GIS-Verschneidung von Informationen zur Lage von Tram- Und U-Bahn-Gleisen	k	Ja; Nein
Bäume	Stehen im 3-m-Radius um Haltung Bäume?; ermittelt über GIS-Verschneidung mit Informationen zu Baumstandorten	k	Ja; Nein
Anzahl Bäume	Anzahl an Bäumen im 3-m-Radius um Haltung; ermittelt über GIS-Verschneidung mit Informationen zu Baumstandorten	n, k	0, 1-2; 3-4; 5-6; >= 7
Baumdichte (longitudinal)	Berechnet aus Quotienten aus Anzahl Bäume (siehe oben) und Länge der Haltung	n, k	0; 0-5; 5-10; >10 [3]
Grundwasser-überdeckung (j/n)	Ist Haltung dauerhaft von Grundwasser überdeckt?; Ermittelt über GIS-Verschneidung mit Daten zum langjährig mittleren Grundwasserstand	k	Ja; Nein
Grundwasser-überdeckung (m)	Höhe der Grundwasser-überdeckung der Haltung bezogen auf die Kanalsohle; Ermittelt über GIS-Verschneidung mit Daten zum langjährig mittleren Grundwasserstand	n, k	< 1 m; 1-2 m, >= 2 m [3], keine
Bodentyp	Bodentyp, in dem der Kanal liegt	k	Aufschüttung, Sand, Sand+Lehm, Andere
Rückstau	Ist Kanal dauerhaft rückgestaut?; betrifft nur Regenkanäle	k	Ja; Nein
Bezirk	Berliner Stadtbezirk, in dem der Kanal liegt	k	12 Berliner Stadtbezirke (Unbekannt: Kanäle im Umland)
Stadtteil	Liegt der Kanal im ehemaligen Ost- oder Westteil der Stadt?	k	Ost; West (Unbekannt: Kanäle im Umland)

Erläuterungen: [1] Typ: n - numerisch; k – kategorisch; [2] die linken Grenzen sind in den Intervall eingeschlossen; [3] die rechten Grenzen sind in den Intervall eingeschlossen

Einige für die Auswertung relevante Variablen lagen nicht direkt vor und mussten aus vorhandenen Informationen berechnet werden:

- *Tiefe:* Berechnet aus der Höhendifferenz zwischen Geländeoberkante und Kanalsohle (Mittelwert über Anfangs- und Endschacht),
- *Überdeckung:* Berechnet aus der Höhendifferenz zwischen Geländeoberkante und Kanaldecke (Mittelwert über Anfangs- und Endschacht); die Kanaldecke wurde durch Addition von Höhe der Kanalsohle und Höhe des Kanals ermittelt;
- *Gefälle:* Berechnet aus der Differenz zwischen Kanalsohle am Anfangs- und Endschacht geteilt durch die Haltungslänge,
- *Alter zum Zeitpunkt der Inspektion:* Berechnet aus Differenz zwischen Inspektions- und Baujahr,
- *Baumdichte (longitudinal):* Berechnet aus dem Quotienten von Baumanzahl und Haltungslänge.

2.3.2 Zielvariable: baulicher Zustand

Zielgröße der statistischen Analyse und der Modellierung ist der bauliche Zustand. Nach dem Berliner Bewertungsmodell werden sechs Zustandsklassen unterschieden, wobei Klasse 6 den sehr guten (schadensfreien) und Klasse 1 den sehr schlechten Zustand repräsentiert. Die Zustandsklasse ist eine kategorische Variable, auch wenn die Werte ordinalskaliert sind (1 bis 6). Aus der Zustandsklasse lässt sich der notwendige Sanierungsbedarf herleiten:

- Zustandsklasse 1: umgehende Schadensbehebung (sofort),
- Zustandsklasse 2: kurzfristige Schadensbehebung (Zeithorizont: 5 Jahre),
- Zustandsklasse 3: mittelfristige Schadensbehebung (Zeithorizont: 10 Jahre),
- Zustandsklasse 4: langfristige Schadensbehebung (Zeithorizont: > 10 Jahre),
- Zustandsklasse 5: Schadensbehebung im Rahmen andere Baumaßnahmen,
- Zustandsklasse 6: ohne Schaden.

Für einige Analysen, insbesondere die Modellierung, wurden die Zustandsklassen in Absprache mit der Abteilung Abwasserentsorgung (BWB-AE) in drei Zustandsbereiche eingestuft (gut, mittel, schlecht, Abbildung 4), die sich am Sanierungsbedarf orientieren.

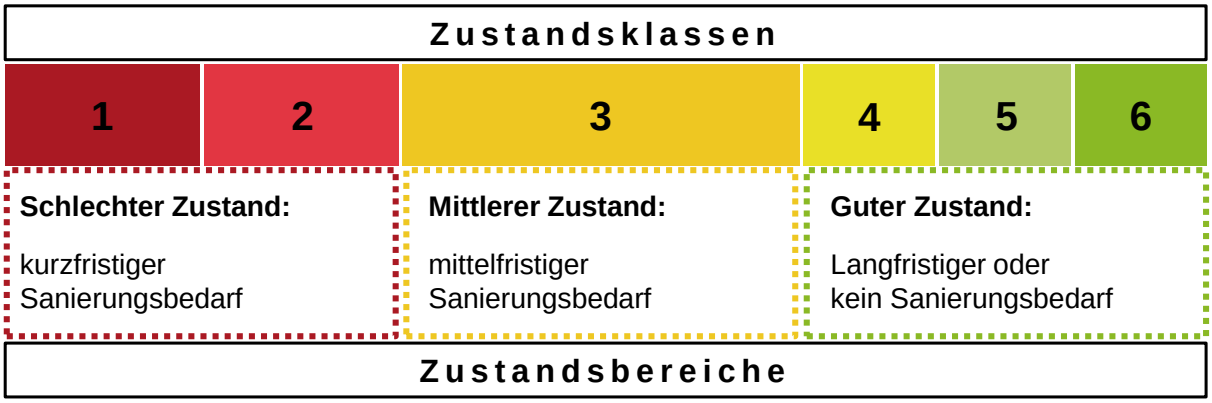


Abbildung 4: Bei den Berliner Wasserbetrieben unterschiedene Zustandsklassen und für die Modellierung aggregierten Zustandsbereiche für die Bewertung der Kanäle.

2.4 Ist-Analyse und Datenverteilungen

Insgesamt standen 115.258 bereinigte Inspektionsdatensätze für 102.858 unterschiedliche Haltungen mit einer Gesamtlänge von 4.303 km zur Verfügung. 12.400 der 115.258 Datensätze beziehen sich auf Zweit-, Dritt-, Viert- oder Fünft-Befahrungen (11%). Im Folgenden

sind die Häufigkeitsverteilungen der einzelnen Variablen dargestellt und erläutert. Die Auswertungen beziehen sich auf alle Inspektionsdatensätze (n = 115.258), wobei mehrfach inspizierte Haltungen mehrfach vertreten sind.

2.4.1 Baujahr, Alter, Material und Abwassertyp

Der älteste inspizierte Kanal war zum Zeitpunkt der Inspektion 140 Jahre alt (Baujahr 1868). Über die Hälfte der Inspektionen (56%) bezieht sich auf Kanäle, die zum Zeitpunkt der Inspektion nicht älter als 50 Jahre waren. Das vorherrschende verbaute Material ist Steinzeug (63%), gefolgt von Beton (18%), Asbestzement (6%) und Beton mit hoher Tragfähigkeit (5%). Die meisten Inspektionen wurden in Schmutzwasserkanälen durchgeführt (49%); danach folgen Regenkanäle (33%) und Mischkanäle (18%). Die tatsächliche Verteilung der Abwassertypen übers gesamte Berliner Kanalnetz, inkl. der nicht inspizierten Kanäle, liegt bei 44% Schmutz-, 37% Regen- und 19% Mischwasser. Aus den absoluten Zahlen lässt sich ableiten, dass die Inspektionsquote von Schmutzwasserkanälen mit 55% etwas höher ist als die von Mischwasser- (45%) und Regenwasserkanälen (44%). Die Datenverteilung für die beschriebenen Variablen ist in Abbildung 5 dargestellt.

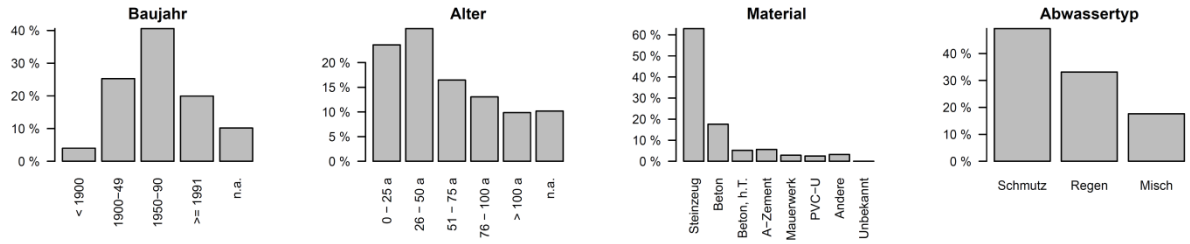


Abbildung 5: Baujahr, Alter, Material und Abwassertyp der inspizierten Haltungen

2.4.2 Profil, Breite, Höhe und Länge

96% der Inspektionen wurden in Kanälen mit Kreisprofil durchgeführt (mit und ohne Vortriebsbauweise). Weitere 3% beziehen sich auf Eiprofile. Maul- oder Sonderprofile kommen nur selten vor (in Summe < 1%). Die Breite der inspizierten Kanäle liegt im Bereich von 100 bis 5.840 mm, wobei sich 74% der Datensätze auf Haltungen < 400 mm beziehen. Da die inspizierten Haltungen zum größten Teil aus Kreisprofilen bestehen, sind die Verteilungen der Kanalhöhe und -breite sehr ähnlich (die maximale Höhe beträgt 3.500 mm). Die Länge der Haltungen beträgt zwischen 0,34 und 306 m, wobei der Schwerpunkt der Daten (41%) im Bereich von 40 bis 60 m Länge liegt. Die Datenverteilung für die beschriebenen Variablen ist in Abbildung 6 dargestellt.

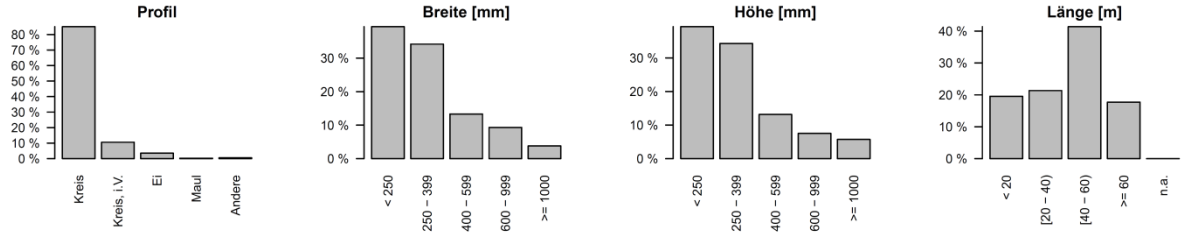


Abbildung 6: Profil, Breite, Höhe und Länge der inspizierten Haltungen

2.4.3 Tiefe, Überdeckung, Gefälle und Straßenklasse

Die Tiefe der inspizierten Kanäle (bezogen auf die Kanalsohle) liegt zwischen 0 und 18,5 m. Die maximale Überdeckung (bezogen auf die Kanaldecke) beträgt 16,7 m, wobei 76% der Kanäle eine Überdeckung zwischen 1 und 3 m aufweisen. Das mittlere Gefälle liegt bei 0,4% (Median). Nur 18% aller Haltungen weisen ein Gefälle ≥ 0,9% auf. Mehr als zwei Drittel aller Haltungen liegen in Nebenstraßen, die nicht durch die Berliner Straßenklassifizierung erfasst sind („Andere“). Von den übrigen Haltungen liegen jeweils die Hälfte in großräumigen oder übergeordneten Straßenverbindungen (Straßenklasse 1 und 2) bzw. in örtlichen Straßenverbindungen und Ergänzungsstraßen (Straßenklasse 3 und 4). Beispiele für die Straßenklassen sind: 1 – Unter den Linden, 2 – Wilhelmstraße, 3 – Französische Straße und 4 – Friedrichstraße. Die Datenverteilung für die beschriebenen Variablen ist in Abbildung 7 dargestellt.

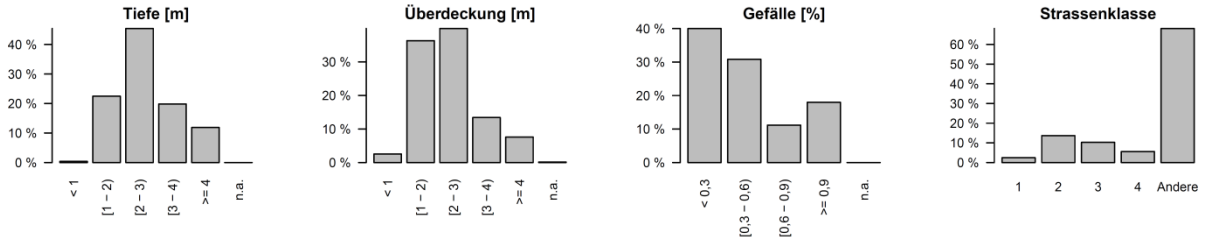


Abbildung 7: Tiefe, Überdeckung, Gefälle und Straßenklasse der inspizierten Haltungen

2.4.4 Schienenverkehr und Bäume

Nur 3% aller inspizierten Haltungen liegen unmittelbar neben U-Bahn- oder Tram-Gleisen (Abstand < 3 m) bzw. werden von diesen gekreuzt. Die gute Hälfte aller inspizierten Haltungen (51%) ist von mindestens einem Baum umgeben (Abstand < 3 m); die meisten davon (70%) jedoch von nicht mehr als vier Bäumen. Im Median aller Haltungen befindet sich alle 59 Meter ein Baum (Baumdichte: 1,7 Bäume / 100 m). Die Datenverteilung für die beschriebenen Variablen ist in Abbildung 8 dargestellt.

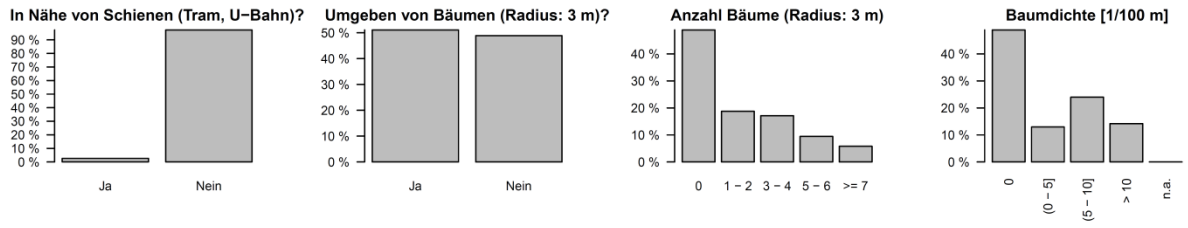


Abbildung 8: Beeinflussung der inspizierten Haltungen durch Schienenverkehr (Tram, U-Bahn) und Bäume (Ja/Nein, Anzahl, Dichte in Längsrichtung der Haltung)

2.4.5 Grundwasserüberdeckung, Bodenart und Rückstau

Nur wenige der inspizierten Haltungen (12%) sind dauerhaft vom Grundwasser überdeckt. Beim Großteil davon (63%) beträgt die Überdeckung < 1 m. Die maximale Überdeckung beträgt 7,33 m. Die häufigste Bodenart sind Aufschüttungen (36%), Sand (37%) und ein Gemisch aus Sand und Lehm (25%). Weniger als 2% aller Haltungen sind durch Rückstau aus dem Gewässer beeinflusst (betrifft potenziell Regenkanäle und Entlastungskanäle des Mischsystems). Die Datenverteilung für die beschriebenen Variablen ist in Abbildung 9 dargestellt.

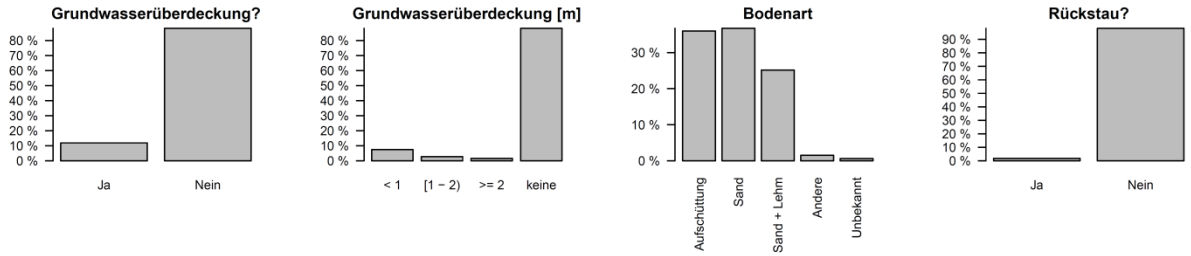


Abbildung 9: Grundwasserüberdeckung (Ja/Nein und Höhe), Bodenart und Rückstau-Beeinflussung der inspizierten Haltungen

2.4.6 Bezirk und Stadtteil

Die meisten der inspizierten Kanäle liegen in den Bezirken Steglitz-Zehlendorf (15%) und Treptow-Köpenick (14%). 46% der inspizierten Kanäle liegen im ehemaligen Ostteil, 54% im ehemaligen Westteil der Stadt. Das entspricht in etwa der tatsächlichen Verteilung über das gesamte Berliner Kanalnetz, inkl. der nicht inspizierten Kanäle (Ost: 46%, West: 53%). Die Datenverteilung für die beschriebenen Variablen ist in Abbildung 10 dargestellt.

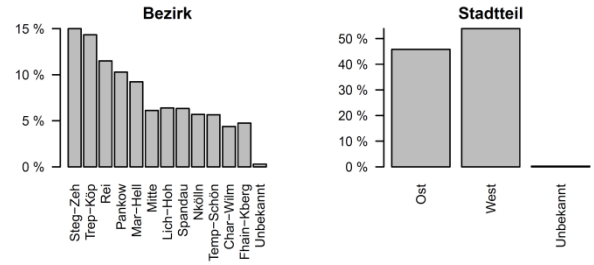


Abbildung 10: Bezirk und Stadtteil der inspizierten Haltungen

2.4.7 Baulicher Zustand

Von den inspizierten Haltungen sind 22% im schlechten oder sehr schlechten Zustand (Zustandsklasse 1 oder 2) und müssten sofort oder kurzfristig saniert bzw. erneuert werden. 24% der Haltungen sind im mittleren Zustand (Zustandsklasse 3) und müssten mittelfristig saniert bzw. erneuert werden (Zeithorizont: 10 Jahre). 54% der Haltungen sind tendenziell im guten Zustand (Zustandsklasse 4-6), d.h. eine Sanierung oder Erneuerung ist maximal langfristig erforderlich. Abbildung 11 zeigt die Zustandsverteilung für alle sechs Zustandsklassen (a) und für die drei aggregierten Zustandsbereiche (b, siehe Kap. 2.3.2).

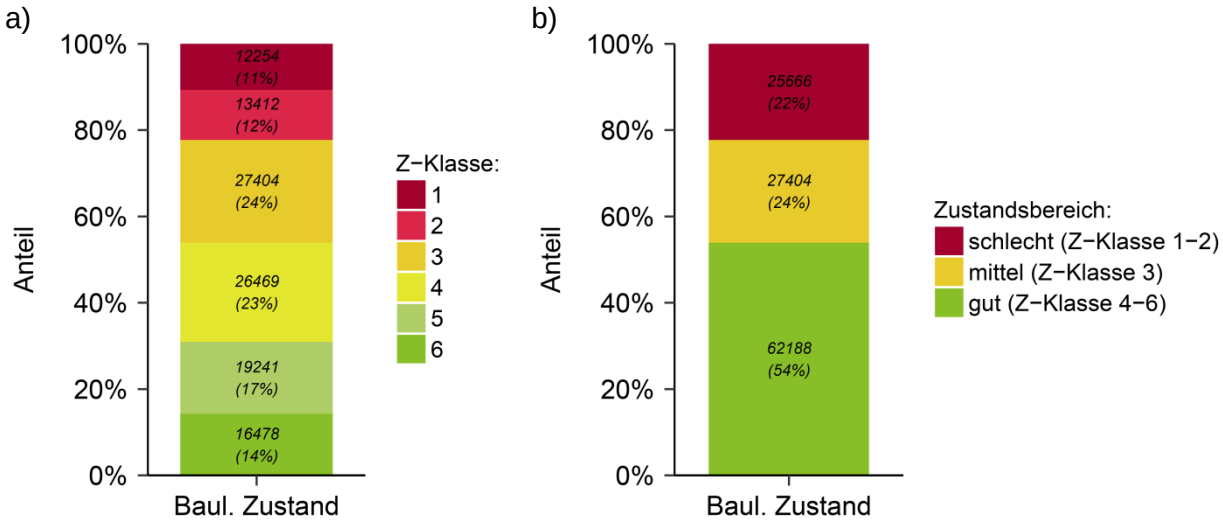


Abbildung 11: Zustandsverteilung über alle inspizierten Haltungen (a: sechs Zustandsklassen der BWB, b: drei aggregierte Zustandsbereiche, Aggregierung nach Absprache mit BWB-AE für den Untersuchungszweck, orientiert am Sanierungsbedarf)

3 Statistische Analyse zum Zustand der Abwasserkanäle

Hauptbestandteile der statistischen Analyse sind:

i) die Untersuchung der Abhängigkeiten der erklärenden Variablen untereinander (z.B. Alter, Material, Länge, etc., Kapitel 3.1),
ii) die Quantifizierung des Einflusses der erklärenden Variablen auf die Zustandsverteilung (Kapitel 3.2),
iii) die Analyse von Art und Häufigkeit der auftretenden Schäden sowie deren Einflussfaktoren (Kapitel 3.3) und
iv) die Quantifizierung der Unsicherheiten bei der Inspektion (Kapitel 3.4).

Die Analysen dienen neben einem besseren Verständnis der Alterungsprozesse und ihrer Einflussfaktoren vor allem der Auswahl an relevanten Eingangsvariablen für die Modellierung (Kap. 4). Die Quantifizierung der Unsicherheiten bei der Inspektion liefert zudem wichtige Erkenntnisse zu den Grenzen der Modellgüte und ist damit bedeutsam für die Interpretation der Modellergebnisse.

3.1 Analyse der Abhängigkeiten der Variablen

3.1.1 Methodisches Vorgehen

3.1.1.1 Korrelationsuntersuchungen nach Spearman

Die Abhängigkeit der numerischen Variablen, z.B. Alter oder Länge (siehe Kap. 2.3.1), wurde mit Hilfe des Rangkorrelationskoeffizienten nach Spearman (r_S) quantifiziert. Im Unterschied zum Korrelationskoeffizienten nach Pearson (r_P) wird die lineare Regression nicht zwischen den Datenpunkten selbst sondern zwischen ihren Rängen berechnet, d.h. es wird der monotone Zusammenhang zweier Variablen gemessen. Der Rangkorrelationskoeffizient nach Spearman r_S ist ein parameterfreies Maß, d.h. es müssen keine Voraussetzungen wie Normalverteilung der Daten erfüllt sein. Die Werte können zwischen -1 und 1 liegen ($r_S < 0$: negative Korrelation, $r_S \sim 0$: keine Korrelation, $r_S > 0$: positive Korrelation).

Abbildung 12 veranschaulicht die Bedeutung der Koeffizienten am Beispiel einer Exponentialfunktion. Die klassische lineare Regression zwischen den Zahlenwerten ergibt einen Pearson-Korrelationskoeffizienten von 0,69 (Abbildung 12, links). Daraus lässt sich ein Bestimmtheitsmaß r^2 von 0,48 berechnen. Wird die lineare Regression nicht zwischen den Zahlenwerten selbst sondern zwischen ihren Rängen berechnet, so ergibt sich für die monoton steigende Exponentialfunktion ein Rangkorrelationskoeffizient von 1,0 (Abbildung 12, rechts).

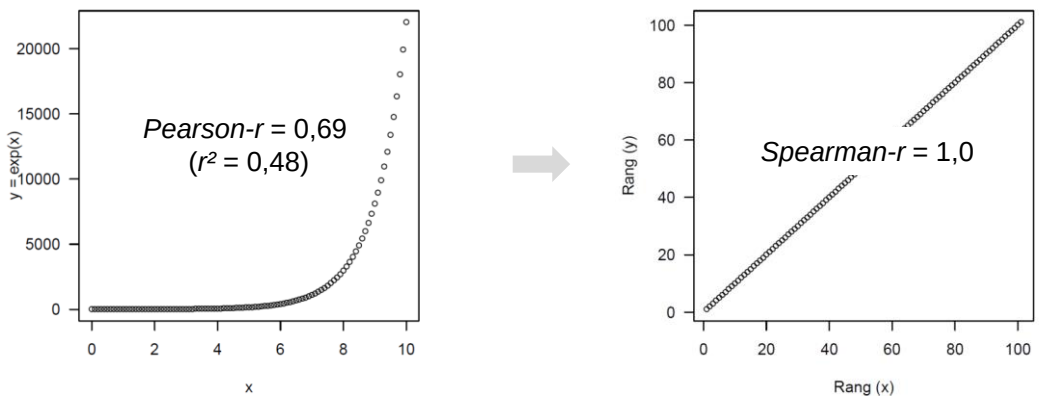


Abbildung 12: Berechnung des Korrelationskoeffizienten nach Pearson (links) und des Rangkorrelationskoeffizienten nach Spearman (rechts) am Beispiel der Exponentialfunktion

Der Spearman-Korrelationskoeffizient wurde für alle 55 möglichen Paare der elf numerischen Variablen (Alter, Baujahr, Breite, Höhe, Länge, Tiefe, Überdeckung, Gefälle, Anzahl Bäume, Baumdichte, Grundwasserüberdeckung) berechnet und dient im Wesentlichen der Identifizierung und dem Ausschluss von stark korrelierten Variablen. Ein Grenzwert von $|r_S| \geq 0{,}95$ wurde festgelegt, ab dem Variablen als sehr stark korreliert gelten und eine der Variablen verworfen wird, da sie keinen zusätzlichen Informationsgewinn erwarten lässt. Mehrfach inspizierte Kanäle wurden mehrfach berücksichtigt.

3.1.1.2 Chi-Quadrat-Unabhängigkeitstest und Cramér's V

Für die Untersuchung der übrigen numerischen und der rein kategorischen Variablen, z.B. Material oder Profil, wurde der Chi-Quadrat-Unabhängigkeitstest durchgeführt (Pearson 1900, Agresti 2007, McHugh 2013). Der Test untersucht, ob zwei Variablen signifikant voneinander abhängen oder ob Unterschiede in der Verteilung zufälliger Natur sind. Für die Analyse wurden die numerischen Variablen zunächst durch Bildung von Klassen in kategorische Variablen überführt (z.B. Länge mit dem Klassen < 20 m; 20-40 m; 40-60 m; ≥ 60 m, siehe Kap. 2.3.1).

Der Chi-Quadrat-Test basiert auf dem Zusammenhang zwischen beobachteten und erwarteten Häufigkeiten in den verschiedenen Merkmalsausprägungen. Dafür wird eine Kontingenz- bzw. Kreuztabelle mit den beobachteten Häufigkeiten von Kombinationen bestimmter Merkmalsausprägungen aufgestellt (Abbildung 13, Schritt 1) und mit den unter Annahme vollständiger Unabhängigkeit erwarteten Häufigkeiten (Abbildung 13, Schritt 2) verglichen. Die erwartete Häufigkeit für jede Zelle der Kreuztabelle ergibt sich aus dem Produkt von Zeilen- und Spaltensumme geteilt durch den gesamten Stichprobenumfang, siehe Formel 1 und Abbildung 13 (Rechenbeispiel in Schritt 2):

$$erwartete\ Häufigkeit = \frac{Zeilensumme * Spaltensumme}{Stichprobenumfang} \qquad \text{Formel 1}$$

Chi-Quadrat (X^2) wird wie folgt berechnet (aufsummiert über jedes Feld der Kreuztabelle):

$$\chi^2 = \sum \frac{(beobachtete\ Häufigkeit - erwartete\ Häufigkeit)^2}{erwartete\ Häufigkeit} \qquad \text{Formel 2}$$

Aus den berechneten X^2-Werten und der kumulierten Verteilungsfunktion der Chi-Quadrat-Verteilung können die Signifikanzwerte *p* abgeleitet werden. Wenn der *p*-Wert unter einem

Signifikanzniveau von 0,05 liegt, so kann von einer statistisch signifikanten Abhängigkeit beider Variablen gesprochen werden. Allerdings misst der *p*-Wert ausschließlich die Wahrscheinlichkeit, dass die Variablen überhaupt statistisch abhängig sind, und hat keine Aussagekraft bezüglich der Stärke des Zusammenhangs (Lin et al. 2013).

Als Maß für die Stärke des Zusammenhangs, die sogenannte Effektstärke, wurde *Cramér's V* (Cramér 1946) verwendet. Die Maßzahl basiert auf der Chi-Quadrat-Statistik (Formel 3) und liegt zwischen 0 (kein Zusammenhang) und 1 (vollständige Abhängigkeit der Variablen).

$$Cramér's\ V = \sqrt{\frac{\chi^2}{n \cdot \min(i-1,\ j-1)}} \qquad \text{Formel 3}$$

X^2 – Chi-Quadrat (Formel 2)
i – Anzahl der Ausprägungen von Variable 1
j – Anzahl der Ausprägungen von Variable 2
n – Anzahl der Datensätze

Cramér's V ist in etwa vergleichbar mit dem für numerische Variablen verwendeten Bestimmtheitsmaß r^2. Nach Cohen (1988) stehen *Cramér's-V*-Werte > 0,5 für einen großen, *Cramér's-V*-Werte zwischen 0,3 und 0,5 für einen mittleren und Werte < 0,3 für einen kleinen Effekt. Chi-Quadrat (X^2) selbst ist als Maß für die Effektstärke ungeeignet, da es vom Stichprobenumfang abhängt und nicht auf das Intervall 0 bis 1 beschränkt ist. Abbildung 13 zeigt die Berechnung von X^2 (Schritt 3) und Cramér's V (Schritt 4) anhand eines Beispiels.

Cramér's V wurde für alle 153 möglichen Paare der 18 kategorischen Variablen (Alter, Material, Abwassertyp, Profil, Breite, Länge, Überdeckung, Gefälle, Straßenklasse, Schienenverkehr (j/n), Bäume (j/n), Anzahl Bäume, GW-Überdeckung (j/n), GW-Überdeckung in m, Bodentyp, Bezirk, Stadtteil, Rückstau (j/n)) berechnet und dient der Identifizierung von Abhängigkeiten und ggf. dem Ausschluss von Variablen. Mehrfach inspizierte Kanäle wurden mehrfach berücksichtigt. Zur Veranschaulichung des Zusammenhangs zwischen den Variablen wurden die *Cramér's-V*-Werte in einer Korrelationsmatrix und in Netzdiagrammen dargestellt.

1. Beobachtete Häufigkeiten

Material

Alter	STZ	S-Beton	∑
Alt	60	20	80
Jung	40	40	80
∑	100	60	160

2. Erwartete Häufigkeiten

Material

Alter	STZ	S-Beton	∑
Alt	50	30	80
Jung	50	30	80
∑	100	60	160

Bsp.: $\frac{80 \cdot 100}{160}$

3. Berechnung von Chi-Quadrat

$$\chi^2 = \sum \frac{(beobachtete\ Häufigk. - erwartete\ Häufigk.)^2}{erwartete\ Häufigk.}$$

$$\chi^2 = \frac{(60-50)^2}{50} + \frac{(20-30)^2}{30} + \frac{(40-50)^2}{50} + \frac{(40-30)^2}{30}$$

$$\chi^2 = 2{,}0 + 3{,}3 + 2{,}0 + 3{,}3 = 10{,}7$$

4. Berechnung von Cramér's V

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(i-1, j-1)}} = \sqrt{\frac{10{,}7}{160 \cdot 1}} = 0{,}26$$

Abbildung 13: Berechnung von *Cramér's V* für ein einfaches Beispiel mit zwei Variablen (Alter und Material) mit je zwei Ausprägungen (Alt / Jung und Steinzeug / Stahlbeton) und insgesamt 160 Datensätzen.

3.1.2 Ergebnisse und Diskussion

Die Korrelationsanalyse der numerischen Variablen zeigt deutliche Zusammenhänge zwischen vier Variablenpaaren (Abbildung 14). Wie erwartet sind das Alter und das Baujahr stark korreliert (r_S = -0,99). Kleinere Abweichungen von der Regressionsgeraden ergeben sich dadurch, dass eine Haltung zum Inspektionszeitpunkt älter als eine früher gebaute aber auch früh inspizierte Haltung sein kann (die Inspektionszeitspanne beträgt 16 Jahre). Ebenfalls stark korreliert sind die Breite und die Höhe (r_S = 0,999), da 96% der Haltungen ein Kreisprofil haben. Auch Tiefe und Überdeckung (r_S = 0,95) sowie die Anzahl an Bäumen und die Baumdichte (r_S = 0,95) sind stark korreliert.

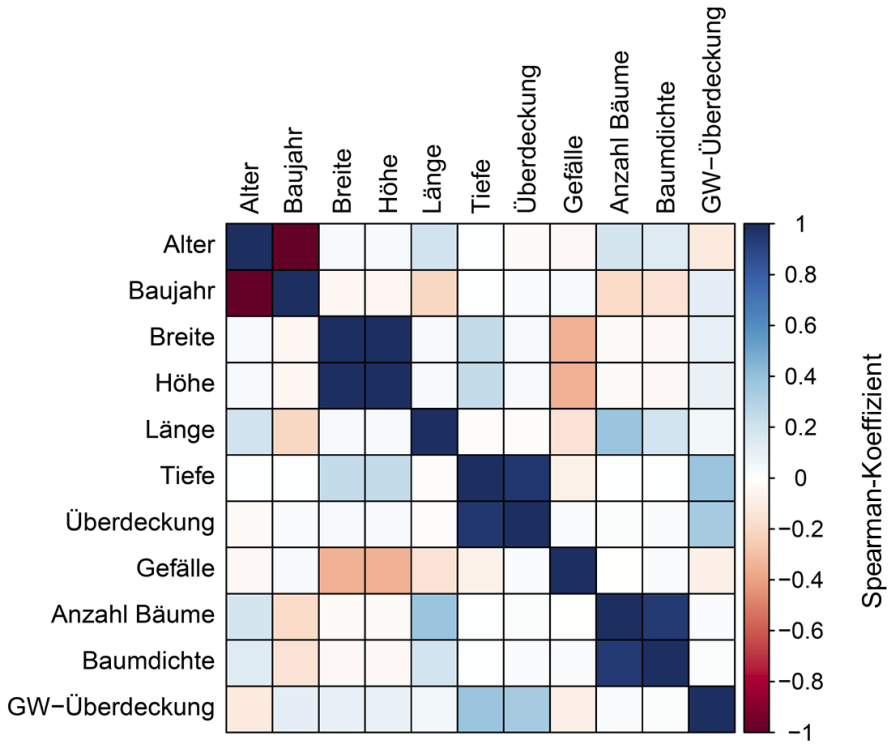


Abbildung 14: Abhängigkeiten aller numerischen Variablen, quantifiziert über den Spearman-Koeffizienten

Zur Reduzierung der Variablen wurde für weitere Analysen jeweils eine der stark korrelierten Variablen, d.h. Baujahr, Höhe, Tiefe und Baumdichte, vernachlässigt. Von den übrigen sieben Variablen gibt es lediglich leichte Korrelationen zwischen dem Gefälle und der Breite (r_S = -0,34) der Anzahl Bäume und der Länge (r_S = 0,38) und der Überdeckung und der Grundwasserüberdeckung (r_S = 0,35) (Abbildung 15).

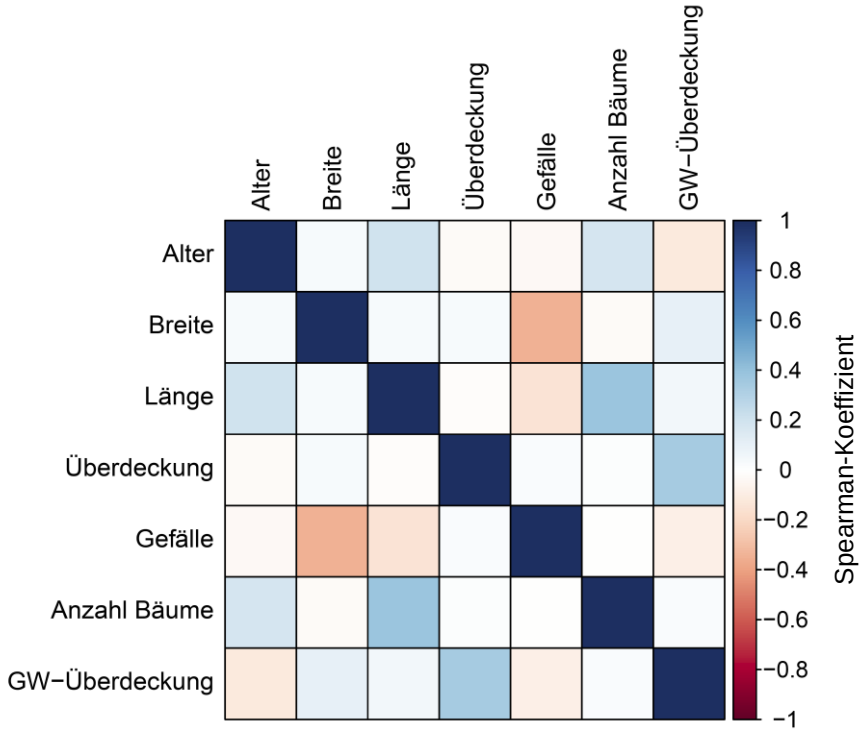


Abbildung 15: Abhängigkeiten der verbleibenden numerischen Variablen, quantifiziert über den Spearman-Koeffizienten

Die 18 verbleibenden Variablen – die sieben ursprünglich numerischen und die elf rein kategorische Variablen – wurden mit dem Chi-Quadrat-Test paarweise auf Unabhängigkeit untersucht. Mit zwei Ausnahmen (Rückstau und Länge sowie Rückstau und Schienenverkehr) liegen alle berechneten *p*-Werte unter dem Signifikanzniveau von 0,05. Das bedeutet, dass fast alle Variablen statistisch signifikant zusammenhängen - ein typisches Phänomen für Daten mit großen Stichprobenumfängen (n > 10.000, Lin et al. 2013).

Die Stärke der Abhängigkeiten wurde über *Cramér's V* quantifiziert. Die Variablenpaare Bäume (j/n) und Anzahl Bäume ($V = 1{,}0$), Grundwasserüberdeckung (j/n) und Grundwasserüberdeckung in Metern ($V = 1{,}0$) sowie Bezirk und Stadtteil ($V = 0{,}97$) sind vollständig korreliert (Abbildung 16). Im Zuge der Analyse des Effektes auf die Zustandsverteilung (Kap. 3.2) wurde je Paar die Variable mit dem geringsten Effekt ausgeschlossen (siehe Kap. 3.2.2).

Stärkere Abhängigkeiten gibt es weiterhin für die Variablenpaare Abwassertyp und Bezirk sowie Abwassertyp und Material (beide $V = 0{,}54$). Die *Cramér's-V*-Werte für Profil und Material ($V = 0{,}40$), Breite und Abwassertyp ($V = 0{,}40$), Alter und Bezirk ($V = 0{,}36$), Abwassertyp und Alter ($V = 0{,}35$), Abwassertyp und Bodentyp ($V = 0{,}32$), Profil und Alter ($V = 0{,}31$) sowie Bodentyp und Bezirk ($V = 0{,}30$) deuten auf mittlere Korrelationen zwischen den Variablen hin. Alle anderen Variablen sind nur gering oder nicht korreliert (z.B. Straßenklasse oder Schienenverkehr). Abbildung 16 zeigt die *Cramér's-V*-Werte in Form einer Korrelationsmatrix. Eine Tabelle mit allen *Cramér's-V*-Werten befindet sich in Anhang A (Tabelle 24).

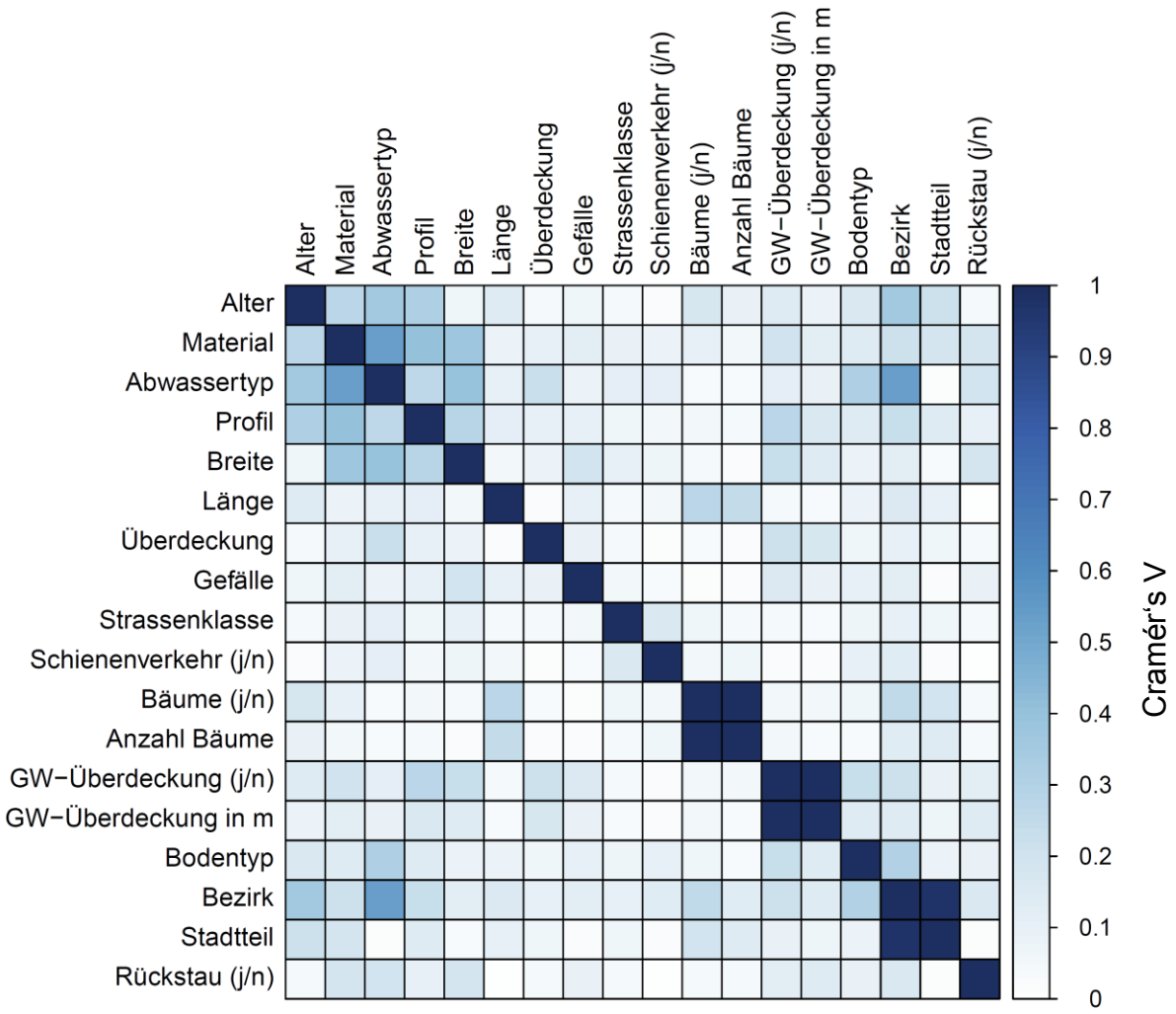


Abbildung 16: Abhängigkeiten der kategorischen Variablen, quantifiziert über *Cramér's V*

Abbildung 17 zeigt die Datenverteilung für die Variablenpaare mit starker (*Cramér's V* > 0,5; Grafiken a und b) und mittlerer Abhängigkeit (*Cramér's V* zwischen 0,3 und 0,5; Grafiken c bis i). Die wichtigsten Zusammenhänge zwischen den Variablen sind im Folgenden aufgelistet:

- *Abwassertyp und Bezirk:* Während Mischkanäle vorwiegend in den Innenstadtbezirken vorkommen, sind Schmutz- und Regenkanäle vor allem in den äußeren Stadtbezirken zu finden (Abbildung 17a);
- *Abwassertyp und Material:* Mischkanäle sind vorwiegend aus Mauerwerk gefertigt, Schmutzkanäle vor allem aus Steinzeug und Regenkanäle meist aus Beton (Abbildung 17b);
- *Profil und Material:* Die meisten Kanäle werden im Kreisprofil gefertigt, lediglich Mauerwerkkanäle sind häufig im Ei-Profil gebaut (Abbildung 17c);
- *Breite und Abwassertyp:* Schmutzkanäle haben tendenziell kleinere Durchmesser als Misch- und Regenkanäle (> 60% sind kleiner als DN 250, Abbildung 17d);
- *Alter und Bezirk:* Entsprechend der historischen Entwicklung Berlins sind die ältesten Kanäle in den Innenstadtbezirken zu finden (Friedrichshain-Kreuzberg, Mitte, Abbildung 17e);

- *Alter und Abwassertyp:* Mischkanäle sind tendenziell älter als Schmutz- und Regenkanäle (Abbildung 17f, siehe auch Zusammenhänge zwischen Bezirk und Abwassertyp bzw. Alter);
- *Abwassertyp und Bodentyp:* In der Umgebung von Mischkanälen sind die Bodentypen Aufschüttungen, Sand sowie Sand und Lehm in etwa zu gleichen Anteilen zu finden, während Aufschüttungen in Gebieten mit Schmutz- und Trennkanalisation kaum vorkommen (Abbildung 17g);
- *Alter und Profil:* Ei-Profile sind tendenziell am ältesten gefolgt von Maul- und Kreisprofilen. Kreis in Vortriebsbauweise (i.V.) kommt erst seit etwa 25 Jahren zum Einsatz (Abbildung 17h);
- *Bodentyp und Bezirk:* Aufschüttungen kommen vor allem in den Innenstadtbezirken (Friedrichshain-Kreuzberg, Mitte) vor, während in den äußeren Stadtbezirken vor allem Sand (z.B. Reinickendorf) oder Sand und Lehm (z.B. Marzahn-Hellersdorf) auftritt (Abbildung 17i, siehe auch Zusammenhänge zwischen Abwassertyp und Bodentyp).

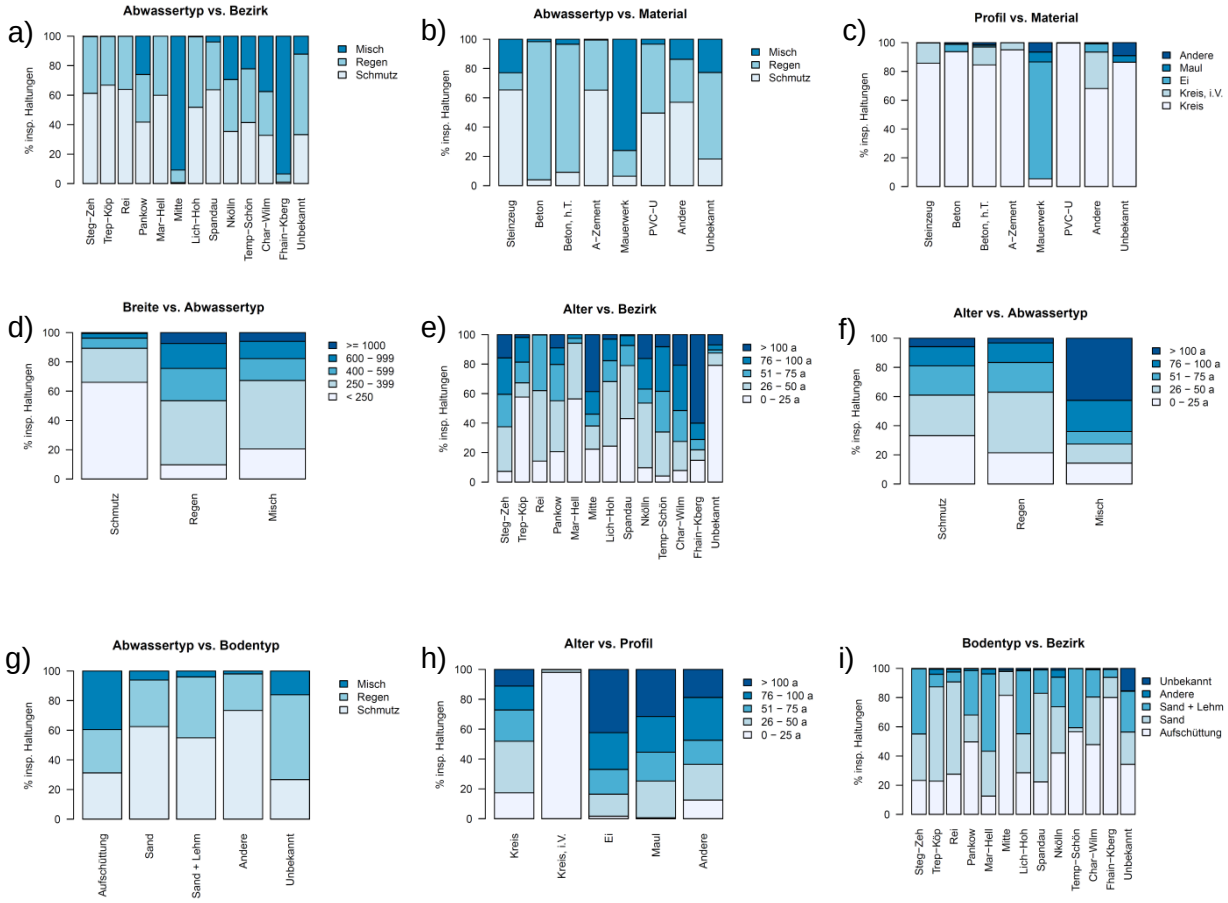


Abbildung 17: Datenverteilung für die neun Variablenpaare mit starker oder mittlerer Abhängigkeit (*Cramér's V* ≥ 0,3)

3.2 Einfluss der Eingangsvariablen auf den Zustand

3.2.1 Methodisches Vorgehen

Analog zu Kap. 3.1.1.2 wurde der Einfluss der erklärenden Variablen auf die Zustandsverteilung untersucht. Das heißt, i) für jede der 18 in Kap. 3.1 identifizierten Variablen wurde ein Chi-Quadrat-Unabhängigkeitstest mit der Zielvariable „baulicher Zustand“ durchgeführt, ii) es wurden die *Cramér's-V*-Werte berechnet und iii) für ein Ranking der Variablen bezüglich des Einflusses auf die Zustandsverteilung verwendet. Mehrfach inspizierte Kanäle wurden mehrfach berücksichtigt. Ziel der Untersuchung ist es einflussreiche Variablen für die Modellierung zu identifizieren und unnötige Variablen auszuschließen. Die Anzahl an Variablen für die Modellierung sollte so klein wie möglich und so groß wie nötig sein.

3.2.2 Ergebnisse und Diskussion

Die Variable mit dem größten Einfluss auf die Zustandsverteilung ist das Alter (*Cramér's V* = 0,27), gefolgt vom Profil (*V* = 0,19), der Länge (*V* = 0,16), der Grundwasserüberdeckung (*V* = 0,15) und des Materials (*V* = 0,15). Die Variablen Gefälle, Straßenklasse und Schienenverkehr haben nahezu keinen Einfluss auf die Zustandsverteilung (*V* < 0,05). Insgesamt liegen alle berechneten Werte unter 0,3, das heißt der Zusammenhang zwischen den erklärenden Variablen und der Zustandsverteilung ist verhältnismäßig gering. Abbildung 18 zeigt das Ranking der 18 Variablen bezüglich des Einflusses auf die Zustandsverteilung. Alle dargestellten Werte sind in Tabelle 24 im Anhang A zusammengefasst.

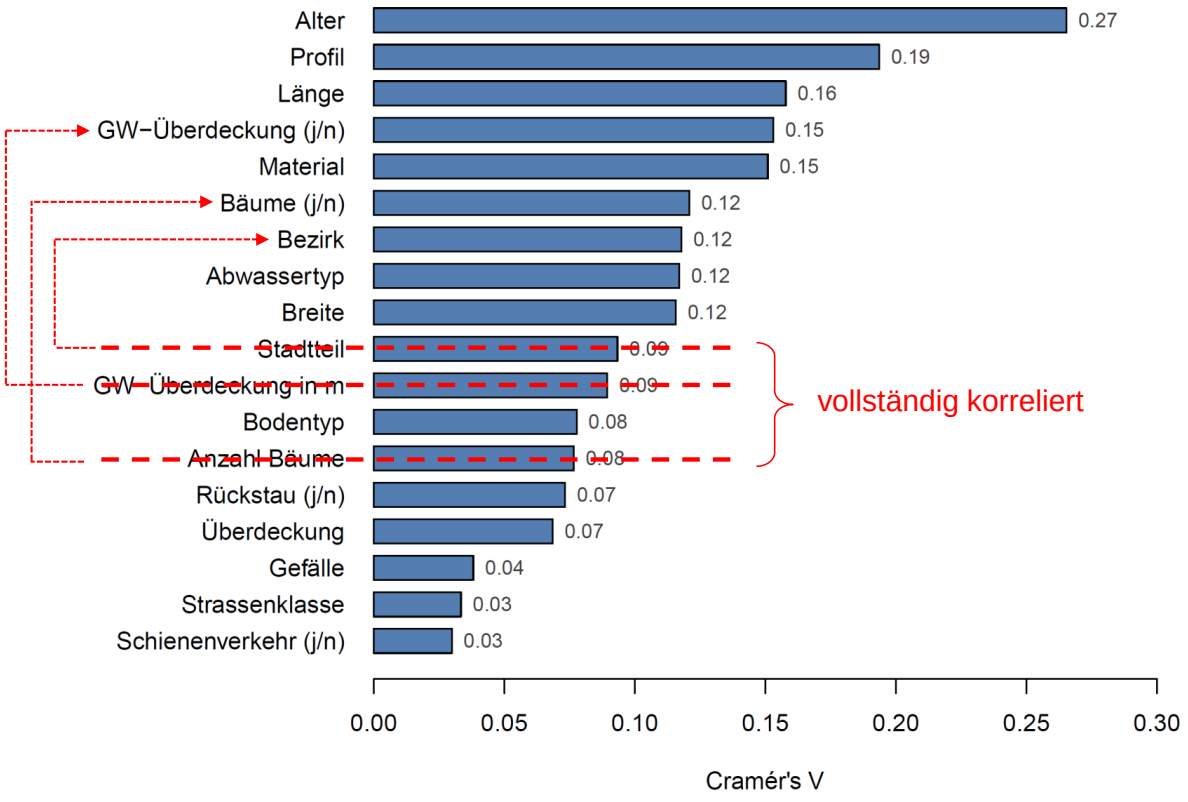


Abbildung 18: Ranking der Eingangsvariablen nach ihrem Effekt auf die Zustandsverteilung, quantifiziert über *Cramér's V*

Von den 18 Variablen gibt es drei Variablen, die vollständig mit anderen korreliert sind (Kap. 3.1.2) und keine zusätzliche Genauigkeit in der Vorhersage der Zustandsklasse erwarten lassen. Dazu zählen der Stadtteil (verknüpft mit dem Bezirk), die Grundwasserüberdeckung in Metern (verknüpft mit der Grundwasserüberdeckung (j/n)) und die Anzahl an Bäumen (verknüpft mit Bäume (j/n)). Da die drei genannten Variablen einen geringeren Zusammenhang mit der Zustandsverteilung haben als ihr jeweiliges Pendant, wurden sie für weitere Analysen verworfen. Darüber hinaus gibt es drei Variablen mit einem vernachlässigbar kleinen Effekt auf die Zustandsverteilung (Gefälle, Straßenklasse und Schienenverkehr; *Cramér's V* < 0,05). Diese Variablen wurden für weitere Analysen ebenfalls vernachlässigt.

Für die zwölf verbleibenden Variablen ist die Zustandsverteilung für die einzelnen Variablenausprägungen in Abbildung 19 (Rang 1 bis 6) und Abbildung 20 (Rang 7 bis 12) dargestellt und wird im Folgenden diskutiert.

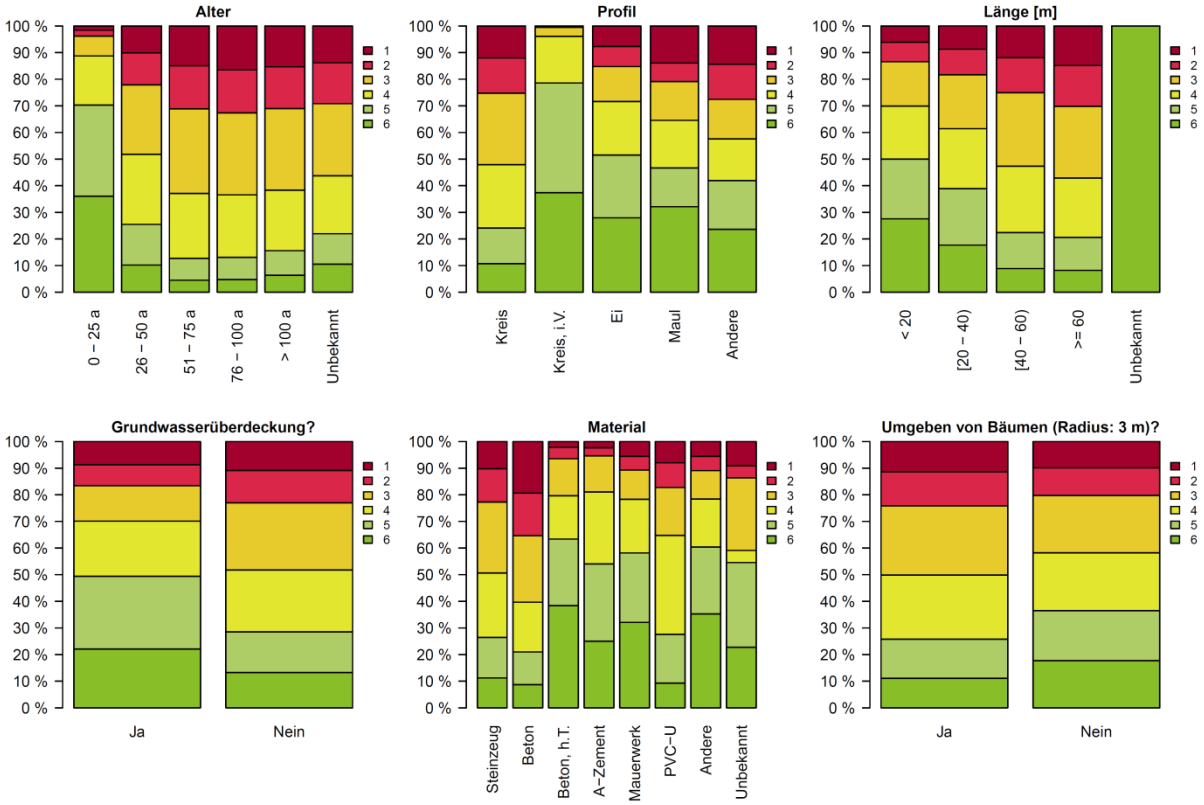


Abbildung 19: Zustandsverteilung für die Variablen Alter, Profil, Länge, Grundwasserüberdeckung, Material und Bäume (Rang 1 bis 6)

Alter: Der Vergleich der Zustandsverteilung der einzelnen Altersklassen zeigt, dass der Anteil an Haltungen im schlechten Zustand innerhalb der ersten 51 bis 75 Jahre tendenziell zunimmt. Danach bleibt die Zustandsverteilung weitestgehend unverändert. Allerdings heißt das nicht, dass sich der Zustand eines Kanals nach 51 bis 75 Jahren nicht mehr verschlechtert. Vielmehr ist es so, dass Kanäle, die nach ihrer Abschreibungsdauer von 50 Jahren nicht mehr im guten Zustand sind, in der Regel zeitnah saniert oder erneuert werden. Daraus ergibt sich eine Verzerrung des Zustandsbildes für alte Kanäle. In der Klasse 0-25 a sind 89% aller Haltungen in Zustandsklasse 4, 5 oder 6, d.h. es gibt kaum Schäden, die kurz- oder mittelfristig behoben werden müssen. Ab der Altersklasse 51-75 a stabilisiert sich der Anteil von Kanälen in Zustandsklasse 4, 5 oder 6 bei etwa 37%. Das heißt auf der anderen Seite, dass knapp zwei Drittel der Haltungen, die älter als 50 Jahre sind, kurz-

(Zustandsklasse 1 und 2) oder mittelfristig (Zustandsklasse 3) saniert oder erneuert werden müssen (Abbildung 19, oben links).

Profil: Eine besondere Häufung von Kanälen im guten Zustand zeigt sich weiterhin bei Kanälen, die im Kreisprofil in Vortriebsbauweise gebaut wurden (96% dieser Haltungen haben Zustandsklasse 4, 5 oder 6, Abbildung 19, oben Mitte). Allerdings kam diese Bauweise fast ausschließlich in den letzten 25 Jahren auf, d.h. die Variable ist vom Alter abhängig (siehe Abbildung 17 h). Unter den anderen Profilarten haben Ei- und Maulprofile trotz ihres vergleichsweise hohen Alters einen allgemein besseren Zustand als beispielsweise Kanäle im Kreisprofil (nur 48% in Zustandsklasse 4, 5 oder 6).

Länge: Mit zunehmender Länge nimmt der Anteil an Kanälen im schlechten Zustand tendenziell zu (Abbildung 19, rechts oben), was sich dadurch begründen lässt, dass die Auftretenswahrscheinlichkeit schwerer Schäden mit zunehmender Haltungslänge zunimmt. Bei der Berliner Bewertungsmethode wird nicht die Schadensdichte entlang der Haltung sondern die Gesamtanzahl und Schwere der Schäden berücksichtigt.

Grundwasserüberdeckung: Unter den Kanälen, die vom Grundwasser überdeckt sind, gibt es einen größeren Anteil an Kanälen im guten Zustand als unter denen, die nicht vom Grundwasser überdeckt sind (70% ggü. 52%). Allerdings zeigt die Grundwasserüberdeckung eine (wenn auch geringe) Korrelation zur (Boden-)Überdeckung (*Cramér's V* = 0,22, Abbildung 16). Vom Grundwasser überdeckte Kanäle liegen im Allgemeinen tiefer unter der Erde als nicht vom Grundwasser überdeckte Kanäle und sind damit weniger anfällig gegenüber mechanischen Belastungen von oben, z.B. durch Verkehr (O'Reilly et al. 1985, Fenner & Sweeting 1999, Davies et al. 2001). Ein weiterer Grund für den schlechteren Zustand bei Kanälen mit geringer Überdeckung kann der variierende Bodenwasserhaushalt sein (Jones 1984).

Material: Es gibt deutliche Unterschiede zwischen den Zustandsverteilungen der einzelnen Materialgruppen. Den höchsten Anteil an Kanälen im guten Zustand gibt es bei den Betonkanälen mit hoher Tragfähigkeit (wandverstärkter Beton, bzw. Stahlbeton). Mauerwerkkanäle (vorwiegend in Mischkanalisation) sind trotz ihres tendenziell hohen Alters (Abbildung 69a, Anhang B) in einem verhältnismäßig guten Zustand. Den größten Anteil an Kanälen im schlechten Zustand haben Betonkanäle (vorwiegend in Regenkanalisation), gefolgt von Steinzeug (vorwiegend in Schmutzkanalisation) und PVC (Schmutz- und Regenkanalisation).

Bäume: In der Umgebung von Bäumen ist der Zustand der Kanäle tendenziell schlechter als ohne Bäume (Anteil an Kanälen im schlechten Zustand (Zustandsklasse 1 oder 2) beträgt 24% ggü. 20%).

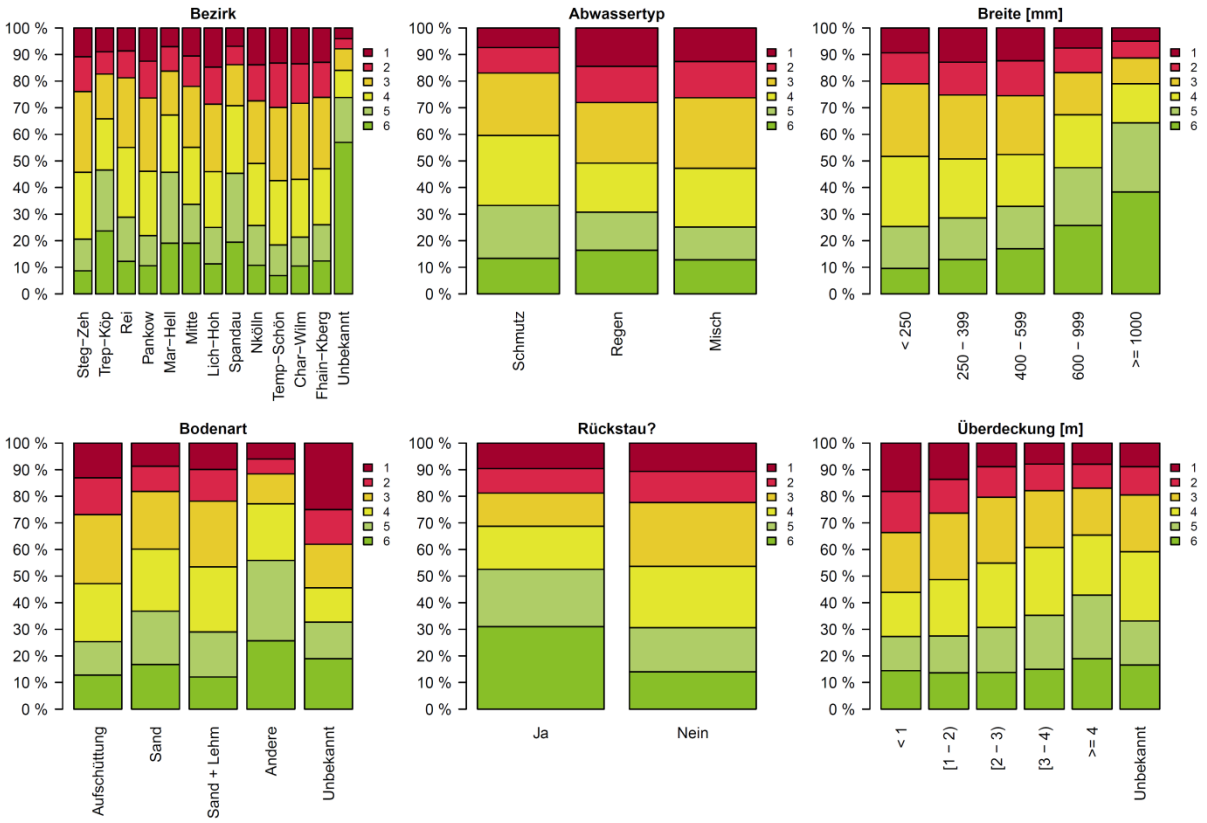


Abbildung 20: Zustandsverteilung für die Variablen Bezirk, Abwassertyp, Breite, Bodenart, Rückstau und Überdeckung (Rang 7 bis 12)

Bezirk: Den besten Zustand haben die Kanäle in den Bezirken Spandau, Marzahn-Hellersdorf und Treptow-Köpenick mit 65% bis 71% im guten Zustand (Zustandsklasse 4, 5 oder 6). Diese Bezirke haben einen hohen Anteil an jungen Kanälen (Abbildung 17e). Am niedrigsten ist der Anteil an Kanälen im guten Zustand in den Bezirken Charlottenburg-Wilmersdorf und Tempelhof-Schöneberg (beide 43%).

Abwassertyp: Die Schmutzwasserkanalisation hat den höchsten Anteil an Kanälen im guten Zustand (60% in Zustandsklasse 4, 5 oder 6) und den geringsten Anteil an Kanälen im schlechten Zustand (17% in Zustandsklasse 1 oder 2). Für die Regen- und Mischwasser-kanalisation liegen die Anteile bei 49% bzw. 47% für den guten und bei 28% bzw. 26% für den schlechten Zustand. Der Anteil an Kanälen im mittleren Zustand ist bei Regenkanälen mit 18% am kleinsten, d.h. die Zustandsverteilung hat hier ihre Schwerpunkte in den Extremen.

Breite: Mit zunehmender Breite nimmt der Anteil an Kanälen im guten Zustand zu (von 52% für DN < 250 auf 79% für DN ≥ 1000), der Anteil an Kanälen im schlechten Zustand nimmt tendenziell ab (25% für DN 250 bis 399, 11% für DN ≥ 1000). Das heißt verkürzt: je kleiner der Kanal, desto schlechter der Zustand.

Bodenart: Kanäle, die in Sand gelagert sind, haben tendenziell einen besseren Zustand als Kanäle, die in Aufschüttungen oder einem Sand-Lehm-Gemisch liegen. Beim Bodentyp Sand befinden sich 60% im guten und 18% im schlechten Zustand. Für Aufschüttungen und Sand-Lehm-Gemische befinden sich 47% bzw. 53% der inspizierten Kanäle im guten Zustand und 27% bzw. 22% im schlechten Zustand.

Rückstau: Rückstaubeeinflusste Kanäle sind tendenziell in einem besseren Zustand als nicht rückstaubeeinflusste Kanäle. Der Anteil an Kanälen im guten Zustand beträgt 69% für rückstaubeeinflusste Kanäle und 54% für nicht rückstaubeeinflusste Kanäle.

Überdeckung: Je höher die Kanäle von Boden bedeckt sind, d.h. je tiefer die Kanäle liegen, desto besser ist tendenziell der Zustand. Der Anteil an Kanälen im guten Zustand (Zustandsklasse 4 bis 6) beträgt für Überdeckungen < 1 m 44% und für Überdeckungen ≥ 4 m 65%.

Die Zustandsverteilung für die drei Variablen mit marginalem Effekt auf die Zustandsverteilung sowie für die drei redundanten Variablen ist in Abbildung 21 zu sehen und wird im Folgenden kurz diskutiert.

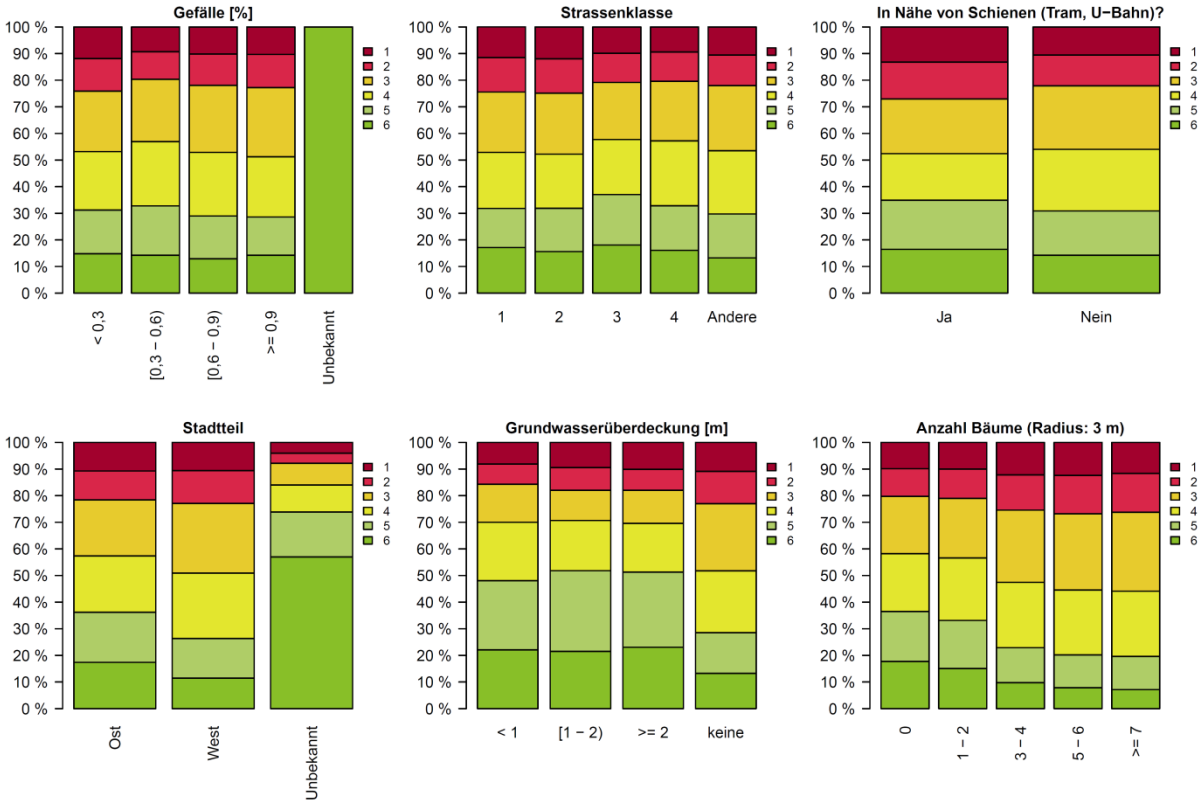


Abbildung 21: Zustandsverteilung der Eingangsvariablen, die aufgrund von geringer Relevanz (oben) oder Abhängigkeiten (unten) nicht weiter untersucht wurden.

Gefälle: Auch wenn in der Gruppe der Kanäle mit geringem Gefälle (< 0,3%) der Anteil an Kanälen im schlechten Zustand (Zustandsklasse 1 und 2) geringfügig höher ist als in den anderen Gruppen (Abbildung 21, oben links), so sind die Unterschiede in der Zustandsverteilung insgesamt marginal.

Straßenklasse: Bis auf eine geringfügige Häufung an Kanälen im schlechten Zustand in der Nähe von sehr stark befahrenen Straßen (Straßenklasse 1 und 2), sind Straßenklasse und Zustandsverteilung kaum korreliert (Abbildung 21, oben Mitte). Das heißt der Zustand der Kanäle ist weitgehend unabhängig vom Befahrungsgrad der Straße, was die Ergebnisse von O'Reilly et al. (1989) bestätigt. Dennoch ist es durchaus möglich, dass der Straßenverkehr an sich - unabhängig von der Verkehrsdichte - auch in Berlin eine wichtige Rolle für den

Alterungsprozess eines Kanals spielt. Dies lässt sich anhand der vorliegenden Daten allerdings nicht abschließend beurteilen.

Schienenverkehr: Der Schienenverkehr hat insgesamt einen geringen Effekt auf die Zustandsverteilung der inspizierten Kanäle (Abbildung 21, oben rechts). Allerdings sind nur ca. 3% aller inspizierten Kanäle überhaupt von Straßenbahn- oder U-Bahn-Gleisen umgeben. Für diese Teilmenge ist der Anteil an Kanälen im schlechten Zustand (Zustandsklasse 1 oder 2) geringfügig höher als in der Gruppe der nicht schienenbeeinflussten Kanäle, der Anteil an Kanälen im guten Zustand (Zustandsklasse 4, 5 oder 6) ist etwa gleich hoch.

Stadtteil: Die Analyse hat gezeigt, dass der Anteil an Kanälen im guten Zustand (Zustandsklasse 4, 5 oder 6) im ehemaligen Ostteil der Stadt etwas größer ist als im ehemaligen Westteil der Stadt (57% ggü. 51%, Abbildung 21, unten links). Dies ist in erster Linie der hohen Sanierungsaktivität im Ostteil Berlins nach der Wiedervereinigung der Stadt geschuldet. Der Anteil an Kanälen im schlechten Zustand (Zustandsklasse 1 oder 2) ist in beiden Teilen etwa gleich hoch (Ost: 22%, West: 23%).

Grundwasserüberdeckung in m: Die Zustandsverteilung unterscheidet sich zwar zwischen den vom Grundwasser überdeckten und den nicht überdeckten Kanälen (siehe Abbildung 19). Wie stark die Kanäle vom Grundwasser überdeckt sind, spielt dagegen keine maßgebende Rolle (Abbildung 21, unten Mitte).

Anzahl Bäume: Nicht nur zwischen den von Bäumen umgebenen und den nicht von Bäumen umgebenen Kanälen gibt es Unterschiede in der Zustandsverteilung. Auch die Anzahl der Bäume im 3-m-Radius um den Kanal hat einen Effekt auf den Zustand. Mit zunehmender Anzahl an Bäumen steigt der Anteil an Kanälen im schlechten Zustand, der Anteil an Kanälen im guten Zustand nimmt ab (Abbildung 21, unten rechts).

3.3 Analyse der Einzelschäden

Die Zustandsklasse einer Haltung hängt von Anzahl und Schwere der einzelnen Schäden ab. Zum besseren Verständnis der Alterungsprozesse wurden - über die Analyse der Zustandsverteilung hinaus - die Häufigkeit sowie die Einflussfaktoren der Einzelschäden untersucht. Tabelle 4 zeigt die nach dem Berliner Schadenskatalog 11 (BWB 2001) unterschiedenen zwölf Hauptschadenstypen.

Tabelle 4: Berliner Schadenstypen nach Schadenskatalog 11 (BWB 2001)

Name	Kodierung	Beschreibung
Undichtigkeiten	U	Sichtbare Undichtigkeiten, die auf keinen anderen erkennbaren Schaden zurückgeführt werden können
Abflusshindernisse	H	Kanalfremde Bestandteile, die auch nach Kanalreinigung eine Einschränkung des Querschnittes darstellen
Verwurzelungen	W	Einwuchs von Wurzeln in den Kanal
Schadhafte Rohrverbindungen	M	Lageabweichungen oder sichtbare Dichtungsschäden im Verbindungsbereich zweier Rohre
Lageabweichungen	L	Horizontale oder vertikale Ausbiegungen von mehr als einem Kanalrohr von der (idealen) Kanalachse
Mechanischer Verschleiß	V	Materialabtrag der Kanalwandung durch mechanischen Einfluss
Korrosion	C	Materialabtrag der Kanalwandung durch chemischen Einfluss
Deformation	D	Sichtbare Deformation des Rohrquerschnittes
Risse, Scherbenbildung	R	Rissbildungen in der Kanalwandung
Rohrbruch, Einsturz	B	Rohrbruch: Fehlen von Wandungsteilen innerhalb des Kanals; Einsturz: sichtbare Gefährdung der statischen Standsicherheit des Kanals
Fehlanschluss	F	Anschluss eines Regen- an einen Schmutzkanal oder eines Schmutz- an einen Regenkanal
Fehlende Rohrwand im Stutzenbereich	E	Rohrwand im Ringraum um den Stutzen fehlt; Wasser tritt ein/aus bzw. Boden ist sichtbar

3.3.1 Methodisches Vorgehen

Zur Untersuchung der Schadenshäufigkeit wurde zum einen die Gesamthäufigkeit der Einzelschäden über alle Inspektionsdatensätze ermittelt (unterschieden nach Schadenstyp).

Zum anderen wurde beurteilt, wie viele Haltungen von den einzelnen Schadenstypen betroffen sind.

Anschließend wurde analog zu Kap. 3.1.1.2 der Einfluss der Eingangsvariablen auf das Auftreten der einzelnen Schadenstypen mittels Chi-Quadrat-Test und Berechnung von *Cramér's V* untersucht. In der Analyse wurden die zwölf Variablen mit der größten Bedeutung für die Zustandsverteilung (siehe Kap.3.2) den zwölf Berliner Haupt-Schadenstypen (Tabelle 4) gegenübergestellt. Die Analyse wurde haltungsbezogen durchgeführt, wobei für jeden Schadenstyp die Merkmale „ja" (Schaden kommt in dieser Haltung mindestens einmal vor) und „nein" (Schaden kommt in dieser Haltung nicht vor) unterschieden wurden. Mehrfachinspektionen wurden mehrfach berücksichtigt. Zur Veranschaulichung des Zusammenhangs zwischen den Variablen und den Schadenstypen wurden die *Cramér's-V*-Werte in einer Korrelationsmatrix dargestellt. Zudem wurden die Verteilungen der Variablen für alle Haltungen, die einen bestimmten Schadenstyp vorweisen, miteinander verglichen und der Verteilung im ganzen Netz gegenübergestellt.

3.3.2 Ergebnisse und Diskussion

Bei den 115.258 ausgewerteten Inspektionsdatensätzen wurde insgesamt 1,4 Mio. Einzelschäden kodiert. Fast 90% aller Schäden entfallen auf die Schadenstypen Verwurzelungen, Schadhafte Rohrverbindungen, Risse und Scherbenbildung, Abfluss-hindernisse und Mechanischer Verschleiß. Fehlanschlüsse, Undichtigkeiten oder Deforma-tionen machen in Summe nur 1% aller Schäden aus und kommen damit nur selten vor. Abbildung 22 zeigt die Gesamthäufigkeit der zwölf Hauptschadenstypen.

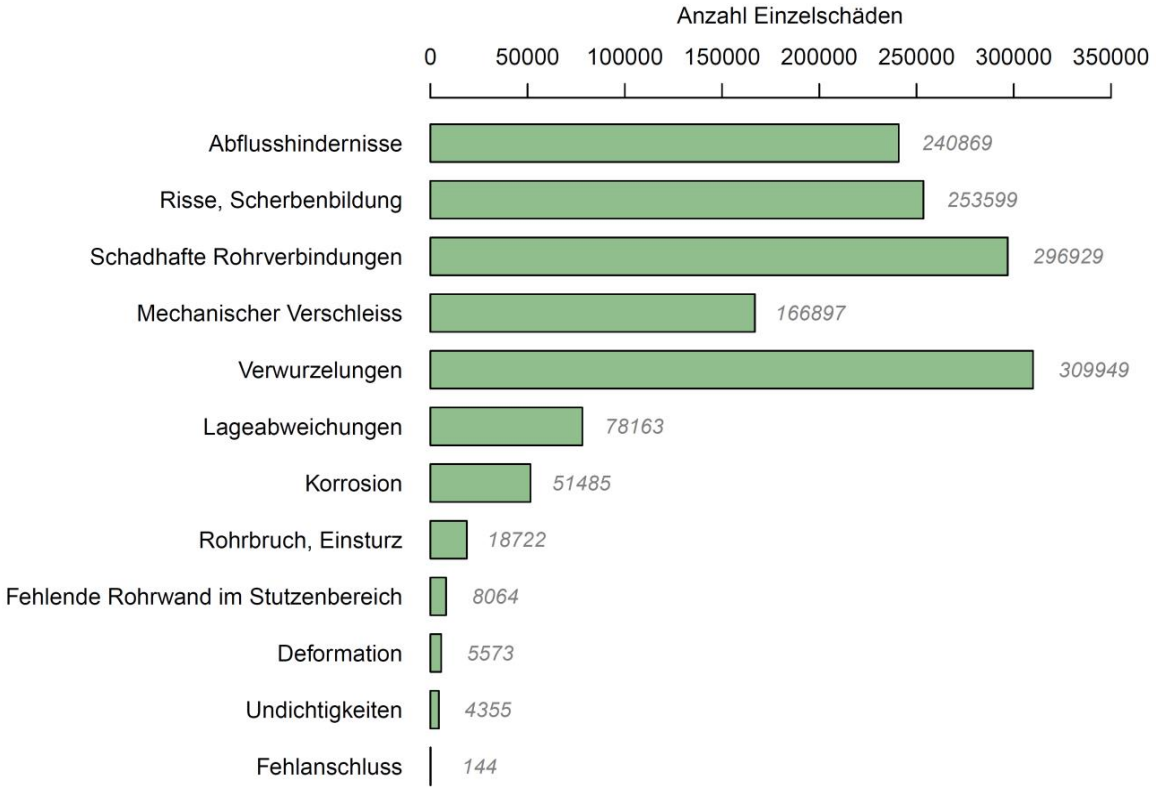


Abbildung 22: Schadenshäufigkeit, differenziert nach Schadenstyp, summiert über alle Inspektionsdatensätze.

Bezogen auf die Haltungen sind Abflusshindernisse der Schadenstyp, von dem die meisten Haltungen betroffen sind (56% aller Haltungen, Abbildung 23). Risse und Scherbenbildung,

Schadhafte Rohrverbindungen, mechanischer Verschleiß und Verwurzelungen sind mit jeweils > 40.000 betroffenen inspizierten Haltungen (mehr als ein Drittel) ebenfalls sehr häufig. Lageabweichungen sind mit etwa 35.000 betroffenen Haltungen (31%) ebenfalls recht häufig. Korrosionsschäden wurden bei 12% aller inspizierten Haltungen (n = 14.030) beobachtet. Deformationen und Undichtigkeiten kommen bei 3% bzw. 1% aller inspizierten Kanäle vor, Fehlanschlüsse sogar nur bei einem aus Tausend. Verwurzelungen treten überdurchschnittlich oft gehäuft in einer Haltung auf (im Mittel 8 Wurzelschäden je schadhafter Haltung, Maximum: 109). Insgesamt weisen 86% aller inspizierten Kanäle mindestens einen Schaden auf.

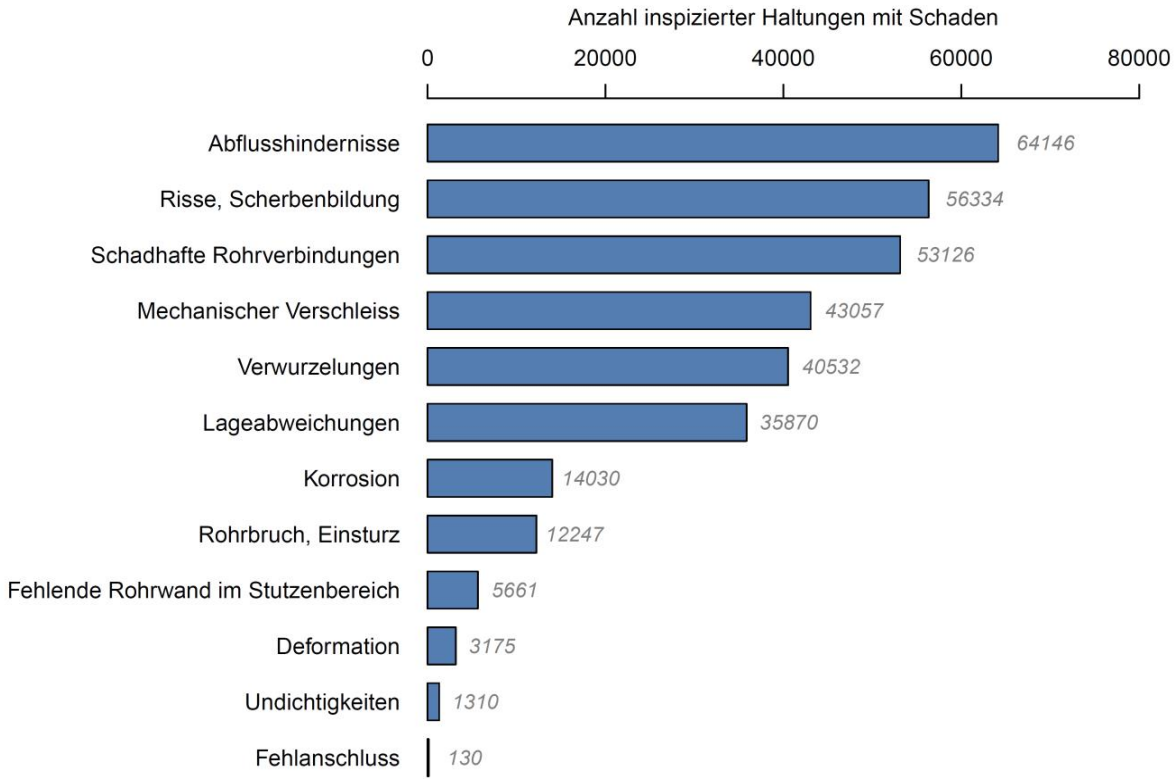


Abbildung 23: Anzahl der schadhaften Haltungen, differenziert nach Schadenstyp

Die wichtigsten Einflussfaktoren für das Auftreten von Schäden sind das Material, das Alter, das Profil, die Breite, der Bezirk, die Länge und der Abwassertyp (mittlere *Cramér's-V*-Werte > 0,1; Abbildung 24, Tabelle 25 in Anhang C). Ein besonders deutlicher Zusammenhang (*Cramér's V* >0,3) wurde für folgende Schadenstypen und Variablen festgestellt:

- Korrosion mit Material (*Cramér's V* = 0,59) und Abwassertyp (*Cramér's V* = 0,38),
- Deformation mit Material (*Cramér's V* = 0,57),
- Risse und Scherbenbildung mit Alter (*Cramér's V* = 0,45), Profil (*Cramér's V* = 0,32) und Material (*Cramér's V* = 0,32),
- Schadhafte Rohrverbindungen mit Material (*Cramér's V* = 0,36) und Breite (*Cramér's V* = 0,35),
- Verwurzelungen mit Alter (*Cramér's V* = 0,41),
- Mechanischer Verschleiß mit Material (*Cramér's V* = 0,32) und
- Abflusshindernisse mit Alter (*Cramér's V* = 0,32).

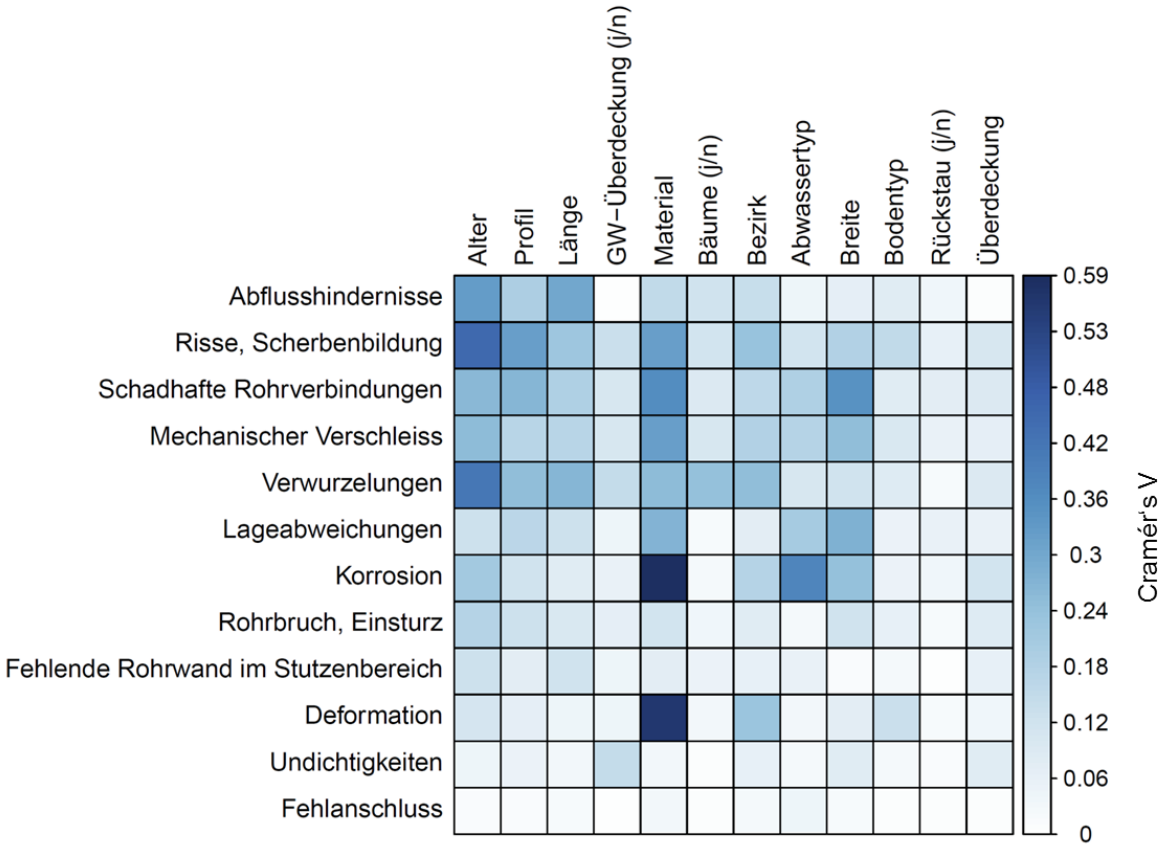


Abbildung 24: Einfluss der erklärenden Variablen auf das Auftreten der einzelnen Schadenstypen

Im Folgenden wird für alle Haltungen mit einem bestimmten Schadenstyp die Datenverteilung für die oben erwähnten Variablen (Material, Alter, Profil, Breite, Bezirk, Länge und Abwassertyp) gezeigt und diskutiert.

Material: Bezüglich des Materials zeigt sich, dass Korrosionsschäden fast ausschließlich bei Betonkanälen (Anteil 77%) und Deformationen größtenteils an PVC-Kanälen (55%) auftreten (Abbildung 25). Unter allen inspizierten Kanälen kommen diese Materialien jedoch nur relativ selten vor (Beton: 18%, PVC-U: 2,5%). Viele andere Schadenstypen, insbesondere mechanischer Verschleiß und schadhafte Rohrverbindungen, treten gehäuft bei Kanälen aus Steinzeug auf.

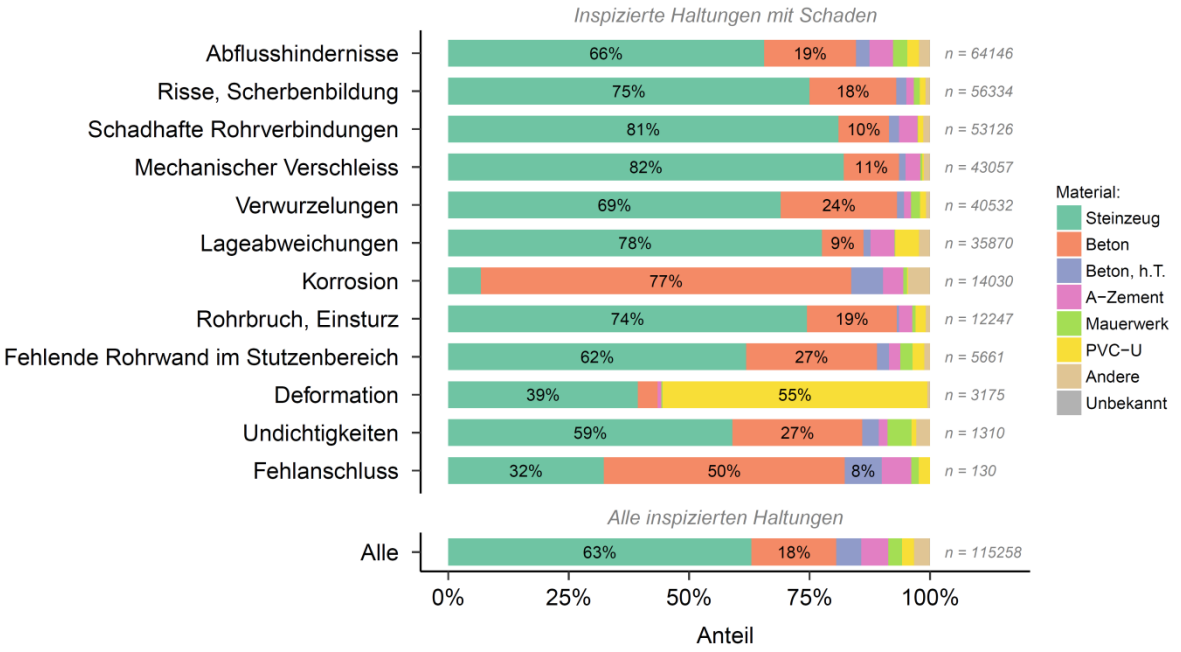


Abbildung 25: Materialverteilung der Haltungen mit einem bestimmten Schaden (oben) und aller inspizierten Haltungen (unten)

Alter: Verwurzelungen, Korrosionsschäden, Rohrbrüche sowie fehlende Rohrwände im Stutzenbereich treten erst ab einem Alter von 25 Jahren in relevantem Ausmaß auf (Abbildung 26). Der Anteil an schadhaften Kanälen ≤ 25 Jahre beträgt nicht mehr als 5% (24% bezogen auf alle inspizierten Kanäle). Kanäle mit Deformationen haben zum Zeitpunkt der Inspektion auffällig oft ein Alter zwischen 26 und 50 Jahren (52%). Grund dafür ist die besondere Häufung von PVC-Kanälen in dieser Altersgruppe - 73% der PVC-Kanäle sind zwischen 26 und 50 Jahre alt (Abbildung 69a, Anhang B) - und die Empfindlichkeit von PVC-Kanälen gegenüber Deformationen (Abbildung 25).

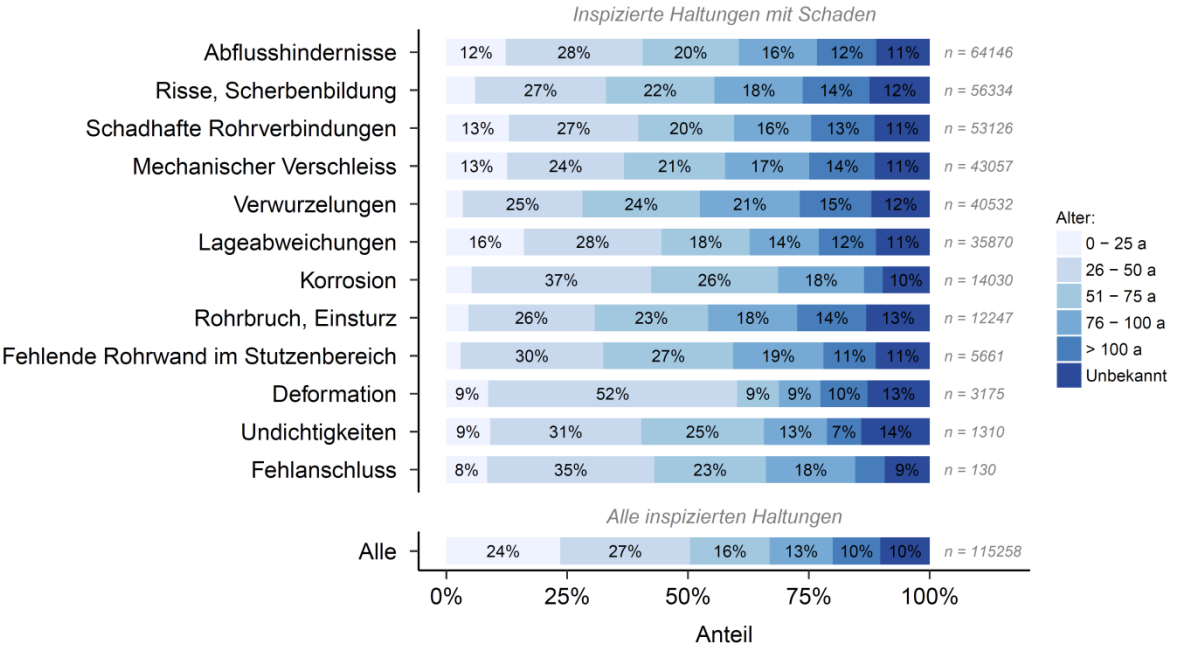


Abbildung 26: Altersverteilung der Haltungen mit einem bestimmten Schaden (oben) und aller inspizierten Haltungen (unten)

Profil: Schäden wie Risse und Scherbenbildung, Verwurzelungen, Rohrbrüche, Einstürze und Deformationen treten bei Kanälen in Vortriebsbauweise aufgrund der dickeren Wandstärke kaum auf (Abbildung 27). Lageabweichungen, Mechanischer Verschleiß und Abflusshindernisse kommen hingegen auch bei Kanälen in Vortriebsbauweise vor, da diese vorwiegend durch bestimmte Betriebs- oder Einbaubedingungen verursacht werden und nicht durch die dickere Rohrwandung verhindert werden können.

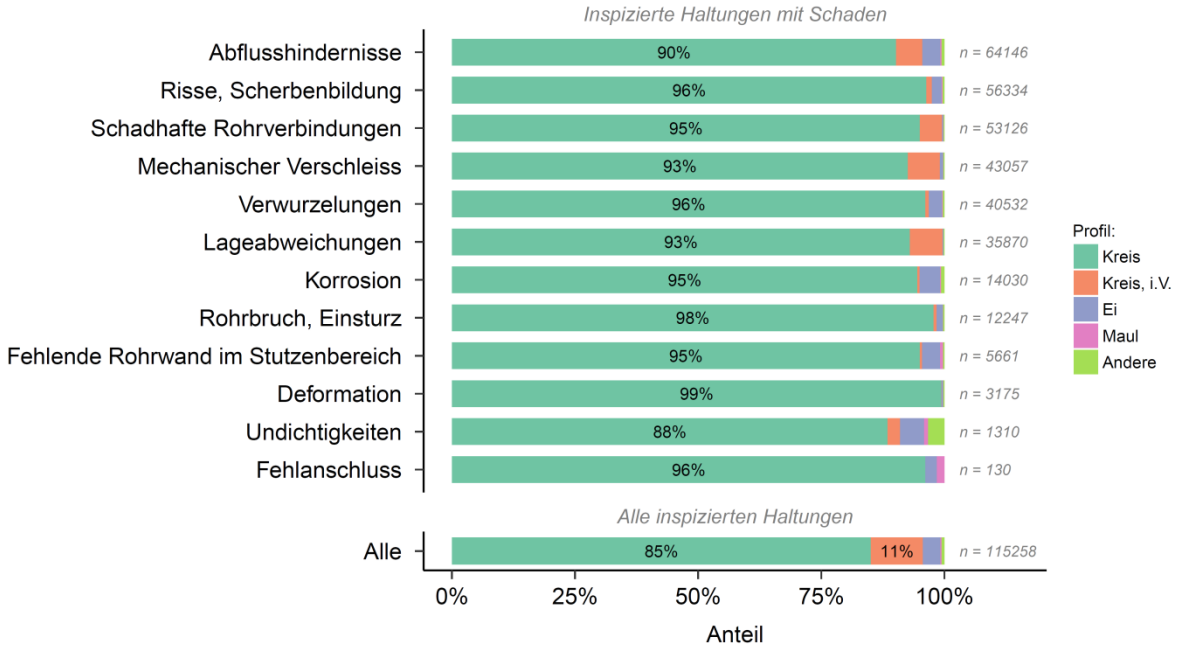


Abbildung 27: Profilverteilung der Haltungen mit einem bestimmten Schaden (oben) und aller inspizierten Haltungen (unten)

Breite: Deformationen, Lageabweichungen, schadhafte Rohrverbindungen sowie Rohrbrüche und Einstürze treten gehäuft bei schmalen Kanälen (DN < 250) auf (Abbildung 28). Korrosionsschäden, Fehlanschlüsse und Undichtigkeiten kommen hingegen vorwiegend bei breiten Kanälen vor. Bei der Interpretation des Effektes auf Deformationen sind Korrelationen zwischen Breite und Material zu berücksichtigen (etwa zwei Drittel aller PVC-Kanäle haben Durchmesser < 250 mm, Abbildung 69i, Anhang B).

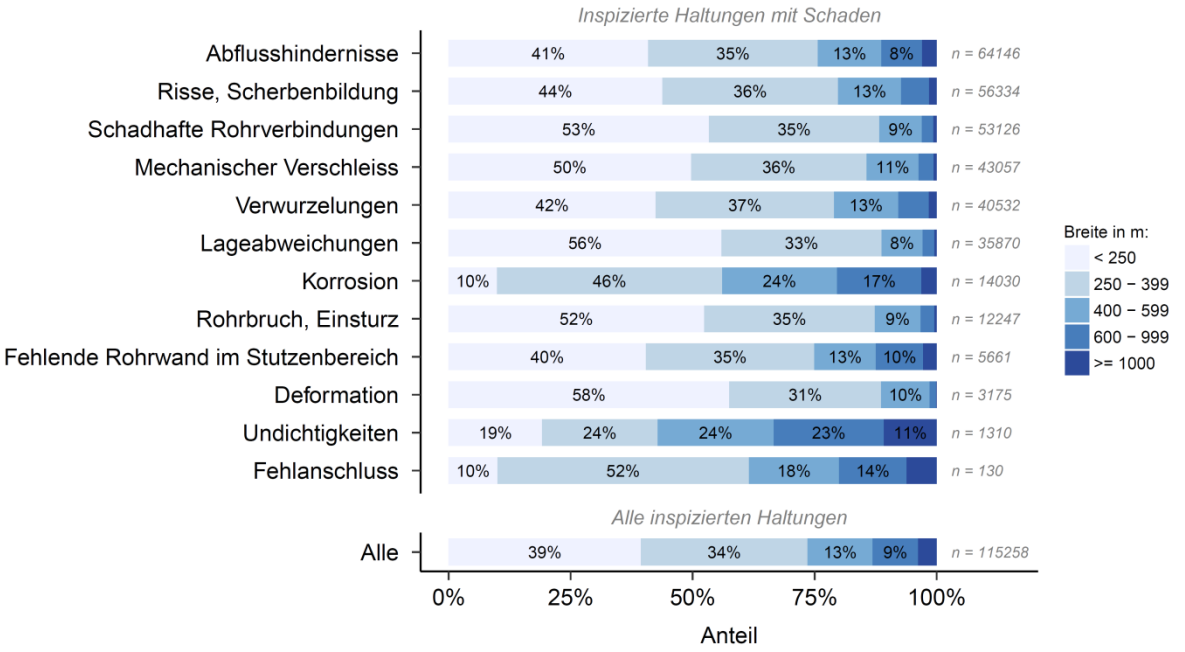


Abbildung 28: Breitenverteilung der Haltungen mit einem bestimmten Schaden (oben) und aller inspizierten Haltungen (unten)

Bezirk: Besonders auffällig ist die Häufung von Deformationsschäden in Marzahn-Hellersdorf (Abbildung 29), was sich mit einer Häufung von PVC-Kanälen in diesem Stadtbezirk begründen lässt. Etwa zwei Drittel aller inspizierten PVC-Kanäle liegen in diesem Bezirk (Abbildung 69g, Anhang B).

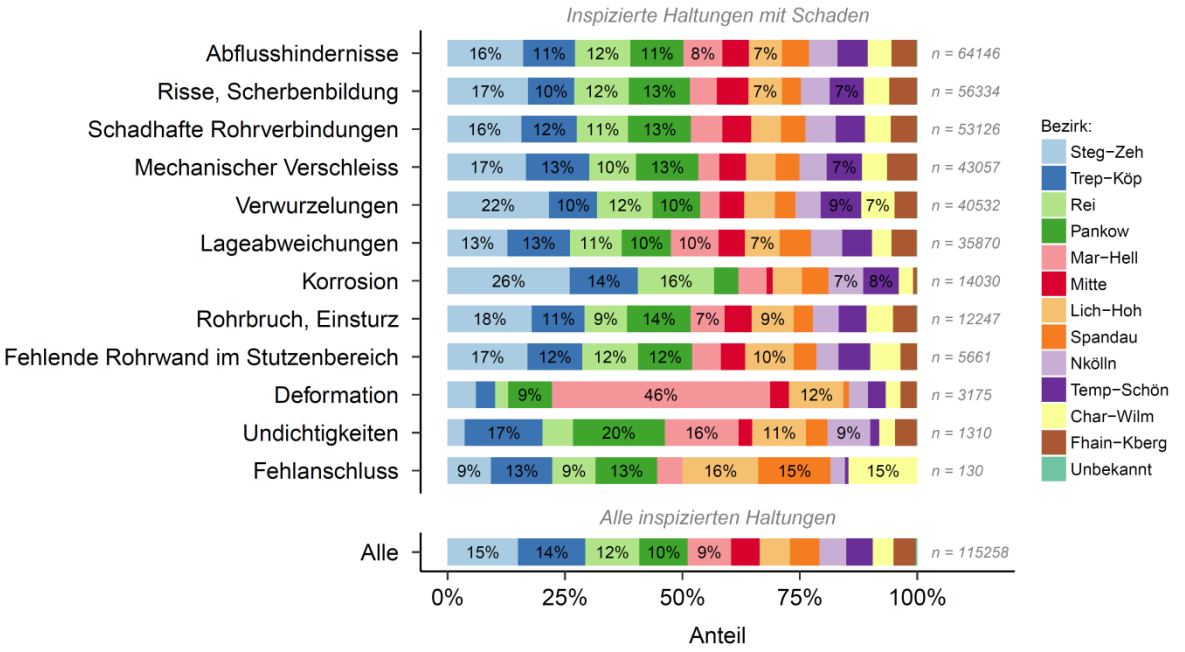


Abbildung 29: Bezirksverteilung der Haltungen mit einem bestimmten Schaden (oben) und aller inspizierten Haltungen (unten)

Länge: Alle Schadenstypen treten gehäuft bei längeren Kanälen (≥ 20 m) auf (Abbildung 30, Vergleich aller Zeilen oben mit der Zeile unten), da die Wahrscheinlichkeit einen Schaden aufzuweisen mit zunehmender Haltungslänge zunimmt. Die Unterschiede zwischen den einzelnen Schadenstypen sind weniger deutlich.

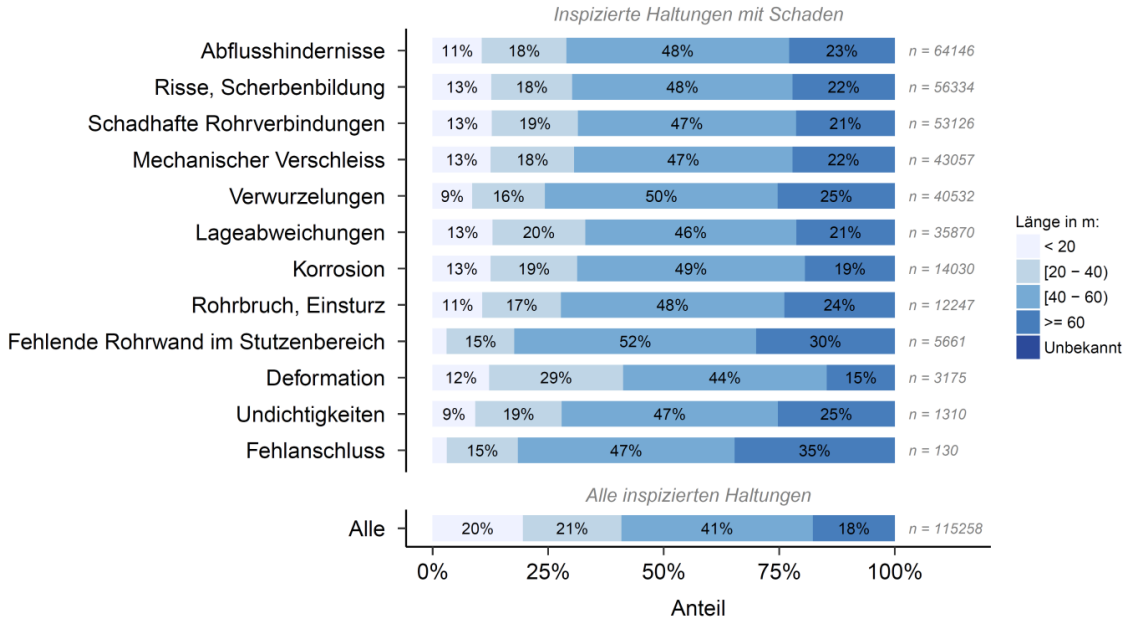


Abbildung 30: Längenverteilung der Haltungen mit einem bestimmten Schaden (oben) und aller inspizierten Haltungen (unten)

Abwassertyp: Korrosionsschäden treten vor allem in Regenkanälen auf (81% der korrodierten Kanäle sind Regenkanäle, Abbildung 31). Allerdings lässt sich daraus nicht schlussfolgern, dass das Medium Regenwasser besonders aggressiv ist. Vielmehr sind 50% aller Regenkanäle aus Beton gefertigt (Abbildung 69e, Anhang B), ein Material was im Allgemeinen korrosionsanfällig ist (siehe Abbildung 25). Für die anderen Abwassertypen liegt der Anteil an Betonkanälen bei < 2%. Fehlanschlüsse werden ebenfalls vorwiegend in Regenkanälen beobachtet, was jedoch nicht heißt, dass sie in Schmutzwasserkanälen nicht vorkommen. Vielmehr sind fehlerhafte Anschlüsse von Regenwasserkanälen ans Schmutzwassernetz aufgrund des unregelmäßigen Niederschlagaufkommens und der geringen sichtbaren Verschmutzungen, z.B. durch Toilettenpapier, allgemein schwieriger zu detektieren.

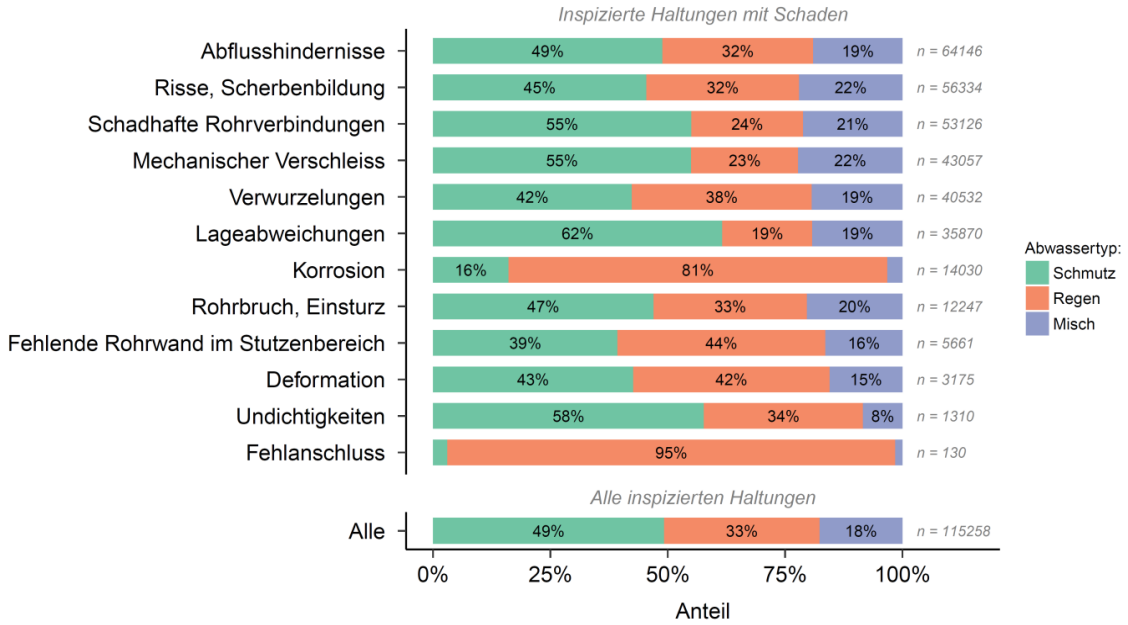


Abbildung 31: Verteilung des Abwassertyps für die Haltungen mit einem bestimmten Schaden (oben) und alle inspizierten Haltungen (unten)

Im Folgenden werden die Verteilungen für die verbleibenden Variablen (Grundwasserüberdeckung, Bäume, Bodentyp, Rückstau, Überdeckung) beschrieben. Die dazugehörigen Abbildungen sind in Anhang C (Abbildung 70) zu finden.

Grundwasserüberdeckung: Undichtigkeiten werden gehäuft bei solchen Kanälen beobachtet, die vom Grundwasser überdeckt sind (Abbildung 70a, Anhang C). Es ist jedoch wahrscheinlich, dass die Häufung vor allem der Tatsache geschuldet ist, dass Undichtigkeiten in nicht vom Grundwasser überdeckten Kanälen bei der Inspektion oft unerkannt bleiben.

Bäume: Wurzelschäden treten vermehrt auf, wenn im Umkreis von 3 m Bäume stehen (zwei Drittel der schadhaften Haltungen, Abbildung 70b in Anhang C), kommen aber auch vor, wenn das nicht der Fall ist (ein Drittel der schadhaften Haltungen vor). Es ist zu vermuten, dass die Wurzeln je nach Baumart auch Abstände > 3 m überwinden können.

Bodentyp: Deformationen kommen gehäuft in Sand-Lehm-Böden vor (58% der schadhaften Haltungen, Abbildung 70c in Anhang C), was jedoch zum großen Teil der Korrelation dieses Bodentyps zum Bezirk und zum Material geschuldet ist. Sand-Lehm-Böden kommen gehäuft in Marzahn Hellersdorf vor (Abbildung 69j, Anhang B), wo PVC-Kanäle besonders häufig sind (Abbildung 69g, Anhang B).

Rückstau: Unter den schadhaften Kanälen ist der Anteil an rückstaubeeinflussten Kanälen ähnlich niedrig wie unter allen inspizierten Kanälen (< 5%, Abbildung 70d in Anhang C). Das heißt der Rückstau hat keinen relevanten Effekt auf die Auftretenswahrscheinlichkeit von Schäden.

Überdeckung: Der Effekt der Überdeckung auf das Auftreten von Schäden ist für die meisten Schadenstypen ebenfalls marginal (Abbildung 70e, Anhang C). Lediglich für Undichtigkeiten ist eine Häufung bei tief liegenden Kanälen zu beobachten. 85% der undichten Kanäle sind mit ≥ 1 m von Boden überdeckt, unter allen inspizierten Haltungen beträgt der Anteil 61%. Begründet werden kann dies vor allem mit einer Korrelation zwischen (Boden-)Überdeckung und Grundwasserüberdeckung. Tief lagernde, d.h. stark vom Boden überdeckte Kanäle, sind häufiger vom Grundwasser überdeckt als Kanäle mit geringer (Boden-)Überdeckung (siehe Abbildung 69d, Anhang B).

3.4 Untersuchung der Unsicherheiten bei der Inspektion

3.4.1 Datenvorbereitung und erste Analysen

Für die Unsicherheitsanalyse wurden die Daten (115.258 Datensätze, siehe Kapitel 2) zunächst nach den in Tabelle 5 gezeigten Kriterien gefiltert.

Tabelle 5: Angewendete Filterschritte und Anzahl der verbleibenden Datensätze für die Unsicherheitsanalyse

Schritt	Filter	Anzahl Datensätze
-	-	115.258
1	Kanalalter verfügbar	103.536
2	Kanalalter > 5 Jahre (Keine Bauabnahme)	97.348
Schritt	**Filter**	**Anzahl Haltungen**
3	Nur Mehrfachbefahrungen	8.946
4	Dauer zwischen Doppelbefahrungen ≤ 5 Jahre	4.695

Etwa 10% der Daten wurden verworfen, weil das Baujahr und damit auch das Alter zum Zeitpunkt der Inspektion unbekannt ist (Tabelle 5, Schritt 1). Weitere 6% wurden verworfen, weil die Inspektion innerhalb der ersten fünf Jahre nach Bau stattfand. Das Mindestalter von fünf Jahren zum Zeitpunkt der Inspektion (Tabelle 5, Schritt 2) wurde festgelegt, um Inspektionen zur Bauabnahme und Inspektionen innerhalb des Gewährleistungszeitraums auszuschließen.

Aus den verbleibenden 97.348 Datensätzen wurden die Haltungen herausgefiltert, die mehrfach inspiziert wurden (Tabelle 5, Schritt 3). Wenn eine Haltung drei Mal befahren wurde, wurden nur die zwei letzten Befahrungen berücksichtigt. Wenn eine Haltung vier Mal befahren wurde, wurden die zwei ersten und die zwei letzten Befahrungen berücksichtigt. Um den Alterungsprozess zwischen den Befahrungen vernachlässigen zu können, wurden weiterhin nur die Haltungen berücksichtigt, bei denen der Zeitraum zwischen den Inspektionen nicht mehr als fünf Jahre beträgt (Tabelle 5, Schritt 4). Der finale Datensatz für die Unsicherheitsanalyse bezieht sich auf 4.695 Haltungen.

Abbildung 32 zeigt den Anteil der Inspektionen pro Inspektionsjahr und pro Haltung. Besonders viele Inspektionen wurden im Zeitraum 2011 bis 2014 durchgeführt. Die verhältnismäßig geringe Anzahl an Inspektionen in den Folgejahren 2015 und 2016 kann durch die verzögerte Bestandsdokumentation erklärt werden, d.h. die Befahrungen waren zum Zeitpunkt der Datenübergabe noch nicht ausgewertet (Abbildung 32, links). 90% der Haltungen wurden nur ein Mal, 9% zwei Mal und 1% mindestens drei Mal inspiziert (Abbildung 32, rechts).

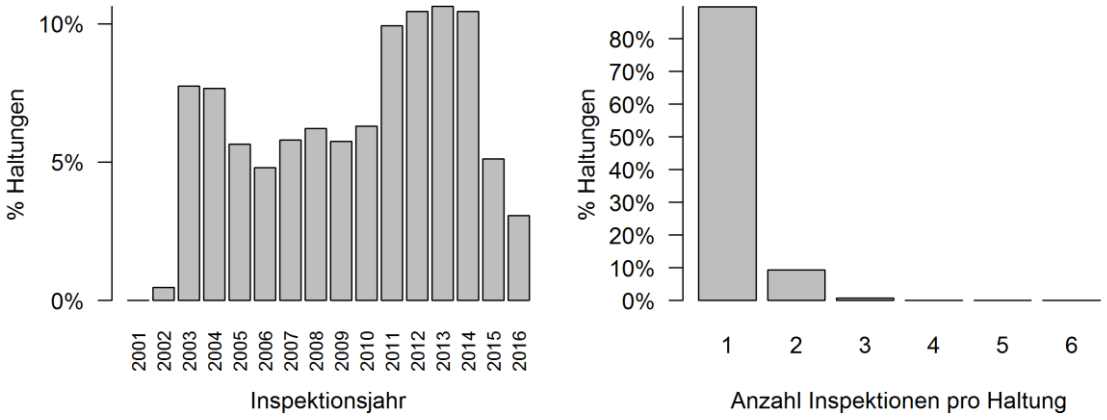


Abbildung 32: Anteil der Inspektionen pro Inspektionsjahr (links) und pro Haltung (rechts). Alle Haltungen wurden zwischen 2001 und 2016 inspiziert. Datensatz nach Anwendung des Filterschrittes Nummer 2 (siehe Tabelle 5)

Abbildung 33 zeigt die Häufigkeitsverteilung der Dauer zwischen den Doppelbefahrungen. Bei etwa der Hälfte der mehrfach befahrenen Haltungen beträgt die Dauer zwischen den Inspektionen höchstens fünf Jahre. Auf Basis dieser Inspektionsdaten wurden die Unsicherheiten der Inspektion beurteilt.

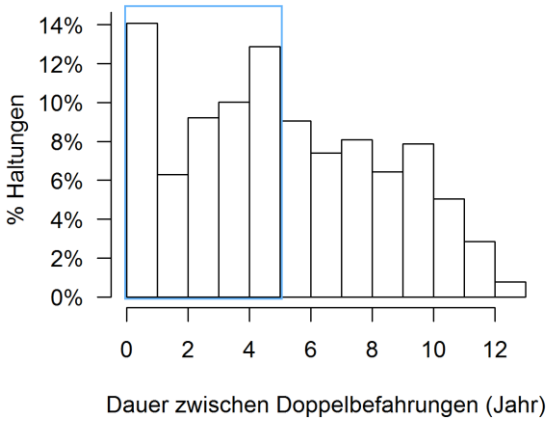


Abbildung 33: Verteilung der Dauer zwischen der Doppelbefahrungen. Datensatz nach Anwendung des Filterschrittes Nummer 3 (siehe Tabelle 1). Das blaue Rechteck zeigt den Datensatz nach Anwendung des Filterschrittes Nummer 4

Bei den Inspektionen kommen im Wesentlichen zwei Kameratypen zum Einsatz: i) Dreh-Schwenkkopf-Kamera (auch als Video-Kamera bezeichnet) und ii) Kugelbildscanner (auch als Panoramo bezeichnet). Während zu Beginn des Inspektionszeitraumes vor allem Dreh-Schwenkkopf-Kameras verwendet wurden, werden seit etwa 2009 verstärkt Kugelbildscanner für die Inspektionen eingesetzt (Abbildung 34). Der Kugelbildscanner enthält ein axial angeordnetes Fisheye-Objektiv mit einem Blickwinkel von über 180°. Im Gegensatz zur Dreh-Schwenkkopf-Kamera werden die Bilder nicht kontinuierlich mittels Videotechnik erfasst, sondern es werden diskrete Fotos in einem definierten Abstand digital aufgenommen. Abbildung 34 zeigt die Verteilung des Kameratyps bzw. des Inspektionszieles pro Jahr. Seit 2014 wird ein Rückgang an Inspektionen mit dem Kugelbildscanner beobachtet, was durch die verzögerte Auswertung und Dokumentation der Inspektionen erklärt werden kann (siehe oben).

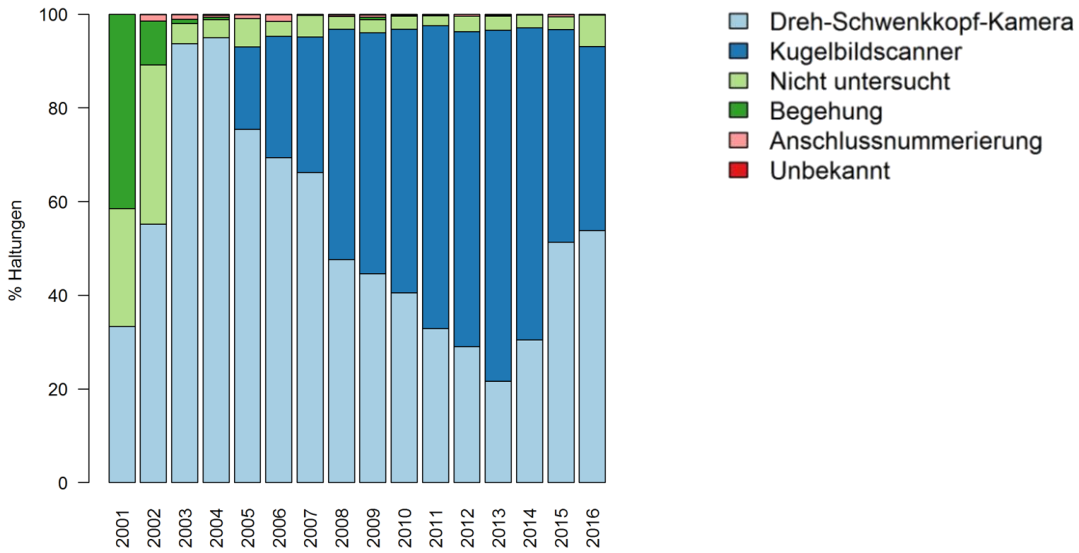


Abbildung 34: Verteilung des Kameratyps bzw. des Inspektionszieles pro Jahr. Datengrundlage vor Filterung (siehe oben).

3.4.2 Abweichungen der Zustandsklasse zwischen den Doppelbefahrungen

Nach Filterung der Daten wurden zunächst die Abweichungen im Inspektionsergebnis zwischen der ersten und der zweiten Befahrung beurteilt. Dabei wurden im ersten Schritt die Abweichungen bezogen auf die sechs Zustandsklassen nach der Berliner Bewertungsmethode analysiert. Im zweiten Schritt wurde die Analyse der Abweichungen mit den drei aggregierten Zustandsbereichen (Kap. 2.3.2) durchgeführt:

- Guter Zustand: langfristiger oder kein Sanierungsbedarf (Zustandsklassen 4, 5 und 6)
- Mittlerer Zustand: mittelfristiger Sanierungsbedarf (Zustandsklasse 3)
- Schlechter Zustand: dringender Sanierungsbedarf (Zustandsklassen 1 und 2)

Die Analyse mit den aggregierten Zustandsbereichen wurde durchgeführt, um den Einfluss der Abweichungen im Inspektionsergebnis auf die Sanierungsentscheidung aufzuzeigen. Zum Beispiel ist eine Abweichung zwischen den Zustandsklassen 6 und 4 nicht so kritisch wie eine Abweichung zwischen den Klassen 3 und 1. In dem ersten Fall würde man bei jeder Inspektion keine Sanierungsmaßnahme planen; im zweiten Fall würde je nach Inspektion die Entscheidung für eine Sanierung anders ausfallen.

Abbildung 35 zeigt die Abweichungen der sechs Zustandsklassen zwischen ersten und zweiten Inspektionen. Daraus lassen sich folgende Ergebnisse ableiten:

- 42% der Haltungen wurden zwei Mal mit der gleichen Zustandsklasse bewertet.
- 58% der Haltungen wurden mit zwei unterschiedlichen Zustandsklassen bewertet.
- 20% der Haltungen wurden bei der zweiten Inspektion um eine Klasse besser bewertet (+1)..
- 17% der Haltungen wurden bei der zweiten Inspektion um eine Klasse schlechter bewertet (-1).

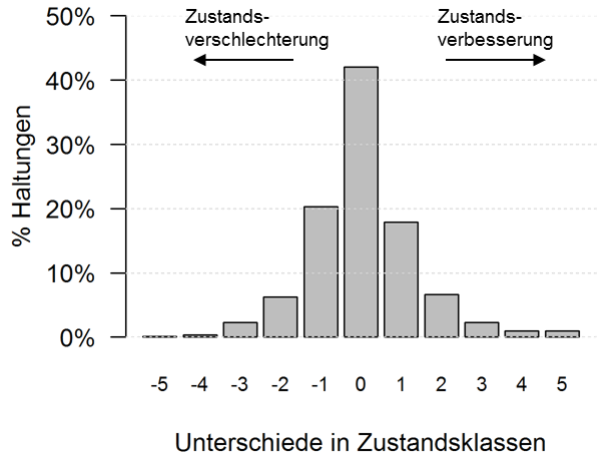


Abbildung 35: Abweichungen zwischen erster und zweiter Inspektion. Datengrundlage nach Anwendung des Filterschrittes 4 (siehe Tabelle 5).

Die in Abbildung 35 dargestellte Verteilung der Abweichungen ist relativ symmetrisch, d.h. der Anteil an Kanälen, die eine Zustandsverbesserung zeigen, ist etwa genauso groß, wie der Anteil an Kanälen, die eine Zustandsverschlechterung zeigen. Daraus lässt sich ableiten, dass Alterungsprozesse in einem Zeitraum von < 5 Jahren keine wesentliche Rolle spielen.

Abbildung 36 zeigt die Abweichungen zwischen Erst- und Zweitinspektion für die drei aggregierten Zustandsbereiche. Die Ergebnisse können folgendermaßen zusammengefasst werden:

- 64% der Haltungen wurden zweimal im gleichen Zustandsbereich inspiziert.
- 36% der Haltungen wurden in zwei unterschiedlichen Zustandsbereichen inspiziert.
- 17% der Haltungen wurden bei der zweiten Inspektion besser bewertet (+1 und +2).
- 19% der Haltungen wurden bei der zweiten Inspektion schlechter bewertet (-1 u. -2).

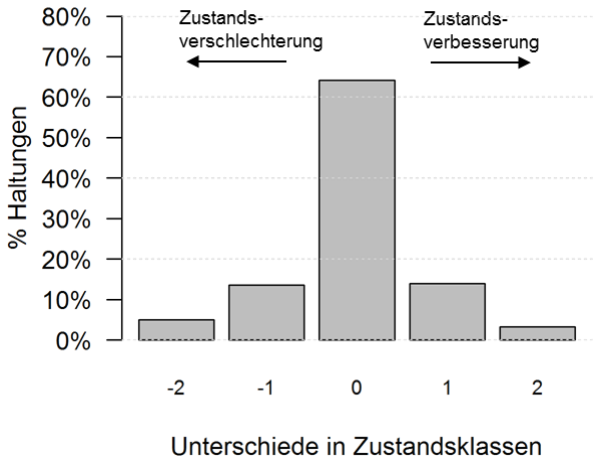


Abbildung 36: Abweichungen zwischen ersten und zweiten Inspektionen mit den drei aggregierten Zustandsbereichen. Datengrundlage nach Anwendung des Filterschrittes 4 (siehe Tabelle 5)

Die Auswertung für die drei aggregierten Zustandsbereiche zeigt viel kleinere Abweichungen als für die sechs Klassen nach der Berliner Bewertungsmethode. Das bedeutet, dass viele Abweichungen zwischen erster und zweiter Inspektion für die Sanierungsentscheidung unbedeutend sind.

3.4.3 Abweichungen der Zustandsklasse mit unterschiedlichen Kameratypen

Für die aggregierten Zustandsbereiche wurde der Einfluss des bei der Befahrung verwendeten Kameratyps auf die Abweichung im Inspektionsergebnis untersucht. Dabei wurden drei Fälle unterschieden:

- Beide Inspektionen wurden mit Dreh-Schwenkkopf-Kamera (Video-Kamera) durchgeführt,
- Beide Inspektionen wurden mit Kugelbildscanner (Panoramo-Foto-Kamera) durchgeführt,
- Die Haltung wurde einmal mit Dreh-Schwenkkopf-Kamera und einmal mit Kugelbildscanner inspiziert.

Tabelle 6 zeigt die Abweichungen zwischen erster und zweiter Inspektion für die drei Fälle. Die Ergebnisse zeigen, dass die Abweichungen zwischen Inspektionen mit Dreh-Schwenkkopf-Kamera in einer ähnlichen Größenordnung sind wie für Inspektionen mit Kugelbildscanner. Das heißt, der Kameratyp hat keinen Einfluss auf die Unsicherheiten der Inspektion.

Tabelle 6: Einfluss der Kameratyp auf die Abweichungen zwischen ersten und zweiten Inspektionen. Datengrundlage nach Anwendung des Filterschrittes 4 (siehe Tabelle 1). Die Dreh-Schwenkkopf-Kamera ist auch als Video-Kamera, der Kugelbildscanner als Panoramo-Foto-Kamera bekannt.

	Abweichung des Zustandes				
	-2	**-1**	**0**	**+1**	**+2**
Nur Dreh-Schwenkkopf-Kamera (n = 2462)	6,4%	13,3%	**64,8%**	12,2%	3,4%
Nur mit Kugelbildscanner (n = 491)	2,2%	15,2%	**63,3%**	17%	2,2%
Dreh-Schwenkkopf-Kamera und Kugelbildscanner (n = 1742)	4,1%	13,4%	**63,3%**	15,5%	3,7%

3.4.4 Bewertung der Unsicherheiten der Kanalzustandsbewertung

3.4.4.1 Methodik

Auf die 4.695 doppelt befahrenen Kanäle wurde eine Optimierungsmethode nach Caradot et al. (2017) zur Ermittlung einer Unsicherheitsmatrix angewendet. Ziel der Untersuchung ist es, in Abhängigkeit vom Inspektionsergebnis Wahrscheinlichkeiten für den tatsächlichen Zustand einer Haltung angeben zu können.

Unter der Annahme, dass jede befahrene Haltung einen *realen* (tatsächlichen) baulichen Zustand hat, der den Sanierungsbedarf beschreibt, ist der reale Zustand definiert als die Zustandsbewertung, die zu der besten Sanierungsentscheidung führen würde. Die beste Schätzung des realen Zustands einer Haltung wäre der durchschnittliche Zustand, der mit einer hohen Anzahl an wiederholten Kamerainspektionen ermittelt worden wäre.

Der reale Zustand einer Haltung ist leider unbekannt, kann aber mit einem *inspizierten* Zustand angenommen werden. Der inspizierte Zustand kann mit dem realen Zustand übereinstimmen. Es ist aber auch möglich, dass der reale Zustand unter- oder überschätzt

wird, da Unsicherheiten jeden Schritt des Zustandsbewertungsverfahrens beeinflussen. Die Unsicherheiten wurden über folgende Schritte quantifiziert.

Schritt 1 - Aufstellen der Unsicherheitsmatrix (P):

Ziel der Analyse ist eine Unsicherheitsmatrix zu ermitteln: diese Unsicherheitsmatrix *P* beschreibt die Wahrscheinlichkeit in einem bestimmten Zustand *i* inspiziert zu sein, wenn eine Haltung im realen Zustand *j* ist. Zum Beispiel: wenn eine Haltung tatsächlich im schlechten Zustand ist, wie hoch ist die Wahrscheinlichkeit, die Haltung im guten oder im schlechten Zustand zu inspizieren? Die Unsicherheitsmatrix ist wie folgt definiert:

$$P = \{P(\beta = i|\alpha = j)\} \qquad \text{Formel 4}$$

Die bedingte Wahrscheinlichkeit $P(\beta = i|\alpha = j)$ beschreibt die Wahrscheinlichkeit im Zustand *i* inspiziert zu sein, wenn eine Haltung im realen Zustand *j* ist.

- Das Element *α* bezeichnet den tatsächlichen Zustand einer Haltung (unbekannt). Die Wahrscheinlichkeit, im tatsächlichen Zustand *j* zu sein, ist $P(\alpha = j)$.
- Das Element *β* bezeichnet den inspizierten Zustand einer Haltung (bekannt). Die Wahrscheinlichkeit, im Zustand *i* inspiziert zu sein, ist $P(\beta = i)$.
- Um die Anzahl von Variablen zu reduzieren und das Gleichungssystem lösen zu können, betrachten wir die drei oben beschriebenen Zustandsbereiche, wobei im Folgenden 1 der gute, 2 der mittlere und 3 der schlechte Zustand ist, der einen dringenden Sanierungsbedarf beschreibt $i, j \in \{1,2,3\}$ (siehe Kap. 3.4.2).

$$P = \begin{vmatrix} P(\beta = 1|\alpha = 1) & P(\beta = 1|\alpha = 2) & P(\beta = 1|\alpha = 3) \\ P(\beta = 2|\alpha = 1) & P(\beta = 2|\alpha = 2) & P(\beta = 2|\alpha = 3) \\ P(\beta = 3|\alpha = 1) & P(\beta = 3|\alpha = 2) & P(\beta = 3|\alpha = 3) \end{vmatrix} \qquad \text{Formel 5}$$

Im Folgenden wird die vereinfachte Schreibweise verwendet:

$$P = \begin{vmatrix} P11 & P12 & P13 \\ P21 & P22 & P23 \\ P31 & P32 & P33 \end{vmatrix} \qquad \text{Formel 6}$$

Schritt 2 - Ermittlung der Anzahl von Haltungen im tatsächlichen Zustand 1, 2, oder 3 (R):

Für eine Reihe von befahrenen Haltungen, kann die unbekannte Anzahl von Haltungen in jedem tatsächlichem Zustand wie folgt ausgedrückt werden:

$$R = \begin{vmatrix} R1 \\ R2 \\ R3 \end{vmatrix} \qquad \text{Formel 7}$$

Schritt 3 - Ermittlung der Anzahl doppelt befahrener Kanäle mit jeweiliger Zustandsbewertung (N):

Wenn Haltungen zweimal befahren wurden, kann eine Matrix *N* mit Ergebnissen der Doppelinspektionen erstellt werden.

$$N = \{N_{ij}\} = \begin{vmatrix} N11 & N21 & N31 \\ N12 & N22 & N32 \\ N13 & N23 & N33 \end{vmatrix} \qquad \text{Formel 8}$$

Jede Zelle der Matrix enthält die Anzahl der doppelbefahren Haltungen, die zuerst im Zustand *i* und dann im Zustand *j* inspiziert wurden. Wenn die Dauer zwischen den wiederholten Inspektionen kurz ist, kann die Zustandsverschlechterung zwischen den zwei Inspektionen vernachlässigt werden. In diesem Fall ist die Matrix (fast) symmetrisch und die wiederholten Inspektionen werden als unabhängig betrachtet.

$$N = \begin{vmatrix} N11 & N21 & N31 \\ N12 & N22 & N32 \\ N13 & N23 & N33 \end{vmatrix} = \begin{vmatrix} N11 & - & - \\ N12 & N22 & - \\ N13 & N23 & N33 \end{vmatrix} \qquad \text{Formel 9}$$

Schritt 4 – Untersuchung der Beziehung zwischen N, R und P:

Die Werte N_{ij} können mit *P* und *R* in einem Gleichungssystem ausgedrückt werden, wobei sich je Zeile der erste Term auf den realen Zustand, der zweite Term auf die erste Inspektion und der dritte Term auf die zweite Inspektion bezieht.

$$N_{Schätzung_{i,j}} \cong \qquad \text{Formel 10}$$
$$R1 \cdot Pi1 \cdot Pj1 +$$
$$R2 \cdot Pi2 \cdot Pj2 +$$
$$R3 \cdot Pi3 \cdot Pj3$$

Die Kanälen, die im Zustand *i* und *j* inspiziert wurden, sind

Kanäle, die tatsächlich im Zustand 1 sind, aber im Zustand *i* und *j* inspiziert wurden und Kanäle, die tatsächlich im Zustand 2 sind, aber im Zustand *i* und *j* inspiziert wurden und Kanäle, die tatsächlich im Zustand 3 sind, aber im Zustand *i* und *j* inspiziert wurden.

- *Schritt 5 - Lösung des Gleichungssystems:*

Das Gleichungssystem enthält 9 Gleichungen. Jede Zelle der Doppelinspektionsmatrix *N* liefert eine Gleichung. Da die Doppelinspektionsmatrix *N* symmetrisch ist, sind die Gleichungen aus dem oberen und unteren Teil der Matrix gleich. Es bedeutet, dass 3 Gleichungen doppelt sind (siehe Formel 9). Das Gleichungssystem enthält dann nur 6 unabhängige Gleichungen.

Das System hat 12 Variablen mit 8 Freiheitsgraden (die Anzahl der Freiheitsgrade eines Gleichungssystems entspricht der Anzahl der unabhängigen Variablen). Dafür sind zunächst folgende Vereinfachungen nötig:

Variablen aus R:

Die Summe der Haltungen in jedem realen Zustand (*R1* + *R2* + *R3*) ist gleich der Summe der doppelt befahrenen Haltungen ($\sum N_{ij}$).

$$R1 + R2 + R3 = \sum N_{ij}$$

Die Anzahl der Haltungen im realen Zustand 3 (*R3*) kann mit *R1* und *R2* und mit der Anzahl von doppelt befahrenen Haltungen $\sum N_{ij}$ ermittelt werden.

$$R3 = \sum N_{ij} - R1 - R2 \qquad \text{Formel 11}$$

Da die Summe der doppelt befahrenen Haltungen $\sum N_{ij}$ bekannt ist, hat *R* 3 Variablen aber nur 2 Freiheitsgrade (2 Variablen sind unabhängig).

Variablen aus P:

Die Summe der Spalten von *P* (3 x 3 Matrix) ist gleich 1, da unabhängig vom tatsächlichen Zustand (in den Spalten) die Wahrscheinlichkeit in einem der 3 Zustände inspiziert zu sein (in den Zeilen), 100% beträgt.

$$P = \begin{vmatrix} P11 & P12 & P13 \\ P21 & P22 & P23 \\ P31 & P32 & P33 \end{vmatrix}$$

$$P11 + P21 + P31 = 1$$

$$P12 + P22 + P32 = 1$$

$$P13 + P23 + P33 = 1$$

Verallgemeinert für die Spalte *i*:

$$P1i + P2i + P3i = 1$$

$$P3i = 1 - P1i - P2i \qquad \text{Formel 12}$$

Die letzte Spalte der Matrix kann mit dem oberen Teil 2 x 3 berechnet werden:

P hat 9 Variablen aber nur 6 Freiheitsgrade (6 Variablen sind unabhängig).

Das System von 6 Gleichungen und 8 Freiheitsgraden, d.h. 8 unabhängigen Variablen kann mathematisch nicht direkt gelöst werden. Es wird daher mit der globalen Optimierungsmethode *ISRES* („Improved Stochastic Ranking Evolution Strategy“, Runarsson und Yao 2005) gelöst. Diese Methode wurde als Derivat-freie Optimierung gewählt (keine Notwendigkeit, den Gradienten der Zielfunktion zu bestimmen). Die Optimierungsmethode minimiert eine Zielfunktion, die mit der Abweichung zwischen der beobachteten Doppelinspektionsmatrix und ihrer Schätzung definiert ist:

$$\min \sum \left(N_{ij} - N_{Schätzung_{i,j}} \right)^2 \qquad \text{Formel 13}$$

Der Algorithmus ermittelt die Variablen, die die Zielfunktion minimieren und die beste Schätzung der Doppelinspektionsmatrix ermöglicht.

Die 8 unabhängigen Variablen sind *P11*, *P21*, *P12*, *P22*, *P13*, *P23*, *R1* und *R2*. Die folgenden Nebenbedingungen wurden definiert.

- $P_{ij} \in [0,1]$ – Die Wahrscheinlichkeit schwankt zwischen 0 und 1.
- $P_{ii} > P_{ji}$ – Die Wahrscheinlichkeit, eine richtige Bewertung zu bekommen, ist höher als die Wahrscheinlichkeit einer falschen Bewertung.

Diese Annahmen unterstützen die Lösung des Gleichungssystems durch das Verwerfen unrealistischer Lösungen.

3.4.4.2 Ergebnisse und Diskussion

Die Optimierungsmethode wurde auf die 4.695 doppelt befahrenen Kanäle unabhängig vom Kameratyp angewendet. Die Analyse wurde für die drei Zustandsbereiche gut, mittel und schlecht durchgeführt, wobei der schlechte Zustand einen dringenden Sanierungsbedarf beschreibt (siehe Kap. 3.4.2). Tabelle 7 zeigt die Doppelinspektionsmatrix *N*, wobei jede Zelle die Anzahl der Haltungen zeigt, die zuerst im Zustand *i* und dann im Zustand *j* bewertet wurden.

Tabelle 7: Doppelinspektionsmatrix *N*

		Doppelinspektionsmatrix *N*		
		Erste Inspektion		
		Gut	**Mittel**	**Schlecht**
Zweite Inspektion	**Gut**	1652	371	196
	Mittel	371	494	274
	Schlecht	196	274	866

Lesebeispiel:

- 494 Haltungen wurden zwei Mal im mittleren Zustand bewertet.
- 274 Haltungen wurden erst im schlechten und dann im mittleren Zustand bewertet.

Die in Kap. 3.4.4.1 beschriebene Optimierungsmethode liefert die folgende Unsicherheits-matrix (Tabelle 8).

Tabelle 8: Unsicherheitsmatrix *P*

		Unsicherheitsmatrix *P*		
		Realer Zustand		
		Gut	**Mittel**	**Schlecht**
Inspizierter Zustand	**Gut**	82%	12%	9%
	Mittel	14%	69%	12%
	Schlecht	4%	19%	79%

Lesebeispiel:

- Wenn eine Haltung tatsächlich im schlechten Zustand ist,
  - Ist die Wahrscheinlichkeit, die Haltung im schlechten Zustand zu bewerten, 79%.
  - Ist die Wahrscheinlichkeit, die Haltung im mittleren Zustand zu bewerten, 12%.
  - Ist die Wahrscheinlichkeit, die Haltung im guten Zustand zu bewerten, 9%.

Die folgenden Ergebnisse können zusammengefasst werden.

- Die Unsicherheit ist relativ gering für Haltungen in gutem Zustand. Wenn eine Haltung im guten Zustand ist (d.h. 4, 5 oder 6 nach der Berliner Bewertungsmethode), liegt die Wahrscheinlichkeit, die Haltung auch in diesem Zustand zu inspizieren, bei 82%. Es gibt relativ geringe Unsicherheiten bei der Bewertung des Zustandes einer Haltung mit wenigen oder minder schweren Mängeln..
- Für Haltungen in schlechtem Zustand liegt die Wahrscheinlichkeit, eine richtige Bewertung zu bekommen, bei 79%. Entsprechend ist die Wahrscheinlichkeit, eine falsche Bewertung zu bekommen, 21%. Die Inspektionen sind geringfügig fehleranfälliger, wenn viele Defekte vorhanden sind.
- Die Unsicherheiten bei der Bewertung von Haltungen im mittleren Zustand sind bedeutend größer als bei Haltungen im guten oder schlechten Zustand. Die Wahrscheinlichkeit, Haltungen im mittleren Zustand richtig zu inspizieren, beträgt 69%. Dies zeigt die Schwierigkeiten bei der Defektbewertung und -quantifizierung in Zusammenhang mit der Kodierung von mittelschweren Schäden.

Die Ergebnisse unterstreichen die hohen Unsicherheiten bei der Zustandsbewertung von Abwasserkanälen. Hauptursachen für die Unsicherheiten sind zum einen die subjektive Bildauswertung (unterschiedliche Interpretation je nach Betrachter, Reinigungsgrad, Lichtverhältnissen, etc.), zum anderen Lücken in der Sanierungsdokumentation. Die Analysen sind nicht nur für die Interpretation der in den vorangegangen Kapiteln gezeigten Ergebnisse wichtig. Sie können auch dafür genutzt werden, die Unsicherheiten der Alterungsmodelle zu erklären. Die Zustandsbewertungen sind die Eingangsdaten der Alterungsmodelle und es ist zu erwarten, dass die Qualität der Prognose von Alterungsmodellen (z.B. zum Investitionsbedarf) stark von der Qualität der Eingangsdaten abhängt.

4 Modellierung des Zustands der Abwasserkanäle

4.1 Untersuchte Modellansätze

Die verwendeten Modellansätze können in statistische Modelle und Modelle des maschinellen Lernens unterschieden werden. Statistische Modelle wie z.B. GompitZ formalisieren Zusammenhänge zwischen den Variablen in Form von mathematischen Gleichungen, um Wahrscheinlichkeiten für den Zustand eines Kanals zu berechnen. Modelle des maschinellen Lernens wie Random Forest, Support Vector Machine und Künstliche Neuronale Netze lernen selbstständig aus den Alterungsmustern und Einflussfaktoren bereits inspizierter Haltungen (Training) und übertragen das Wissen auf andere Haltungen, für die dann Vorhersagen getroffen werden können. Die im Rahmen von SEMA-Berlin untersuchten Modellansätze werden in den folgenden Abschnitten beschrieben.

4.1.1 GompitZ

Das Modell GompitZ (Le Gat 2008) basiert auf der Theorie der Überlebensanalyse (Elandt-Johnson und Johnson 1980) und Markow-Ketten (Markow 2006). Die Eingangsdaten des Modells sind (i) die Zustandsklasse (Kap. 2.3.2) und (ii) die Einflussfaktoren für jede einzelne Haltung (Kap. 2.3.1). Die Einflussfaktoren können kategorische oder numerische Variablen sein.

Zunächst werden die Haltungen einzelnen Kohorten zugeordnet. Kohorten sind Gruppen von Haltungen mit gleichen Merkmalen, z.B. gleichem Material und gleichem Abwassertyp. Für die Kohortenbildung werden die kategorischen Variablen verwendet, wobei numerische Variablen durch Klassifizierung in kategorische überführt werden können (z.B. Länge < 20 m; 20-40 m; 40-60 m; ≥ 60 m). Während des Kalibrierungsprozesses werden für jede einzelne Kohorte Zustandsübergangsfunktionen (Überlebenskurven) ermittelt. Innerhalb einer Kohorte stellen die Überlebenskurven die mittlere Alterung des Kanalnetzes über die Zeit dar: sie definieren den Anteil an Haltungen, die in einem bestimmten Alter noch überlebt haben. Die Überlebensfunktionen haben die mathematische Form einer Gompertz-Verteilung (Abbildung 37) und werden mit der Maximum-Likelihood-Methode (Le Gat 2008) kalibriert. Die Form der Überlebenskurven kann durch die numerischen Variablen, hier Kovariaten genannt, beeinflusst werden. Für *n* Zustandsklassen gibt es *n-1* Überlebenskurven.

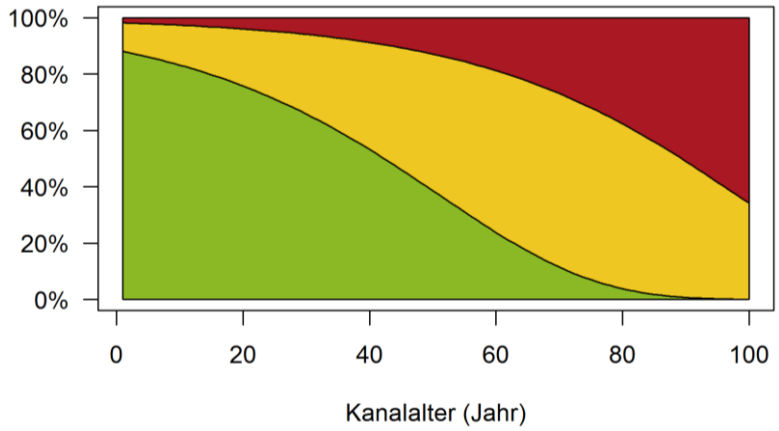


Abbildung 37. Beispiel von zwei Gompertz-Überlebenskurven für drei Zustandsklassen/-bereiche

Mit den kalibrierten Überlebenskurven kann für jede Haltung und für jeden Zeitpunkt die Wahrscheinlichkeit berechnet werden, in einem bestimmten Zustand zu sein. Auf diese Weise lässt sich der zukünftige Zustand der einzelnen Haltungen simulieren. Für die Vorhersage werden zwei Fälle unterschieden:

Nicht-inspizierte Haltungen: Für nicht inspizierte Haltungen werden die Vorhersagen direkt aus den Überlebenskurven für die jeweilige Kohorte abgeleitet. Für eine Vorhersage im Jahr *T* geben die Überlebensfunktionen *ÜF* einen Wahrscheinlichkeitsvektor mit den Wahrscheinlichkeiten P_i, in jedem Zustand i zu sein (Formel 14).

$$P(T) = \big(P_1(T), P_2(T), P_3(T)\big) = (\text{Ü}F_1(T), \text{Ü}F_2(T) - \text{Ü}F_1(T), 100 - \text{Ü}F_2(T)) \qquad \text{Formel 14}$$

Wenn mehrere Haltungen simuliert werden, wird die Zustandsverteilung der Haltungen als Mittelwert der einzelnen Wahrscheinlichkeiten berechnet.

Inspizierte Haltungen: Für bereits inspizierte Haltungen, für die der Zustand zum Zeitpunkt der letzten Inspektion bekannt ist, können die Überlebenskurven nicht direkt für die Vorhersage verwendet werden. Zum Beispiel wenn eine Haltung im Jahr *T* im Zustand „1“ inspiziert wurde, würde der Wahrscheinlichkeitsvektor (Formel 14) wie folgt aussehen:

$$P(T) = (1, 0, 0)$$

Im Jahr der Inspektion *T* ist die Haltung mit einer Wahrscheinlichkeit von 100% im Zustand „1“. Die Zustandswahrscheinlichkeiten wurden im Jahr *T* neu initialisiert. In diesem Fall wird die Methode der Markow-Kette angewendet. Der Wahrscheinlichkeitsvektor zum Zeitpunkt *T+1* hängt vom Wahrscheinlichkeitsvektor zum Zeitpunkt *T* und der Übergangswahrscheinlichkeitsmatrix *Q(T+1)* ab.

$$P(T+1) = P(T) \times Q(T+1) \qquad \text{Formel 15}$$

Die Übergangswahrscheinlichkeitsmatrix kann mathematisch aus den Gompertz-Überlebenskurven berechnet werden (siehe detaillierte Berechnung in Le Gat, 2008). Die Markow-Kette für eine Vorhersage zur Zeit *T+n* ist wie folgt definiert (Formel 16):

$$P(T+n) = P(T) \times Q(T+1) \times Q(T+1) \times Q(T+2) \times \ldots \times Q(T+n)$$

$$P(T+n) = P(T) \times \prod_{1}^{n} Q(T+i) \qquad \text{Formel 16}$$

4.1.2 Random Forest

Random Forest (RF) ist ein Klassifikationsverfahren aus dem Bereich des Maschinellen Lernens. Es besteht aus einer Vielzahl unterschiedlicher Entscheidungsbäume, die zu einem Ensemble, d.h. „Wald“ (engl.: „forest“), kombiniert werden. Ziel des Random Forest bzw. der zugrunde liegenden Entscheidungsbäume ist es, Eigenschaften in den Daten zu ermitteln, die eine gute Klassifizierung der Daten erlauben. Für ein besseres Verständnis von Random Forest wird im Folgenden zunächst das Konzept der Entscheidungsbäume erläutert.

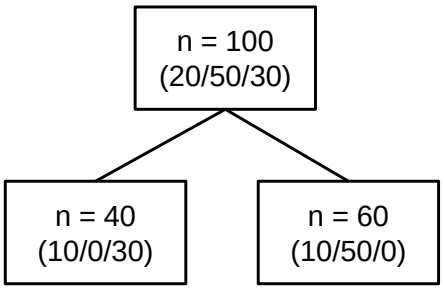
Entscheidungsbäume sind baumartige Strukturen, die bestimmte, hierarchisch aufgebaute Entscheidungsregeln beinhalten und veranschaulichen. Ein Baum besteht aus Knoten und

Verzweigungen. Jeder Knoten repräsentiert einen Entscheidungsregel zu einem bestimmten Merkmal (z.B. Alter ≤ oder > 25 Jahre). Die Verzweigungen stellen die Ergebnisse der Entscheidung dar, d.h. sie sammeln alle Datensätze, die das jeweilige Kriterium erfüllen, in neuen Knoten. Die Anzahl an Knoten und Verzweigungen eines Baumes ist durch die Parametrisierung des Entscheidungsbaumes bzw. des Random Forest vorgegeben (Parameter: *nodesize*; definiert die minimale Anzahl an Datensätzen in einem Knoten). Die letzten Knoten des Entscheidungsbaumes stellen Blätter dar, für die sich möglichst eindeutige Klassifikationsmuster ergeben sollen. Im Idealfall gehören alle Daten eines Blattes zur selben Klasse.

Für jede Entscheidungsregel des Baumes werden die Variable und die entsprechenden Ausprägungen (für kategorische Variablen) bzw. der entsprechende Schwellenwert (für numerische Variablen) ausgewählt, mit denen sich die Daten am besten klassifizieren lassen. Als Maß für die „Reinheit" eines Knotens wird der Gini-Index (Harvey et al. 2014) verwendet (Formel 17):

$$Gini = 1 - \sum_{i=1}^{j} {p_i}^2 \qquad \text{Formel 17}$$

wobei p_i der Anteil an Datensätzen der Klasse *i* in dem jeweiligen Knoten (von insgesamt *j* Klassen) ist. Je kleiner der Gini-Index, desto besser die Klassifizierung. Die Entscheidung, ob und wenn ja, mit welcher Variable und welchen Ausprägungen bzw. welchem Schwellenwert die Verzweigung vorgenommen wird, erfolgt aus dem Vergleich der Gini-Indices vor und nach der Verzweigung. Für die beiden Knoten nach der Verzweigung wird der gewichtete Mittelwert beider Gini-Indices verwendet. Wenn sich keine weitere Reduzierung erreichen lässt oder die minimale Anzahl an Datensätzen in einem Knoten (Parameter: *nodesize)* unterschritten werden würde, so wird keine Verzweigung vorgenommen. Ein Berechnungsbeispiel ist in Abbildung 38 dargestellt.



$$Gini\ (vorher) = 1 - \left(\frac{20}{100}\right)^2 - \left(\frac{50}{100}\right)^2 - \left(\frac{30}{100}\right)^2 = 0{,}62$$

$$Gini\ (nachher) = \frac{40}{100} \times \left(1 - \left(\frac{10}{40}\right)^2 - \left(\frac{0}{40}\right)^2 - \left(\frac{30}{40}\right)^2\right) + \frac{60}{100} \times \left(1 - \left(\frac{10}{60}\right)^2 - \left(\frac{50}{60}\right)^2 - \left(\frac{0}{60}\right)^2\right) = 0{,}32$$

Abbildung 38: Berechnungsbeispiel für den Gini-Index. Im Beispiel führt die Verzweigung zu einer Verringerung des Gini-Indexes und damit zu einer besseren Klassifizierung der Daten.

Entscheidungsbäume können auf verschiedenen Algorithmen basieren, die sich hinsichtlich der Anzahl an möglichen Ästen je Verzweigung und der verwendeten Maßzahl für die „Reinheit" der Knoten unterscheiden. In der vorliegenden Arbeit wurde der CART-Algorithmus (Classification and Regression Trees, Breiman et al. 1984) verwendet. Er ist ein binäres Klassifikationsverfahren, d.h. an jedem Knoten gibt es zwei Verzweigungen. Diaz-Uriarte et al. (2006) haben gezeigt, dass nicht-binäre Algorithmen mit mehr als zwei möglichen Verzweigungen je Knoten (z.B. ID3, C4.5, etc.) trotz der höheren Komplexizität in der Regel keine bessere Klassifizierung erreichen.

Abbildung 39 zeigt ein einfaches Beispiel für einen Entscheidungsbaum aus unserer Studie.

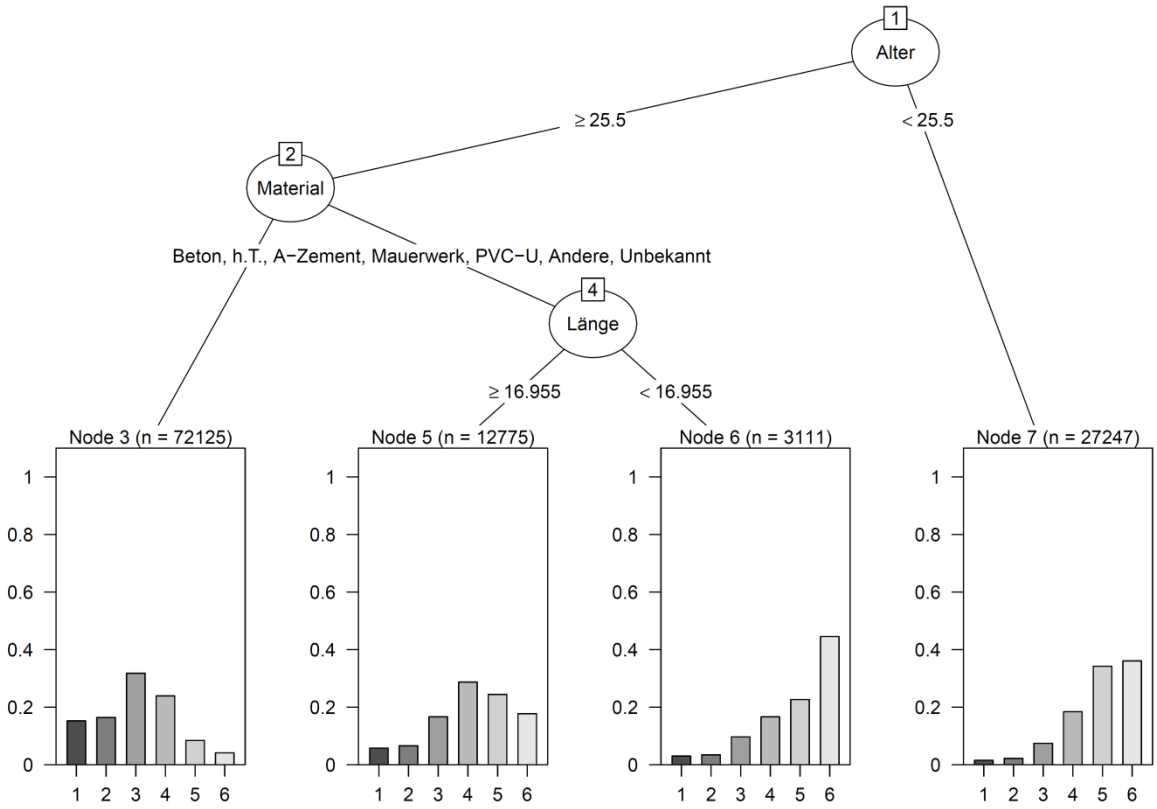


Abbildung 39: Darstellung eines einfachen Entscheidungsbaums zur Klassifizierung der Haltungen nach ihrer Zustandsklasse. Die Balkengrafiken zeigen die Zustandsverteilungen für jedes Blatt (Node 3, 5, 6 und 7; Knoten hier als „Node“ bezeichnet).

Entscheidungsbäume haben den Nachteil, dass sie relativ instabil sind, d.h. bereits kleine Änderungen in den Daten können den Baum stark verändern (Hastie et al. 2008). Da sie außerdem nicht auf einem probabilistischen Modell beruhen, d.h. keine Wahrscheinlichkeitsverteilungsfunktion berücksichtigen, ist es nicht möglich ein Konfidenzintervall (Vertrauensbereich) für die Vorhersage anzugeben (Rokach et al. 2008).

Random-Forest-Modelle, die aus einer Vielzahl an Entscheidungsbäumen bestehen, sind deutlich robuster und erlauben eine genauere Vorhersage als einfache Entscheidungsbäume. Dies gilt besonders für große Datenmengen mit vielen Variablen (Liaw and Wiener 2002). Durch den Ensemble-Ansatz können zudem Konfidenzintervalle für die Vorhersage angegeben werden. Die Größe des Ensembles, d.h. die Anzahl der Bäume, kann über den Parameter *ntrees* gesteuert werden.

Wie alle „lernenden“ Modellansätze werden Random-Forest-Modelle zunächst an Trainingsdaten trainiert, d.h. die Entscheidungsbäume werden wie oben beschrieben aufgebaut. Anschließend werden die den Bäumen zugrunde liegenden Entscheidungsregeln (von der „Wurzel“ bis zum „Blatt“) auf unabhängige Testdaten angewandt. Die Vorhersage des Random-Forest-Modells entspricht dem häufigsten Klassifikationsergebnis aller Bäume des Ensembles. Über festgelegte Güteindikatoren wird die vorhergesagte mit der tatsächlichen Klasse verglichen und die Generalisierbarkeit des Modells beurteilt. Über Gewichtungsfaktoren (w_1 bis w_n) kann bewirkt werden, dass die Klassifikationsergebnisse für eine bestimmte Klasse mehrfach gezählt und damit stärker gewichtet werden.

Zwei „Zufallselemente“ führen dazu, dass sich die Bäume eines Random Forests unterscheiden und der Ensemble-Ansatz gewährleistet ist. Zum einen wird für den Aufbau jedes Baumes ein zufällig ausgewählter Teil der Trainingsdaten verwendet (standardmäßig 63,2%). Zum anderen wird an jedem Knoten eine zufällige Auswahl an Variablen für die nächste Verzweigung bereitgestellt. Beide Größen können über zu wählende Modellparameter *out-of-bag-Rate* und *mtry* gesteuert werden.

4.1.3 Support Vector Machine

Support Vector Machine (SVM) ist ein maschinelles Lern- und Klassifikationsverfahren eingeführt von Vapnik uns Chervonenkis (1974), basierend auf Arbeiten von Fisher (1936) und Rosenblatt (1958). Dabei wird für eine Menge von Trainingsobjekten (hier: Daten zu Kanälen und Inspektionen), für die jeweils bekannt ist, welcher Klasse (hier: Zustandsklasse) sie angehören, jedes Objekt durch einen Vektor in einem Vektorraum repräsentiert. Die Anzahl der Dimensionen des Vektorraums entspricht der Anzahl an Variablen, d.h. bei Berücksichtigung von Alter, Länge und Breite ist der Vektorraum dreidimensional. Aufgabe der Support Vector Machine ist es, in diesen Raum eine Hyperebene[2] (engl.: *hyperplane*) einzupassen, die als Trennfläche fungiert und die Trainingsobjekte in zwei Klassen teilt. Dabei wird der Abstand (engl.: *margin*) derjenigen Vektoren, die der Hyperebene am nächsten liegen, der sogenannten Stützvektoren (engl.: *support vectors*), über das Optimierungsverfahren der Lagrange-Multiplikatoren (Smith 2004) maximiert. Dieser breite, leere Rand soll später dafür sorgen, dass auch Objekte, die nicht genau den Trainingsobjekten entsprechen, möglichst zuverlässig klassifiziert werden. Abbildung 40 veranschaulicht die Trennung der Daten durch eine Hyperebene am Bespiel eines zweidimensionalen Parameterraums.

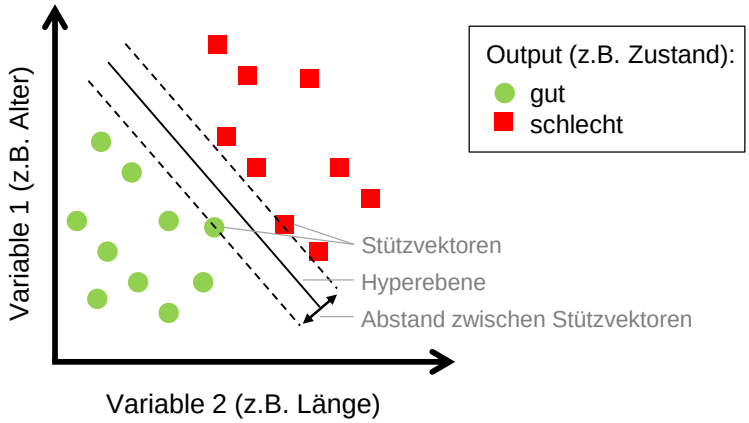


Abbildung 40: Veranschaulichung von Hyperebene, Stützvektoren und Abstand für einen zweidimensionalen Parameterraum

Eine Hyperebene kann nicht „verbogen“ werden, sodass eine saubere Trennung durch die Hyperebene nur dann möglich ist, wenn die Objekte linear trennbar sind. Diese Bedingung ist für reale Trainingsobjektmengen im Allgemeinen nicht erfüllt. Support Vector Machines verwenden im Fall nichtlinear trennbarer Daten den sogenannten Kernel-Trick, um eine nichtlineare Klassengrenze einzuziehen.

[2] Eine Hyperebene ist eine Verallgemeinerung des Konzepts der Ebene auf Räume beliebiger Dimension. Im zweidimensionalen Parameterraum würde man vereinfachend von einer Linie, im dreidimensionalen Raum von einer Ebene sprechen.

Die Idee hinter dem Kernel-Trick ist, den Vektorraum und damit auch die darin befindlichen Trainingsvektoren in einen höherdimensionalen Raum zu überführen. In einem Raum mit genügend hoher Dimensionsanzahl wird auch die komplexeste Vektormenge linear trennbar. In diesem höherdimensionalen Raum wird nun die trennende Hyperebene bestimmt (Abbildung 41). Bei der Rücktransformation in den niedrigerdimensionalen Raum wird die lineare Hyperebene zu einer nichtlinearen, unter Umständen sogar nicht zusammenhängenden Hyperfläche, die die Trainingsvektoren sauber in zwei Klassen trennt. Beim Kernel-Trick kommt in der Regel die Gaußsche Kernel-Funktion, auch Radiale Basisfunktion (RBF) genannt, zum Einsatz.

Sowohl lineare als auch nichtlineare Support Vector Machines lassen sich durch zusätzliche Parameter flexibler gestalten. Der Parameter *C,* eine Konstante im Regularisierungsterm der Lagrange-Formel*,* ist ein Toleranzfaktor, der es erlaubt, einzelne Objekte falsch zu klassifizieren. Ein großes *C* bedeutet einen kleinen Abstand zwischen Stützvektoren und eine geringe Falschklassifizierung für Trainingsdaten (kleines *C*: vice versa). Je größer *C*, desto weniger toleriert das Modell eine falsche Klassifizierung für die Trainingsdaten. Die Trainingsdaten werden besser klassifiziert aber es besteht die Gefahr einer Überanpassung des Modells.

Der Parameter *sigma*, eine Konstante der Gaußschen Kernel-Funktion, definiert, wie weit der Einfluss eines einzelnen Trainingsbeispiels reicht (große *sigma*-Werte: kleine Breite der Gaußschen Kernel-Funktion und kleine Reichweite, und vice versa). Es kommt zwar zu keiner Falschklassifizierung bei den Trainingsdaten. Es besteht jedoch die Gefahr einer Überanpassung des Modells, weil die Trainingsdaten zu exakt abgebildet werden.

Darüber hinaus gibt es Gewichtungsfaktoren (w_1, bis w_n), mit denen die Ungleichverteilung der Daten ausgeglichen werden kann (siehe Random Forest, Kap. 4.1.2).

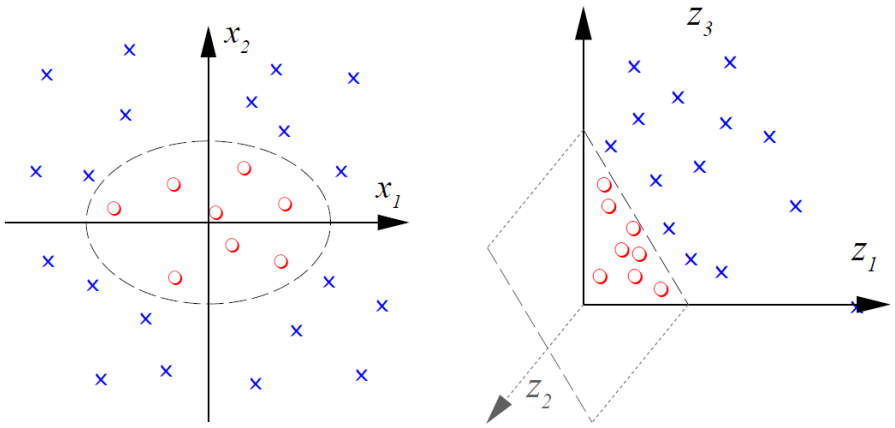


Abbildung 41: Beispiel einer Transformation aus zwei Dimensionen in einem neuen Raum mit drei Dimensionen. Die Punkte und Kreuze zeigen den Output (z.B. Zustand) für jedes Objekt. Die Transformation erlaubt die lineare Trennung der Punkte durch die dargestellte Hyperebene.

Support Vector Machines sind binäre Klassifikatoren, d.h. sie unterscheiden nur zwei Ausprägungen der Zielvariable, z.B. männlich/weiblich, gut/schlecht, etc. Um die Methode auch für Zielvariablen mit mehr als zwei Ausprägungen anwenden zu können, z.B. zur Vorhersage von sechs Zustandsklassen oder drei Zustandsbereichen, muss das Problem in mehrere „binäre“ Probleme zerlegt werden. Dafür gibt es verschiedene Verfahren.

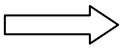
Bei der im Rahmen dieser Studie verwendeten, sogenannten „One-against-one“ Methode wird für jede Kombination zweier Klassen ein eigenes SVM-Modell aufgebaut, das darauf trainiert wird, die Objekte beider Klassen zu trennen. Für drei Klassen werden drei binäre Modelle aufgebaut:

- Modell A: enthält nur Objekte der Klassen 1 und 2
- Modell B: enthält nur Objekte der Klassen 2 und 3
- Modell C: enthält nur Objekte der Klassen 1 und 3

Die Vorhersage (Klasse) eines Objektes wird durch das Modell vorgegeben, das die sicherste Prognose liefert (d.h. den größten Abstand zur Hyperebene).

Modelle des Typs „Support Vector Machine“ können ausschließlich mit numerischen Variablen umgehen. Um dennoch kategorischen Variablen als Eingangsvariablen nutzen zu können, werden diese in sogenannte „Dummy“-Variablen umgewandelt. Dabei wird die kategorische Variable mit *i* Ausprägungen in *i* numerische Variablen aufgeteilt, wobei jede dieser Variablen die Werte 1 (Ja, Ausprägung liegt vor) oder 0 (Nein, Ausprägung liegt nicht vor) enthalten kann. Das Vorgehen ist in Abbildung 42 dargestellt.

Haltung	**Abwassertyp**
1	Schmutz
2	Misch
3	Regen
4	Misch
5	Schmutz
…	…



Haltung	**Abwassertyp = Schmutz?**	**Abwassertyp = Regen?**	**Abwassertyp = Misch?**
1	1	0	0
2	0	0	1
3	0	1	0
4	0	0	1
5	1	0	0
…	…	…	…

Abbildung 42: Umwandlung einer kategorischen Variable in numerische „Dummy“-Variablen am Beispiel des Abwassertyps

4.1.4 Künstliche Neuronale Netze

Künstliche Neuronale Netze (KNN) sind maschinelle Lernverfahren nach dem Vorbild des menschlichen Gehirns, die erstmals von D.O. Hebb (1949) vorgestellt wurden. Wesentlich weiterentwickelt wurde der Modellansatz in den letzten 20 Jahren (Schmidhuber et al. 2015, Shin et al. 2016, Zhang et al. 2018). Künstliche Neuronale Netze werden für Klassifizierungs- oder Regressionsaufgaben aus verschiedensten Fachbereichen wie Autonomes Fahren (Broggi et al. 2007), Bilderkennung (Krizhevsky et al. 2012), Spracherkennung (Hinton et al. 2012), Medizin (Shannon et al. 2003) und Siedlungsentwässerung (Tran et al. 2006) verwendet.

Künstliche Neuronale Netze bestehen aus künstlichen Neuronen, die in sogenannten Layern angeordnet sind und über künstliche Synapsen verbunden sind. Es gibt im Allgemeinen drei Typen von Neuronen: i) Input-Neuronen, ii) versteckte Neuronen und iii) Output-Neuronen. Die Input-Neuronen im sogenannten *input layer* repräsentieren die numerischen Eingangsvariablen des Modells, wobei jedes Neuron eine Variable repräsentiert. Die Output-Neuronen im sogenannten *output layer* repräsentieren die Wahrscheinlichkeit, in einem bestimmten

Zustand zu sein. Für drei Zustandsbereiche gibt es drei Output-Neuronen. Die versteckten Neuronen im sogenannten *hidden layer* sind Verbindungseinheiten zwischen den Input- und Output-Neuronen. Abbildung 43 zeigt eine schematische Darstellung eines Künstlichen Neuronalen Netzes mit einem *hidden layer*.

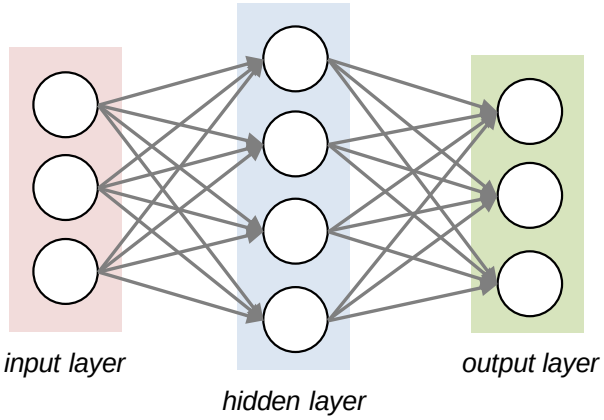


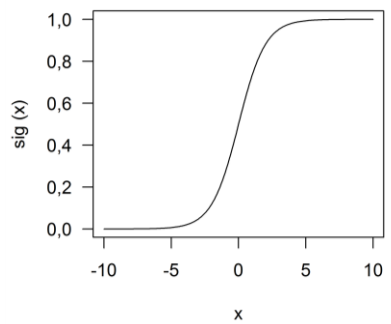
Abbildung 43: Schema eines Künstlichen Neuronalen Netzes mit einem *hidden layer*

Jedes Neuron hat einen Wert und jede Verbindung zwischen zwei Neuronen hat ein Gewicht. Der Wert eines Neurons hängt von den Werten der Neuronen des vorangehenden Layers und den Gewichten der Verbindungen zwischen dem vorangehenden Layer und dem Neuron ab. Lediglich die Werte der Input-Neuronen werden durch die Werte der Variablen vorgegeben. Aus den Werten der Neuronen und den Gewichten der Verbindungen werden unter Anwendung einer sogenannten Aktivierungsfunktion die Neuronenwerte der *hidden layers* und die Wahrscheinlichkeiten des Output-Layers berechnet (Formel 18).

$$n_{i,j} = f\left[\sum_{k=1}^{N_{j-1}} (n_{k,j-1} * w_{k,j-1}^{i,j})\right] \qquad \text{Formel 18}$$

$n_{i,j}$ ist der gesuchte Wert des Neurons *i* im Layer *j,*
f ist die Aktivierungsfunktion,
N_{j-1} ist die Anzahl an Neuronen des vorangehenden Layers,
$n_{k,j-1}$ sind die Werte der Neuronen des vorangehenden Layers,
$w_{k,j-1}^{i,j}$ sind die Gewichte der Verbindungen von jedem dieser Neuronen zum Neuron *i* des Layer *j*.

Die Aktivierungsfunktion hat das Ziel, den Output jedes Neurons zu „glätten“. Zu den am häufigsten verwendeten Aktivierungsfunktionen zählt die sigmoidal-Kurve (Abbildung 44).

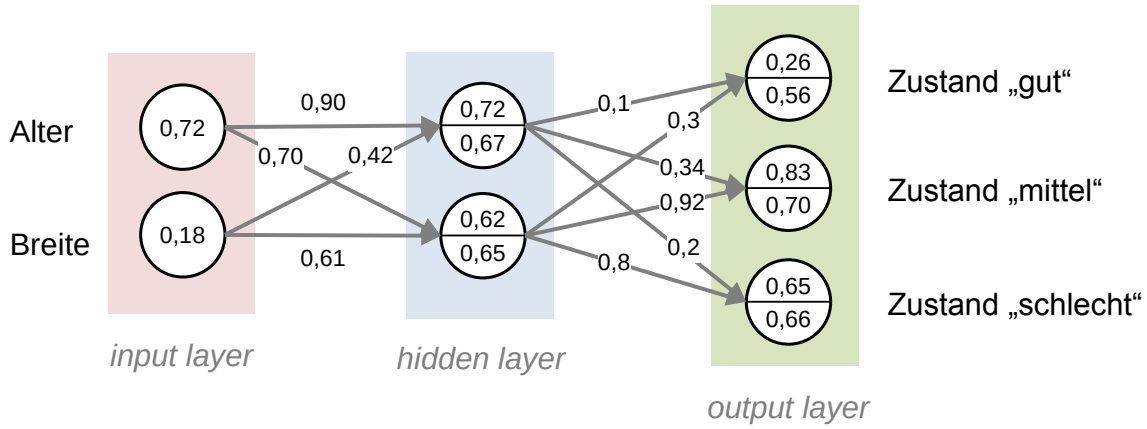


$$sig(x) = \frac{1}{1 + e^{-x}}$$

Abbildung 44: sigmoide Aktivierungsfunktion für Künstliche Neuronale Netze

Die Vorhersage für eine bestimmte Kombination an Eingangsvariablen wird aus dem Output-Neuron, mit der größten Wahrscheinlichkeit abgeleitet. Die Berechnung der Neuronenwerte unter Anwendung von Formel 18 wird beispielhaft in Abbildung 45 erläutert. Das Verfahren wird als „*Feed Forward Propagation*“ bezeichnet.

Rechenbeispiel - „Feed Forward Propagation“:



Herleitung der Werte: Zuerst werden die Werte der versteckten Neuronen berechnet, z.B. für das erste versteckte Neuron: $x_1 = 0{,}72 * 0{,}9 + 0{,}18 * 0{,}42 = 0{,}72$. Dann wird die Aktivierungsfunktion angewandt: $y_1 = sig(x_1) = 0{,}67$. Anschließend werden die Werte der Output-Neuronen berechnet, z.B. für das erste Output-Neuron: $x_1 = 0{,}67 * 0{,}1 + 0{,}65 * 0{,}3 = 0{,}26$. Unter Anwendung der Aktivierungsfunktion ergibt sich: $y_1 = sig(x_1) = 0{,}56$. Die Werte der anderen Neuronen werden analog berechnet. Für das Beispiel würde der Zustand „mittel“ vorhergesagt werden, da das zweite Output—Neuron den höchsten Wert hat ($y_2 = 0{,}70$).

Abbildung 45: Beispiel für die Berechnung der Neuronenwerte unter Anwendung der sig-Aktivierungsfunktion

Die Gewichte der Neuronenverbindungen werden zunächst zufällig gewählt („Initialisierung“). Ziel des Trainings ist es, die Gewichte so zu optimieren, dass es eine möglichst große Übereinstimmung zwischen der Vorhersage und den Inspektionsergebnissen gibt. Für die Optimierung der Gewichte wird häufig die sogenannte Gradientenverfahren („*Gradient Descent*“, Hastie et al. 2008) verwendet. Die Methode beruht auf der schrittweisen Suche des Minimums einer Fehlerfunktion, die die Abweichung zwischen Vorhersage und Inspektion beschreibt. Für jeden Schritt wird eine Teilmenge der Trainingsdaten verwendet, bis alle Trainingsdaten einmal verwendet wurden.

Heute existieren eine Vielzahl verschiedener Varianten Künstlicher Neuronaler Netze. Im Rahmen dieser Studie wurde der ELM-Algorithmus („Extreme Learning Machine“, Huang et al. 2006) für ein sogenanntes „Single Hidden Layer Feedforward Neural Network“ verwendet. Ein Neuronales Netz dieses Typs besteht aus nur einem versteckten Layer zwischen Input- und Output-Layer. Die Parameter dieses Modells sind die Anzahl an versteckten Neuronen (*nhid*) und die Wahl der Aktivierungsfunktion (*actfun*).

Ähnlich den Support Vector Machines können Künstliche Neuronale Netze ausschließlich mit numerischen Variablen umgehen, d.h. die kategorischen Variablen werden wie in Kap. 4.1.3 beschrieben in „Dummy“-Variablen umgewandelt. Darüber hinaus müssen alle Variablen auf den Intervall 0 bis 1 skaliert werden, damit jede Variable denselben Wertebereich hat. Dies geschieht über Formel 19. Ein Alter von 101 Jahren bei einer beobachten Altersspanne von 0 bis 140 Jahren ergibt einen normalisierten Wert von 0,72.

$$x_{[0;1]} = \frac{x - min}{max - min}$$ Formel 19

4.2 Methodisches Vorgehen und Bewertungskriterien

Ziel der Modellierung ist es, den Zustand der Kanäle anhand von baulichen und betrieblichen Eigenschaften sowie Umweltfaktoren vorherzusagen. Zwölf Variablen wurden in Kapitel 3.2.2 als relevant für die Zustandsverteilung des Netzes identifiziert: Alter, Profil, Länge, Grundwasserüberdeckung, Material, Bäume, Bezirk, Abwassertyp, Breite, Bodentyp, Rückstau und Überdeckung. Sie wurden als mögliche Eingangsvariablen der Modelle untersucht. Zielvariable ist der bauliche Zustand, wobei für die Vorhersage die sechs Zustandsklassen zu drei Zustandsbereichen („gut“, „mittel“, „schlecht“) zusammengefasst wurden. Eine nähere Erläuterung zu den Eingangs- und zur Zielvariable ist in Kap. 2.3 zu finden.

4.2.1 Datenfilterung

Bevor die Modelle getestet wurden, wurde der bereinigte und für die statistische Analyse eingesetzte Datensatz (siehe Kap. 2.2) nach den in Tabelle 9 aufgeführten Kriterien gefiltert. Ziel der Filterung war es, Datensätze mit fehlenden Informationen und sehr stark unterrepräsentierte Gruppen zu entfernen. Solche Datensätze können zwar für das Prozessverständnis interessante Informationen beinhalten (Kapitel 3), sind aber für die Modellierung unbrauchbar. Insgesamt wurden, zusätzlich zu den in Kap. 2.2 beschriebenen Filterschritten, nochmals 17.711 von 115.258 Datensätzen entfernt (15%). Nach Filterung bleiben noch 97.547 Datensätze für die Modellierung erhalten.

Tabelle 9: Filterschritte für die Vorbereitung der Daten für die Modellierung

Filterschritte	Anzahl Datensätze [1]
Ausschluss aller Kanäle mit unbekanntem Alter	11.722
Ausschluss aller Kanäle mit unbekannter Länge	1
Ausschluss aller Kanäle mit unbekannter Überdeckung	169
Ausschluss aller Kanäle, die nicht den vier Hauptprofilen (Kreis, Kreis i.V., Ei, Maul) zugeordnet werden können	627
Ausschluss aller Kanäle mit Materialien, die nicht den sechs Hauptgruppen (Steinzeug, Beton, Beton h.T., A-Zement, Mauerwerk, PVC-U) zugeordnet werden können	3.771
Ausschluss aller Kanäle mit unbekanntem Bezirk	344
Ausschluss aller Kanäle mit Bodenarten, die nicht den drei Hauptgruppen (Aufschüttungen, Sand, Sand + Lehm) zugeordnet werden können	2.482

[1] *Einige Datensätze erfüllen mehrere Filterkriterien. Daher ist die Anzahl der insgesamt gefilterten Datensätze kleiner als die Summe der Spalte „Anzahl Datensätze (je Kriterium)“.*

Die Untersuchung der Modellgüte wurde im Wesentlichen nach folgenden Schritten durchgeführt:

1. Aufteilung der Daten in Trainings- und Testdaten
2. Training: Bestimmung der Modellparameter und Modellaufbau
3. Test: Beurteilung der Genauigkeit der Vorhersage

Die einzelnen Schritte für die Untersuchung der Modellgüte werden im Folgenden detaillierter erläutert.

4.2.2 Aufteilung der Daten in Trainings- und Testdaten

Die 97547 Datensätze wurden im Verhältnis 60:40 zufällig in Trainings- und Testdaten aufgeteilt. An den Trainingsdaten (n = 58.528) werden die Modelle trainiert und parametrisiert. An den Testdaten (n = 39.019) wird die Güte der Modelle geprüft. Die Aufteilung der Daten wurde einmal vorgenommen, anschließend wurden für alle Modellansätze die gleichen Trainings- und Testdaten verwendet.

Um sicherzustellen, dass weder Trainings- noch Testdaten Muster aufweisen, die nicht den allgemeinen Netzeigenschaften entsprechen, wurden zunächst die Verteilungen aller Eingangsvariablen beider Gruppen miteinander verglichen. Der Vergleich zeigt, dass sich Trainings- und Testdaten hinsichtlich der Datenverteilungen nicht unterscheiden (Abbildung 46).

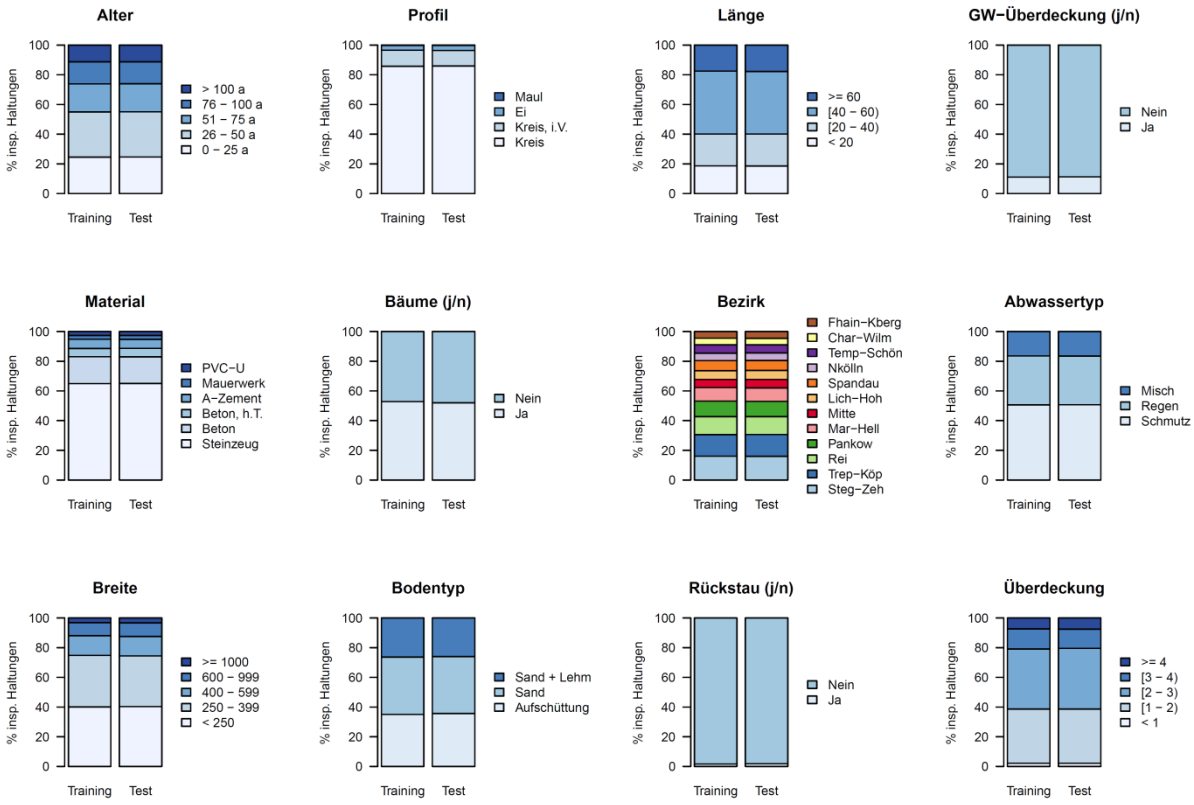


Abbildung 46: Datenverteilungen für Trainingsdaten (n = 58.528) und Testdaten (n = 39.019)

4.2.3 Training der Modelle

Beim Training der Modelle wurden zunächst je Modell Parameterbereiche festgelegt, für die die Modelle getestet wurden (Tabelle 10). Ziel ist es den Parametersatz zu finden, für den das Modell die beste Vorhersage bietet. Die Parametersuche wurde über das Verfahren der Kreuzvalidierung durchgeführt, eine häufig verwendete Methode um eine Überanpassung des Modells („overfitting“) zu vermeiden (Pados & Papantoni-Kazakos 1994). Dabei wurden die Trainingsdaten (n = 58.528) zunächst zufällig in k gleich große Teilmengen aufgeteilt. Anschließend wurde das Modell k mal mit jeweils k-1 der k Teilmengen trainiert, die jeweils verbleibende Teilmenge wurde für die Validierung (vorläufiger Test des Modells) verwendet. Jede der k Teilmengen bildet einmal den Validierungsdatensatz. Die jeweils anderen k-1 Teilmengen bilden die Trainingsdaten. Bei jeder der k Validierungen wurden anhand des Vergleiches von prognostizierter und inspizierter Zustandsklasse die in Kap. 4.2.5 beschriebenen Güteindikatoren berechnet. Der Mittelwert der Güteindikatoren wird dazu verwendet, die Generalisierbarkeit der Parameterschätzung zu beurteilen und den Parametersatz mit der besten Vorhersagegenauigkeit zu bestimmen. Die Anzahl der Kreuzvalidierungen in dieser Studie beträgt je nach Rechenaufwand k = 5 (für Random Forest, Support Vector Machine, Künstliche Neuronale Netze) bzw. k = 3 (für GompitZ). Der gewählte Wert richtet sich nach dem Rechenaufwand. Die Datenaufteilung bei der Kreuzvalidierung ist Abbildung 47 veranschaulicht.

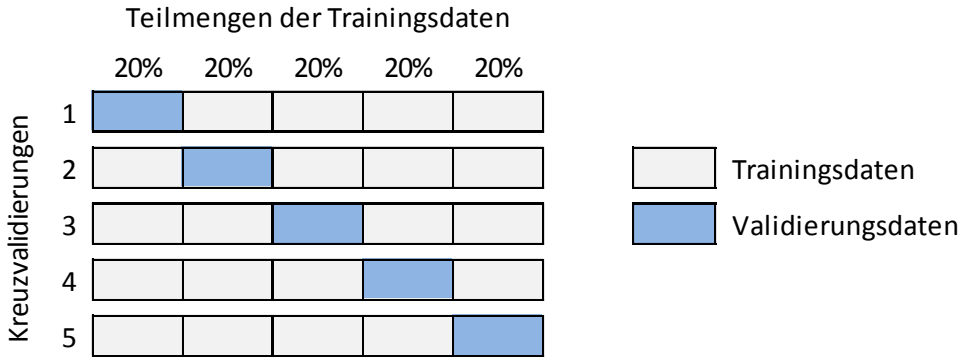


Abbildung 47: Aufteilung der Daten bei der Kreuzvalidierung (Beispiel: 5-fache Kreuzvalidierung)

Die Parametersuche wurde für Random Forest und Support Vector Machine in zwei Schritten durchgeführt:

1. Zufällige Variation aller Parameter (gleichzeitig) über festgelegten Parameterbereich (ca. 1.000 Stichproben) und Visualisierung des Effektes auf Bewertungsindikatoren (siehe Kap. 4.2.5) → Fixierung der Parameter, die klares Optimum zeigen (Reduzierung des „Parameterraums“),
2. Gitter-Suche für verbleibende Parameter (ca. 100 Kombinationen) und Visualisierung des Effektes auf Bewertungsindikatoren (siehe Kap. 4.2.5) → Fixierung der verbleibenden Parameter.

Durch die zweistufige Parametersuche mit Fixierung von Parametern nach dem ersten Schritt konnte die Rechenzeit gegenüber einer vollständigen Gittersuche mit Berücksichtigung aller Parameter und großer Wertebereiche deutlich reduziert werden.

Für Support Vector Machine wurde die Parametersuche aufgrund des hohen Rechenaufwands nur für eine Teilmenge von 10.000 Datensätzen durchgeführt (8.000 für

Training, 2.000 für Validierung, 5-fache Kreuzvalidierung). Der ermittelte Parametersatz wurde später verifiziert, indem das Modell insgesamt dreimal mit unterschiedlichen Teilmengen aus 8.000 Datensätzen trainiert und anschließend getestet wurde.

Für Künstliche Neuronale Netze, für die es nur zwei Modellparameter gibt (Kap. 4.1.4), wurde auf den Schritt der zufälligen Parametervariation (Schritt 1, oben) verzichtet und eine vereinfachte Gitter-Suche mit 40 Parameterkombinationen durchgeführt. Um eine Überschreitung der Rechnerkapazität zu vermeiden, wurde die Anzahl der Trainingsdaten um 20% auf 46.822 Datensätze reduziert.

Für GompitZ besteht die Parametersuche im Wesentlichen aus der Bildung von Kohorten (siehe Kap. 4.1.1) und der Kalibrierung der Überlebensfunktionen. Für die Bildung der Kohorten wurden verschiedene Kombinationen aus zwei bis fünf kategorischen Variablen untersucht. D.h. die Kanäle wurden nach den Merkmalen dieser Variablen gruppiert. Anschließend wurden unter Berücksichtigung des Alters die Überlebensfunktionen kalibriert und die Modellgüte beurteilt. Auch andere numerische Variablen, wie z.B. die Länge oder die Breite des Kanals, wurden bei der Kalibrierung berücksichtigt, führten aber nicht zu einer Verbesserung der Modellergebnisse.

Bei Daten, in denen die verschiedenen Klassen unterschiedlich häufig vorkommen, besteht die Gefahr, dass das Modell bevorzugt auf die am stärksten vertretene Klasse trainiert wird und die Prognosequalität für die unterrepräsentierten Klassen ungenügend ist (Japkowicz & Stephen 2002, Maalouf & Trafalis 2011). Um dies zu vermeiden, kann das Verfahren des „Upsampling“ oder „Downsampling“ angewandt werden (Barandela et al. 2004). Dabei wird eine definierte Anzahl zufällig ausgewählter Datensätze der unterrepräsentierten Klassen mehrfach für das Training verwendet („Upsampling“) oder verworfen („Downsampling“), so dass am Ende alle Klassen gleich häufig in den Trainingsdaten vertreten sind. Beide Ansätze wurden untersucht, führten aber nicht zu einer Verbesserung der Modellergebnisse.

In Absprache mit BWB wurden drei Zustandsbereiche festgelegt, die sich am Sanierungsbedarf der Haltung orientieren (kurz-, mittel-, langfristig) und die bei der Vorhersage unterschieden werden sollen (siehe Kap. 2.3.2). Beim Modelltraining wurden zwei Ansätze zur Aggregierung der sechs Zustandsklassen stichpunktartig geprüft: i) Aggregierung der Klassen vor der Modellierung, d.h. in den Eingangsdaten und ii) Aggregierung der Klassen nach der Modellierung, d.h. die Vorhersage wird zunächst für sechs Klassen gemacht, anschließend werden die Ergebnisse der Vorhersage aggregiert. Für alle Modelle lieferte die Aggregierung der Klassen in den Eingangsdaten (i) die besseren Ergebnisse und wurde weiterverfolgt. Tabelle 10 zeigt die Konfigurationen beim Training der vier Modelle.

Tabelle 10: Konfigurationen bei Kalibrierung / Training der untersuchten Modelle

Modellansatz	Variablen-typ [1]	Untersuchte Parameterbereiche [2]	Anzahl Datensätze	Anzahl KV [3]
GompitZ	k + n	Keine (Parameter werden im Rahmen der Kalibrierung bestimmt, siehe Kap. 4.1.1)	58.528	3
Random Forest	k + n	*ntrees*: 1–700; *nodesize*: 4–1808; *mtry*: 1–12; *w1*: 0,2–3; *w2*: 0,2–3; *w3*: 1	58.528	5
Support Vector Machine	n	*C: 0,1 – 700; sigma: 0,001 – 100; w1*: 1; *w2: 0,1 – 3; w3: 0,1 – 3*	10.000 [4]	5
Künstliche Neuronale Netze	n	*actfun*: sig, tansig, radbas, purelin; *nhid*: 100–1000	46.822	5

Erläuterungen: [1] Mögliche Variablentypen sind kategorische (Typ: k) und numerische Variablen (Typ: n). Für Künstliche Neuronale Netze und Support Vector Machine sind nur numerische Variablen zulässig, d.h. die kategorischen Variablen mussten wie oben beschrieben in numerische Variablen überführt werden. [2] Die Bedeutung der einzelnen Modellparameter ist in den Kapiteln 4.1.1 bis 4.1.4 erläutert. [3] KV = Kreuzvalidierungen. [4] 10.000 Datensätze für die Parametersuche mit Kreuzvalidierung, 8.000 Datensätze für den finalen Modellaufbau.

Im Anschluss an die Untersuchung des Einflusses der Modellparameter wurden die Modelle für die jeweils beste Konfiguration abschließend mit allen Trainingsdaten (n = 58.528) kalibriert bzw. trainiert. Ausnahmen stellen die Modelle Support Vector Machine (SVM) und Künstliche Neuronale Netze (KNN) dar, die aufgrund des hohen Rechenaufwands nur für 8.000 (SVM) bzw. 46.822 Datensätze (KNN) trainiert wurden. Für SVM wurde durch dreifache Wiederholung von Training und Test verifiziert, dass das Modell trotz der verhältnismäßig kleinen Menge an Trainingsdaten stabile Ergebnisse liefert.

4.2.4 Test der Modelle

Für die Beurteilung der Vorhersagequalität wurden alle Modelle auf die Testdaten (n = 39.019) angewandt und die prognostizierte mit der inspizierten Zustandsklasse verglichen. Die Bewertung erfolgte über die im nachfolgenden Unterkapitel beschriebenen Güte- bzw. Bewertungsindikatoren.

Für ein statistisches Modell (GompitZ) und ein Modell für maschinelles Lernen (Random Forest) wurde zudem untersucht, inwiefern die Modellgüte von der für die Kalibrierung bzw. das Training zur Verfügung stehenden Datenmenge abhängt. Dafür wurden aus den Trainingsdaten zufällig n Datensätze gezogen und das Modell damit trainiert. Anschließend wurde an den Testdaten (Kap. 4.2.2) die Modellgüte geprüft. Folgende Datenumfänge n für das Training wurden untersucht: 150, 400, 1.000, 3.000, 8.000, 20.000 und 58.000. Für jede dieser Datenumfänge wurden 50 Wiederholungen durchgeführt, d.h. das Modell wurde 50 mal mit zufällig ausgewählten n Datensätzen trainiert und mit den Testdaten getestet (Monte-Carlo-Simulation).

4.2.5 Bewertungsindikatoren

Die Modellierung hat im Wesentlichen zwei unterschiedliche Zielsetzungen: i) Vorhersage der Zustandsverteilung auf Netzebene und ii) Vorhersage des Zustands jeder einzelnen Haltung. Während die Simulation auf Netzebene (i) vor allem für die Unterstützung von mittel- bis langfristigen Sanierungs- und Investitionsstrategien relevant ist, kann die Simulation auf Haltungsebene (ii) für die Festlegung von Inspektionsstrategien genutzt werden. Für beide Bereiche wurden gemeinsam mit den Berliner Wasserbetrieben jeweils sechs Bewertungsindikatoren definiert, die die relevanten Stärken und Schwächen der Modelle aufzeigen sollen. Die Bewertung basiert in allen Fällen auf dem Vergleich der Modellergebnisse (prognostizierte Zustandsklasse) mit dem bekannten Zustand der Kanäle aus den Kamerainspektionen (inspizierte Zustandsklasse).

Die *Bewertungsindikatoren auf Netzebene* quantifizieren die Abweichungen in den Häufigkeitsverteilungen von prognostizierter und inspizierter Zustandsklasse, i) für alle Haltungen unabhängig vom Alter und ii) für alle Haltungen der Altersklasse 51 bis 75 Jahre. Die Altersklasse 51 bis 75 Jahre entspricht der Lebensdauer von Beton- und Steinzeugkanälen, d.h. die Kanäle müssten bei Erreichen dieses Alters theoretisch saniert oder erneuert werden.

K1, *K2* und *K3* sind die Indikatoren für die absoluten Abweichungen der Anteile aller Haltungen im guten (*K1*), mittleren (*K2*) oder schlechten Zustand (*K3*) zwischen Prognose und Inspektion, unabhängig vom Alter der Haltung. *K4*, *K5* und *K6* sind äquivalente Indikatoren, beschränkt auf die Altersklasse 51 bis 75 Jahre. Die Indikatoren liegen im Wertebereich von -100% (sehr starke Überschätzung der Häufigkeit) bis 100% (sehr starke Unterschätzung der Häufigkeit). Je näher die Indikatoren bei 0 liegen, desto geringer die Abweichungen zwischen Inspektion und Prognose und desto besser das Modell. Formel 20 zeigt die allgemeine Berechnungsformel für die sechs Indikatoren.

$$K1 \dots 6 = \left(\frac{n_{insp\_Zustand\_i}}{n_{gesamt}} - \frac{n_{prog\_Zustand\_i}}{n_{gesamt}} \right) * 100 \qquad \text{Formel 20}$$

n_insp_Zustand_i ist die Anzahl der Haltungen, die im Zustand i inspiziert wurden. *n_prog_Zustand_i* die Anzahl der Haltungen, die im Zustand i prognostiziert wurden. *n_gesamt* ist die Anzahl aller Haltungen.

Darüber hinaus wurden die Bewertungsindikatoren *K1* bis *K6* zu dem aggregierten Indikator K_{Netz} zusammengefasst (Formel 21). Dieser Indikator entspricht der Wurzel des mittleren Fehlerquadrats (engl. *root mean square error, RMSE*) der sechs Indikatoren. Durch das Quadrieren der einzelnen Fehler werden tendenziell größere Abweichungen bei einzelnen Indikatoren stärker gewichtet als geringe Abweichungen bei allen Indikatoren.

$$K_{Netz} = \sqrt{\frac{K1^2 + K2^2 + K3^2 + K4^2 + K5^2 + K6^2}{6}} \qquad \text{Formel 21}$$

Die *Bewertungsindikatoren auf Haltungsebene* (*K7* bis *K12*) leiten sich aus der Kreuztabelle ab, die die Anzahl der Haltungen mit ihren jeweils inspizierten und prognostizieren Klassen kreuzweise gegenüberstellt. Die Indikatoren beurteilen nicht die Zustandsverteilung über das gesamte Netz sondern, ob für jede einzelne Haltung die Prognose dem Inspektionsergebnis entspricht. Ein Modell kann nämlich in der Lage sein, die Zustandsverteilung über das

gesamte Netz richtig zu simulieren, dabei jedoch die falschen Haltungen in den jeweiligen Zuständen prognostizieren, d.h. Haltungen die im schlechten Zustand inspiziert wurden, fälschlicherweise im guten Zustand prognostizieren und umgekehrt.

K7, *K8* und *K9* sind Indikatoren für die sogenannte Trefferquote des Modells bezüglich der Haltungen im guten (*K7*), mittleren (*K8*) und schlechten Zustand (*K9*). Sie bewerten, wie viele der Haltungen in einem bestimmten Zustand auch in diesem prognostiziert wurden. Die Indikatoren liegen im Wertebereich zwischen 0 und 100%. Je höher die Trefferquote, desto besser das Modell. Die Indikatoren werden wie folgt berechnet:

$$K7, K8, K9 = \frac{n_{insp\_Zustand\_i\_prog\_Zustand\_i}}{n_{insp\_Zustand\_i}} * 100 \qquad \text{Formel 22}$$

n_insp_Zustand_i_prog_Zustand_i ist die Anzahl der Haltungen, die im Zustand *i* (gut, mittel oder schlecht) inspiziert und auch im selben Zustand prognostiziert wurden.

K10 und *K11* sind Maßzahlen für den Anteil an Haltungen, die im mittleren (K10) bzw. schlechten Zustand (K11) inspiziert aber fälschlicherweise im guten Zustand prognostiziert wurden (Falsch-negativ-Fehlerquote). Diese Haltungen würde man bei einer modellgestützten Inspektionsstrategie, die auf Haltungen im schlechten Zustand abzielt, verpassen. Die Indikatoren liegen im Wertebereich zwischen 0 und 100%. Je niedriger die Falsch-negativ-Fehlerquote, desto weniger Haltungen werden durch das Modell als zu gut bewertet. Die Indikatoren werden wie folgt berechnet:

$$K10 = \frac{n_{insp\_mittel\_prog\_gut}}{n_{insp\_mittel}} * 100 \qquad \text{Formel 23}$$

$$K11 = \frac{n_{insp\_schlecht\_prog\_gut}}{n_{insp\_schlecht}} * 100 \qquad \text{Formel 24}$$

n_insp_mittel_prog_gut ist die Anzahl der Haltungen, die im mittleren Zustand inspiziert aber im guten Zustand prognostiziert wurden. Andere Formelzeichen analog.

K12 ist eine Maßzahl für den Anteil an Haltungen, die im guten Zustand inspiziert aber fälschlicherweise im schlechten Zustand prognostiziert wurden (Falsch-positiv-Fehlerquote, Formel 25). Diese Haltungen würde man bei einer modellgestützten Inspektionsstrategie, die auf Haltungen im schlechten Zustand abzielt, umsonst inspizieren. Der Indikator liegt im Wertebereich zwischen 0 und 100%. Je geringer die Falsch-positiv-Fehlerquote, desto weniger Haltungen werden durch das Modell als zu schlecht bewertet.

$$K12 = \frac{n_{insp\_gut\_prog\_schlecht}}{n_{prog\_schlecht}} * 100 \qquad \text{Formel 25}$$

Die verschiedenen Bewertungsindikatoren auf Haltungsebene wurden zu dem aggregierten Indikator $K_{Haltung}$ zusammengefasst (Formel 26).

$$K_{Haltung} = \sqrt{\frac{(100 - K7)^2 + (100 - K8)^2 + (100 - K9)^2 + K10^2 + K11^2 + K12^2}{6}} \qquad \text{Formel 26}$$

Tabelle 11 zeigt eine Übersicht der Bewertungsindikatoren auf Netz- und Haltungsebene.

Tabelle 11: Übersicht der Indikatoren auf Netz- und Haltungsebene.

Indikatoren	Erläuterung	Wertebereich / Optimum
Netzebene		
K1, K2, K3	Abweichung des Anteils an Haltungen im guten, mittleren bzw. schlechten Zustand unter Berücksichtigung aller Haltungen	Wertebereich: [-100%, 100%]; Optimum: 0%
K4, K5, K6	Abweichung des Anteils an Haltungen im guten, mittleren bw. schlechten Zustand unter Berücksichtigung der Haltungen im Alter zwischen 51 und 75 Jahren	Wertebereich: [-100%, 100%]; Optimum: 0%
K_{Netz}	Wurzel des mittleren Fehlerquadrats aus *K1* bis *K6*	Wertebereich: [0%, 100%]; Optimum: 0%
Haltungsebene		
K7, K8, K9	Anzahl korrekter Prognosen im guten, mittleren bzw. schlechten Zustand geteilt durch Anzahl Inspektionen im guten, mittleren bzw. schlechten Zustand (Trefferquote)	Wertebereich: [0%, 100%]; Optimum: 100%
K10, K11	Anzahl an Haltungen, die im mittleren bzw. schlechten Zustand inspiziert aber fälschlicherweise als gut prognostiziert wurden (Falsch-Negativ-Fehlerquote)	Wertebereich: [0%, 100%]; Optimum: 0%
K12	Anzahl an Haltungen, die im guten Zustand inspiziert, aber fälschlicherweise als schlecht prognostiziert wurden (Falsch-Positiv-Fehlerquote)	Wertebereich: [0%, 100%]; Optimum: 0%
$K_{Haltung}$	Wurzel des mittleren Fehlerquadrats aus *K7* bis *K12*	Wertebereich: [0%, 100%]; Optimum: 0%

4.2.6 Simulation der Zustandsentwicklung

Im Anschluss an die Modellbewertung wurde mit einem statistischen Modell (GompitZ) und einem Modell des maschinellen Lernens (Random Forest) beispielhaft die Entwicklung des Netzzustandes über die nächsten 50 Jahre (Zeitraum: 2017 bis 2066) prognostiziert. Mögliche Sanierungsstrategien wurden dabei nicht berücksichtigt. Für die Prognose wurden die Testdaten für die Modellbewertung verwendet, wobei mehrfach inspizierte Kanäle nur einmal berücksichtigt wurden (37.324 Datensätze).

4.3 Modellbewertung

4.3.1 GompitZ

4.3.1.1 Bildung der Kohorten und Kalibrierung der Überlebensfunktionen

Die Untersuchung verschiedener Varianten der Kohortenbildung zeigte, dass die Modellgüte von GompitZ von der Auswahl der berücksichtigten Variablen abhängt. Die geringsten Abweichungen auf Netzebene zwischen Prognose und Inspektionsergebnis werden für die Kohortenbildung mit fünf Variablen (Material, Profil, Bäume (j/n), Bodentyp, Bezirk) ohne Verwendung zusätzlicher Kovariaten für die Kalibrierung erzielt (*K_Netz* = 1,39; Mittelwert

der dreifachen Kreuzvalidierung). Dennoch lassen sich auch für einfache Modelle, bei denen die Kohorten nur entsprechend der Merkmalsausprägungen zweier Variablen (z.B. Material und Profil) gebildet wurden, relativ geringe Abweichungen zwischen Prognose und Inspektionsergebnis erreichen (*K_Netz* < 2,0; Tabelle 12). Die Verwendung von Kovariaten für die Kalibrierung führte zu keiner Verbesserung der Modellergebnisse.

Tabelle 12: Zusammengefasste Ergebnisse (*K_Netz* und *K_Haltung*) für die untersuchten Varianten der Kohortenbildung

Variablen für Kohortenbildung	Anz.Koh.	Kovariaten	K_Netz	K_Haltung
Profil + Abwassertyp	2	-	2,31	52,54
Bäume (j/n) + Abwassertyp	2	-	3,17	53,69
GW-Überdeckung (j/n) + Abwassertyp	2	-	2,76	53,57
Rückstau (j/n) + Abwassertyp	2	-	3,48	53,69
Bodentyp + Abwassertyp	2	-	3,07	54,01
Bezirk + Abwassertyp	2	-	2,30	52,62
Material + Abwassertyp	2	-	2,03	52,07
Bäume (j/n) + Profil	2	-	3,64	53,46
GW-Überdeckung (j/n) + Profil	2	-	3,93	53,41
Rückstau (j/n) + Profil	2	-	3,16	53,23
Bodentyp + Profil	2	-	3,88	53,56
Bezirk + Profil	2	-	2,17	52,88
Material + Profil	2	-	1,94	52,05
GW-Überdeckung (j/n) + Bäume (j/n)	2	-	3,94	54,79
Rückstau (j/n) + Bäume (j/n)	2	-	4,70	54,85
Bodentyp + Bäume (j/n)	2	-	4,03	54,94
Bezirk + Bäume (j/n)	2	-	2,48	53,62
Material + Bäume (j/n)	2	-	2,01	52,18
Rückstau (j/n) + GW-Überdeckung (j/n)	2	-	3,98	54,67
Bodentyp + GW-Überdeckung (j/n)	2	-	3,86	54,23
Bezirk + GW-Überdeckung (j/n)	2	-	2,62	53,35
Material + GW-Überdeckung (j/n)	2	-	2,43	52,59
Bodentyp + Rückstau (j/n)	2	-	3,83	54,45
Bezirk + Rückstau (j/n)	2	-	2,71	53,71
Material + Rückstau (j/n)	2	-	2,05	52,60
Bezirk + Bodentyp	2	-	2,09	53,61

Variablen für Kohortenbildung	Anz.Koh.	Kovariaten	K_Netz	K_Haltung
Material + Bodentyp	2	-	1,98	52,02
Material + Bezirk	2	-	1,49	51,82
Material + Profil + Bäume (j/n) + Bodentyp	4	Länge + Breite	1,46	50,52
Material + Profil + GW-Überdeckung (j/n)	3	Länge	2,25	51,92
Material + Profil + Bäume (j/n) + Bodentyp + Bezirk	***5***	***-***	***1,39***	***50,75***
Material + Profil + Bäume (j/n) + Bodentyp + Bezirk	5	Länge + Breite	1,42	49,9

Nach dem Test verschiedener Varianten wurden folgende fünf kategorische Variablen für die Kohortenbildung ausgewählt: Material, Profil, Bäume (j/n), Bodenart und Bezirk. Die Kombinationen aller Merkmale dieser fünf Variablen würden zu insgesamt 618, zum Teil sehr kleinen Kohorten führen. Um innerhalb einer Kohorte genügend Datensätze für eine zuverlässige Kalibrierung der Überlebensfunktionen zu haben, wurden Kohorten mit weniger als 100 Haltungen mit ähnlichen Kohorten zusammengelegt. Zum Beispiel, wurden alle Kanäle mit dem Profil „Kreis im Vortrieb“ unabhängig von den anderen Variablen in einer einzigen Kohorte gesammelt. Des Weiteren wurden alle Kanäle, die aus insgesamt nur wenig vertretenen Materialien, d.h. Asbestzement, Mauerwerk, Beton mit hoher Tragfähigkeit oder PVC, bestehen, unabhängig von den anderen Merkmalen in jeweils einer Kohorte gesammelt. Der finale Datensatz enthält 59 Kohorten. Eine vollständige Auflistung aller Kohorten befindet sich in Tabelle 26 in Anhang D.

Für jede dieser Kohorten wurden Überlebensfunktionen kalibriert, die sich zum Teil deutlich unterscheiden. Folgende Alterungsmuster können aus den Überlebenskurven abgeleitet werden:

- *Bezirk:* Kanäle in Reinickendorf, Treptow-Köpenick, Spandau und Marzahn-Hellersdorf zeigen eine relativ schnelle Alterung, während Kanäle in Friedrichshain-Kreuzberg, Charlottenburg-Wilmersdorf und Pankow eine relativ langsame Alterung zeigen.
- *Material:* Kanäle aus Beton zeigen eine verhältnismäßig schnelle Alterung. Dagegen altern Kanäle aus Asbestzement, Beton mit hoher Tragfähigkeit und PVC-U nur langsam. Es wird jedoch darauf hingewiesen, dass fast alle Kanäle aus diesen Materialen weniger als 50 Jahre alt sind und die angenommene Alterungsdynamik über dieses Alter hinaus auf einer Extrapolation der Überlebenskurven basiert.
- *Profil:* Im Vortrieb gebaute Kanäle mit Kreisprofil zeigen eine sehr langsame Alterung. Ähnlich wie die oben genannten Materialien ist die Vortriebsbauweise jedoch sehr jung (≤ 25 Jahre). Prognosen für ein höheres Alter basieren auf einer Extrapolation der Überlebenskurven.
- *Bäume:* Kanäle mit Bäumen in einem Abstand von 3 m zeigen eine geringfügig schnellere Alterung als Kanäle ohne Bäume in ihrer Umgebung.
- *Bodenart:* Kanäle, die von Sand umgeben sind, zeigen eine relativ schnelle Alterung.

Abbildung 48 zeigt die ermittelten Überlebenskurven für ausgewählte Kohorten mit unterschiedlicher Alterungsdynamik. Die Überlebenskurven aller Kohorten sind in Abbildung 71 in Anhang D zu sehen. Es ist zu beachten, dass die Überlebenskurven bisher für

Altersbereiche, die über dem beobachteten Alter der jeweiligen Kohorte liegen, extrapoliert werden. Daher sind bei langfristigen Prognosen erhöhte Unsicherheiten zu erwarten. Für die Simulation von langfristigen Investitionsstrategien müsste diskutiert werden, wie mit der Extrapolation in diesen Altersbereichen umzugehen ist.

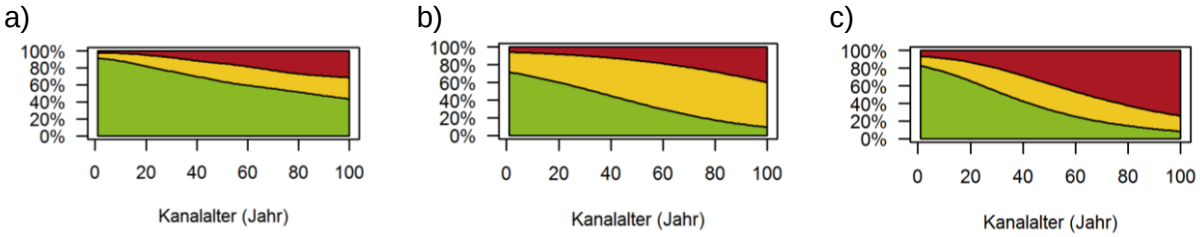


Abbildung 48: Überlebenskurven für ausgewählte Kohorten: a) Material: Beton mit hoher Tragfähigkeit; b) Bezirk: Reinickendorf, Material: Steinzeug, Profil: Kreisprofil, Bäume: Ja, Bodenart: Sand-Lehm; c) Bezirk: Tempelhof-Schöneberg, Material: Beton. Für Altersbereiche über dem beobachteten Alter der Kanäle der jeweiligen Kohorten wurden die Überlebenskurven extrapoliert.

4.3.1.2 Bewertung der Modellgüte

Netzebene: Das Modell GompitZ liefert für die Testdaten (n = 39.019) eine sehr genaue Vorhersage auf Netzebene. Die Abweichungen zwischen Prognose und Inspektionsergebnis liegen für alle drei Zustandsbereiche unter 1%. Dies gilt sowohl für die gesamte Stichprobenmenge unabhängig vom Alter ($K1$ = 0,4, $K2$ = 0,0, $K3$ = -0,4) als auch für die Haltungen in der Altersklasse 51 bis 75 a ($K4$ = -0,4, $K5$ = -0,3, $K6$ = 0,7). Damit ist das Modell in der Lage, die beobachtete Zustandsverteilung nahezu exakt wiederzugeben. Die berechneten Abweichungen sind in Abbildung 49 dargestellt, die Zahlenwerte der Bewertungsindikatoren sind in Tabelle 14 zusammengefasst.

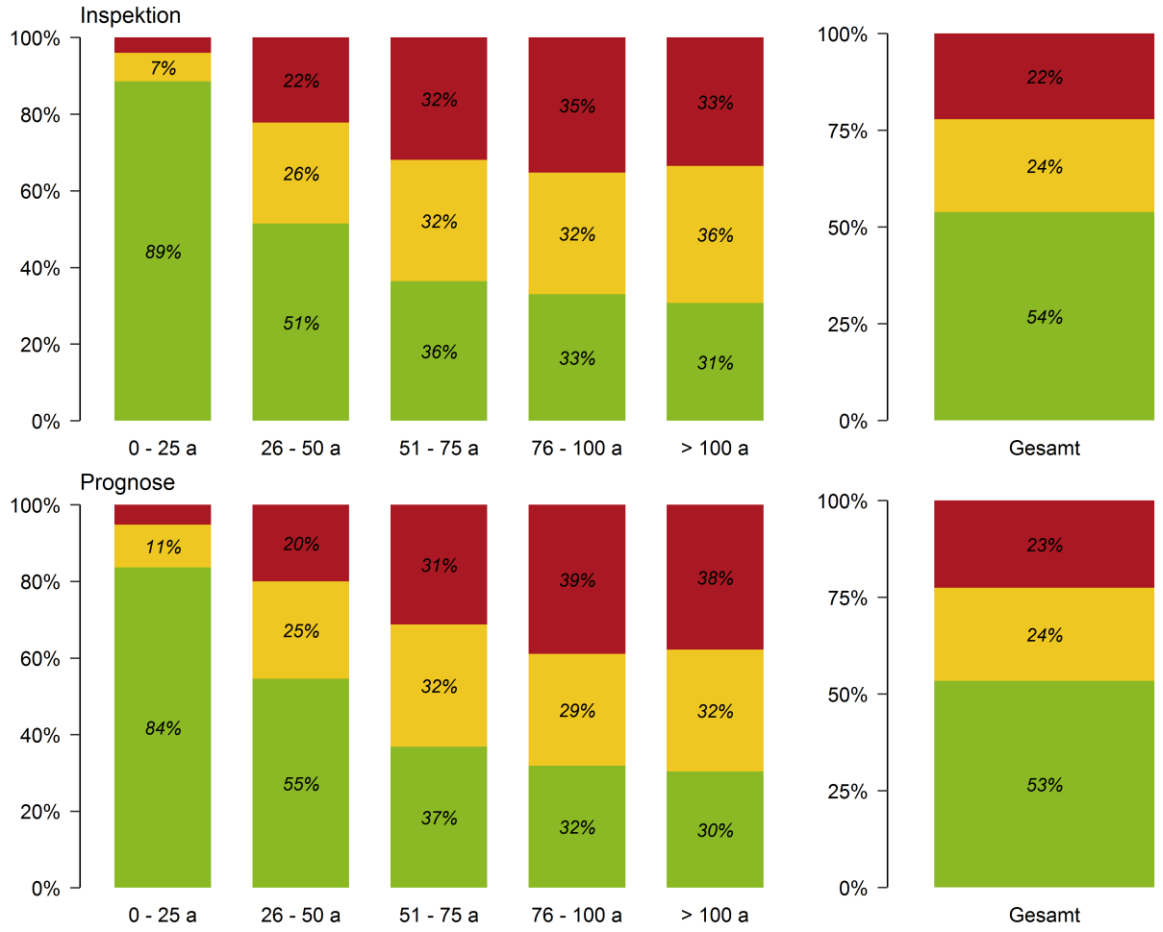


Abbildung 49: Abweichungen auf Netzebene zwischen Inspektion (oben) und Prognose mit GompitZ (unten). Die drei Farben repräsentieren die drei aggregierten Zustandsbereiche „gut“ (grün; Zustandsklasse 4, 5 und 6), „mittel“ (gelb; Zustandsklasse 3) und „schlecht“ (rot; Zustandsklasse 1 und 2)

Haltungsebene: Auf Haltungsebene liefert das Modell deutlich schlechtere Ergebnisse. Die Trefferquoten sind insbesondere für die Haltungen im mittleren (*K8*) und schlechten (*K9*) Zustand sehr niedrig und im Bereich eines sogenannten Zufallsmodells, das für jeden der drei Zustandsbereiche eine Trefferquote von 33% liefen würde. Lediglich für den guten Zustand (*K7*) werden etwa zwei von drei Haltungen richtig bewertet (*K7 = 64%*). Die Fehlerquoten, sowohl bei der Überschätzung (Falsch-Negativ-Raten, *K10* und *K11*) als auch bei der Unterschätzung des Zustands (Falsch-Positiv-Rate, *K12*), liegen im Bereich von 40%. Für die Identifizierung prioritärer Haltungen, z.B. für Inspektionen oder Sanierungsprogramme, ist das Modell daher ungeeignet. Die Kreuztabelle für Prognose und Inspektionsergebnisse ist in Tabelle 13 dargestellt. Tabelle 14 zeigt die Bewertungsindikatoren auf Haltungsebene.

Tabelle 13: Kreuztabelle mit den Häufigkeiten der drei Zustandsbereiche für Inspektion (Zeilen) und Prognose mit GompitZ (Spalten)

		Prognose			
		Gut	Mittel	Schlecht	Summe Inspektionen
Inspektion	Gut	12616	3837	3134	19587
	Mittel	3760	2585	2406	8751
	Schlecht	3062	2324	2662	8048
Summe Prognose		19438	8746	8202	

Tabelle 14: Bewertungsindikatoren für GompitZ

Indikatoren auf Netzebene						
K1 (Ziel → ±0)	**K2** (Ziel → ±0)	**K3** (Ziel → ±0)	**K4** (Ziel → ±0)	**K5** (Ziel → ±0)	**K6** (Ziel → ±0)	**K_Netz** (Ziel → min)
0,4	0,0	-0,4	-0,4	-0,3	0,7	0,4
Indikatoren auf Haltungsebene						
K7 (Ziel → max)	**K8** (Ziel → max)	**K9** (Ziel → max)	**K10** (Ziel → min)	**K11** (Ziel → min)	**K12** (Ziel → min)	**K_Haltung** (Ziel → min)
64,4	29,5	33,1	43,0	38,1	38,2	50,8

4.3.1.3 Einfluss der Datenmenge auf Modellgüte

Im Rahmen einer Monte-Carlo-Simulation wurde der Einfluss der für die Kalibrierung zur Verfügung stehenden Datenmenge auf die Modellgüte untersucht. Die Untersuchung wurde für ein vereinfachtes Modell mit 12 Kohorten für die 12 Berliner Stadtbezirke durchgeführt. Die Untersuchung konnte nicht für das Modell mit 59 Kohorten durchgeführt werden, weil für geringe Datenumfänge in einigen Kohorten zu wenige Haltungen vertreten wären, um bei der Kalibrierung eine Konvergenz zu erreichen.

Die Untersuchung unterschiedlicher Datenumfänge für die Modellkalibrierung hat gezeigt, dass sich für die Bewertungsindikatoren auf Netzebene (*K1* bis *K6,* siehe Abbildung 50) bereits für sehr wenige Stichproben (ab $n = 3000$) ein gutes Ergebnis erzielen lässt. Die Mittelwerte der Bewertungsindikatoren für die 50 Wiederholungen von Training und Test liegen bei 0%, die Schwankungsbreite beträgt etwa +/-2%. Durch eine Erhöhung der Stichprobenumfänge kann die Schwankungsbreite der Modellgüte weiter reduziert werden. Auch für noch niedrigere Stichprobenumfänge, z.B. $n = 400$, geben die Modelle im Mittel eine gute Vorhersage, allerdings ist die Schwankungsbreite der Ergebnisse mit +/-10% sehr hoch.

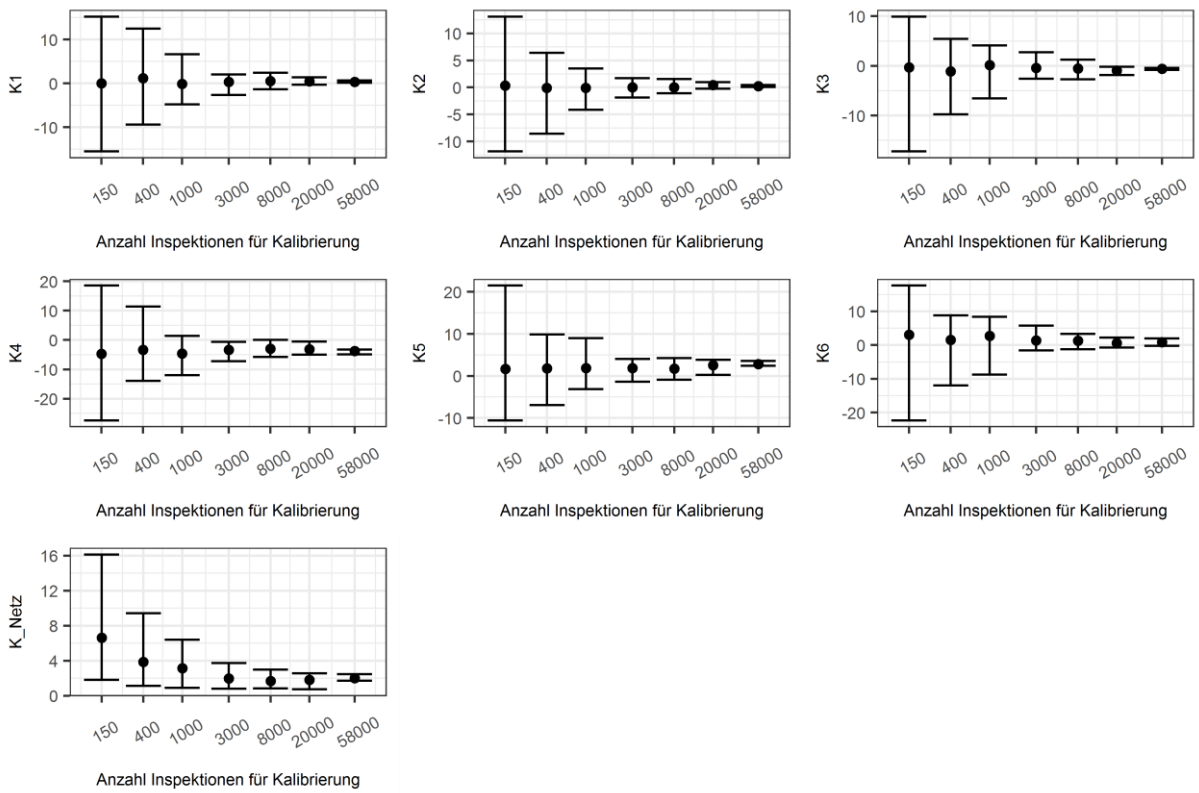


Abbildung 50: Einfluss der für die Kalibrierung verwendeten Datenmenge auf die Modellgüte von GompitZ auf Netzebene (Indikatoren *K1* bis *K6* und *K_Netz*). Dargestellt sind die Bewertungsindikatoren mit Mittelwert, Minimum und Maximum der 50 Wiederholungen je Stichprobenumfang.

Der Einfluss der Datenmenge auf die Vorhersage auf Haltungsebene (*K7* bis *K12,* siehe Abbildung 51) ist ähnlich wie auf Netzebene. Die Mittelwerte der Bewertungsindikatoren für die 50 verschiedenen Modellläufe sind weitestgehend unabhängig vom Stichprobenumfang. Allerdings ist die Schwankungsbreite der Ergebnisse mit über +/-4% für Stichprobenumfänge bis *n* = 3000 verhältnismäßig hoch.

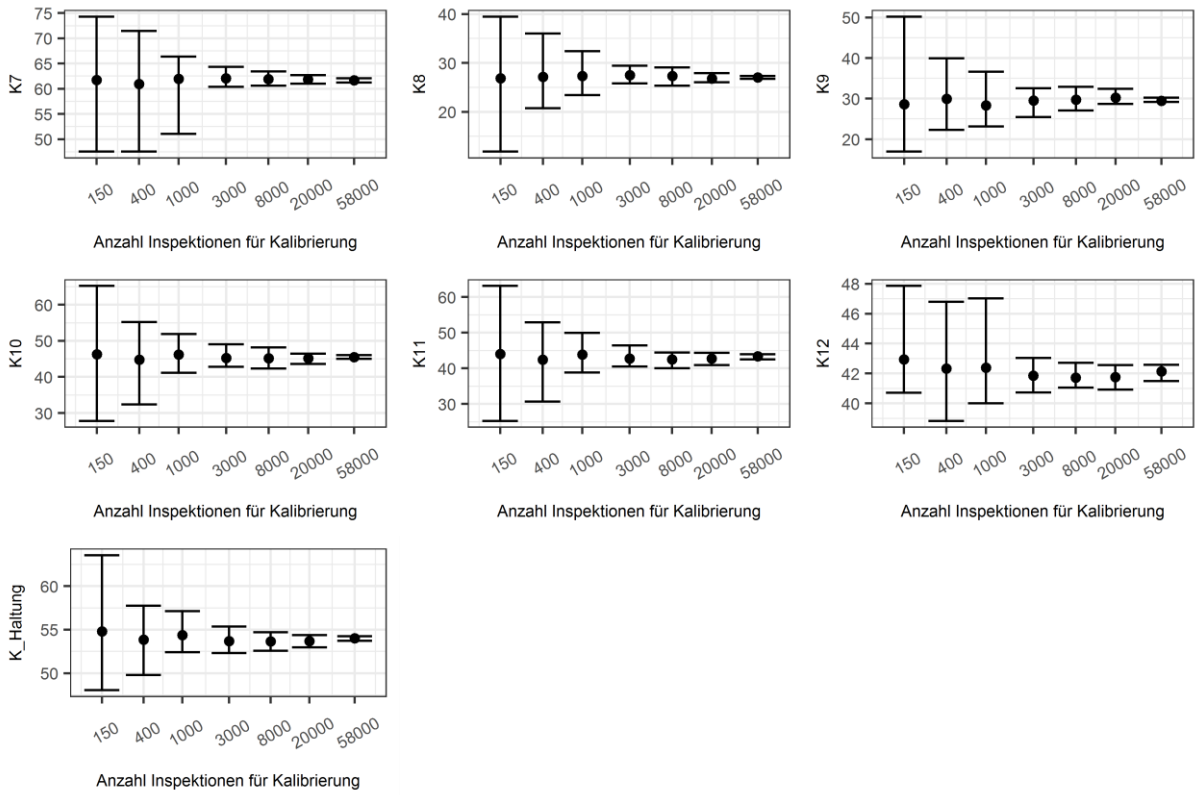


Abbildung 51: Einfluss der für die Kalibrierung verwendeten Datenmenge auf die Modellgüte von GompitZ auf Haltungsebene (Indikatoren *K7* bis *K12* und *K_Haltung*). Dargestellt sind die Bewertungsindikatoren mit Mittelwert, Minimum und Maximum der 50 Wiederholungen je Stichprobenumfang.

4.3.2 Random Forest

4.3.2.1 Einfluss der Modellparameter

Die Bestimmung der Modellparameter für Random Forest erfolgte wie in Kap. 4.2.3 beschrieben in zwei Schritten. Die zufällige Variation aller Modellparameter innerhalb festgelegter Grenzen (Schritt 1, siehe Kap. 4.2.3) zeigte, dass die Gewichtungsfaktoren *w1* und *w2* einen starken Einfluss auf die Modellgüte haben. Trotz der gleichzeitigen Variation der anderen Parameter (*mtry, nodesize, ntrees*) zeigten sich klare Optima von *w1* = 2,0 und *w2* = 1,0 für die Vorhersage auf Netzebene (Abbildung 52, links) und *w1* = 1,0 und *w2* = 0,8 für die Vorhersage auf Haltungsebene (Abbildung 52, Mitte und rechts). Zwar werden die niedrigsten Werte für *K_Haltung* eigentlich bei *w2* = 1,0 beobachtet. Durch die leichte Reduktion auf *w2* = 0,8 ergibt sich jedoch eine deutlich verbesserte Trefferquote für Kanäle im schlechten Zustand (*K9,* Abbildung 52, unten rechts), der wichtigste Indikator für die Priorisierung von Haltungen für Inspektionen. Die Untersuchung hat weiter gezeigt, dass die Anzahl der Entscheidungsbäume (Parameter *ntrees*) keinen deutlichen Einfluss auf die Modellgüte hat. Der Parameter wurde für weitere Analysen auf einen Wert von 100 festgesetzt. Die Abbildungen zu allen untersuchten Parametern und Güteindikatoren sind in Anhang E (Abbildung 72 bis Abbildung 76) zu finden.

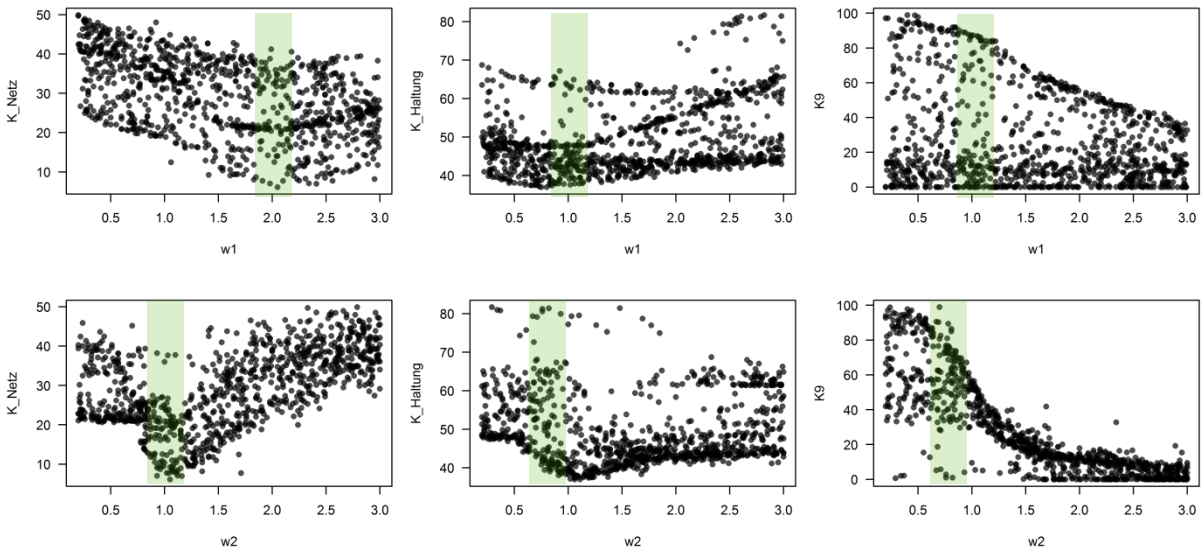


Abbildung 52: Einfluss des Gewichtungsfaktors *w1* (oben) und *w2* (unten) auf ausgewählte Güteindikatoren für die Vorhersage mit Random Forest auf Netzebene (*K_Netz*, links) und Haltungsebene (*K_Haltung* und *K9*, Mitte und rechts). Optimale Parameterbereiche sind farblich hervorgehoben.

Mit den wie oben beschriebenen fixierten Werten für die Parameter *w1*, *w2* und *ntrees* wurde systematisch das Optimum für die übrigen Parameter *mtry* und *nodesize* gesucht (Schritt 2: Gitter-Suche, siehe Kap. 4.2.3). Für die Vorhersage auf Haltungsebene werden die tendenziell besten Ergebnisse für große *mtry*-Werte und kleine *nodesize*-Werte erzielt (Abbildung 53). Für die Vorhersage auf Haltungsebene werden die besten Ergebnisse ebenfalls für große *mtry*-Werte erzielt, für *nodesize* sind Werte zwischen 20 und 90 optimal (gemessen an *K_Haltung*, Abbildung 54). Für weitere Untersuchungen wurde der Parameter *mtry* auf Werte von 10 bzw. 11 und der Parameter *nodesize* auf Werte von 7 bzw. 55 (jeweils auf Netz- bzw. Haltungsebene) festgelegt.

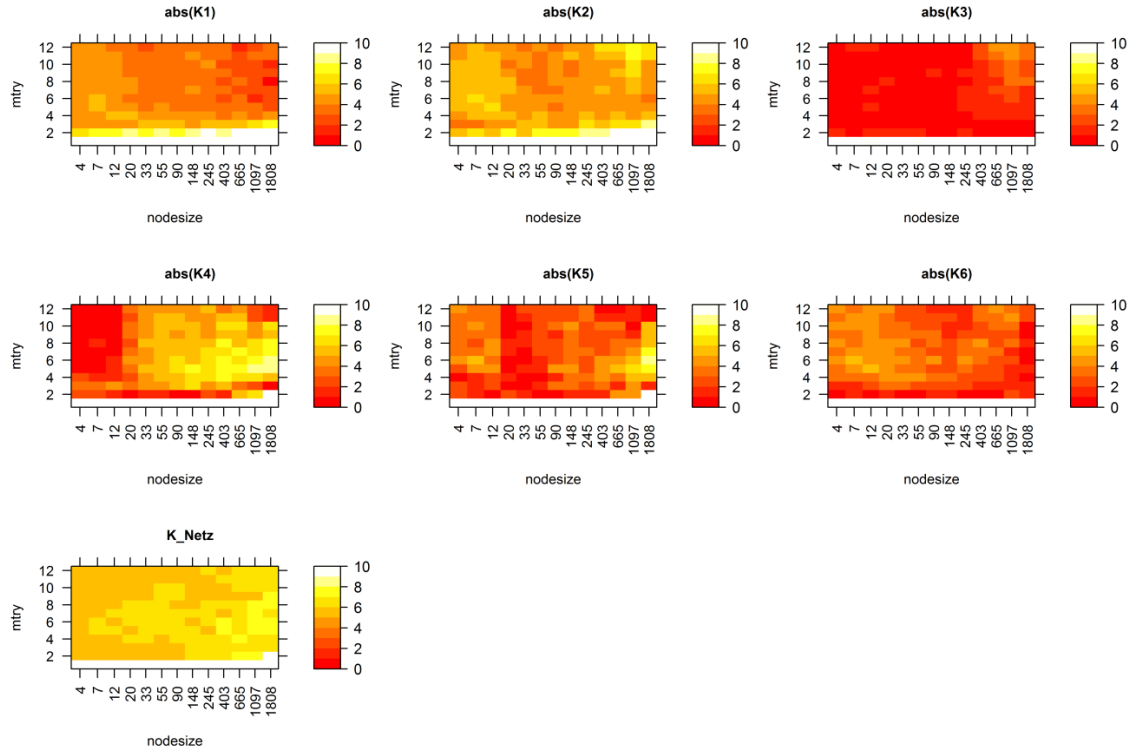


Abbildung 53: Einfluss der Parameter *nodesize* und *mtry* auf die Modellgüte von Random Forest auf Netzebene (Indikatoren *K1* bis *K6* und *K_Netz*). Fixierte Parameter: *w1* = 2,0; *w2* = 1,0; ntrees = 100.

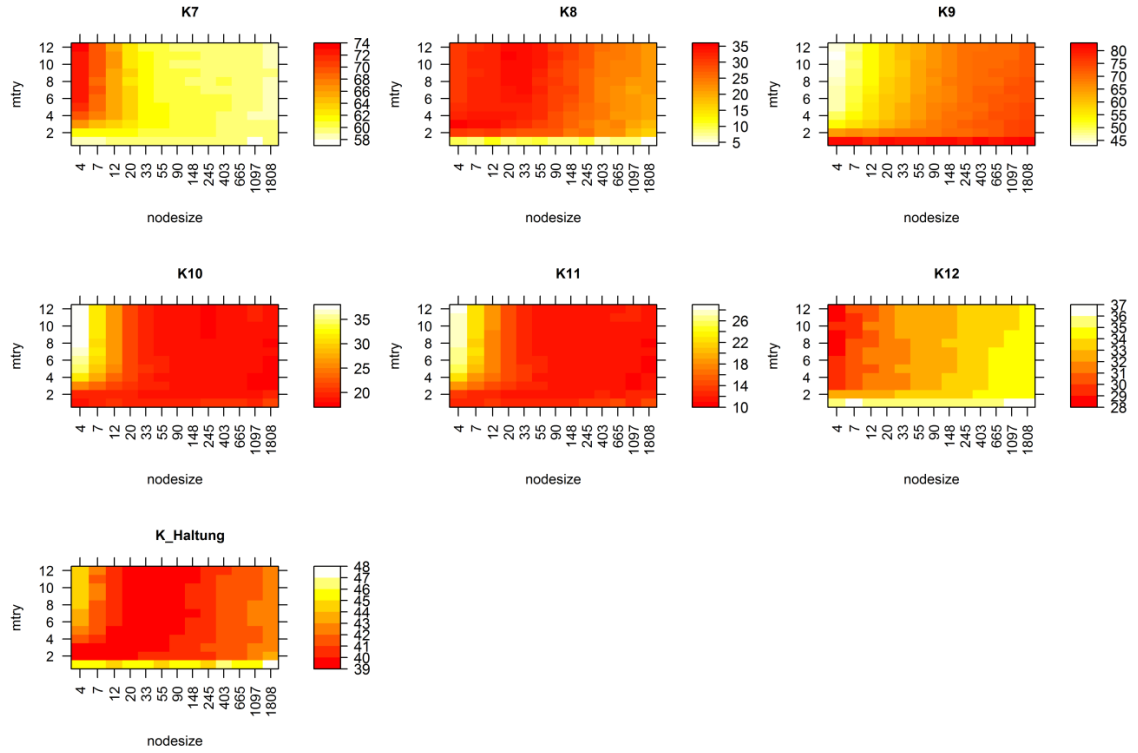


Abbildung 54: Einfluss der Parameter *nodesize* und *mtry* auf die Modellgüte von Random Forest auf Haltungsebene (Indikatoren *K7* bis *K12* und *K_Haltung*) Fixierte Parameter: *w1* = 0,8; *w2* = 1,0; *ntrees* = 100.

Aufgrund der unterschiedlichen Optima ergeben sich für die Vorhersage auf Netz- und auf Haltungsebene unterschiedliche Modellparametrisierungen, die in Tabelle 15 zusammengefasst sind. Das für die Vorhersage auf Netzebene parametrisierte Modell wird im Folgenden als Modell A (oder RF_A), das für die Vorhersage auf Haltungsebene parametrisierte Modell als Modell B (oder RF_B) bezeichnet.

Tabelle 15: Modellparameter für Modell A (für Simulation auf Netzebene) und Modell B (für Simulation auf Haltungsebene) basierend auf Random Forest

Modell-parameter	Modell A	Modell B
ntrees	100	100
nodesize	7	55
mtry	10	11
w1	2,0	1,0
w2	1,0	0,8
w3	1,0	1,0

Die Parameter wurden mit besonderem Augenmerk auf die aggregierten Indikatoren *K_Netz* und *K_Haltung* sowie die Trefferquote für die Haltungen im schlechten Zustand *K9* gewählt. Die Untersuchungen zeigen jedoch, dass je nach Zielstellung unterschiedliche Parameterwerte gewählt werden sollten. Wenn z.B. die Priorität der Vorhersage auf einer geringen Falsch-Positiv-Fehlerquote (Indikator *K12*) liegen soll, d.h. möglichst wenige Kanäle sollen als schlechter prognostiziert werden als sie eigentlich sind, empfehlen sich *nodesize*-Werte von 4. Abbildung 53 und Abbildung 54 zeigen den Einfluss der Parameterwerte auf die verschiedenen Güteindikatoren.

4.3.2.2 Bewertung der Modellgüte

Netzebene (Modell A): Die allgemeinen Abweichungen für alle drei Zustandsbereiche liegen unter 4% (Indikatoren *K1*, *K2* und *K3*). Auch für die Altersklasse 51 bis 75 Jahre (*K4*, *K5* und *K6*) werden ähnlich gute Ergebnisse erzielt. Damit ist das Modell in der Lage, die Zustandsverteilung im Kanalnetz prinzipiell richtig abzubilden. Der Indikator *K_Netz*, der die sechs Indikatoren auf Netzebene zusammenfasst, beträgt 2,3%. Die berechneten Abweichungen sind in Abbildung 55 dargestellt, die Zahlenwerte der Bewertungsindikatoren sind in Tabelle 17 zusammengefasst.

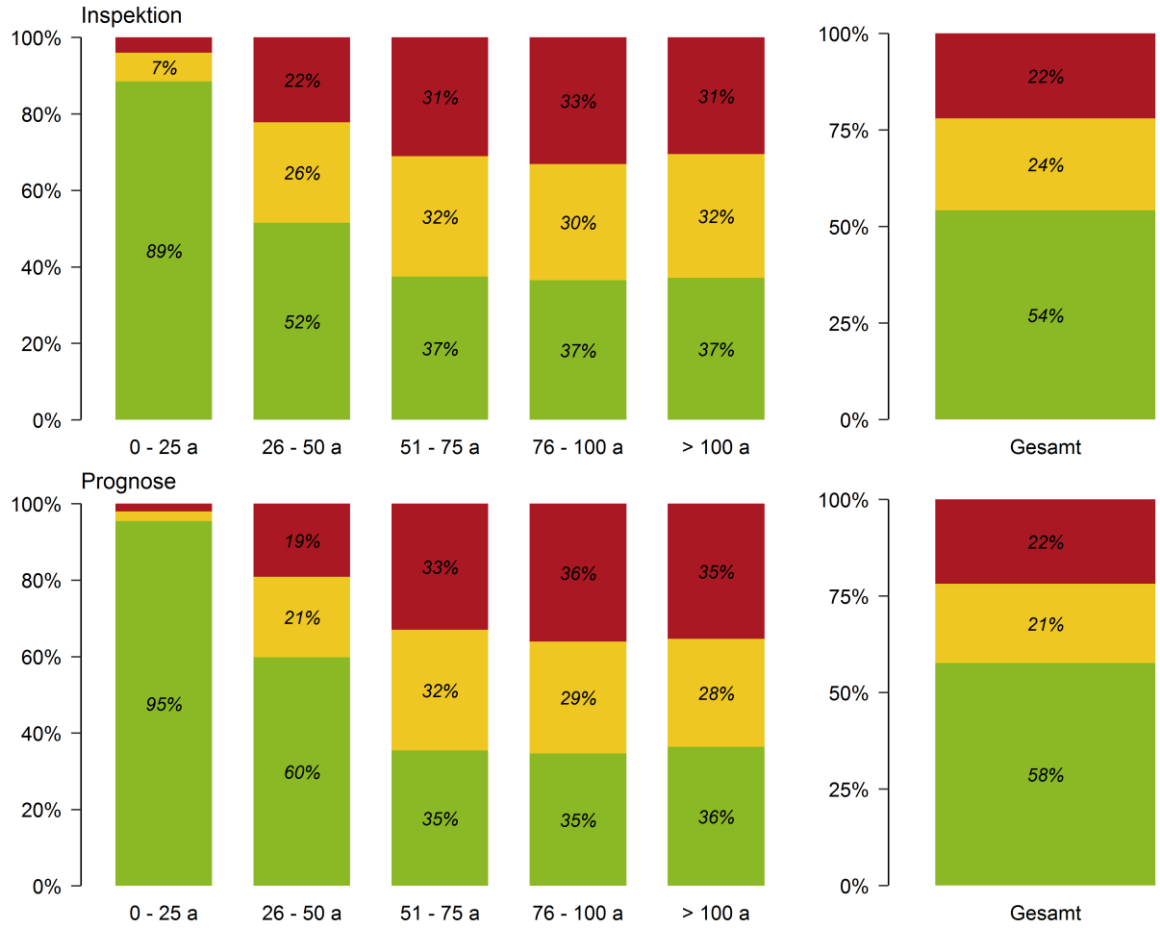


Abbildung 55: Abweichungen auf Netzebene zwischen Inspektion (oben) und Prognose mit Random Forest (Modell A, unten). Die drei Farben repräsentieren die drei aggregierten Zustandsbereiche „gut“ (grün; Zustandsklasse 4, 5 und 6), „mittel“ (gelb; Zustandsklasse 3) und „schlecht“ (rot; Zustandsklasse 1 und 2)

Haltungsebene (Modell B): Auch auf Haltungsebene liefert das Modell gute Ergebnisse. Die Trefferquoten für den guten (*K7*) und schlechten Zustand (*K9*) liegen bei 64% bzw. 67%, d.h. zwei von drei Kanälen im guten oder schlechten Zustand werden von dem Modell richtig prognostiziert. Vor dem Hintergrund der Unsicherheiten bei der Zustandsbewertung durch Kamerainspektionen, die im Bereich von 70 bis 80% liegen (Kap. 3.4), sind diese Trefferquoten als hoch zu bewerten. Die Trefferquote für den mittleren Zustand liegt zwar nur bei 40% (*K8*), sie ist aber dennoch höher als die eines Zufallsmodells (Wahrscheinlichkeit je Zustand: 33%). Die Falsch-Negativ-Rate für Haltungen im mittleren (*K10*) und schlechten Zustand (*K11*) können als niedrig bezeichnet werden. Nur 17% der Haltungen im mittleren und 10% der Haltungen im schlechten Zustand werden durch das Modell fälschlicherweise als gut prognostiziert. Umgekehrt sind 28% der Haltungen, die im schlechten Zustand prognostiziert wurden, eigentlich in gutem Zustand (*K12,* Falsch-Positiv-Rate). Der Indikator *K_Haltung*, der die sechs Indikatoren auf Haltungsebene zusammenfasst, liegt bei 34,6%. Die Kreuztabelle für Prognose und Inspektionsergebnisse ist in Tabelle 16 dargestellt. Tabelle 17 zeigt die Bewertungsindikatoren auf Haltungsebene.

Tabelle 16: Kreuztabelle mit den Häufigkeiten der drei Zustandsbereiche für Inspektion (Zeilen) und Prognose (Spalten) mit Random Forest (Modell B)

		Prognose			
		Gut	Mittel	Schlecht	Summe Inspektionen
Inspektion	Gut	13525	3771	3832	21128
	Mittel	1593	3728	4000	9321
	Schlecht	817	2040	5713	8570
Summe Prognose		15935	9539	13545	

Tabelle 17: Bewertungsindikatoren für Random Forest

Indikatoren auf Netzebene (für Modell A)						
K1 (Ziel → ±0)	**K2** (Ziel → ±0)	**K3** (Ziel → ±0)	**K4** (Ziel → ±0)	**K5** (Ziel → ±0)	**K6** (Ziel → ±0)	**K_Netz** (Ziel → min)
-3,5	3,4	0,1	2,0	-0,1	-1,9	2,3
Indikatoren auf Haltungsebene (für Modell (B)						
K7 (Ziel → max)	**K8** (Ziel → max)	**K9** (Ziel → max)	**K10** (Ziel → min)	**K11** (Ziel → min)	**K12** (Ziel → min)	**K_Haltung** (Ziel → min)
64,0	40,0	66,7	17,1	9,5	28,3	34,6

Die Untersuchungen haben gezeigt, dass es für beide Ebenen (Netz- und Haltungsebene) geeignete Random-Forest-Modelle gibt. Allerdings liefert das Modell, das für die Vorhersage auf Netzebene optimiert wurde (Modell A), eine relativ ungenaue Vorhersage auf Haltungsebene (*K_Haltung*: 41,2). Umgekehrt bildet das für die Vorhersage auf Haltungsebene optimierte Modell B die Zustandsverteilung im Netz deutlich schlechter ab (*K_Netz*: 15,7) als das Modell A.

4.3.2.3 Einfluss der Datenmenge auf Modellgüte

Die Untersuchung unterschiedlicher Datenumfänge für das Modelltraining hat gezeigt, dass sich für die Bewertungsindikatoren auf Netzebene (*K1* bis *K6* und *K_Netz*) bereits für wenige Stichproben (ab $n = 3000$) ein relativ gutes Ergebnis erzielen lässt. Allerdings beträgt die Schwankungsbreite für die 50 Wiederholungen von Training und Test für $n = 3000$ je nach Indikator noch zwischen +/-4% und +/-8%. Das heißt, je nachdem welche 3000 Datensätze für das Training verwendet werden, kann sich die Vorhersagequalität deutlich unterscheiden. Durch eine Erhöhung der Stichprobenumfänge kann die Schwankungsbreite der Modellgüte und damit die Verlässlichkeit des Modells stetig reduziert werden (Abbildung 56).

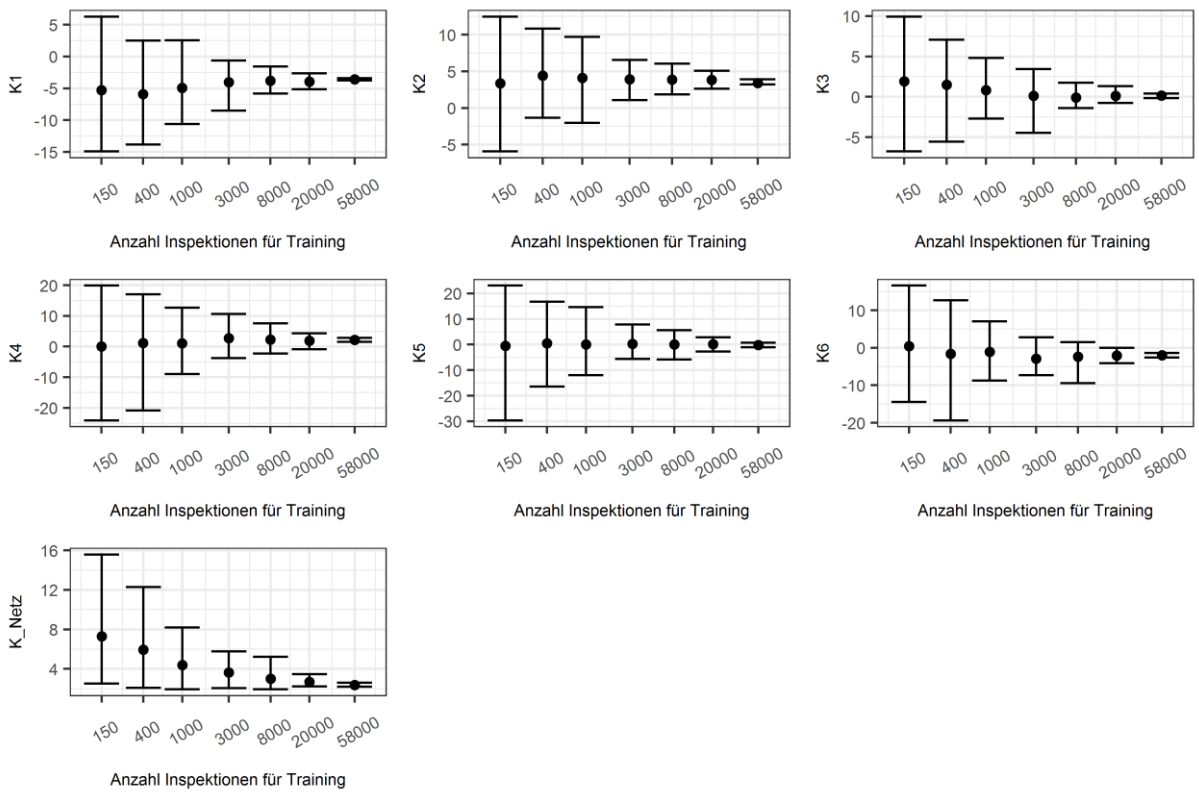


Abbildung 56: Einfluss der für das Training verwendeten Datenmenge auf die Modellgüte von Random Forest (Modell A) auf Netzebene (Indikatoren *K1* bis *K6* und *K_Netz*). Dargestellt sind die Bewertungsindikatoren mit Mittelwert, Minimum und Maximum der 50 Wiederholungen je Stichprobenumfang.

Für die Vorhersage auf Haltungsebene (*K7* bis *K12* und *K_Haltung*) verbessert sich das Ergebnis ebenfalls mit zunehmendem Datenumfang, d.h. die Trefferquoten erhöhen sich und die Fehlerquoten verringern sich, ohne dass ein Grenzwert erreicht wird (Abbildung 57). Sowohl der Mittelwert der Vorhersagen als auch die Schwankungsbreiten verbessern sich mit zunehmendem Stichprobenumfang deutlich. Das heißt, dass das Modell immer weiter trainiert werden kann und sich die Vorhersagequalität durch Berücksichtigung weiterer Inspektionsergebnisse wahrscheinlich verbessern würde. Dies ist ein wesentlicher Unterschied zum statistischen Modell GompitZ, bei dem sich die Vorhersagequalität für mehr als 3000 Datensätze kaum verbessert.

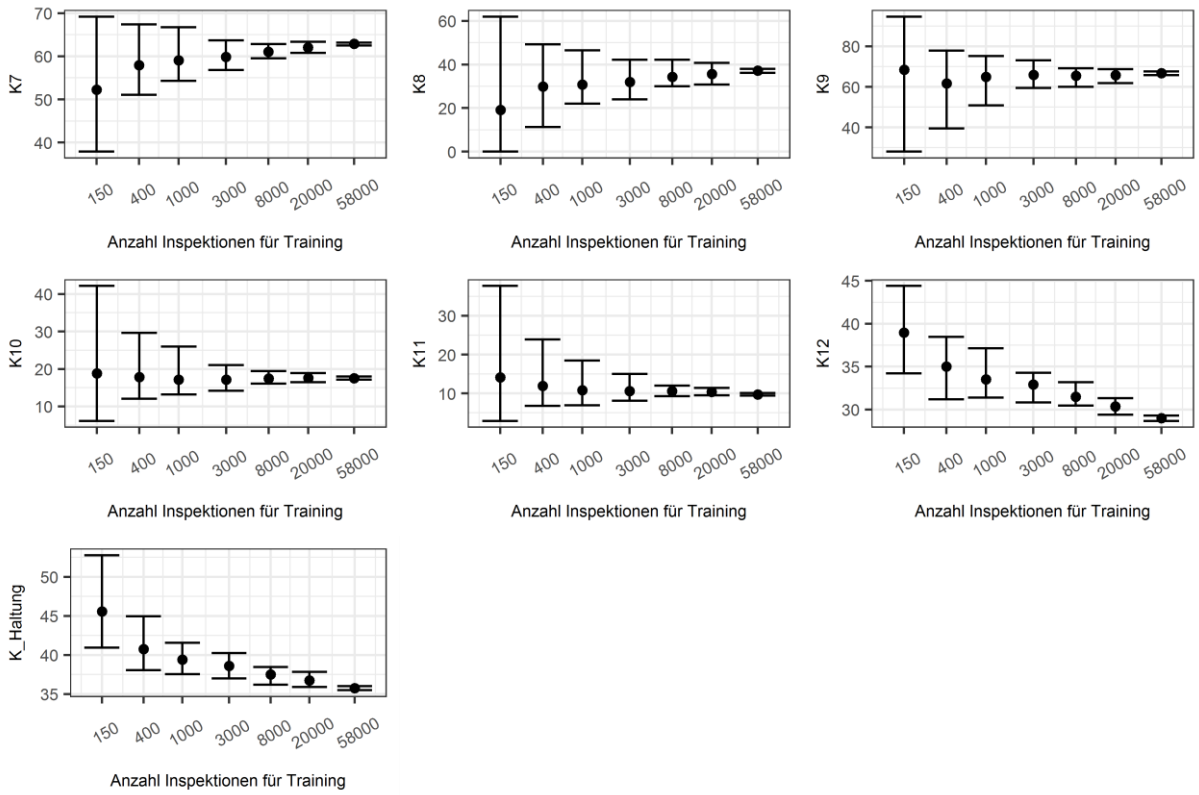


Abbildung 57: Einfluss der für das Training verwendeten Datenmenge auf die Modellgüte von Random Forest (Modell B) auf Haltungsebene (Indikatoren *K7* bis *K12* und *K_Haltung*). Dargestellt sind die Bewertungsindikatoren mit Mittelwert, Minimum und Maximum der 50 Wiederholungen je Stichprobenumfang.

4.3.3 Support Vector Machine

4.3.3.1 Einfluss der Modellparameter

Die Bestimmung der Modellparameter für Support Vector Machine erfolgte ebenfalls wie in Kap. 4.2.3 beschrieben in zwei Schritten. Die zufällige Variation aller Modellparameter innerhalb festgelegter Grenzen (Schritt 1, siehe Kap. 4.2.3) zeigte, dass die Gewichtungsfaktoren *w2* und *w3* einen starken Einfluss auf die Modellgüte haben. Trotz der gleichzeitigen Variation der anderen Parameter (*C* und *sigma*) zeigten sich deutliche Optima von *w2* = 1,4 und *w3* = 1,8 für die Vorhersage auf Netz- und Haltungsebene (Abbildung 58). Die gewählten Werte stellen eine Kompromisslösung dar, da sich die verschiedenen Indikatoren bezüglich der Parameterwirkung zum Teil gegensätzlich verhalten. Die Abbildungen zu allen untersuchten Parametern und Güteindikatoren sind in Anhang E (Abbildung 77 bis Abbildung 80) zu finden.

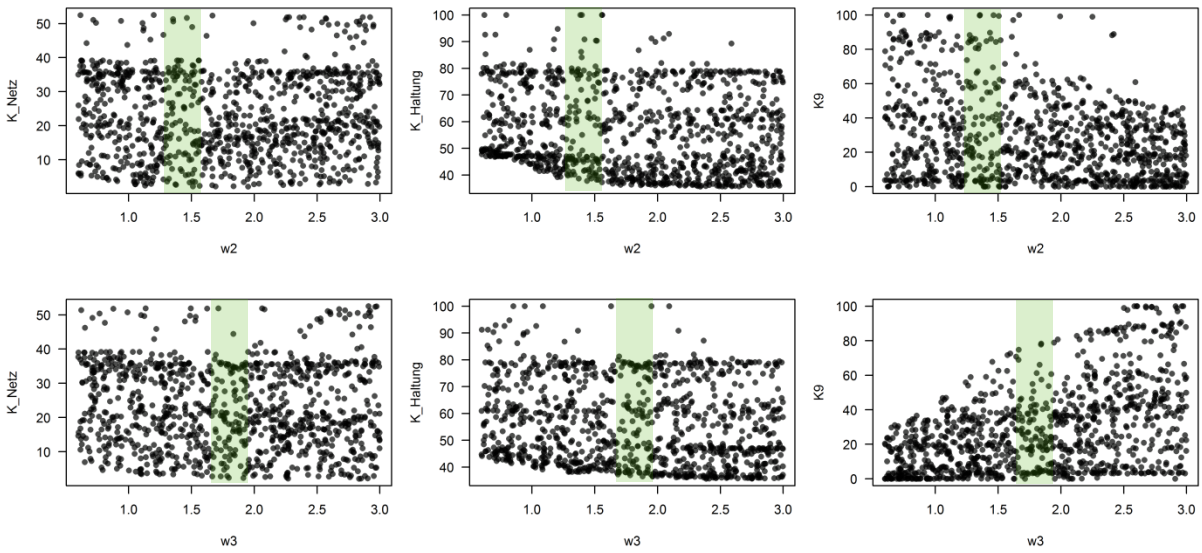


Abbildung 58: Einfluss des Gewichtungsfaktors *w2* (oben) und *w3* (unten) auf ausgewählte Güteindikatoren für die Vorhersage auf Netz- und Haltungsebene (*K_Netz, K_Haltung* und *K9*). Optimale Parameterbereiche sind farblich hervorgehoben.

Mit den wie oben beschriebenen fixierten Werten für die Parameter *w2*, *w3* wurde systematisch das Optimum für die übrigen Parameter *C* und *sigma* gesucht (Schritt 2: Gitter-Suche, siehe Kap. 4.2.3). Für die Vorhersage auf Haltungsebene werden die tendenziell besten Ergebnisse für *sigma*-Werte zwischen 0,1 und 1,0 erzielt (Abbildung 59). Für die Vorhersage auf Haltungsebene werden die besten Ergebnisse für *sigma*-Werte zwischen 0,01 und 0,1 erzielt (Abbildung 60). Der Parameter *C* hat nur einen nur relativ geringen Einfluss auf die Güteindikatoren. Es wurde der häufigste Parameterwert unter den besten zehn Ergebnissen ausgewählt. Für weitere Untersuchungen wurde der Parameter *sigma* auf Werte von 0,4 bzw. 0,07 und der Parameter *C* auf Werte von 400 bzw. 10 (jeweils auf Netz- bzw. Haltungsebene) festgelegt.

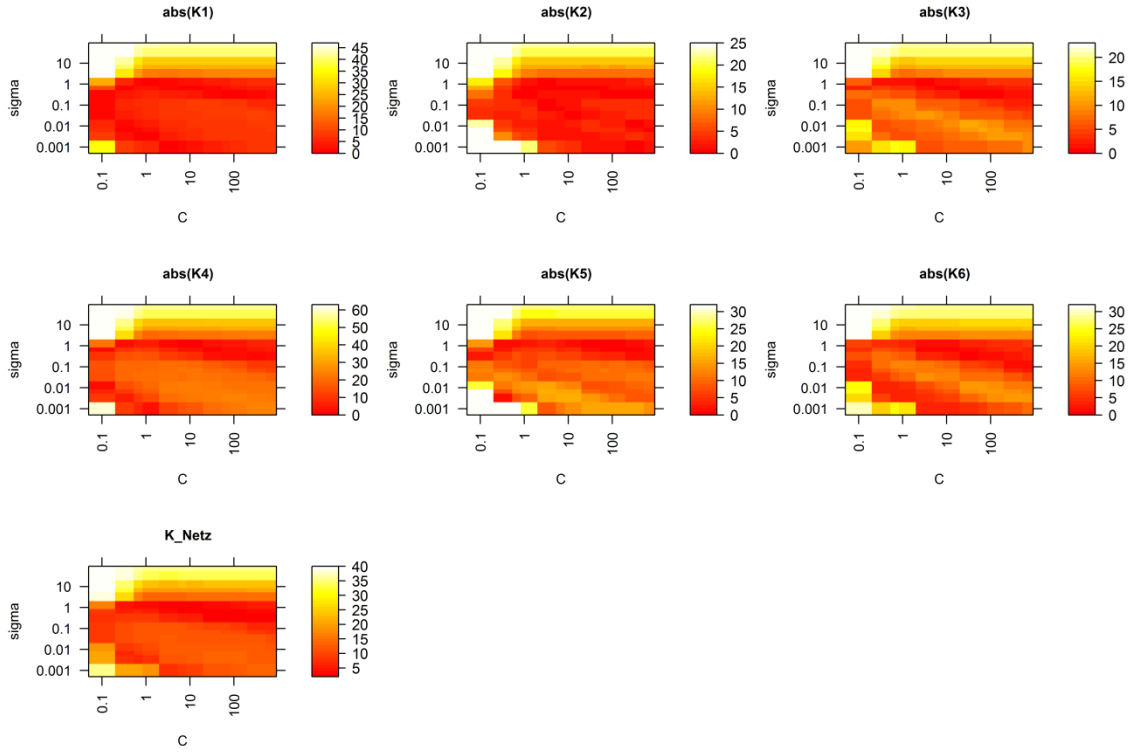


Abbildung 59: Einfluss der Parameter *C* und *sigma* auf die Modellgüte von Support Vector Machine auf Netzebene (Indikatoren *K1* bis *K6* und *K_Netz*) . Fixierte Parameter: für *w2* = 1,4; *w3* = 1,8.

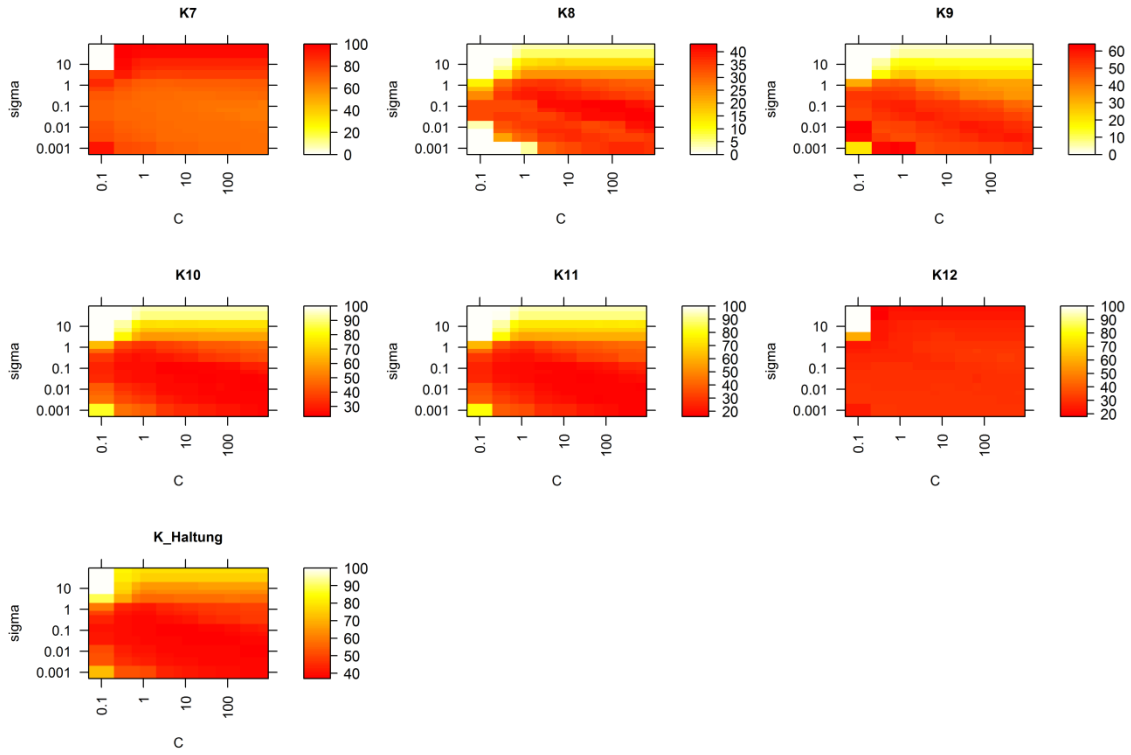


Abbildung 60: Einfluss der Parameter *C* und *sigma* auf die Modellgüte von Support Vector Machine auf Haltungsebene (Indikatoren *K7* bis *K12* und *K_Haltung*). Fixierte Parameter: für *w2* = 1,4; *w3* = 1,8.

Aufgrund der unterschiedlichen Optima von *C* und *sigma* ergeben sich für die Vorhersage auf Netz- und auf Haltungsebene unterschiedliche Modellparametrisierungen, die in Tabelle 18 zusammengefasst sind. Das für die Vorhersage auf Netzebene parametrisierte Modell wird im Folgenden als Modell A (oder SVM_A), das für die Vorhersage auf Haltungsebene parametrisierte Modell als Modell B (oder SVM_B) bezeichnet.

Tabelle 18: Modellparameter für Modell A (für Simulation auf Netzebene) und Modell B (für Simulation auf Haltungsebene) basierend auf Support Vector Machine

Modell-parameter	Modell A	Modell B
C	400	10
sigma	0,4	0,07
w1	1,0	1,0
w2	1,4	1,4
w3	1,8	1,8

4.3.3.2 Bewertung der Modellgüte

Netzebene (Modell A): Die allgemeinen Abweichungen für alle drei Zustandsbereiche liegen bei unter 1% (Indikatoren *K1*, *K2* und *K3*). Für die Altersklasse 51 bis 75 Jahre (*K4*, *K5* und *K6*) liegen die Abweichungen bei maximal 3%. Damit ist das Modell in der Lage, die Zustandsverteilung im Kanalnetz prinzipiell richtig abzubilden. Der Indikator *K_Netz*, der die sechs Indikatoren auf Netzebene zusammenfasst, beträgt 1,6%. Die berechneten Abweichungen sind in Abbildung 61 dargestellt, die Zahlenwerte der Bewertungsindikatoren sind in Tabelle 20 zusammengefasst.

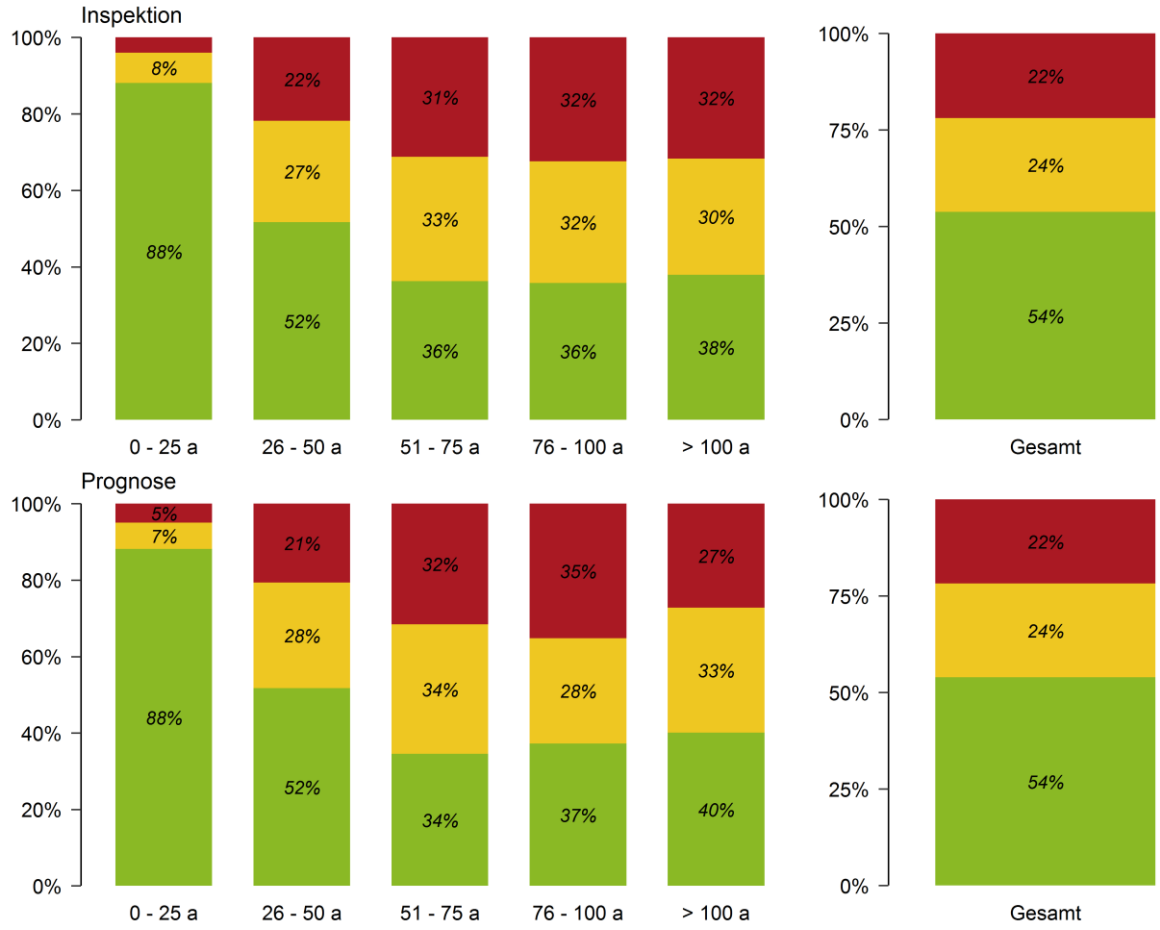


Abbildung 61: Abweichungen auf Netzebene zwischen Inspektion (oben) und Prognose mit Support Vector Machine (Modell A, unten). Die drei Farben repräsentieren die drei aggregierten Zustandsbereiche „gut“ (grün; Zustandsklasse 4, 5 und 6), „mittel“ (gelb; Zustandsklasse 3) und „schlecht“ (rot; Zustandsklasse 1 und 2)

Haltungsebene (Modell B): Die Trefferquoten für den guten (*K7*) und schlechten Zustand (*K9*) liegen bei 66% und 56%, d.h. etwa zwei Drittel der Kanäle im guten und etwas mehr als die Hälfte der Kanäle im schlechten Zustand werden von dem Modell richtig prognostiziert. Für die Kanäle im mittleren Zustand beträgt die Trefferquote 39%. Die Trefferquoten für alle Zustandsbereiche sind deutlich besser als die von einem Zufallsmodell (Wahrscheinlichkeit je Zustand: 33%). Die Falsch-Negativ-Rate für Haltungen im mittleren (*K10*) und schlechten Zustand (*K11*) können als moderat bezeichnet werden. 23% der Haltungen im mittleren und 15% der Haltungen im schlechten Zustand werden durch das Modell fälschlicherweise als gut prognostiziert. Umgekehrt sind 30% der Haltungen, die im schlechten Zustand prognostiziert wurden, eigentlich in gutem Zustand (*K12,* Falsch-Positiv-Rate). Der Indikator *K_Haltung*, der die sechs Indikatoren auf Haltungsebene zusammenfasst, liegt bei 37,7%. Die Kreuztabelle für Prognose und Inspektionsergebnisse ist in Tabelle 19 dargestellt. Tabelle 20 zeigt die Bewertungsindikatoren auf Haltungsebene.

Tabelle 19: Kreuztabelle mit den Häufigkeiten der drei Zustandsbereiche für Inspektion (Zeilen) und Prognose (Spalten) mit Support Vector Machine (Modell B)

		Prognose			
		Gut	Mittel	Schlecht	Summe Inspektionen
Inspektion	Gut	13848	3798	3482	21128
	Mittel	2158	3623	3540	9321
	Schlecht	1312	2469	4789	8570
Summe Prognose		17318	9890	11811	

Tabelle 20: Bewertungsindikatoren für Support Vector Machine

Indikatoren auf Netzebene (für Modell A)						
K1 (Ziel → ±0)	**K2** (Ziel → ±0)	**K3** (Ziel → ±0)	**K4** (Ziel → ±0)	**K5** (Ziel → ±0)	**K6** (Ziel → ±0)	**K_Netz** (Ziel → min)
0,1	-0,3	0,2	3,1	-2,4	-0,7	1,6
Indikatoren auf Haltungsebene (für Modell (B)						
K7 (Ziel → max)	**K8** (Ziel → max)	**K9** (Ziel → max)	**K10** (Ziel → min)	**K11** (Ziel → min)	**K12** (Ziel → min)	**K_Haltung** (Ziel → min)
65,5	38,9	55,9	23,2	15,3	29,5	37,7

Die Untersuchungen haben gezeigt, dass Support Vector Machine die Zustandsverteilung im Netz sehr gut wiedergeben kann. Auch auf Haltungsebene liefert das Modell passable Ergebnisse. Allerdings erfordern beide Ziele unterschiedliche Modellparametrisierungen. Ein Modell, das für die Vorhersage auf Netzebene optimiert wurde (Modell A), liefert auf Haltungsebene keine guten Ergebnisse (*K_Haltung*: 47,1). Umgekehrt sind für ein Modell, das für die Vorhersage auf Haltungsebene optimiert wurde (Modell B) die Abweichungen auf Netzebene relativ hoch (*K_Netz: 12,8).*

Wie in Kap. 4.2.3 beschrieben, wurde das Modell aufgrund des hohen Rechenaufwands nur für eine Teilmenge von 10.000 Daten trainiert. Der anschließende Modelltest und die Berechnung der Güteindikatoren erfolgten für alle 58.528 Testdaten. Durch eine zweimalige Wiederholung des Vorgehens (Training für 10.000 zufällig gewählte Trainingsdaten, Test für alle Testdaten) wurden die Modellergebnisse bestätigt. Die Standardabweichungen der je drei Ergebnisse für *K_Netz* und *K_Haltung* beträgt 0,54 und 0,69.

4.3.4 Künstliche Neuronale Netze

4.3.4.1 Einfluss der Modellparameter

Wie in Kap. 4.2.3 beschrieben wurde für Künstliche Neuronale Netze eine vereinfachte Gitter-Suche mit 40 Parameterkombinationen durchgeführt (5-fache Kreuzvalidierung). Untersucht wurden vier verschiedene Aktivierungsfunktionen (Parameter *actfun*, siehe Kap.

4.1.4) und verschiedene Werte für die Anzahl versteckter Neuronen (Parameter *nhid*). Die Ergebnisse zeigen zum einen, dass der Einfluss der Aktivierungsfunktion auf die Güteindikatoren marginal ist (Ausnahme: *actfun* „purelin“). Zum anderen verbessern sich die Modellergebnisse mit zunehmender Anzahl an versteckten Neuronen (Abbildung 62 und Abbildung 63). *nhid*-Werte > 1000 konnten aufgrund des hohen Rechenaufwands beim Training des Modells nicht untersucht werden.

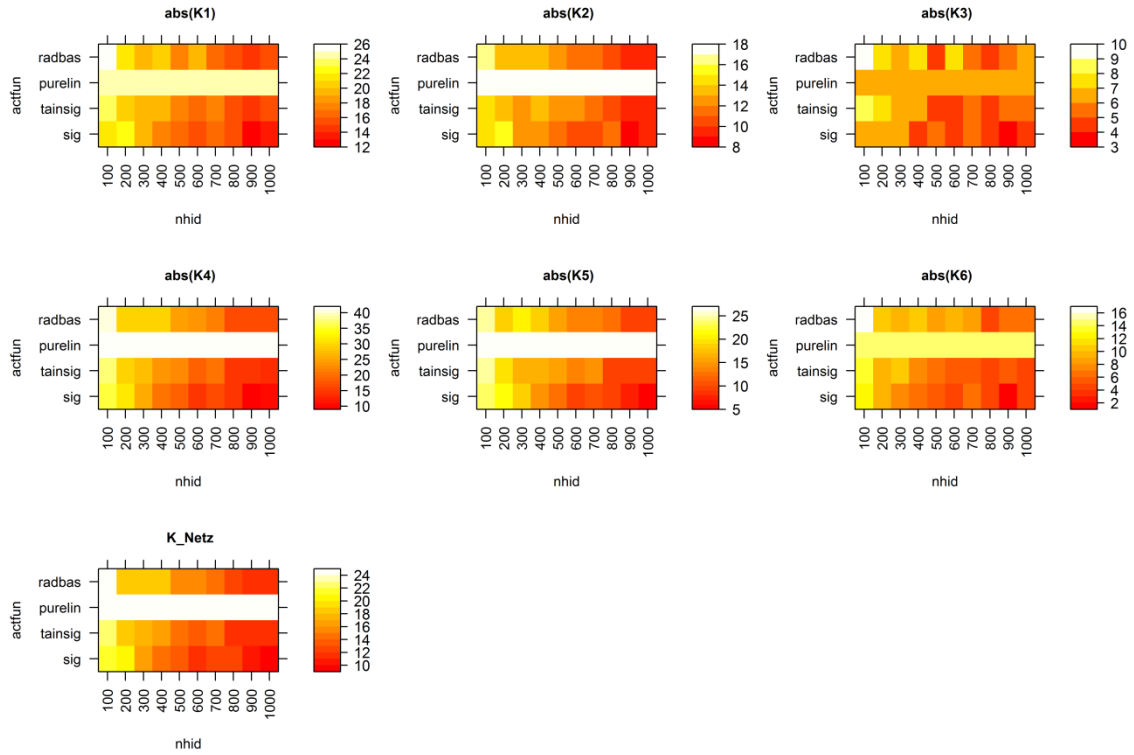


Abbildung 62: Einfluss der Parameter *nhid* und *actfun* auf die Modellgüte des Künstlichen Neuronalen Netzes auf Netzebene (Indikatoren *K1* bis *K6* und *K_Netz*)

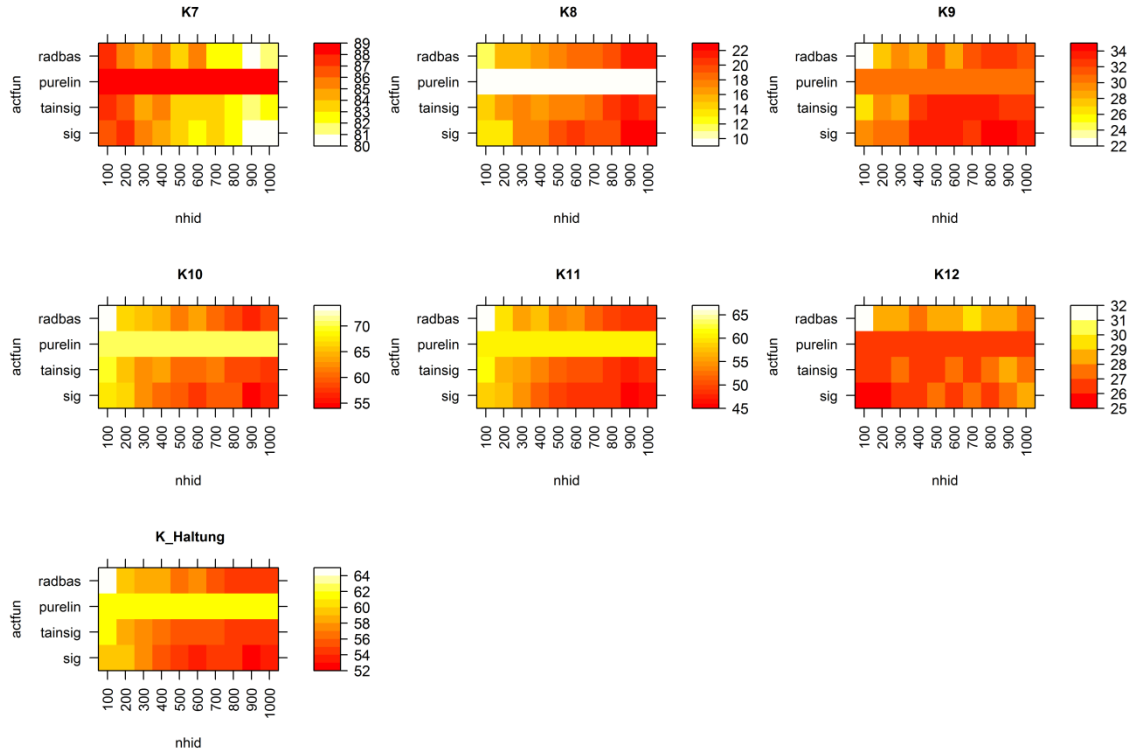


Abbildung 63: Einfluss der Parameter *nhid* und *actfun* auf die Modellgüte des Künstlichen Neuronalen Netzes auf Haltungsebene (Indikatoren *K7* bis *K12* und *K_Haltung*)

Für weitere Untersuchungen wurde die sigmoide Aktivierungsfunktion „sig“ und ein *nhid*-Wert von 1000 verwendet, da diese Parameterwerte im Rahmen der Kreuzvalidierung die besten Ergebnisse lieferten.

4.3.4.2 Bewertung der Modellgüte

Netzebene: Die allgemeinen Abweichungen für die drei Zustandsbereiche liegen im Bereich von 5 bis 12% (Indikatoren *K1*, *K2* und *K3*). Insbesondere für die jungen (≤ 50 a) und sehr alten Kanäle (> 100 a) wird der Anteil an Haltungen im guten Zustand deutlich überschätzt (Abbildung 64). Für die Altersklasse 51 bis 75 Jahre (*K4*, *K5* und *K6*) werden bessere Ergebnisse erzielt (Abweichungen < 4%). Der Indikator *K_Netz*, der die sechs Indikatoren auf Netzebene zusammenfasst, liegt bei 6,1%. Damit ist das Modell nur sehr bedingt in der Lage, die Zustandsverteilung im Kanalnetz wiederzugeben. Die berechneten Abweichungen sind in Abbildung 64 dargestellt, die Zahlenwerte der Bewertungsindikatoren sind in Tabelle 22 (oben) zusammengefasst.

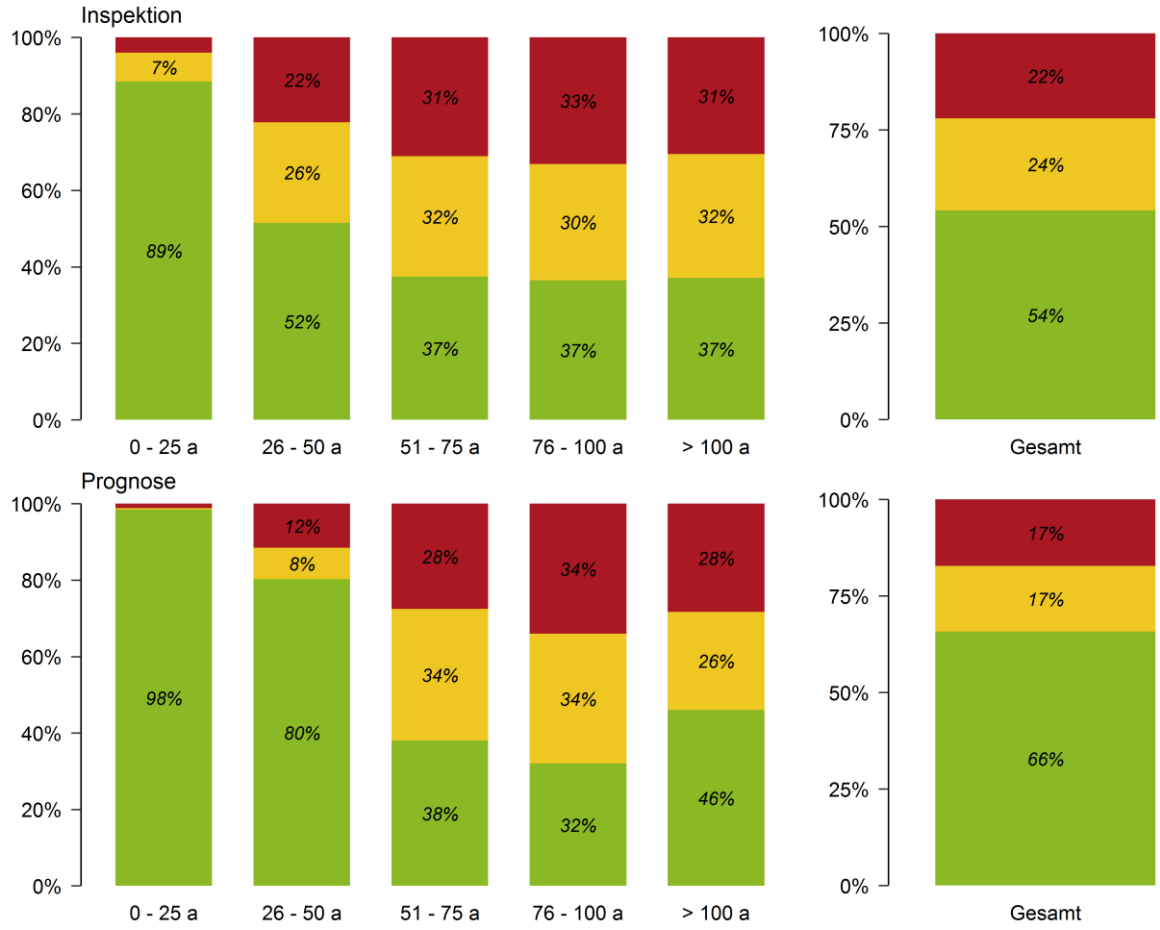


Abbildung 64: Abweichungen auf Netzebene zwischen Inspektion (oben) und Prognose mit einem Künstlichen Neuronalen Netz (unten). Die drei Farben repräsentieren die drei aggregierten Zustandsbereiche „gut“ (grün; Zustandsklasse 4, 5 und 6), „mittel“ (gelb; Zustandsklasse 3) und „schlecht“ (rot; Zustandsklasse 1 und 2)

Haltungsebene: Entsprechend der allgemeinen Überschätzung des Anteils an Kanälen im guten Zustand (siehe oben), ist die Trefferquote für Kanäle im guten Zustand (*K7*) mit 82% sehr hoch. Auf der anderen Seite werden jedoch 52% der Kanäle im mittleren und 40% der Kanäle im schlechten Zustand fälschlicherweise als gut prognostiziert (Falsch-Negativ-Fehlerquoten *K10* und *K11*). Die Trefferquoten für den mittleren (*K8*) und schlechten Zustand (*K9*) sind mit 28% und 38% sehr gering (etwa im Bereich eines Zufallsmodells). Für die Identifizierung prioritärer Haltungen, z.B. für Inspektionen oder Sanierungsprogramme, ist das Modell daher ungeeignet. Der Indikator *K_Haltung*, der die sechs Indikatoren auf Haltungsebene zusammenfasst, liegt bei 48,7%. Die Kreuztabelle für Prognose und Inspektionsergebnisse ist in Tabelle 24 dargestellt. Tabelle 22 zeigt die Bewertungsindikatoren auf Haltungsebene.

Tabelle 21: Kreuztabelle mit den Häufigkeiten der drei Zustandsbereiche für Inspektion (Zeilen) und Prognose (Spalten) mit Random Forest (Modell B)

		Prognose			
		Gut	Mittel	Schlecht	Summe Inspektionen
Inspektion	Gut	17381	2198	1549	21128
	Mittel	4824	2576	1921	9321
	Schlecht	3459	1869	3242	8570
Summe Prognose		25664	6643	6712	

Tabelle 22: Bewertungsindikatoren für Künstliches Neuronales Netz

Indikatoren auf Netzebene						
K1 (Ziel → ±0)	**K2** (Ziel → ±0)	**K3** (Ziel → ±0)	**K4** (Ziel → ±0)	**K5** (Ziel → ±0)	**K6** (Ziel → ±0)	**K_Netz** (Ziel → min)
-11,6	6,9	4,8	-0,6	-2,9	3,5	6,1
Indikatoren auf Haltungsebene						
K7 (Ziel → max)	**K8** (Ziel → max)	**K9** (Ziel → max)	**K10** (Ziel → min)	**K11** (Ziel → min)	**K12** (Ziel → min)	**K_Haltung** (Ziel → min)
82,3	27,6	37,8	51,8	40,4	23,1	48,7

Das untersuchte Künstliche Neuronale Netz liefert weder auf Netz- noch auf Haltungsebene sehr gute Ergebnisse. Ein mögliches Hindernis könnte das „Rauschen“ in den Daten sein, d.h. es gibt Kanäle, die die gleichen oder ähnliche Eigenschaften aufweisen aber dennoch mit unterschiedlichen Zuständen bewertet wurden, und daher nur schwer zu klassifizieren sind (Zhu und Wu, 2004). Dennoch heißt das nicht, das Künstliche Neuronale Netze für die vorliegende Fragestellung allgemein ungeeignet sind. Es gibt eine Vielzahl von Varianten der Künstlichen Neuronalen Netze, die im Rahmen dieser Studie aufgrund der hohen Modellkomplexizität und des hohen Rechenaufwands nicht vollständig untersucht werden konnten.

4.3.5 Vergleich der Modelle

Die untersuchten Modelle werden im Folgenden abschließend hinsichtlich ihrer Modellgüte, ihrer Modellkomplexizität und ihrem Rechenaufwand miteinander verglichen.

Modellgüte: Die Modelle GompitZ, Support Vector Machine und Random Forest erzielen mit Abweichungen im Bereich von 0 bis 3% gute bis sehr gute Ergebnisse bezüglich der Simulation der Zustandsverteilung auf Netzebene. Das untersuchte Künstliche Neuronale Netz gibt die Zustandsverteilung im Netz nur sehr ungenau wieder (Abbildung 65). Auf Haltungsebene liefern Random Forest und Support Vector Machine gute Ergebnisse. Die Trefferquoten kommen zumindest für den guten und schlechten Zustand in die Nähe der

Inspektionsgenauigkeit (69 bis 82%, siehe Kap. 3.4.4.2). Random Forest erreicht mit seiner Trefferquote für Kanäle im schlechten Zustand etwa 85% der Inspektionsgenauigkeit. Das Künstliche Neuronale Netz und GompitZ sind für Vorhersagen auf Haltungsebene hingegen ungeeignet (Abbildung 66). Tabelle 27 in Anhang F fasst die Ergebnisse tabellarisch zusammen.

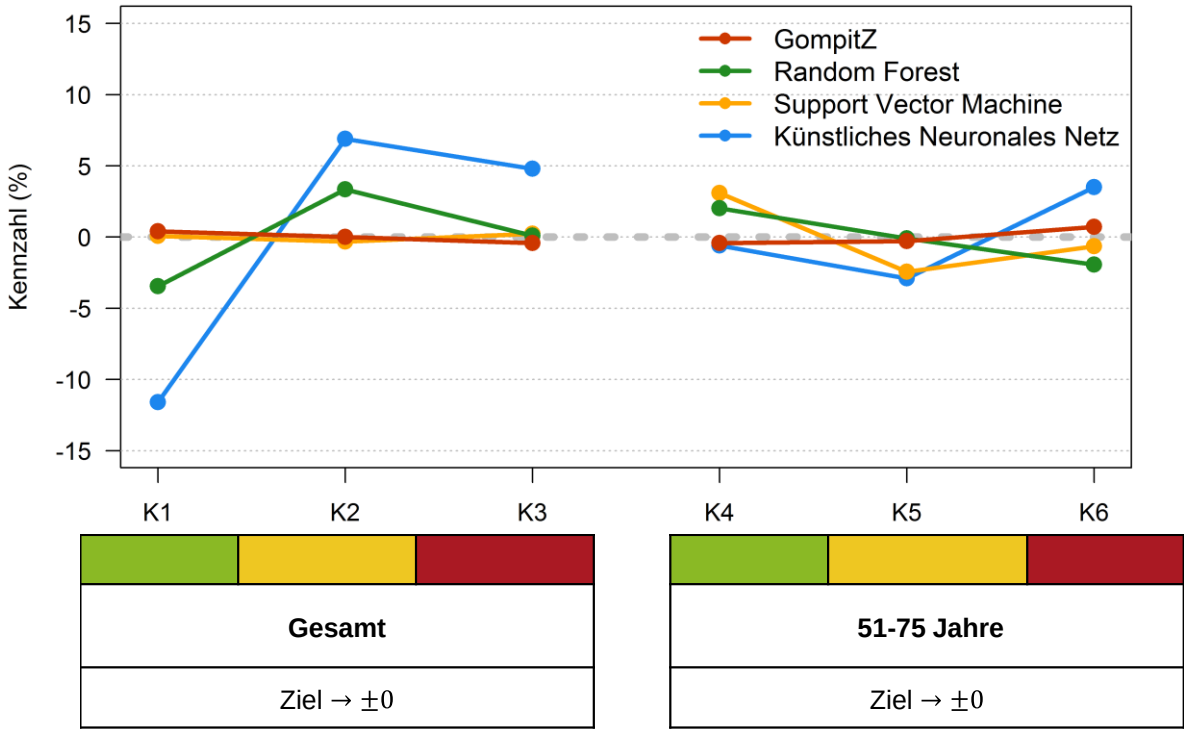


Abbildung 65: Modellgüte auf Netzebene für die vier untersuchten Modellansätze. Erläuterung der Bewertungsindikatoren in Kap. 4.2.5.

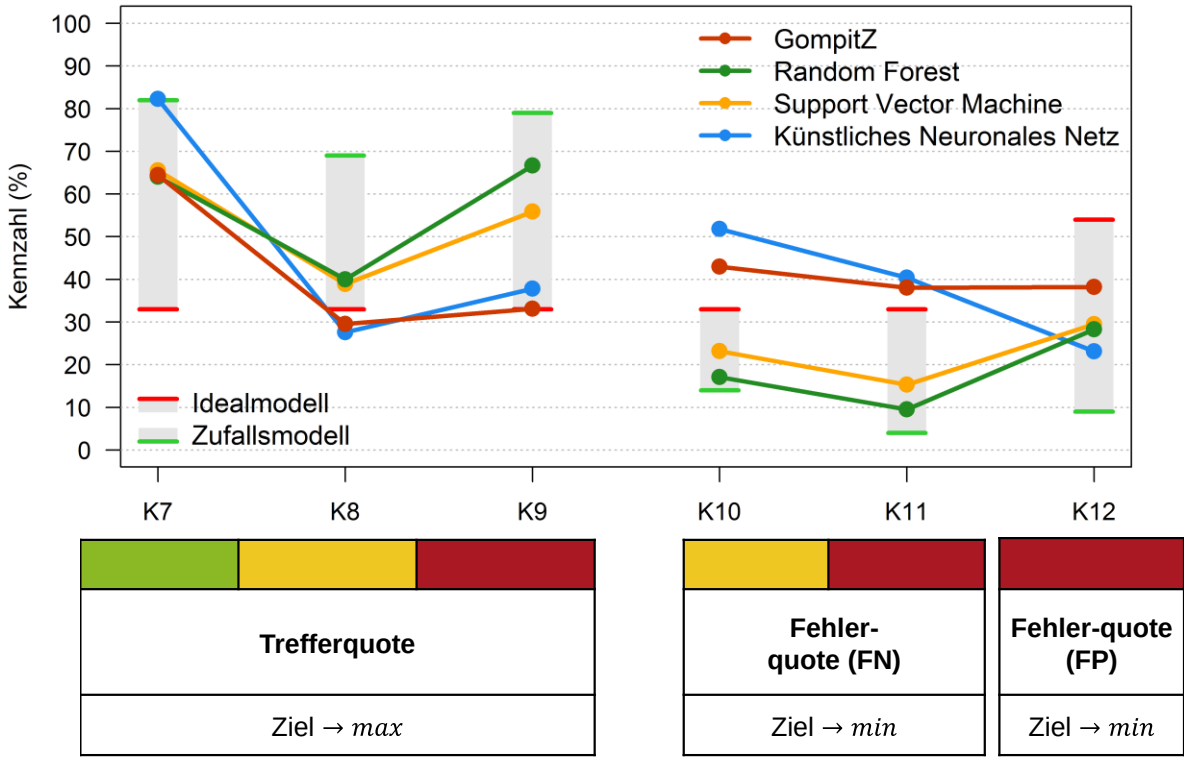


Abbildung 66: Modellgüte auf Haltungseebene für die vier untersuchten Modellansätze. Erläuterung der Bewertungsindikatoren in Kap. 4.2.5.

Modellkomplexizität: Die Bewertung der Modellkomplexizität ist sehr subjektiv. Vermutlich aber können Random Forest und GompitZ als die transparentesten Modellansätze bezeichnet werden. Die Struktur des Modells lässt sich für beide Ansätze gut visualisieren, für GompitZ über die Überlebenskurven (siehe Abbildung 37, Kap. 4.1.1), für Random Forest über die Entscheidungsbäume (siehe Abbildung 39, Kap. 4.1.2), und entsprechend gut erklären. Die angewandten mathematischen Verfahren sind zudem vergleichsweise einfach. Support Vector Machine ist bezogen auf die eingesetzte Mathematik deutlich komplexer (Kernel-Trick, Lagrange-Multiplikator, 4.1.3). Das Künstliche Neuronale Netz ist vermutlich der komplexeste der untersuchten Modellansätze, da die Anzahl der Neuronen sehr groß ist und das Verfahren der Gewichtsoptimierung („Gradient Descent“) relativ kompliziert ist (4.1.4).

Rechenaufwand: Die Rechenzeiten für die Simulation sind im Allgemeinen gering und liegen für alle Modellansätze im Sekundenbereich (Tabelle 23). Die Dauer für den Aufbau der Modelle (Training) unterscheidet sich je nach berücksichtigter Datenmenge. Besonders aufwändig ist das Training für Support Vector Machine und Künstliche Neuronale Netze, wobei durch die Reduzierung der Menge an Trainingsdaten von 58.528 auf 8.000 (Support Vector Machine) bzw. 46.822 (Künstliche Neuronale Netze) die Dauer auf < 3 min reduziert werden kann.

Einschränkungen der Modellergebnisse: Aufgrund des sehr hohen Rechenaufwands und Speicherbedarfs beim Training der Künstlichen Neuronalen Netze, waren die Möglichkeiten der Untersuchung hier beschränkt. Das heißt jedoch nicht, dass solche Modelle prinzipiell nicht für die vorliegende Fragestellung geeignet sind. Für eine tiefergehende Untersuchung Künstlicher Neuronaler Netze, z.B. sogenannter „Deep Neural Networks“ (Schmidhuber

2015, Zhang et al. 2018), müsste ein größerer zeitlicher Rahmen und zusätzliche Computer-Hardware vorgesehen werden.

Tabelle 23: Rechenzeiten für Modellaufbau (Training) und Simulation mit den vier untersuchten Modellansätzen (ermittelt mit PC-Konfiguration: i7-2600 CPU; 3,4 GHz, 8 bzw. 12 GB RAM)

Modell	Dauer Modellaufbau [1]	Dauer Simulation [2]
GompitZ	107 s	2 s
Random Forest (Modell A)	92 s	1 s
Random Forest (Modell B)	27 s	1 s
Support Vector Machine (Modell A)	34 s	26 s
Support Vector Machine (Modell B)	15 s	36 s
Künstliche Neuronale Netze	173 s	5 s

Erläuterungen: [1] Anzahl Datensätze für Modellaufbau: 58.528 für GompitZ und Random Forest, 8.000 für Support Vector Machine, 46.822 für Künstliche Neuronale Netze; [2] Anzahl Datensätze für Simulation: 39.019 für alle Modelle. Bei Simulationen für mehrere Jahre, z.B. im Rahmen von Prognosen zur zukünftigen Entwicklung, erhöht sich die Simulationsdauer proportional zur Anzahl der Jahre.

4.4 Simulation der Zustandsentwicklung

Wie in Kap. 4.2.6 beschrieben wurde mit den Modellen GompitZ und Random Forest die Entwicklung der Zustandsverteilung des Netzes über die nächsten 50 Jahre prognostiziert. Für GompitZ wurden die in Kap. 4.3.1.1 beschriebenen Kohorten und Überlebenskurven verwendet. Für Random Forest wurde das in Kap. 4.3.2.1 beschriebene Modell A verwendet, welches für die Vorhersage auf Netzebene optimiert wurde.

Beide Modelle simulieren eine kontinuierliche Verschlechterung des Zustands der Kanäle. Der Prognose mit GompitZ zufolge erhöht sich der Anteil an Kanälen im schlechten Zustand bis ins Jahr 2066 kontinuierlich von 24 auf 39%. Der Anteil an Kanälen im guten Zustand verringert sich von 52 auf 38%. Der Anteil an Kanälen im mittleren Zustand bleibt mit etwa 24% konstant. Die Alterung der Kanalisation bzw. Zustandsverschlechterung verläuft über die 50 Jahre mit ähnlicher Geschwindigkeit. Jedes Jahr gehen 0,2 bis 0,3% der Kanäle in den schlechten Zustand über. Abbildung 67 zeigt die mit GompitZ prognostizierte Zustandsentwicklung für die Jahre 2017 bis 2066.

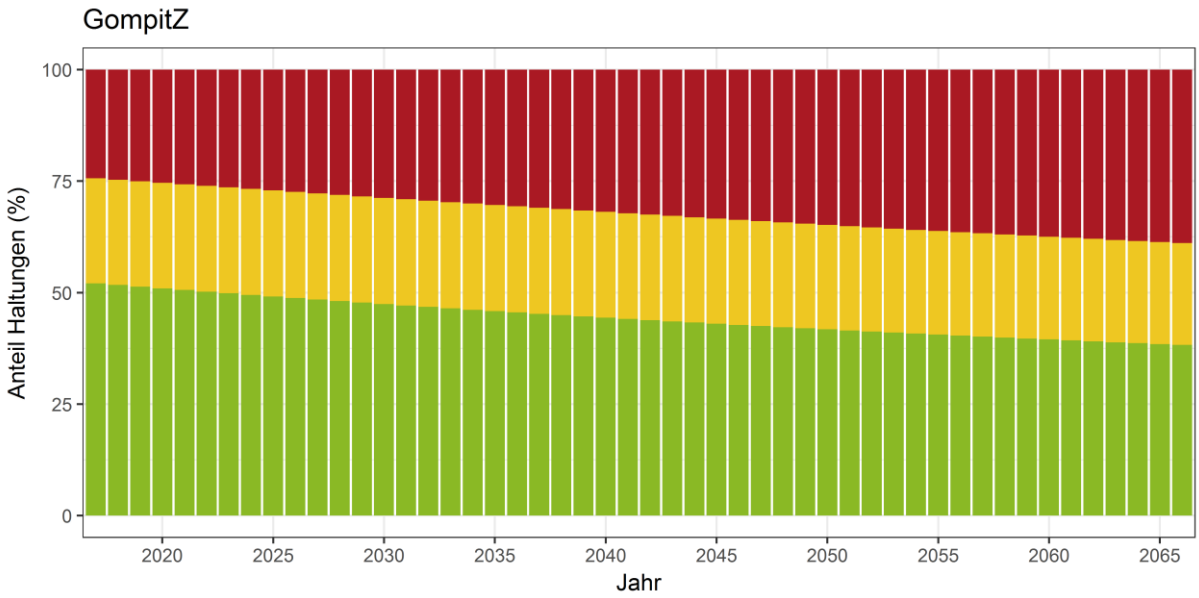


Abbildung 67: Prognose der Zustandsentwicklung von 2017 bis 2066 mit GompitZ

Der Prognose mit Random Forest zufolge verschlechtert sich der Zustand der Kanalisation in den ersten fünf Jahren der Simulation etwas schneller als bei GompitZ. Etwa 0,5% der Kanäle gehen in diesem Zeitraum jährlich in den schlechten Zustand über. Danach flachen die Kurven ab, so dass nach 50 Jahren nur noch etwa 0,2% der Kanäle pro Jahr in den schlechten Zustand übergehen. Die simulierte Zustandsverteilung nach 50 Jahren ähnelt der von Gompitz (siehe oben). Der Anteil an Kanälen im schlechten Zustand erhöht sich bis ins Jahr 2066 von 26 auf 38%. Der Anteil an Kanälen im guten Zustand verringert sich von 56 auf 43%. Der Anteil an Kanälen im mittleren Zustand bleibt mit etwa 20% unverändert. Abbildung 68 zeigt die mit Random Forest (Modell A) prognostizierte Zustandsentwicklung für die Jahre 2017 bis 2066.

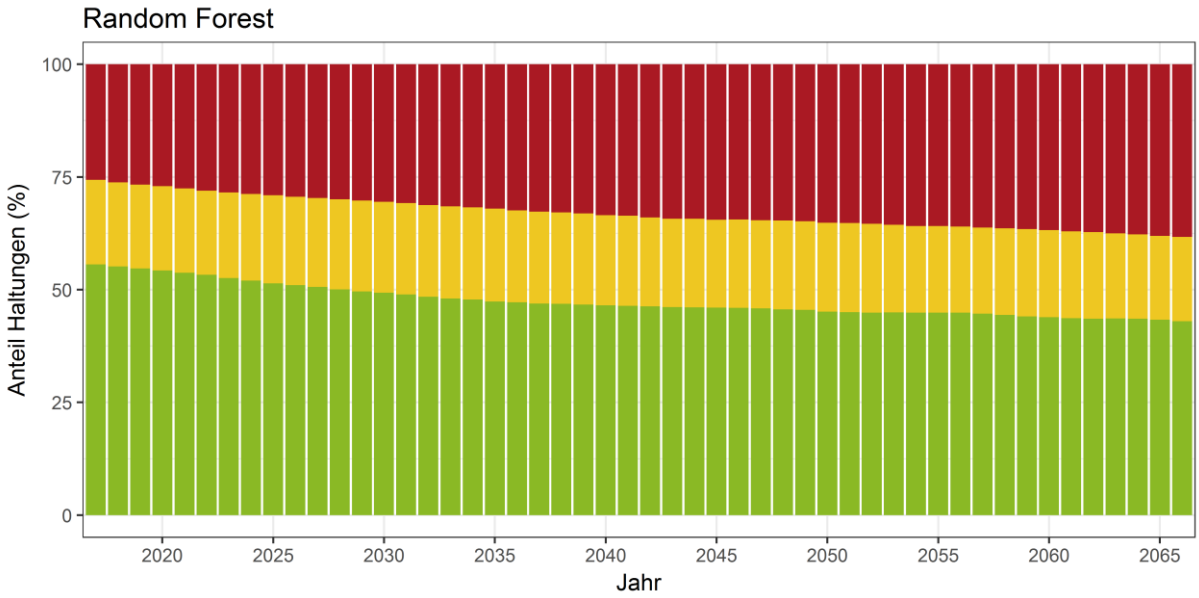


Abbildung 68: Prognose der Zustandsentwicklung von 2017 bis 2066 mit Random Forest (Modell A)

Bei der Interpretation der Prognoseergebnisse sollte beachtet werden, dass die Modelle auf Daten kalibriert bzw. trainiert wurden, in denen alte schadhafte Kanäle stark unterrepräsentiert sind. Die Kanäle, die bereits ihre erwartete Lebensdauer überschritten haben und schwere Schäden aufweisen, wurden bereits erneuert (siehe Kapitel 3.2) und können in den Modellen nicht mehr berücksichtigt werden. Aufgrund dieser verzerrten Zustandsverteilung (v.a. für Kanäle mit Alter > 75 a) ist zu erwarten, dass die bisherigen Prognosen zu optimistisch ausfallen. Es ist grundsätzlich zu diskutieren und zu testen, wie diese Verzerrung in den Daten (Bias) korrigiert werden kann und wie der Alterungsprozess nach 75 Jahren aussehen sollte, um zu möglichst realistischen Prognosen zu kommen. Andernfalls besteht die Gefahr, dass das auf Grundlage der Prognosen eigeplante Budget nicht ausreicht, um den erwünschten Kanalzustand zu erreichen bzw. zu erhalten. Darüber hinaus sollten die Annahmen zur Extrapolation der Überlebenskurven über das bekannte Alter der Kanäle hinaus (Kap. 4.3.1.1) kritisch hinterfragt werden, insbesondere für langfristige Prognosen.

Bezüglich Random Forest ist zu berücksichtigen, dass das Modell die aus den Trainingsdaten erlernten Alterungsmuster anwendet und nur solche Anteile an Kanälen im schlechten Zustand simulieren kann, die es in den vorliegenden Trainingsdaten „sieht“. D.h. die Prognose wird nie schlechter sein als die beobachtete Zustandsverteilung in den Trainingsdaten. Aus diesem Grund ist der Einsatz der untersuchten Modelle des maschinellen Lernens für langfristige Prognosen (über den beobachteten Altersbereich der Trainingsdaten hinaus) nicht geeignet.

5 Fazit

5.1 Zusammenfassung

Basierend auf mehr als 140.000 Inspektionsdatensätzen aus den Jahren 2001 bis 2016 wurden im Rahmen des Forschungsvorhabens SEMA-Berlin Analysen zum baulichen Zustand der Abwasserkanalisation durchgeführt und Alterungsmodelle entwickelt und getestet. Dabei wurden die individuellen Kanaleigenschaften und Umgebungsfaktoren der Stadt Berlin berücksichtigt.

Die Zustandsanalyse zeigt, dass zum aktuellen Zeitpunkt mehr als die Hälfte der inspizierten Kanäle in einem guten Zustand sind und höchstens langfristig saniert werden müssten (Zeithorizont > 10 Jahre). Auf der anderen Seite befinden sich 22% der inspizierten Kanäle in einem schlechten Zustand, d.h. diese Kanäle müssten sofort oder kurzfristig saniert werden (Zeithorizont < 5 Jahre).

Den statistisch stärksten Effekt auf die Zustandsverteilung hat das Alter gefolgt vom Profil und der Länge des Kanals. Auch die Grundwasserüberdeckung, das Material, die Nähe zu Bäumen, der Bezirk, der Abwassertyp, die Breite des Kanals, der Bodentyp, die Beeinflussung durch Rückstau und die Bodenüberdeckung zeigen einen Zusammenhang zum Zustand der Kanäle. Insgesamt wurden zwölf relevante Einflussfaktoren identifiziert. Der Einfluss des Gefälles, des Straßenverkehrs und des Schienenverkehrs ist marginal und kann vernachlässigt werden.

Unter den Einflussfaktoren wurden Abhängigkeiten festgestellt - insbesondere zwischen i) dem Bezirk und dem Abwassertyp, ii) dem Abwassertyp und dem Material und iii) dem Material und dem Profil. Auch das Alter ist mit vielen anderen Faktoren verknüpft, z.B. dem Bezirk, dem Abwassertyp und dem Profil. Das heißt, dass ein Teil des Effektes, den eine Variable auf die Zustandsverteilung hat, auch durch eine andere Variable erklärt werden kann und nicht auf einem kausalen Zusammenhang beruht. Vollständig abhängige Variablen wie der Stadtteil (abhängig vom Bezirk) wurden im Vorfeld ausgeschlossen, um die Modelle nicht unnötig komplex zu machen.

Insgesamt wurde bei 86% aller Inspektionen mindestens ein Schaden festgestellt. Die am häufigsten beobachteten Schadenstypen sind i) Abflusshindernisse, ii) Risse und Scherbenbildung, iii) Schadhafte Rohrverbindungen, iv) Mechanischer Verschleiß und v) Verwurzelungen, die bei mindestens jeder dritten Inspektion festgestellt wurden. Die wichtigsten Faktoren für das Auftreten dieser Schäden sind das Material, das Alter, das Profil, die Breite, der Bezirk, die Länge und der Abwassertyp. Für Wurzelschäden spielen Bäume nachweislich eine Rolle. Nicht alle beobachteten Schäden führen zu einer deutlichen Verschlechterung des Zustands und müssen unmittelbar saniert werden. Der Sanierungsbedarf einer Haltung hängt vor allem von der Schwere des jeweiligen Schadens, z.B. der Breite und Länge von Rissen, ab, was im Rahmen dieser Arbeit nicht umfassen analysiert werden konnte.

Basierend auf Mehrfachinspektionen derselben Haltung innerhalb eines Zeitraums von höchstens fünf Jahren wurden die Unsicherheiten bei der Zustandsbewertung ermittelt. In etwa einem Drittel aller Fälle weichen die Ergebnisse von Erst- und Zweitinspektion voneinander ab, d.h. die Klassifizierung der Haltung in die drei Zustandsbereiche („gut“, „mittel“, „schlecht“) wurde unterschiedlich vorgenommen. Die Wahrscheinlichkeit, den

Zustand bei der Inspektion richtig einzuschätzen, unterscheidet sich je nach Zustandsbereich und liegt für Kanäle im guten Zustand bei 82%, für Kanäle im mittleren Zustand bei 69% und für Kanäle im schlechten Zustand bei 79%. Für Kanäle mit keinen oder sehr starken Schäden ist die Bewertung auf Grundlage von Kamerainspektionen sicherer als die von Kanälen mit mittelschweren Schäden (z.B. tiefe oder Länge eines Risses), deren Beurteilung stark vom Betrachter, den Lichtverhältnissen oder dem Reinigungsgrad des Kanals abhängt.

Auf Basis der Berliner Daten zu den Kanaleigenschaften und Umweltfaktoren sowie der Ergebnisse aus der Zustandsanalyse wurden ein statistisches Modell und drei Modelle des Maschinellen Lernens entwickelt und bewertet. Für die Bewertung wurden gemeinsam mit den Berliner Wasserbetrieben insgesamt zwölf Bewertungsindikatoren für die Güte auf Netz- und Haltungsebene definiert.

Durch das statistische Modell GompitZ kann die Zustandsverteilung des Netzes mit einer Genauigkeit von 99% simuliert werden. Das Modell kann damit als Grundlage für Prognosen zur zukünftigen Zustandsentwicklung des Kanalnetzes verwendet werden und birgt ein großes Potenzial für die strategische Investitionssteuerung (siehe Ausblick).

Durch ein Modell des Maschinellen Lernens (Random Forest) können Kanäle im schlechten Zustand mit einer Trefferquote von 67% detektiert werden, d.h. zwei von drei Kanälen im schlechten Zustand werden von dem Modell richtig prognostiziert. Das entspricht 85% der Inspektionsgenauigkeit. Die Ergebnisse können dafür genutzt werden, den Zustand nicht inspizierter Kanäle vorherzusagen und Haltungen für Inspektionen oder Sanierungen zu priorisieren. Durch eine modellgestützte Inspektionsstrategie ließen sich drei Mal mehr Kanäle im schlechten Zustand finden als durch „zufällige“ Inspektionen.

Untersuchungen zum Einfluss der Datenmenge auf das Vorhersageergebnis zeigen zum einen, dass bereits 3000 Inspektionsdatensätze für den Modellaufbau ausreichen, um eine relativ genaue Vorhersage der Zustandsverteilung auf Netzebene zu erhalten (statistisches Modell „GompitZ“). Auf der anderen Seite würde sich die Modellgüte auf Haltungsebene wahrscheinlich weiter verbessern lassen, wenn das Modell über die verwendeten 58.528 Datensätze hinaus mit zusätzlichen Inspektionsdaten trainiert werden könnte (Modell des maschinellen Lernens „Random Forest“). Die Fähigkeit aus weiteren Daten zu lernen und die Prognosequalität damit zu verbessern ist der Hauptunterschied zwischen Modellen des maschinellen Lernens und statistischen Modellen.

Erste Prognosen zur Zustandsentwicklung des Netzes zeigen, dass jedes Jahr etwa 0,3% der Kanäle in den schlechten Zustand übergehen und kurzfristig saniert werden müssten. Bei den Prognosen ist jedoch zu beachten, dass die Modelle auf Daten kalibriert bzw. trainiert wurden, in denen alte schadhafte Kanäle stark unterrepräsentiert sind, da diese bereits saniert wurden und aus der Betrachtung herausfallen. Durch diese modelltechnische Verzerrung (Bias) fallen die Prognosen zu optimistisch aus.

Für die praktische Anwendung wird empfohlen, das Modell GompitZ für Vorhersagen auf Netzebene und das Modell Random Forest für Vorhersagen auf Haltungsebene zu verwenden. Die Komplexität der Modelle ist vergleichsweise gering, d.h. die Modellstruktur lässt sich gut visualisieren und erklären. Die Rechenzeit ist für alle untersuchten Modellansätze sehr gering und spielt für die Wahl der Modelle keine Rolle.

5.2 Offene Fragen

Aus den vorgestellten Arbeiten haben sich folgende offene Fragen ergeben, die für eine praktische Umsetzung tiefergehend untersucht werden sollten:

- Wie kann man mit der Verzerrung der Zustandsverteilung bei hohem Alter umgehen?
- Wie können Prognosen für Altersbereiche erstellt werden, die bisher nie oder kaum beobachtet wurden? Ist die Extrapolation des Alterungsverhaltens weit über beobachteten Altersbereich plausibel? Können Betriebserfahrungen für die Extrapolation berücksichtigt werden?
- Kann die Komplexität der Modelle des Maschinellen Lernens weiter vereinfacht werden, z.B. durch die Reduzierung der Eingangsvariablen, ohne an Genauigkeit zu verlieren?
- Sind die Modellergebnisse auf Haltungsebene physikalisch plausibel? Wie können Überanpassungen des Modells vermieden werden?
- Wie können Ergebnisse einer eventuellen Zweit-, Dritt- oder Viertinspektion für die Simulation auf Haltungsebene berücksichtigt werden?
- Wie ist das Alterungsverhalten reparierter und renovierter Kanäle (Liner) und wie kann das in Modellen berücksichtigt werden?

5.3 Ausblick

Die Ergebnisse zeigen ein großes Potenzial für die Festlegung zukünftiger Inspektions- und Sanierungsprogramme und die strategische Investitionssteuerung.

Konkret könnten die entwickelten Modelle des Maschinellen Lernens dafür eingesetzt werden, Gebiete oder Haltungen im schlechten Zustand zu erkennen und für Inspektionen oder Sanierungen zu priorisieren. Durch Verknüpfung der Modellprognosen mit einer Methode zur Quantifizierung des Gefährdungspotenzials von Gebieten ließe sich außerdem das Risiko von Schäden, beispielsweise durch Rohrbrüche, reduzieren.

Das untersuchte statistische Modell steht bereit, um darauf aufbauend ein Simulationswerkzeug für die strategische Investitionssteuerung zu entwickeln. Mit einem solchen Werkzeug könnte der Einfluss langfristiger Investitionsszenarien auf die Zustandsentwicklung des Kanalnetzes untersucht werden. Es ließe sich ermitteln, mit welchen Sanierungsstrategien der Zustand der Kanalisation am effizientesten erhalten bzw. verbessert werden kann. In diesem Zusammenhang ist die Frage zu beantworten, ob und inwieweit Synergien mit anderen Infrastrukturbereichen (Straßenbau, Stromnetze, etc.) zu effizienteren Sanierungsstrategien führen.

Die erzielten Erkenntnisse zur Unsicherheit des Inspektionsergebnisses ließen sich für die Quantifizierung von Unsicherheiten bei der Modellprognose und der Investitionsvorhersage nutzen. Darauf aufbauend könnten Schwankungsbreiten für die Sanierungskosten zur Erreichung oder Erhaltung eines bestimmten Kanalzustands angegeben werden.

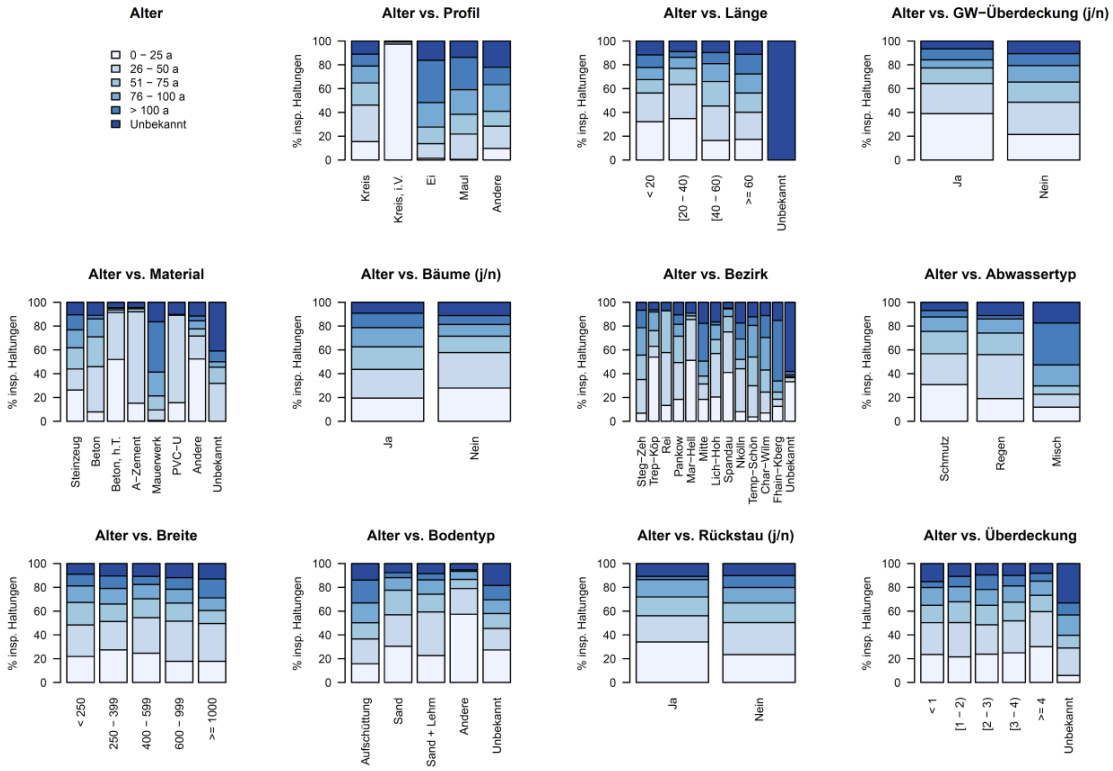
Anhang A: Zustandsverteilung und Einflussfaktoren

Tabelle 24: *Cramér's-V*-Werte für alle Variablen untereinander und kombiniert mit der Zustandsklasse

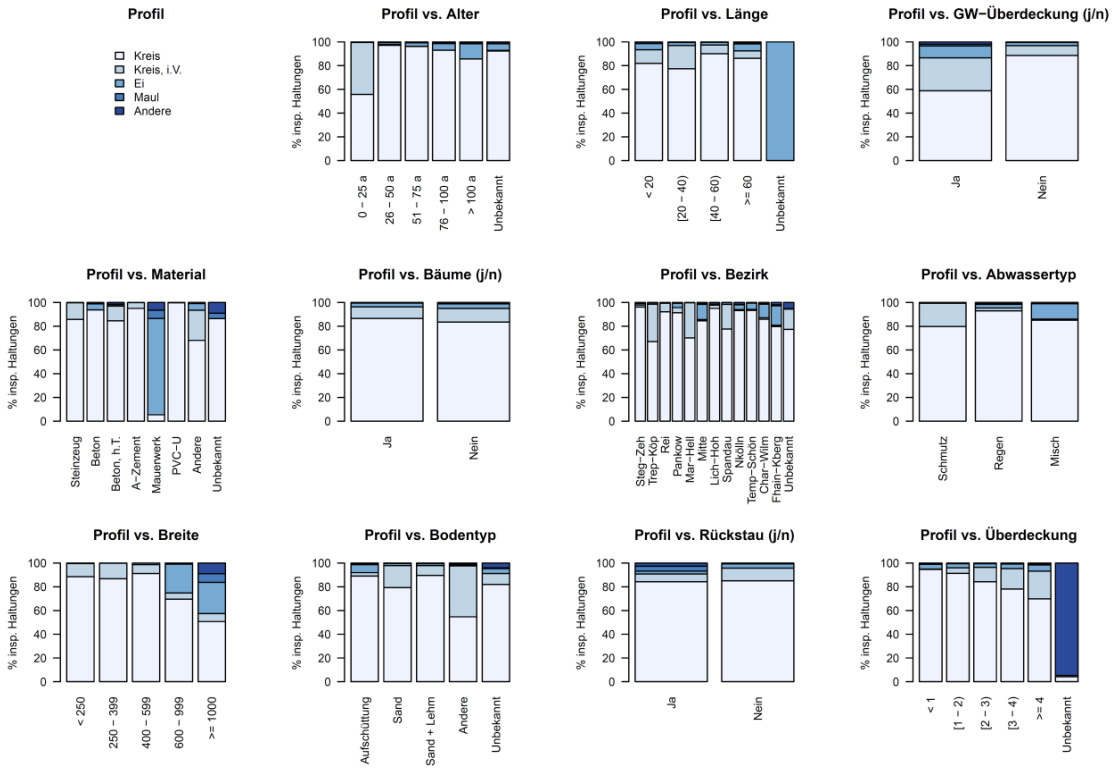
	Alter	Material	Abwassertyp	Profil	Breite	Länge	Überdeckung	Gefälle	Strassenklasse	Schienenverkehr (j/n)	Bäume (j/n)	Anzahl Bäume	GW-Überdeckung (j/n)	GW-Überdeckung in m	Bodentyp	Bezirk	Stadtteil	Rückstau (j/n)	Zustandsklasse
Alter	1,00	0,27	0,35	0,31	0,06	0,15	0,05	0,07	0,05	0,02	0,18	0,10	0,14	0,09	0,17	0,36	0,22	0,05	0,27
Material	0,27	1,00	0,54	0,40	0,37	0,09	0,10	0,13	0,10	0,08	0,11	0,06	0,20	0,12	0,14	0,21	0,18	0,19	0,15
Abwassertyp	0,35	0,54	1,00	0,27	0,40	0,10	0,22	0,08	0,12	0,11	0,03	0,03	0,11	0,09	0,32	0,54	0,02	0,19	0,12
Profil	0,31	0,40	0,27	1,00	0,29	0,11	0,10	0,10	0,06	0,05	0,05	0,04	0,28	0,17	0,14	0,24	0,15	0,11	0,19
Breite	0,06	0,37	0,40	0,29	1,00	0,05	0,09	0,19	0,11	0,08	0,05	0,02	0,24	0,14	0,08	0,12	0,04	0,19	0,12
Länge	0,15	0,09	0,10	0,11	0,05	1,00	0,02	0,10	0,05	0,06	0,27	0,25	0,05	0,03	0,08	0,15	0,11	0,01	0,16
Überdeckung	0,05	0,10	0,22	0,10	0,09	0,02	1,00	0,09	0,04	0,02	0,03	0,02	0,22	0,17	0,06	0,11	0,07	0,05	0,07
Gefälle	0,07	0,13	0,08	0,10	0,19	0,10	0,09	1,00	0,05	0,03	0,02	0,02	0,16	0,09	0,10	0,13	0,02	0,09	0,04
Strassenklasse	0,05	0,10	0,12	0,06	0,11	0,05	0,04	0,05	1,00	0,17	0,07	0,05	0,05	0,04	0,08	0,10	0,07	0,05	0,03
Schienenverkehr (j/n)	0,02	0,08	0,11	0,05	0,08	0,06	0,02	0,03	0,17	1,00	0,06	0,07	0,02	0,03	0,10	0,13	0,03	0,00	0,03
Bäume (j/n)	0,18	0,11	0,03	0,05	0,05	0,27	0,03	0,02	0,07	0,06	1,00	1,00	0,05	0,05	0,06	0,25	0,19	0,04	0,12
Anzahl Bäume	0,10	0,06	0,03	0,04	0,02	0,25	0,02	0,02	0,05	0,07	1,00	1,00	0,05	0,03	0,04	0,14	0,14	0,04	0,08
GW-Überdeckung (j/n)	0,14	0,20	0,11	0,28	0,24	0,05	0,22	0,16	0,05	0,02	0,05	0,05	1,00	1,00	0,23	0,21	0,10	0,13	0,15
GW-Überdeckung in m	0,09	0,12	0,09	0,17	0,14	0,03	0,17	0,09	0,04	0,03	0,05	0,03	1,00	1,00	0,14	0,14	0,07	0,14	0,09
Bodentyp	0,17	0,14	0,32	0,14	0,08	0,08	0,06	0,10	0,08	0,10	0,06	0,04	0,23	0,14	1,00	0,30	0,08	0,09	0,08
Bezirk	0,36	0,21	0,54	0,24	0,12	0,15	0,11	0,13	0,10	0,13	0,25	0,14	0,21	0,14	0,30	1,00	0,97	0,17	0,12
Stadtteil	0,22	0,18	0,02	0,15	0,04	0,11	0,07	0,02	0,07	0,03	0,19	0,14	0,10	0,07	0,08	0,97	1,00	0,01	0,09
Rückstau (j/n)	0,05	0,19	0,19	0,11	0,19	0,01	0,05	0,09	0,05	0,00	0,04	0,04	0,13	0,14	0,09	0,17	0,01	1,00	0,07
Zustandsklasse	0,27	0,15	0,12	0,19	0,12	0,16	0,07	0,04	0,03	0,03	0,12	0,08	0,15	0,09	0,08	0,12	0,09	0,07	1,00

Anhang B: Abhängigkeiten der wichtigsten Variablen

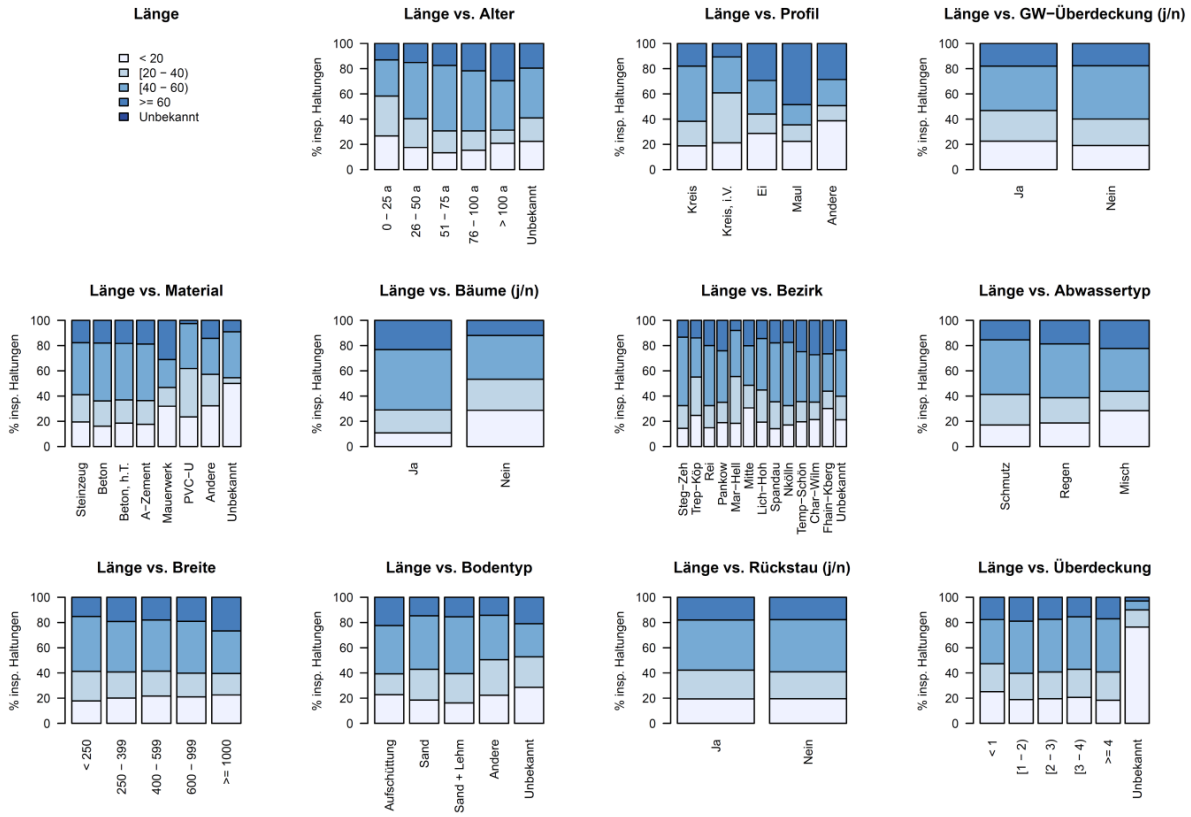
a) Alter



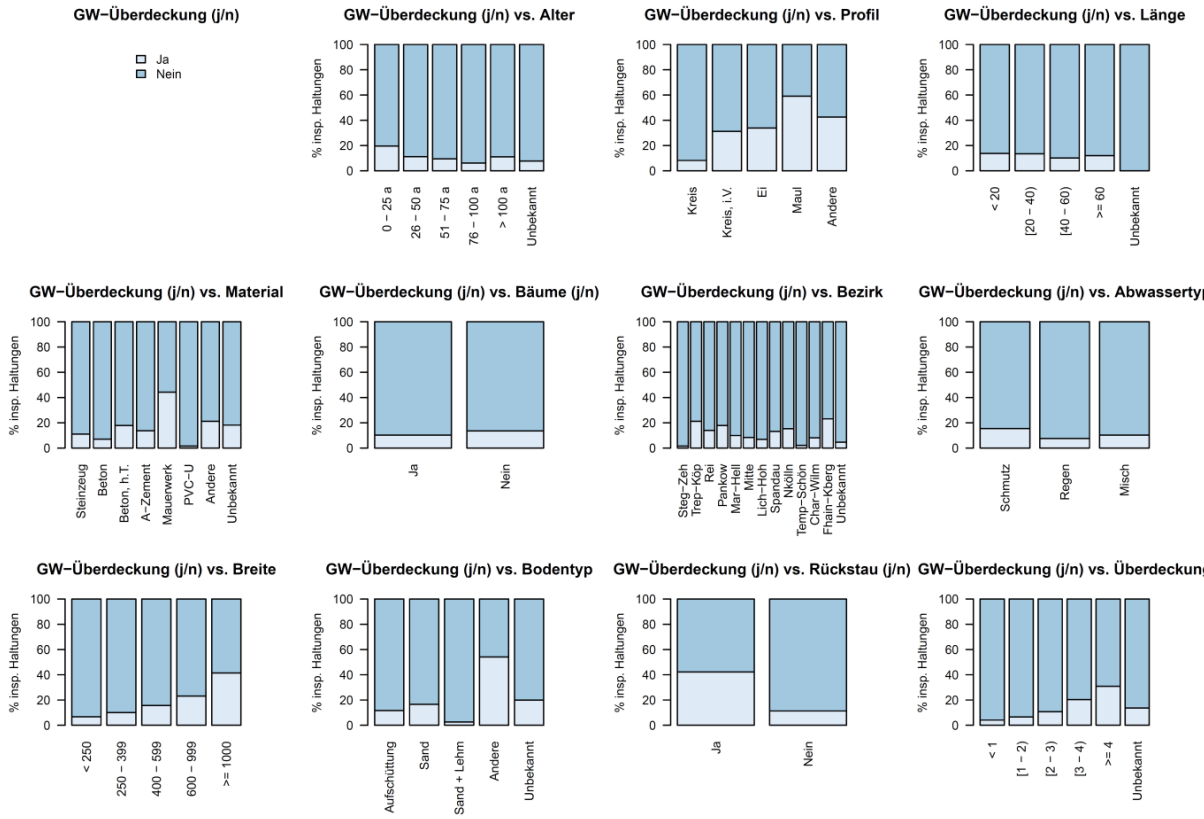
b) Profil



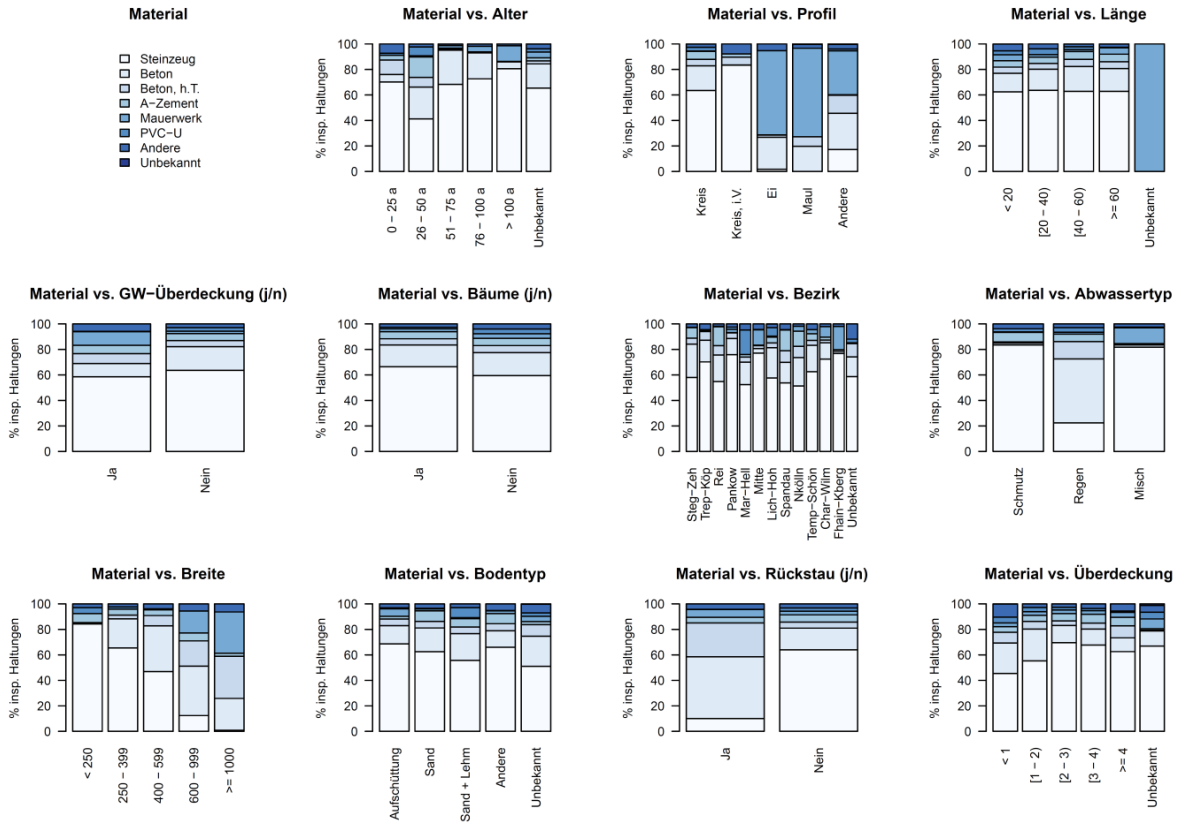
c) Länge



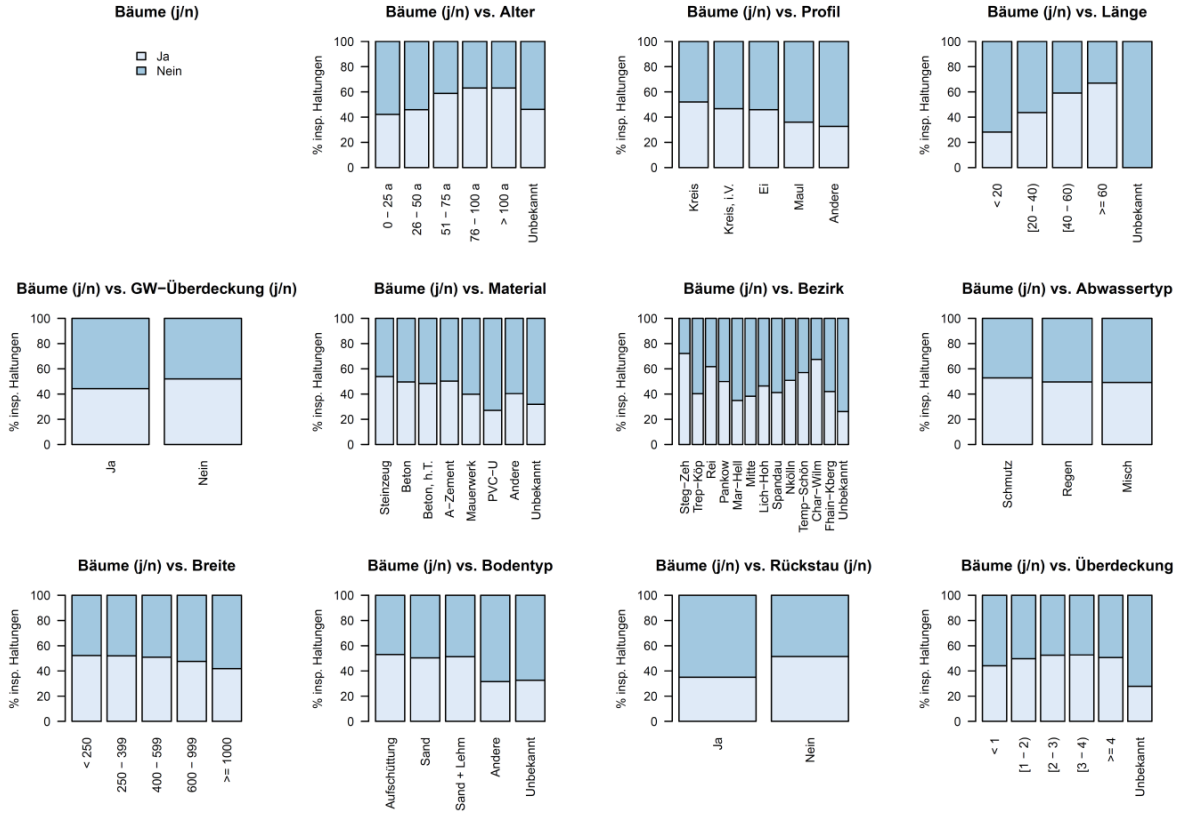
d) Grundwasserüberdeckung



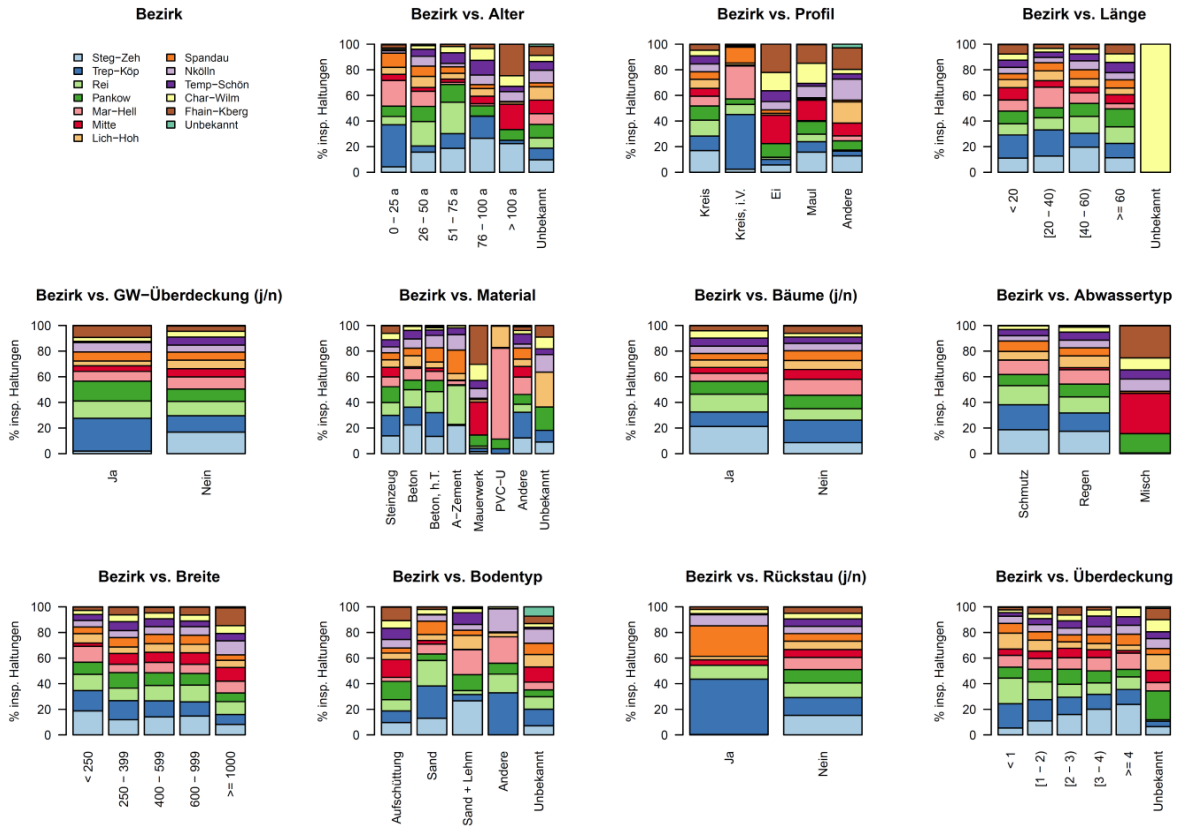
e) Material



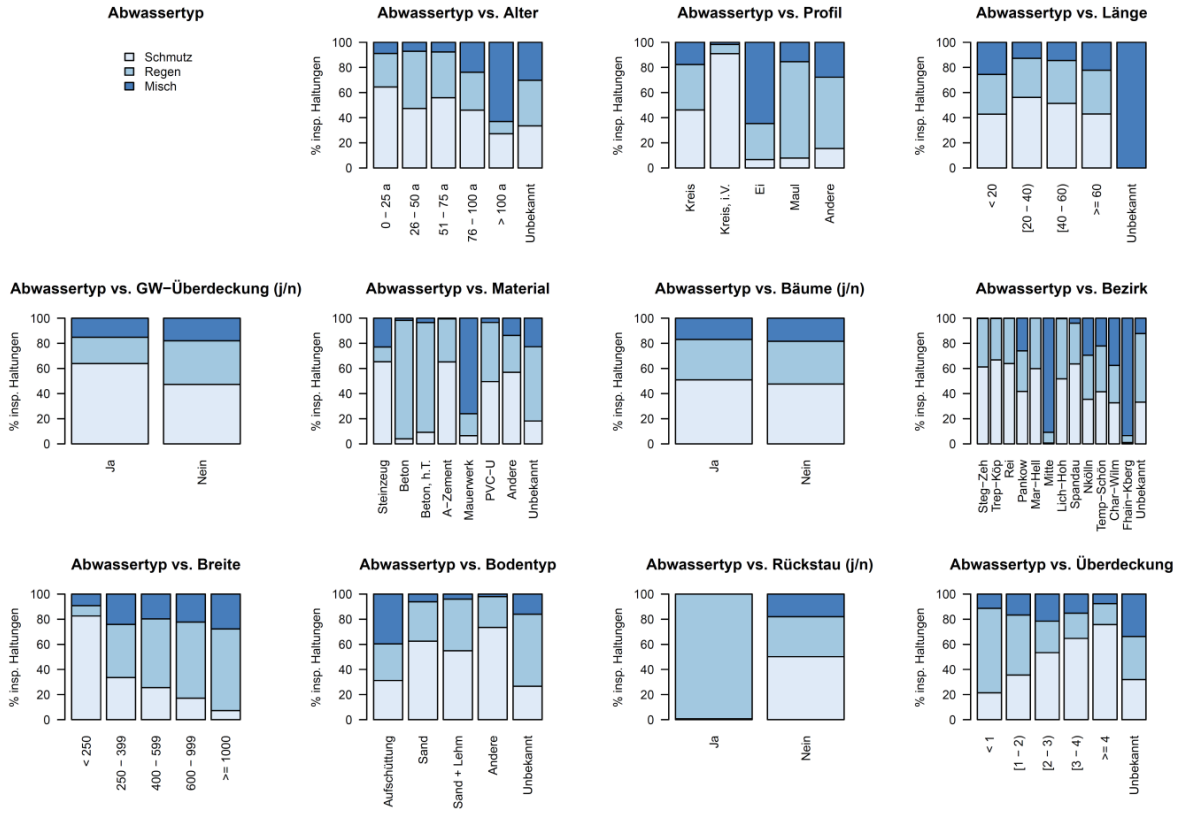
f) Bäume



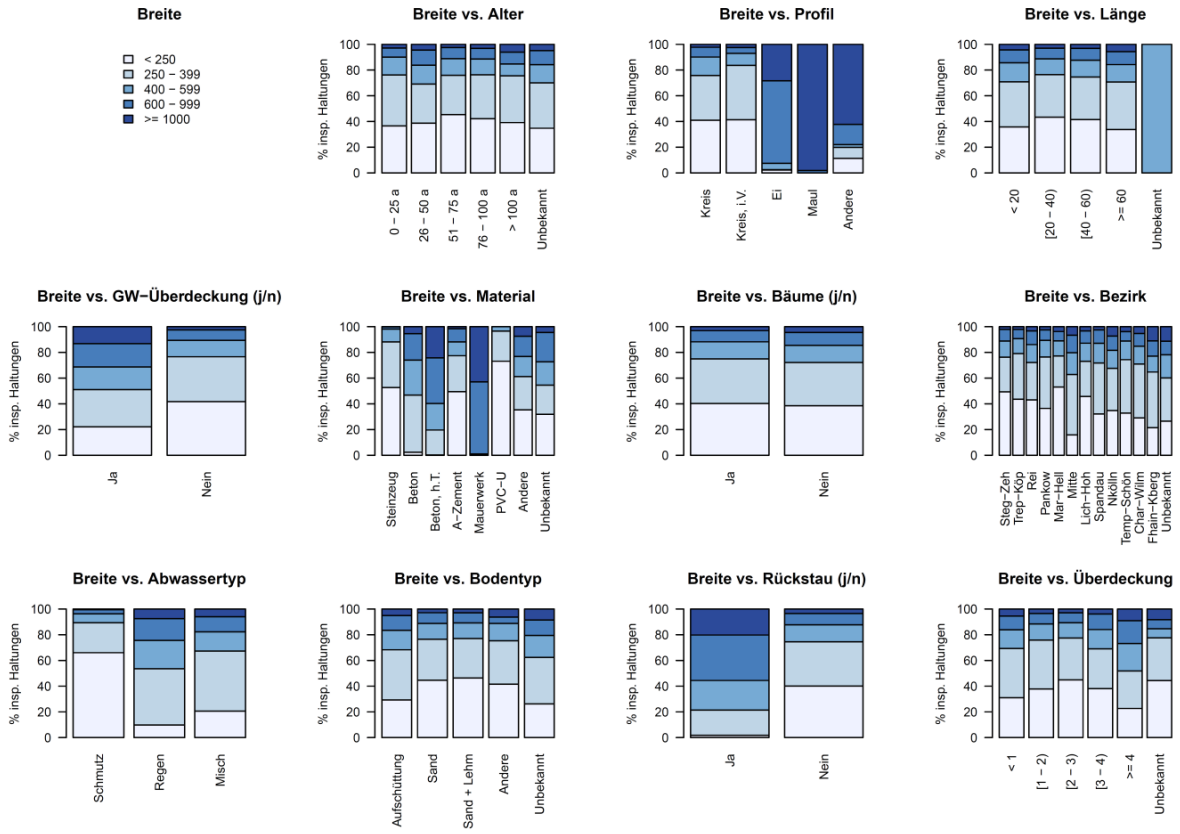
g) Bezirk



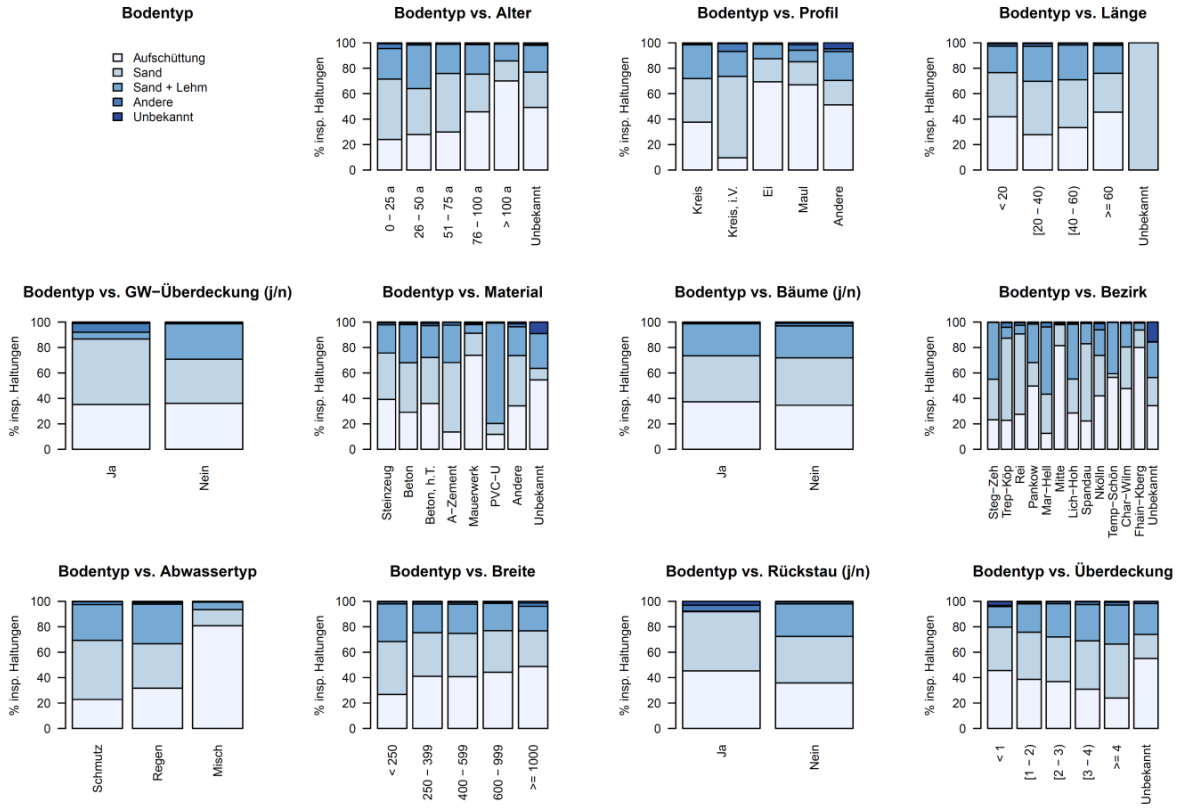
h) Abwassertyp



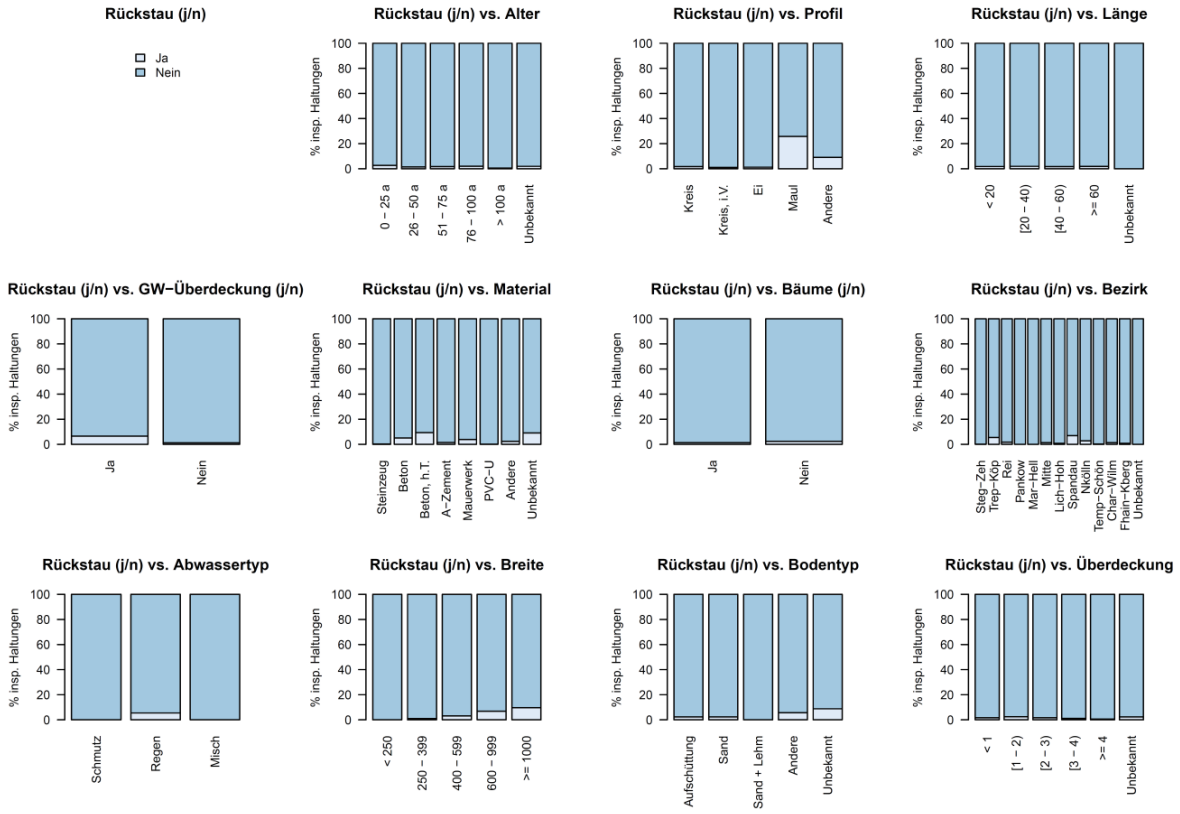
i) Breite



j) Bodentyp



k) Rückstau



l) Überdeckung

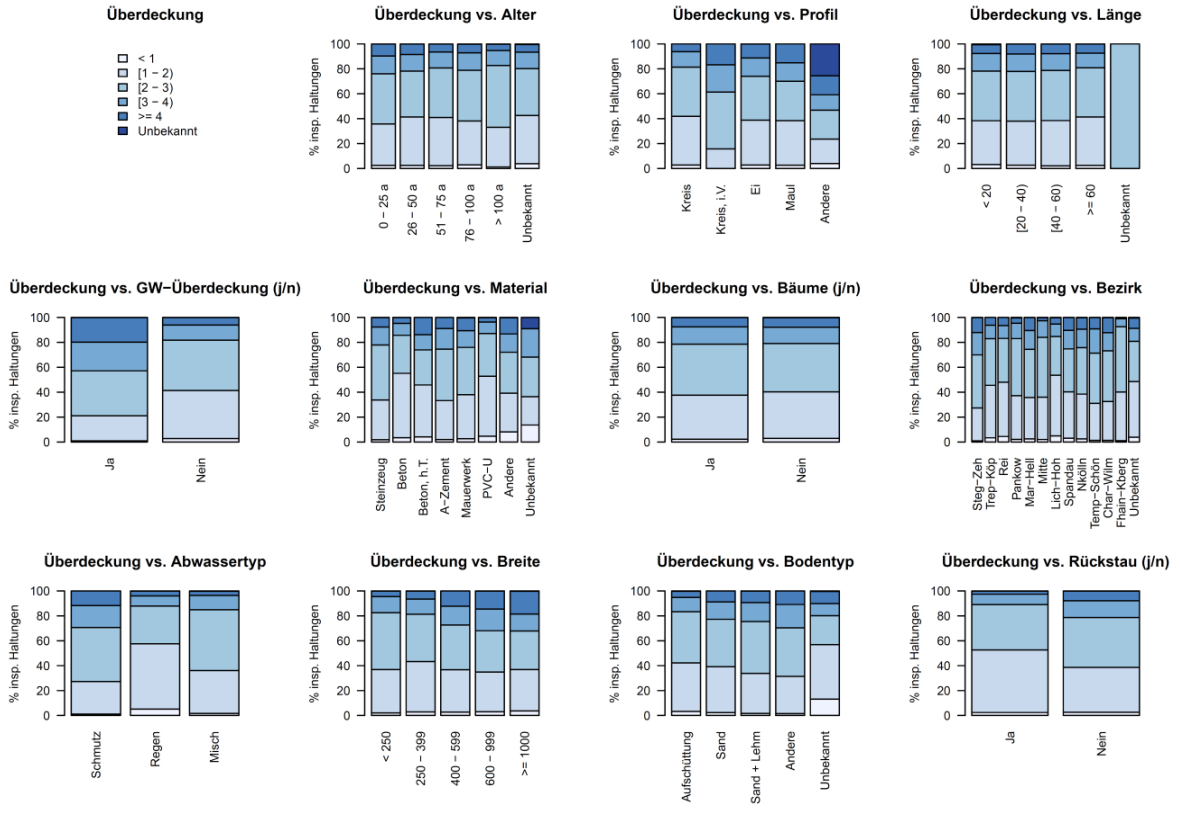


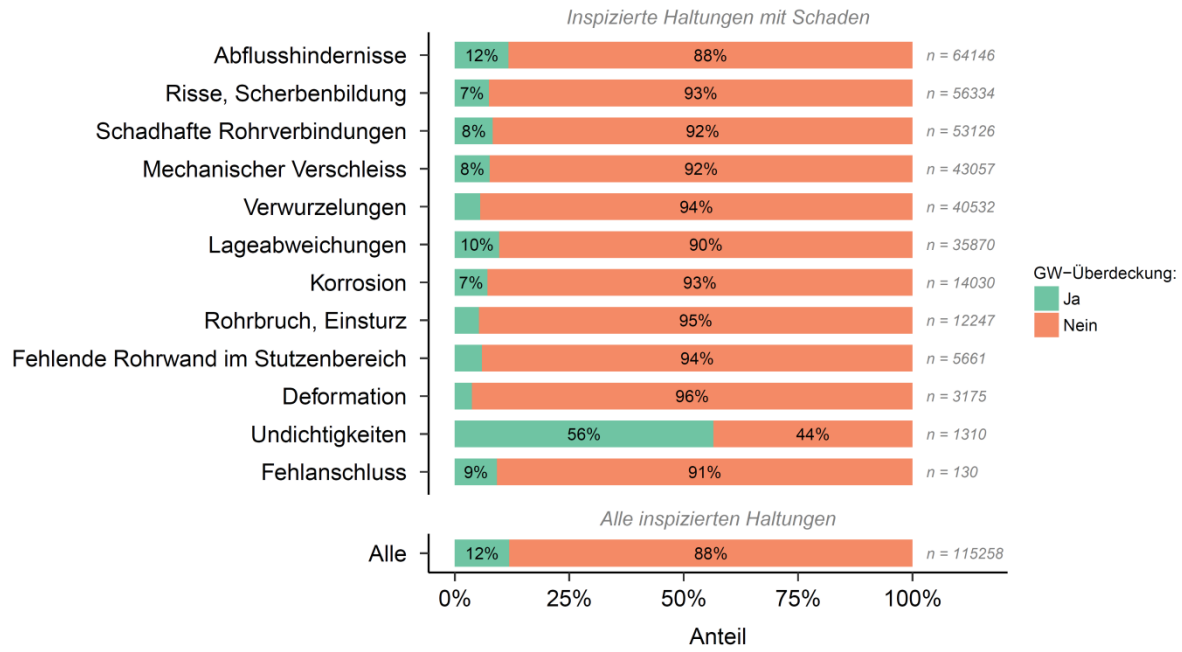
Abbildung 69: Datenverteilungen der Variablen untereinander

Anhang C: Schadenstypen und ihre Einflussfaktoren

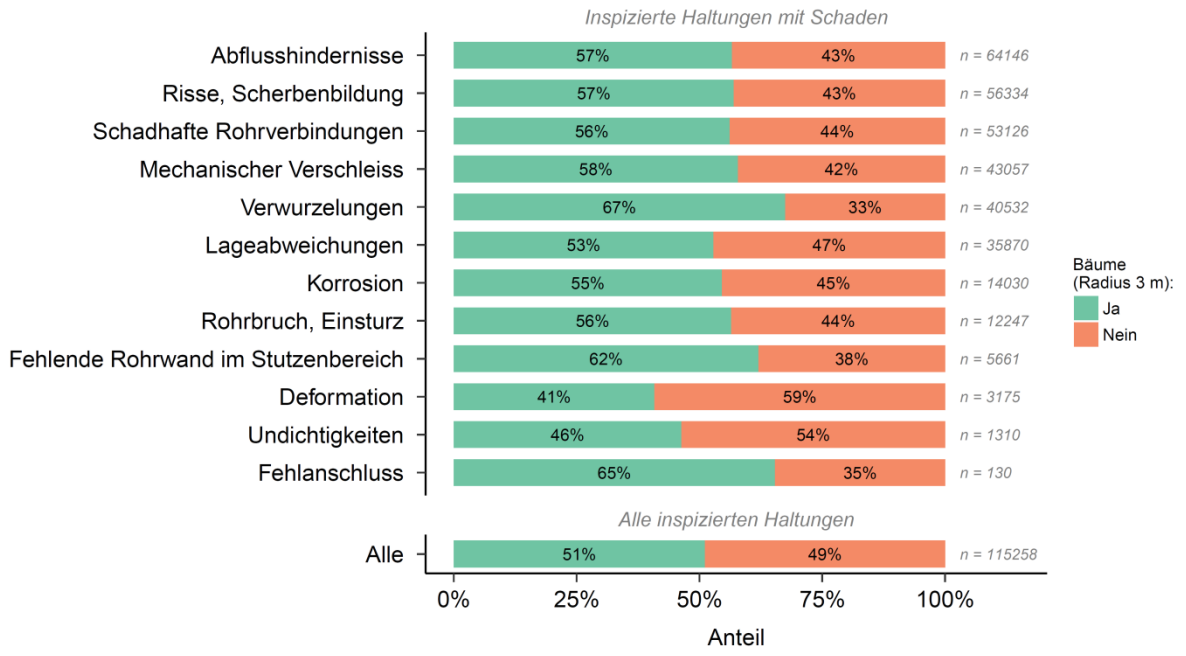
Tabelle 25: *Cramér's-V*-Werte für die wichtigsten Variablen und die zwölf Hauptschadenstypen

	Alter	Profil	Länge	GW-Überd. (j/n)	Material	Bäume (j/n)	Bezirk	Abwasser-typ	Breite	Bodentyp	Rückstau (j/n)	Überd.
Abflusshindernisse	0,32	0,19	0,30	0,01	0,15	0,12	0,14	0,04	0,07	0,08	0,04	0,01
Risse, Scherbenbildung	0,45	0,32	0,22	0,13	0,32	0,11	0,24	0,12	0,18	0,15	0,06	0,10
Schadhafte Rohrverbindungen	0,26	0,26	0,19	0,10	0,36	0,09	0,16	0,19	0,35	0,08	0,07	0,09
Mechanischer Verschleiss	0,25	0,17	0,17	0,10	0,32	0,10	0,18	0,17	0,25	0,10	0,06	0,07
Verwurzelungen	0,41	0,24	0,26	0,14	0,25	0,24	0,24	0,10	0,12	0,09	0,02	0,09
Lageabweichungen	0,13	0,16	0,13	0,05	0,27	0,02	0,07	0,20	0,28	0,05	0,06	0,06
Korrosion	0,21	0,12	0,08	0,05	0,59	0,03	0,18	0,38	0,24	0,05	0,04	0,11
Rohrbruch, Einsturz	0,17	0,12	0,10	0,07	0,12	0,04	0,08	0,02	0,12	0,06	0,02	0,09
Fehlende Rohrwand im Stutzenbereich	0,13	0,08	0,12	0,04	0,07	0,05	0,06	0,05	0,01	0,03	0,00	0,06
Deformation	0,11	0,07	0,04	0,04	0,57	0,03	0,23	0,03	0,08	0,13	0,02	0,04
Undichtigkeiten	0,04	0,05	0,03	0,15	0,04	0,01	0,06	0,03	0,08	0,03	0,01	0,08
Fehlanschluss	0,02	0,01	0,02	0,00	0,03	0,01	0,03	0,04	0,02	0,01	0,01	0,01
Mittelwert	*0,21*	*0,15*	*0,14*	*0,07*	*0,26*	*0,07*	*0,14*	*0,12*	*0,15*	*0,07*	*0,03*	*0,07*

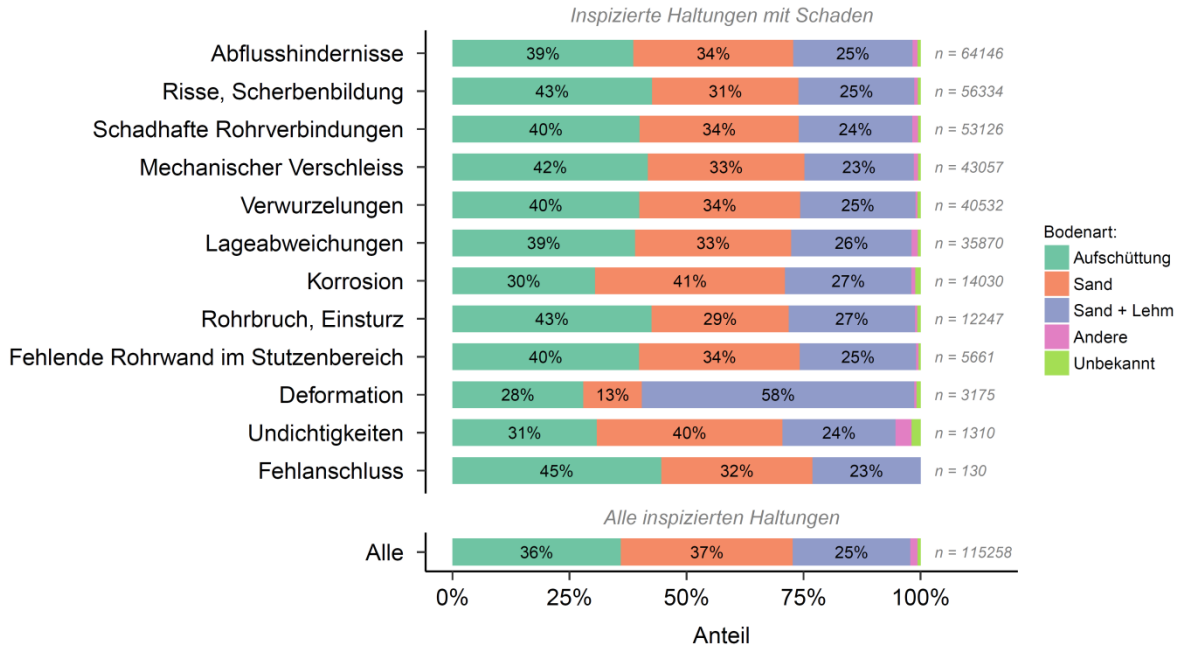
a) Grundwasserüberdeckung (j/n)



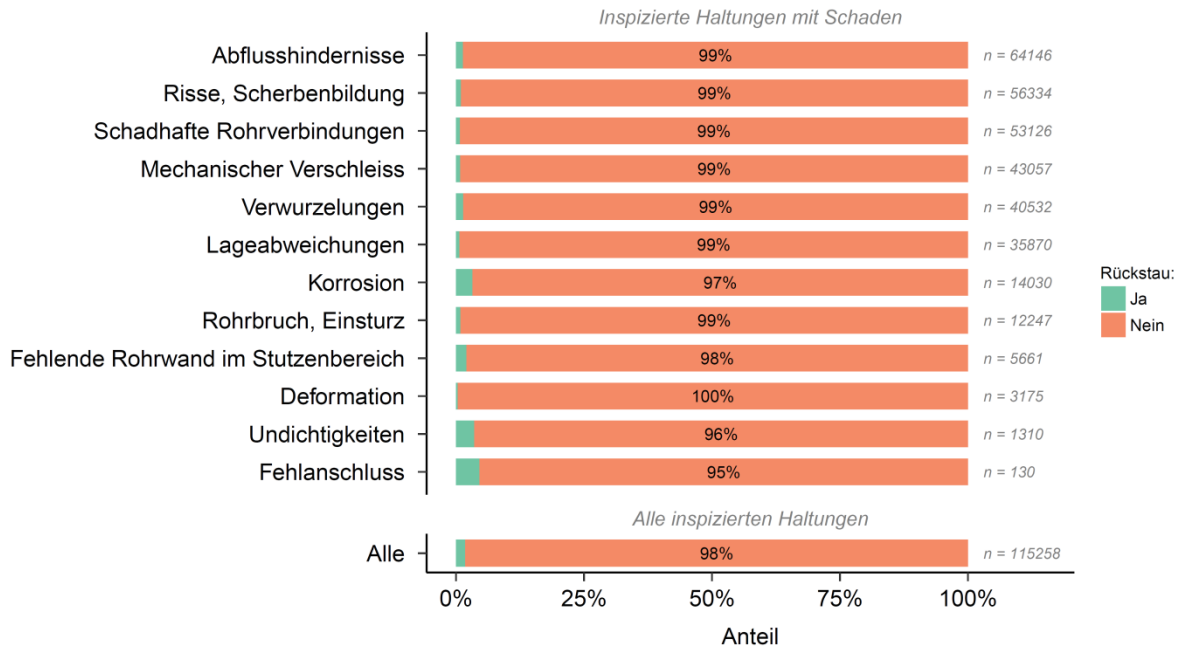
b) Bäume (j/n)



c) Bodentyp



d) Rückstau



e) Überdeckung

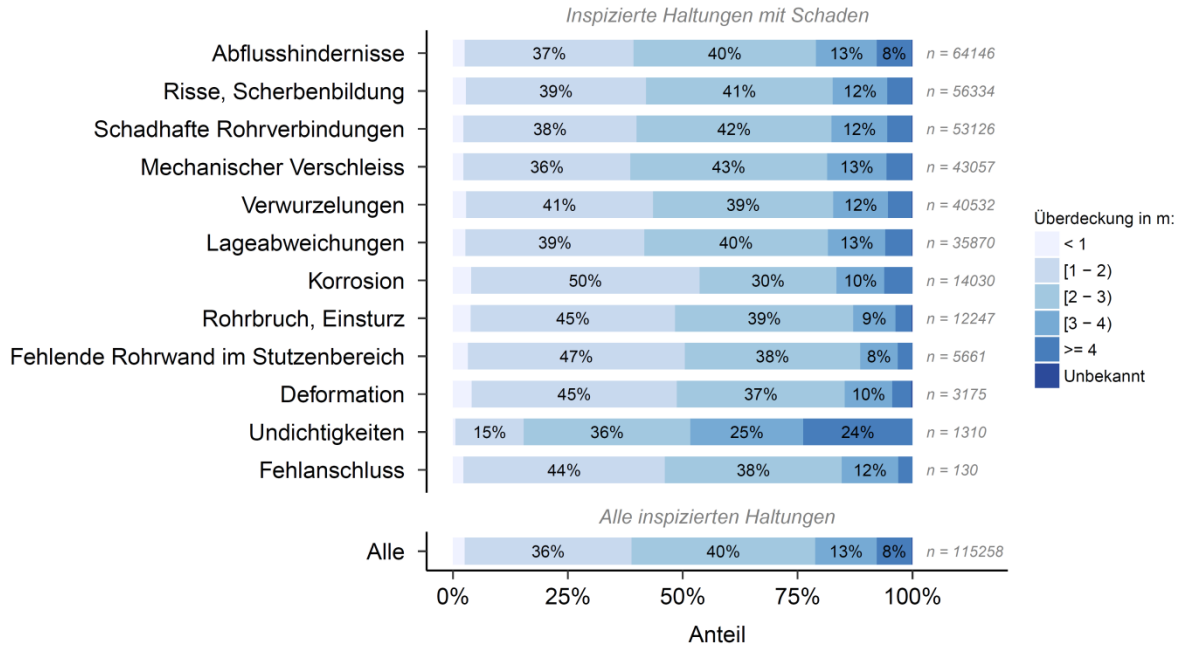


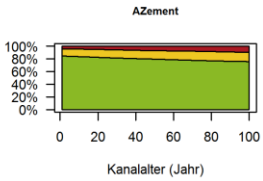
Abbildung 70: Datenverteilung für Grundwasserüberdeckung (a), Bäume (b), Bodentyp (c), Rückstau (d) und Überdeckung (e) für die Haltungen mit einem bestimmten Schaden (jeweils oben) und alle inspizierten Haltungen (jeweils unten)

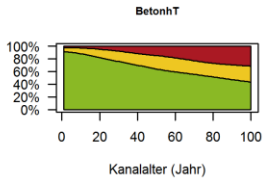
Anhang D: GompitZ: Kohorten und Überlebenskurven

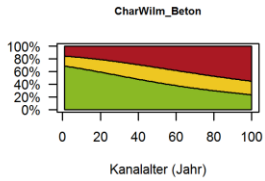
Tabelle 26. Übersicht über die 59 Kohorten, die unter Berücksichtigung der Variablen „Bezirk“, „Material“, „Profil“, „Baum (Ja/Nein)“ und „Bodenart“ gebildet wurden.

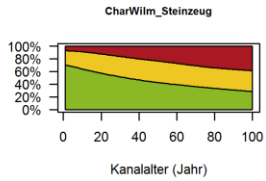
Nr.	Bezirk	Material	Profil	Baum (j/n)	Bodenart	Anzahl Halt.	Abbildungsbezeichnung
1	alle	alle	Kreis. i.V.	alle	alle	5717	Kreisi.v.
2	Trep-Köp	Steinzeug	alle	alle	alle	3717	TrepKöp_Steinzeug
3	alle	A-Zement	alle	alle	alle	3649	AZement
4	alle	Beton, h.T.	alle	alle	alle	3254	BetonhT
5	Mitte	alle	alle	alle	alle	2657	Mitte
6	Temp-Schön	Steinzeug	alle	alle	alle	2135	Temp-Schön_Steinzeug
7	Fhain-Kberg	alle	alle	alle	alle	2085	FhainKberg
8	alle	Mauerwerk	alle	alle	alle	2066	Mauerwerk
9	Char-Wilm	Steinzeug	alle	alle	alle	1901	CharWilm_Steinzeug
10	Steg-Zeh	Steinzeug	Kreis	Ja	Sand-Lehm	1854	StegZeh_Steinzeug _Kreis_Ja_SandLehm
11	Pankow	Steinzeug	Kreis	Ja	Aufschüttung	1568	Pankow_Steinzeug _Kreis_Ja_Aufschüttung
12	Rei	Steinzeug	Kreis	Ja	Sand	1565	Rei_Steinzeug _Kreis_Ja_Sand
13	alle	PVC-U	alle	alle	alle	1552	PVCU
14	Rei	Beton	alle	alle	alle	1484	Rei_Beton
15	Trep-Köp	Beton	alle	alle	alle	1422	TrepKöp_Beton
16	Steg-Zeh	Steinzeug	Kreis	Ja	Sand	1258	StegZeh_Steinzeug _Kreis_Ja_Sand
17	Rei	Steinzeug	Kreis	Nein	alle	1227	Rei_Steinzeug_Kreis_Nein
18	Pankow	Steinzeug	Kreis	Nein	Aufschüttung	985	Pankow_Steinzeug _Kreis_Nein_Aufschüttung
19	Steg-Zeh	Steinzeug	Kreis	Ja	Aufschüttung	922	StegZeh_Steinzeug _Kreis_Ja_Aufschüttung
20	Mar-Hell	Beton	alle	alle	alle	907	MarHell_Beton
21	Lich-Hoh	Beton	alle	alle	alle	795	LichHoh_Beton
22	Steg-Zeh	Beton	Kreis	Ja	Sand-Lehm	777	StegZeh_Beton_Kreis _Ja_SandLehm
23	Mar-Hell	Steinzeug	Kreis	Nein	Sand-Lehm	694	MarHell_Steinzeug _Kreis_Nein_SandLehm
24	Pankow	Beton	alle	alle	alle	679	Pankow_Beton
25	Spandau	Steinzeug	Kreis	Ja	alle	679	Spandau_Steinzeug_Kreis_Ja
26	Spandau	Steinzeug	Kreis	Nein	alle	669	Spandau_Steinzeug _Kreis_Nein
27	Nkölln	Beton	alle	alle	alle	653	Nkölln_Beton
28	Steg-Zeh	Steinzeug	Kreis	Nein	Sand-Lehm	643	StegZeh_Steinzeug _Kreis_Nein_SandLehm
29	Spandau	Beton	alle	alle	alle	630	Spandau_Beton
30	Rei	Steinzeug	Kreis	Ja	Aufschüttung	626	Rei_Steinzeug_Kreis _Ja_Aufschüttung
31	Temp-Schön	Beton	alle	alle	alle	624	Temp-Schön_Beton
32	Pankow	Steinzeug	Kreis	Nein	Sand-Lehm	621	Pankow_Steinzeug _Kreis_Nein_SandLehm
33	Pankow	Steinzeug	Kreis	Ja	Sand-Lehm	573	Pankow_Steinzeug _Kreis_Ja_SandLehm
34	Steg-Zeh	Beton	Kreis	Ja	Sand	569	StegZeh_Beton _Kreis_Ja_Sand

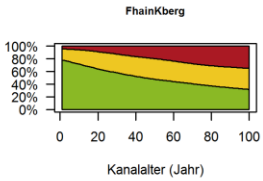
Nr.	Bezirk	Material	Profil	Baum (j/n)	Bodenart	Anzahl Halt.	Abbildungsbezeichnung
35	Nkölln	Steinzeug	Kreis	Ja	Aufschüttung	544	Nkölln_Steinzeug _Kreis_Ja_Aufschüttung
36	Pankow	Steinzeug	Kreis	Nein	Sand	439	Pankow_Steinzeug _Kreis_Nein_Sand
37	Pankow	Steinzeug	Kreis	Ja	Sand	428	Pankow_Steinzeug _Kreis_Ja_Sand
38	Mar-Hell	Steinzeug	Kreis	Ja	alle	426	MarHell_Steinzeug_Kreis_Ja
39	Steg-Zeh	Steinzeug	Kreis	Nein	Sand	416	StegZeh_Steinzeug _Kreis_Nein_Sand
40	Lich-Hoh	Steinzeug	Kreis	Ja	Sand-Lehm	392	LichHoh_Steinzeug _Kreis_Ja_SandLehm
41	Lich-Hoh	Steinzeug	Kreis	Nein	Sand-Lehm	391	LichHoh_Steinzeug _Kreis_Nein_SandLehm
42	Steg-Zeh	Beton	Kreis	Ja	Aufschüttung	374	StegZeh_Beton_Kreis _Ja_Aufschüttung
43	Lich-Hoh	Steinzeug	Kreis	Ja	Aufschüttung	369	LichHoh_Steinzeug _Kreis_Ja_Aufschüttung
44	Steg-Zeh	Steinzeug	Kreis	Nein	Aufschüttung	366	StegZeh_Steinzeug _Kreis_Nein_Aufschüttung
45	Lich-Hoh	Steinzeug	Kreis	Nein	Sand	351	LichHoh_Steinzeug _Kreis_Nein_Sand
46	Steg-Zeh	Beton	Kreis	Nein	Sand-Lehm	347	StegZeh_Beton_Kreis _Nein_SandLehm
47	Lich-Hoh	Steinzeug	Kreis	Ja	Sand	255	LichHoh_Steinzeug _Kreis_Ja_Sand
48	Char-Wilm	Beton	alle	alle	alle	253	CharWilm_Beton
49	Lich-Hoh	Steinzeug	Kreis	Nein	Aufschüttung	234	LichHoh_Steinzeug _Kreis_Nein_Aufschüttung
50	Nkölln	Steinzeug	Kreis	Ja	Sand	219	Nkölln_Steinzeug _Kreis_Ja_Sand
51	Steg-Zeh	Beton	Kreis	Nein	Sand	219	StegZeh_Beton _Kreis_Nein_Sand
52	Nkölln	Steinzeug	Kreis	Nein	Aufschüttung	218	Nkölln_Steinzeug _Kreis_Nein_Aufschüttung
53	Rei	Steinzeug	Kreis	Ja	Sand-Lehm	192	Rei_Steinzeug _Kreis_Ja_SandLehm
54	Nkölln	Steinzeug	Kreis	Nein	Sand	189	Nkölln_Steinzeug _Kreis_Nein_Sand
55	Mar-Hell	Steinzeug	Kreis	Nein	Aufschüttung	186	MarHell_Steinzeug _Kreis_Nein_Aufschüttung
56	Steg-Zeh	Beton	Kreis	Nein	Aufschüttung	177	StegZeh_Beton_Kreis _Nein_Aufschüttung
57	Mar-Hell	Steinzeug	Kreis	Nein	Sand	158	MarHell_Steinzeug_Kreis _Nein_Sand
58	Nkölln	Steinzeug	Kreis	Nein	Sand-Lehm	122	Nkölln_Steinzeug _Kreis_Nein_SandLehm
59	Nkölln	Steinzeug	Kreis	Ja	Sand-Lehm	106	Nkölln_Steinzeug _Kreis_Ja_SandLehm

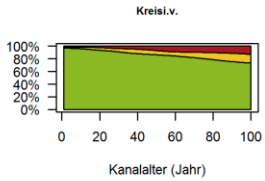

AZement
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

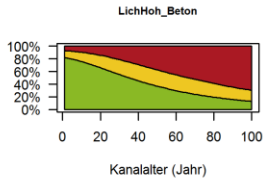

BetonhT
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

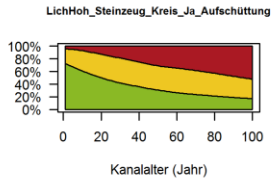

CharWilm_Beton
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

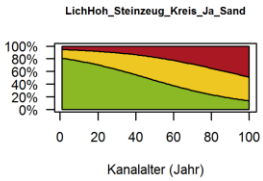

CharWilm_Steinzeug
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

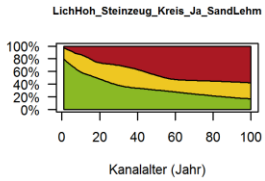

FhainKberg
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

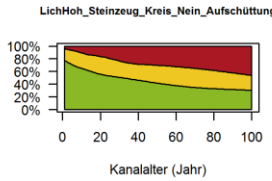

Kreisi.v.
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

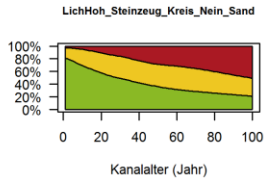

LichHoh_Beton
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

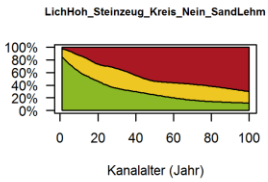

LichHoh_Steinzeug_Kreis_Ja_Aufschüttung
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

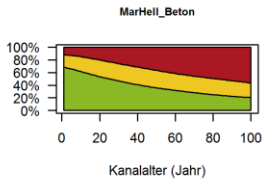

LichHoh_Steinzeug_Kreis_Ja_Sand
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

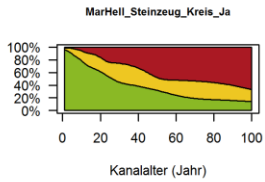

LichHoh_Steinzeug_Kreis_Ja_SandLehm
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

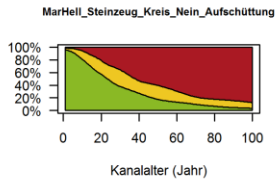

LichHoh_Steinzeug_Kreis_Nein_Aufschüttung
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

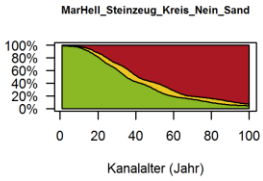

LichHoh_Steinzeug_Kreis_Nein_Sand
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

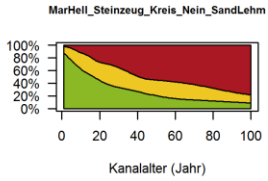

LichHoh_Steinzeug_Kreis_Nein_SandLehm
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

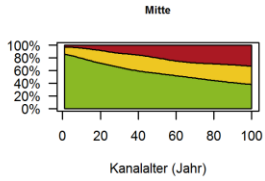

MarHell_Beton
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

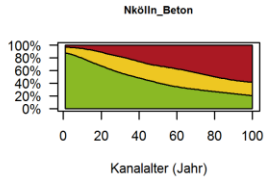

MarHell_Steinzeug_Kreis_Ja
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

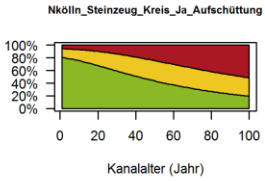

MarHell_Steinzeug_Kreis_Nein_Aufschüttung
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

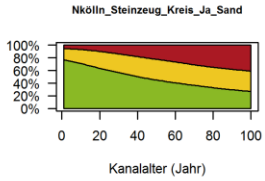

MarHell_Steinzeug_Kreis_Nein_Sand
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

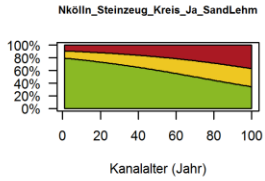

MarHell_Steinzeug_Kreis_Nein_SandLehm
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

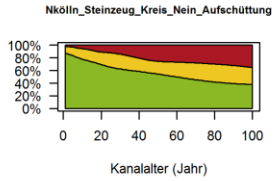

Mitte
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

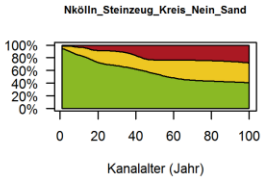

Nkölln_Beton
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

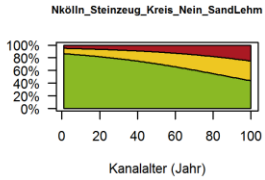

Nkölln_Steinzeug_Kreis_Ja_Aufschüttung
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

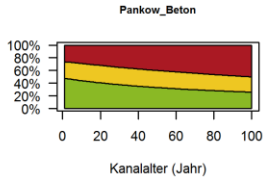

Nkölln_Steinzeug_Kreis_Ja_Sand
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

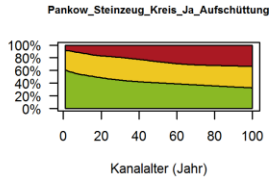

Nkölln_Steinzeug_Kreis_Ja_SandLehm
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

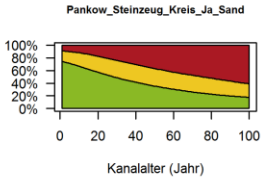

Nkölln_Steinzeug_Kreis_Nein_Aufschüttung
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

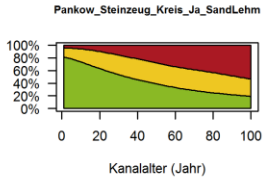

Nkölln_Steinzeug_Kreis_Nein_Sand
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

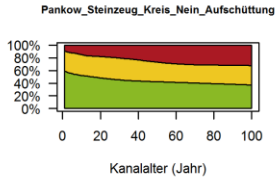

Nkölln_Steinzeug_Kreis_Nein_SandLehm
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

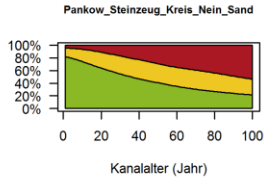

Pankow_Beton
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)


Pankow_Steinzeug_Kreis_Ja_Aufschüttung
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)


Pankow_Steinzeug_Kreis_Ja_Sand
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)


Pankow_Steinzeug_Kreis_Ja_SandLehm
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)


Pankow_Steinzeug_Kreis_Nein_Aufschüttung
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)


Pankow_Steinzeug_Kreis_Nein_Sand
100%
80%
60%
40%
20%
0%
0 20 40 60 80 100
Kanalalter (Jahr)

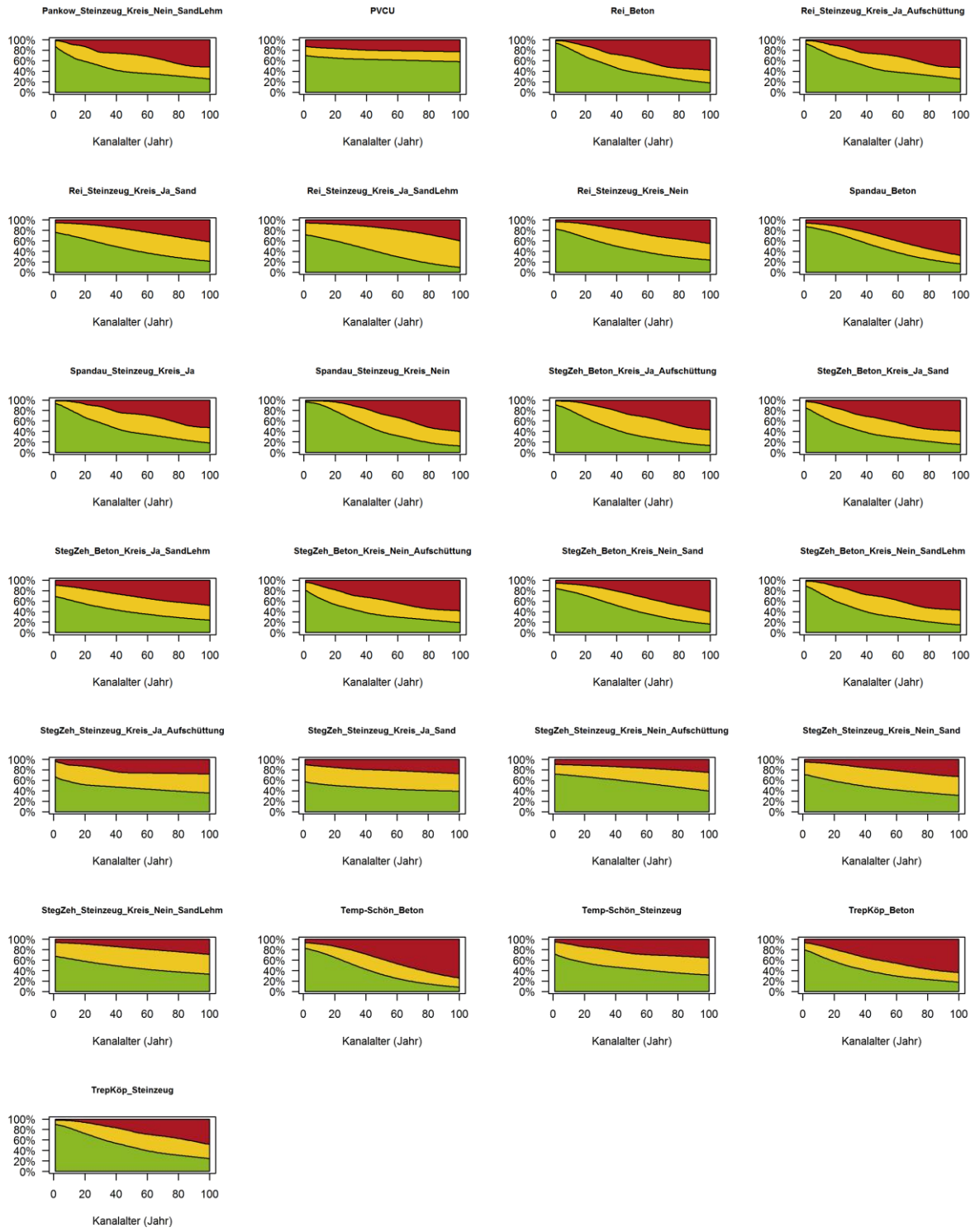


Abbildung 71. Darstellung der Überlebenskurven für alle Kohorten (nur für 57 von 59 Kohorten). Für zwei Kohorten („Mauerwerk“; „StegZeh_Steinzeug_Kreis_Ja_SandLehm“) war die Kalibrierung nicht erfolgreich (keine Konvergenz bei der Maximum-Likelihood-Methode). Für diese Kohorten wurden die Überlebenskurven der nächsten konvergierenden Kohorte übernommen.

Anhang E: Einfluss der Parameter auf Modellgüte

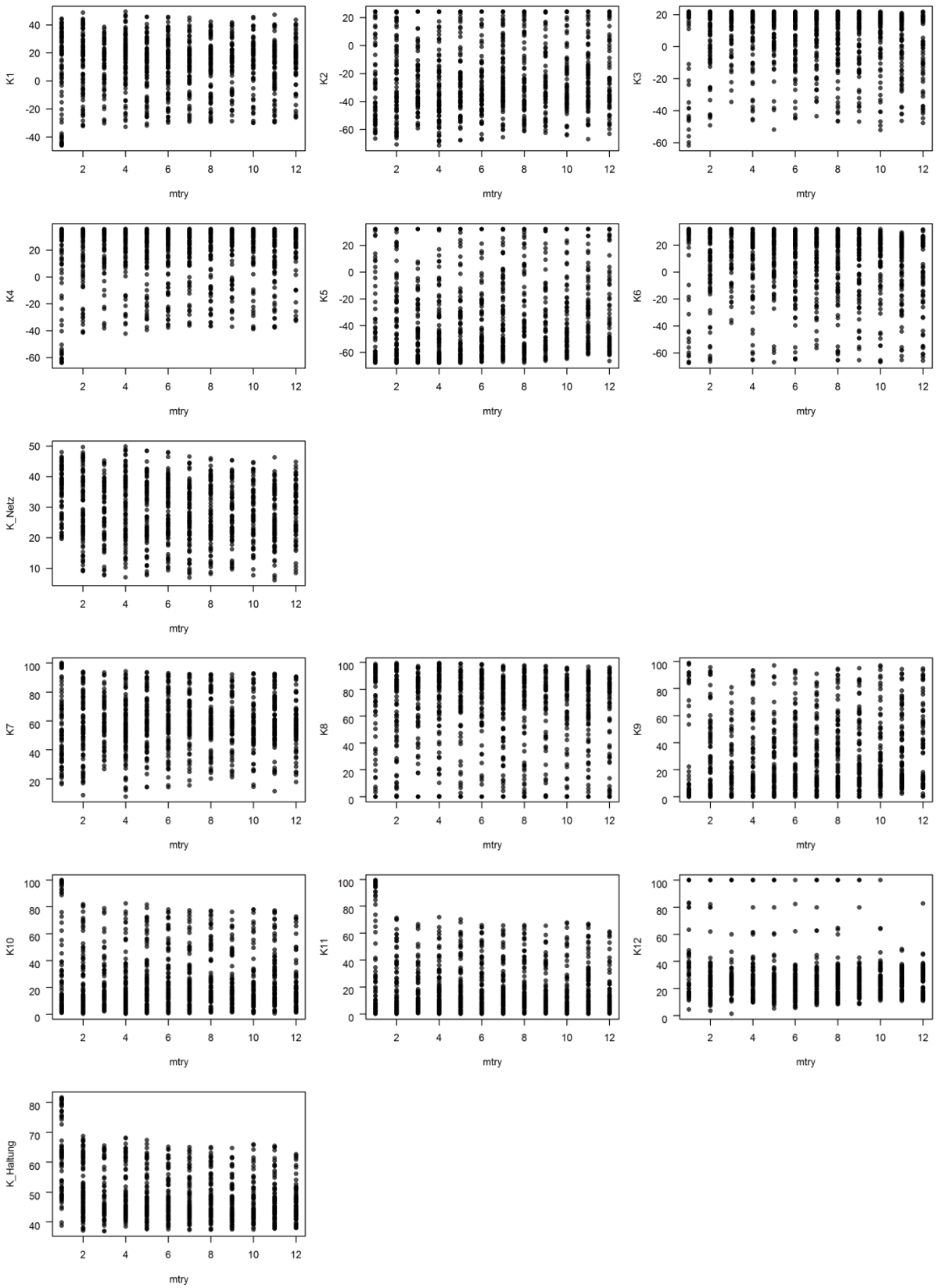


Abbildung 72: Random Forest - Einfluss des Parameters *mtry* auf die Güteindikatoren auf Netz- und Haltungsebene (Parameterwerte für *mtry*, *nodesize*, *ntrees*, *w1* und *w2* zufällig variiert).

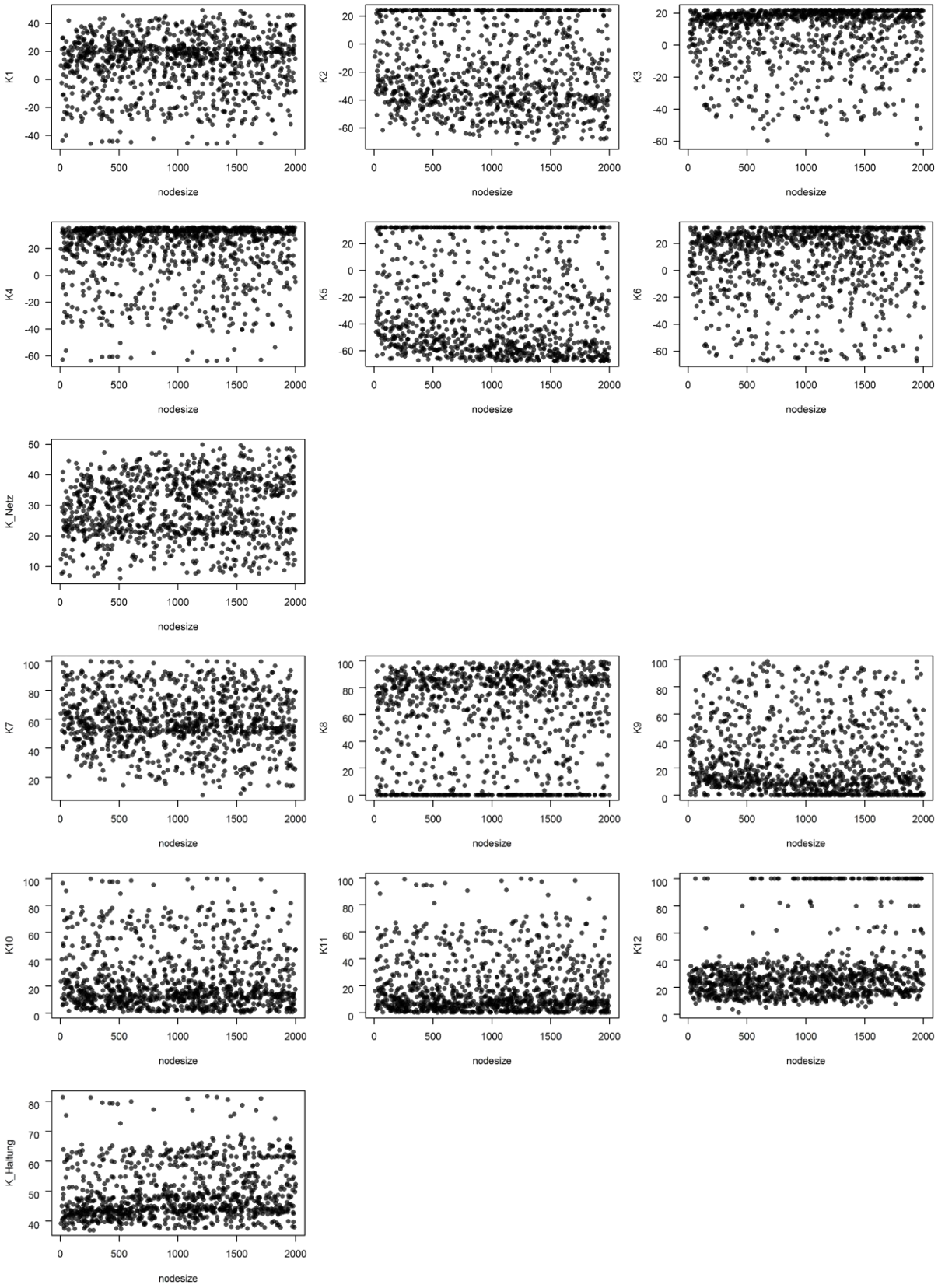


Abbildung 73: Random Forest - Einfluss des Parameters *nodesize* auf die Güteindikatoren auf Netz- und Haltungsebene (Parameterwerte für *mtry*, *nodesize*, *ntrees*, *w1* und *w2* zufällig variiert).

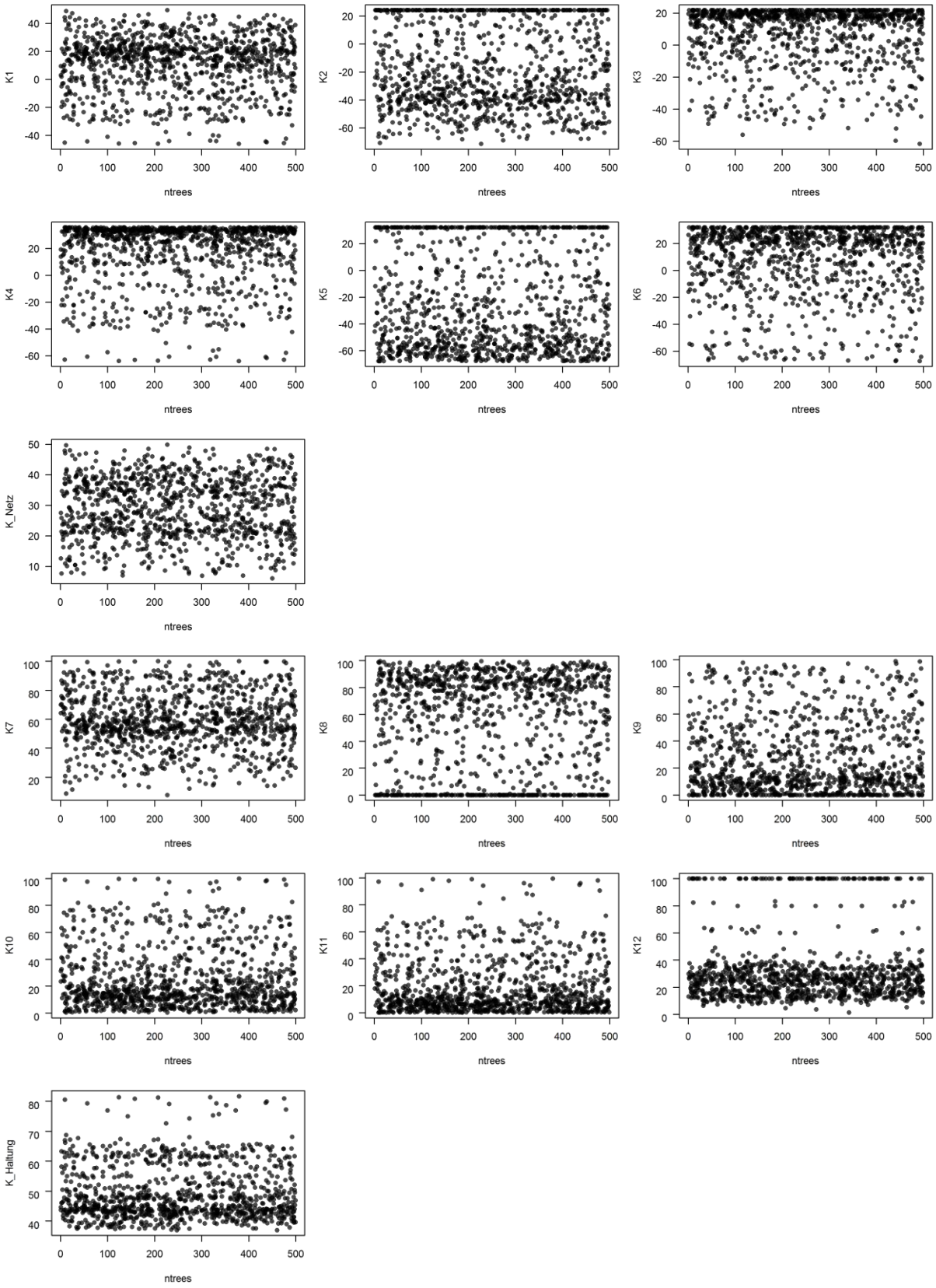


Abbildung 74: Random Forest - Einfluss des Parameters *ntrees* auf die Güteindikatoren auf Netz- und Haltungsebene (Parameterwerte für *mtry*, *nodesize*, *ntrees*, *w1* und *w2* zufällig variiert).

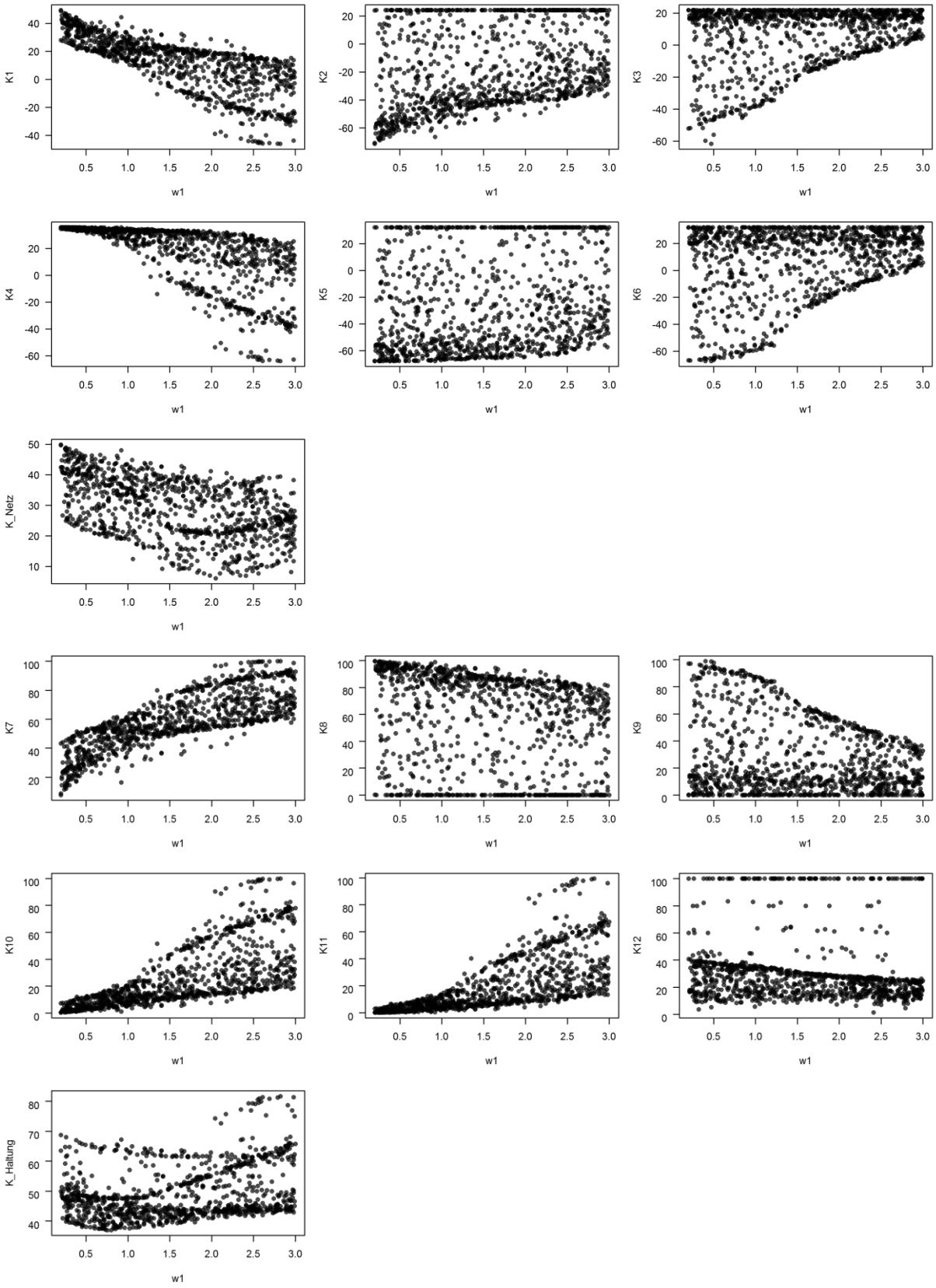


Abbildung 75: Random Forest - Einfluss des Parameters *w1* auf die Güteindikatoren auf Netz- und Haltungsebene (Parameterwerte für *mtry*, *nodesize*, *ntrees*, *w1* und *w2* zufällig variiert).

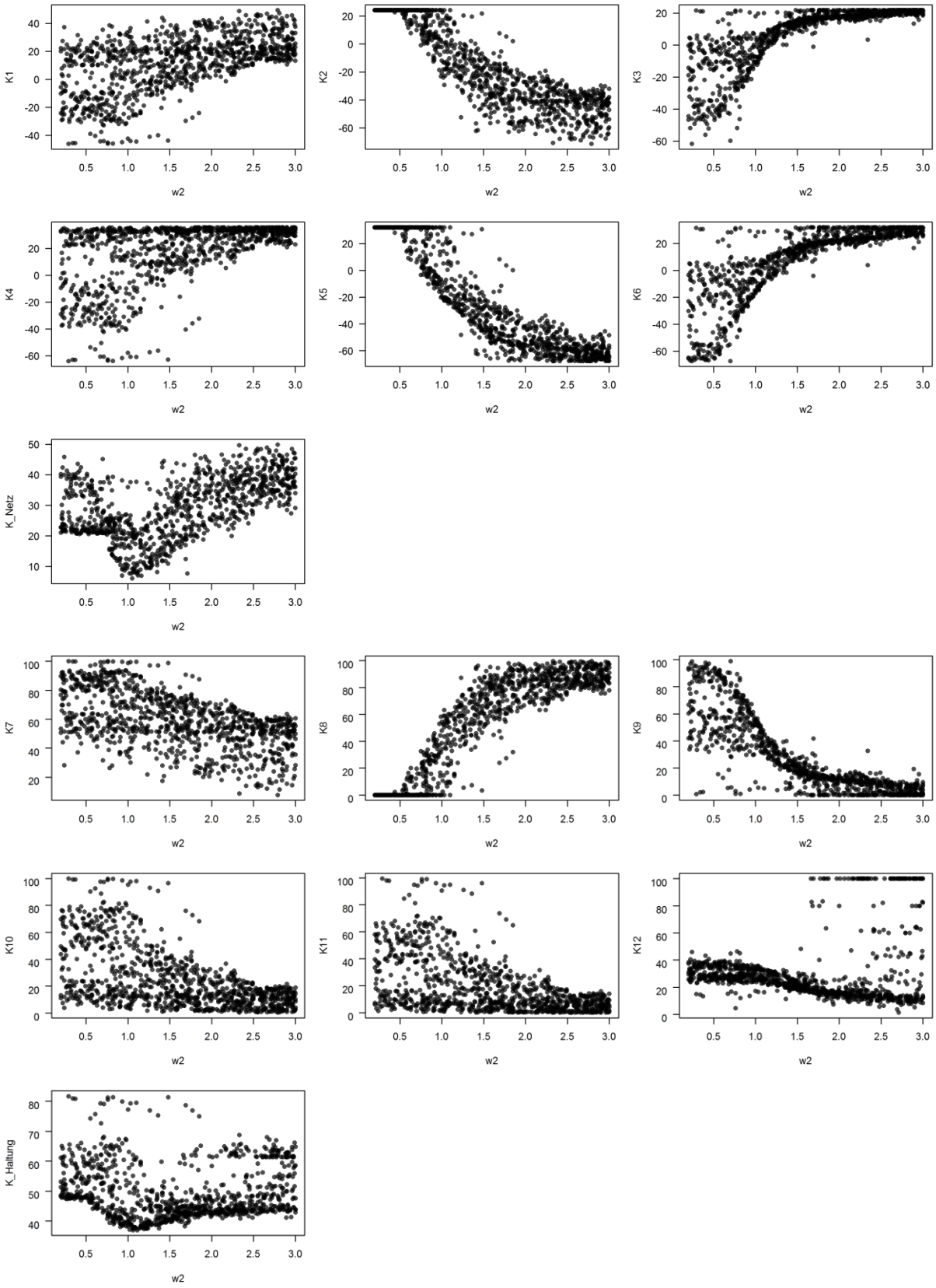


Abbildung 76: Random Forest - Einfluss des Parameters *w2* auf die Güteindikatoren auf Netz- und Haltungsebene (Parameterwerte für *mtry*, *nodesize*, *ntrees*, *w1* und *w2* zufällig variiert).

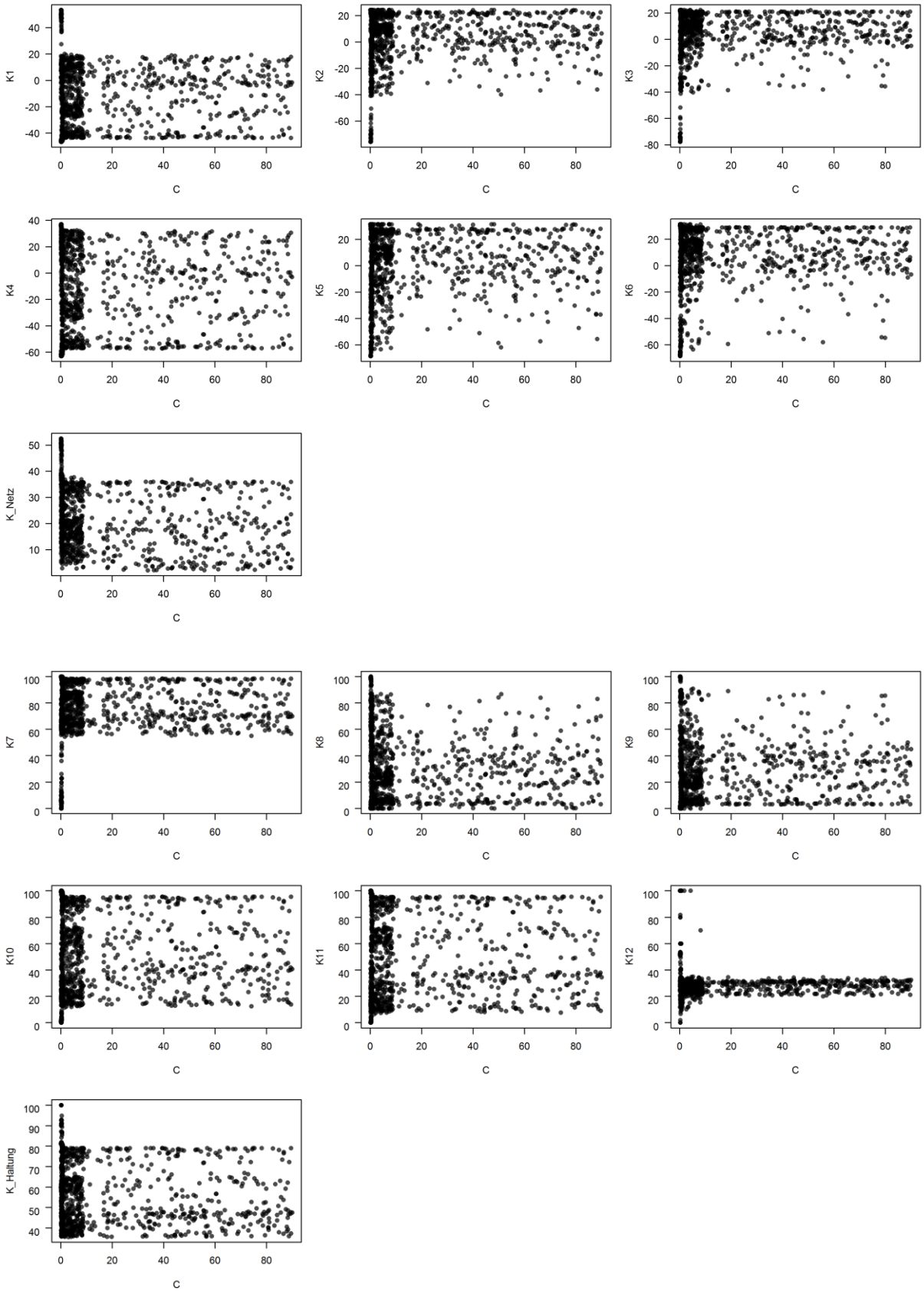


Abbildung 77: Support Vector Machine - Einfluss des Parameters *C* auf die Güteindikatoren auf Netz- und Haltungsebene (Parameterwerte für *C*, *sigma*, *w2* und *w3* zufällig variiert).

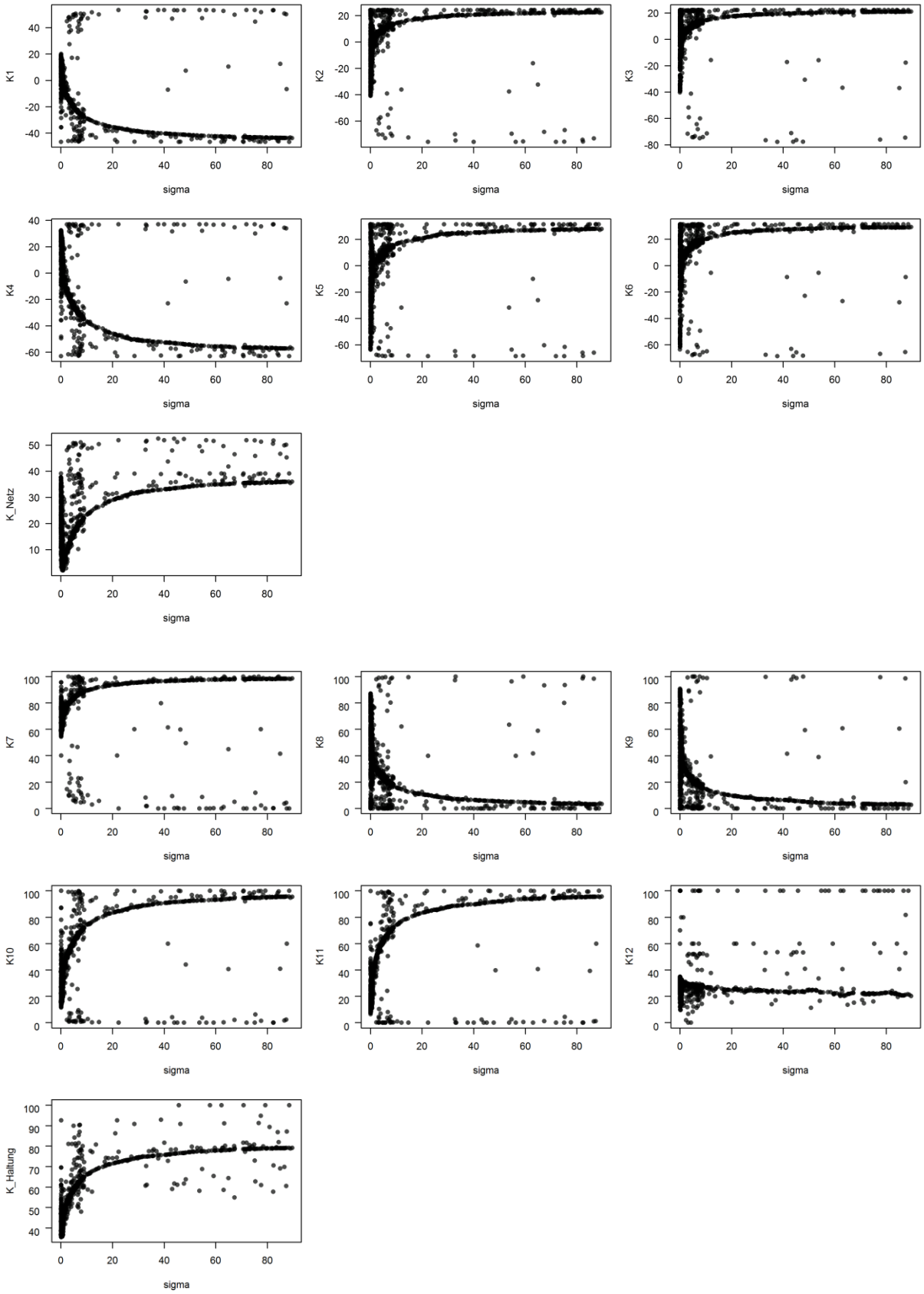


Abbildung 78: Support Vector Machine - Einfluss des Parameters *sigma* auf die Güteindikatoren auf Netz- und Haltungsebene (Parameterwerte für *C*, *sigma*, *w2* und *w3* zufällig variiert).

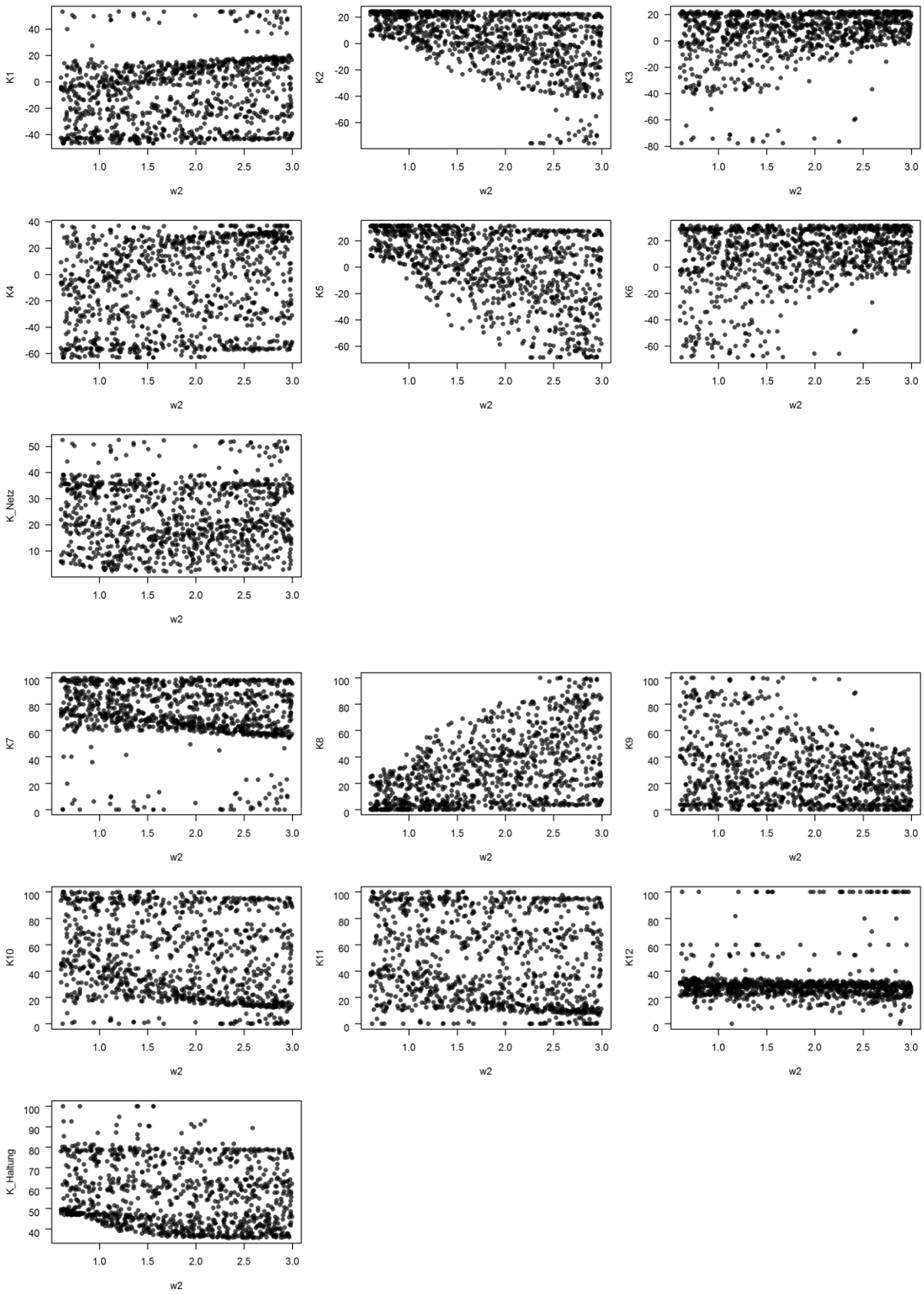


Abbildung 79: Support Vector Machine - Einfluss des Parameters *w2* auf die Güteindikatoren auf Netz- und Haltungsebene (Parameterwerte für *C*, *sigma*, *w2* und *w3* zufällig variiert).

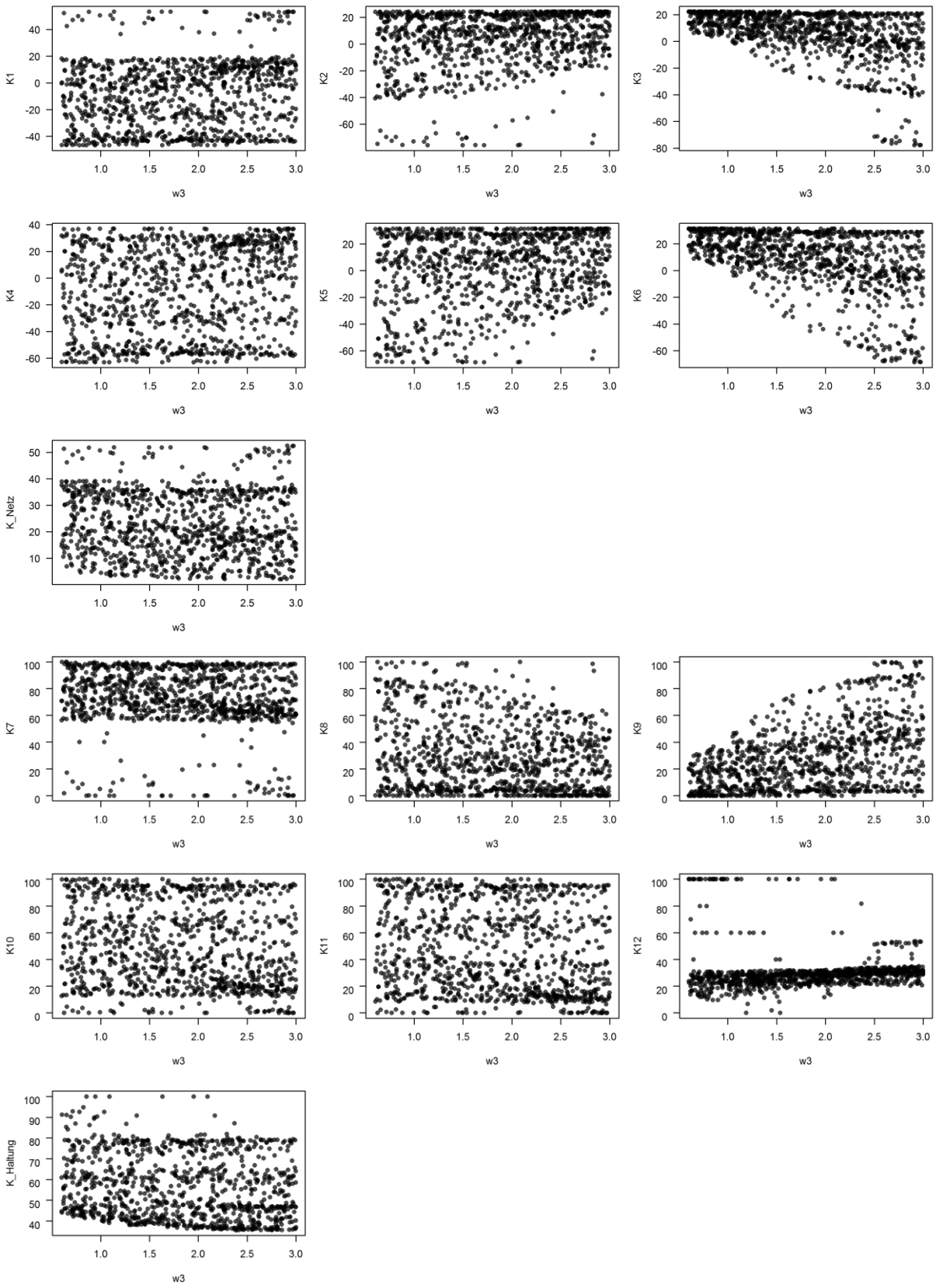


Abbildung 80: Support Vector Machine - Einfluss des Parameters *w3* auf die Güteindikatoren auf Netz- und Haltungsebene (Parameterwerte für *C*, *sigma*, *w2* und *w3* zufällig variiert).

Anhang F: Zusammengefasste Modellergebnisse

Tabelle 27: Bewertungsindikatoren für alle untersuchten Modellansätze im Vergleich. Für Random Forest (RF) und Support Vector Machine (SVM) wurden für Netz- und Haltungsebene unterschiedliche Modellparametrisierungen verwendet. KNN = Künstliche Neuronale Netze.

Indikatoren auf Netzebene							
Modell	**K1** (Ziel → ±0)	**K2** (Ziel → ±0)	**K3** (Ziel → ±0)	**K4** (Ziel → ±0)	**K5** (Ziel → ±0)	**K6** (Ziel → ±0)	**K_Netz** (Ziel → min)
GompitZ	0,4	0,0	-0,4	-0,4	-0,3	0,7	0,4
RF (A)	-3,5	3,4	0,1	2,0	-0,1	-1,9	2,3
SVM (A)	0,1	-0,3	0,2	3,1	-2,4	-0,7	1,6
KNN	-0,5	1,3	-0,8	17,6	-17,0	-0,6	10,0
Indikatoren auf Haltungsebene							
Modell	**K7** (Ziel → max)	**K8** (Ziel → max)	**K9** (Ziel → max)	**K10** (Ziel → min)	**K11** (Ziel → min)	**K12** (Ziel → min)	**K_Haltung** (Ziel → min)
GompitZ	64,4	29,5	33,1	43,0	38,1	38,2	50,8
RF (B)	64,0	40,0	66,7	17,1	9,5	28,3	34,6
SVM (B)	65,5	38,9	55,9	23,2	15,3	29,5	37,7
KNN	73,7	35,1	41,3	36,0	28,5	29,0	43,4

Tabellenverzeichnis

Tabelle 1: Filterschritte zur Entfernung nicht zu bewertender Datensätze 7

Tabelle 2: Auflistung der entfernten Datenbankeinträge 8

Tabelle 3: Eingangsvariablen für die statistische Analyse und die Modellierung 9

Tabelle 4: Berliner Schadenstypen nach Schadenskatalog 11 (BWB 2001) 30

Tabelle 5: Angewendete Filterschritte und Anzahl der verbleibenden Datensätze für die Unsicherheitsanalyse 39

Tabelle 6: Einfluss der Kameratyp auf die Abweichungen zwischen ersten und zweiten Inspektionen. Datengrundlage nach Anwendung des Filterschrittes 4 (siehe Tabelle 1). Die Dreh-Schwenkkopf-Kamera ist auch als Video-Kamera, der Kugelbildscanner als Panoramo-Foto-Kamera bekannt. 43

Tabelle 7: Doppelinspektionsmatrix *N* 47

Tabelle 8: Unsicherheitsmatrix *P* 48

Tabelle 9: Filterschritte für die Vorbereitung der Daten für die Modellierung 58

Tabelle 10: Konfigurationen bei Kalibrierung / Training der untersuchten Modelle 62

Tabelle 11: Übersicht der Indikatoren auf Netz- und Haltungsebene 65

Tabelle 12: Zusammengefasste Ergebnisse (*K_Netz* und *K_Haltung*) für die untersuchten Varianten der Kohortenbildung 66

Tabelle 13: Kreuztabelle mit den Häufigkeiten der drei Zustandsbereiche für Inspektion (Zeilen) und Prognose mit GompitZ (Spalten) 70

Tabelle 14: Bewertungsindikatoren für GompitZ 70

Tabelle 15: Modellparameter für Modell A (für Simulation auf Netzebene) und Modell B (für Simulation auf Haltungsebene) basierend auf Random Forest 75

Tabelle 16: Kreuztabelle mit den Häufigkeiten der drei Zustandsbereiche für Inspektion (Zeilen) und Prognose (Spalten) mit Random Forest (Modell B) 77

Tabelle 17: Bewertungsindikatoren für Random Forest 77

Tabelle 18: Modellparameter für Modell A (für Simulation auf Netzebene) und Modell B (für Simulation auf Haltungsebene) basierend auf Support Vector Machine 82

Tabelle 19: Kreuztabelle mit den Häufigkeiten der drei Zustandsbereiche für Inspektion (Zeilen) und Prognose (Spalten) mit Support Vector Machine (Modell B) 84

Tabelle 20: Bewertungsindikatoren für Support Vector Machine 84

Tabelle 21: Kreuztabelle mit den Häufigkeiten der drei Zustandsbereiche für Inspektion (Zeilen) und Prognose (Spalten) mit Random Forest (Modell B) 88

Tabelle 22: Bewertungsindikatoren für Künstliches Neuronales Netz 88

Tabelle 23: Rechenzeiten für Modellaufbau (Training) und Simulation mit den vier untersuchten Modellansätzen (ermittelt mit PC-Konfiguration: i7-2600 CPU; 3,4 GHz, 8 bzw. 12 GB RAM) 91

Tabelle 24: *Cramér's-V*-Werte für alle Variablen untereinander und kombiniert mit der Zustandsklasse 97

Tabelle 25: *Cramér's-V*-Werte für die wichtigsten Variablen und die zwölf Hauptschadenstypen ... 104

Tabelle 26. Übersicht über die 59 Kohorten, die unter Berücksichtigung der Variablen „Bezirk", „Material", „Profil", „Baum (Ja/Nein)" und „Bodenart" gebildet wurden. 107

Tabelle 27: Bewertungsindikatoren für alle untersuchten Modellansätze im Vergleich. Für Random Forest (RF) und Support Vector Machine (SVM) wurden für Netz- und Haltungsebene unterschiedliche Modellparametrisierungen verwendet. KNN = Künstliche Neuronale Netze. 120

Abbildungsverzeichnis

Abbildung 1: Der Hobrechtplan zur Kanalisation von Berlin, 1871 (aus Bärthel 2003) 1

Abbildung 2: Definition des Sanierungsbedarfs außerhalb der Wasserschutzgebiete nach Schadensklassen 3

Abbildung 3: Mögliche Schadensbilder in einem Kanal (nicht klassifiziert) 3

Abbildung 4: Bei den Berliner Wasserbetrieben unterschiedene Zustandsklassen und für die Modellierung aggregierten Zustandsbereiche für die Bewertung der Kanäle. 11

Abbildung 5: Baujahr, Alter, Material und Abwassertyp der inspizierten Haltungen 12

Abbildung 6: Profil, Breite, Höhe und Länge der inspizierten Haltungen 12

Abbildung 7: Tiefe, Überdeckung, Gefälle und Straßenklasse der inspizierten Haltungen 13

Abbildung 8: Beeinflussung der inspizierten Haltungen durch Schienenverkehr (Tram, U-Bahn) und Bäume (Ja/Nein, Anzahl, Dichte in Längsrichtung der Haltung) 13

Abbildung 9: Grundwasserüberdeckung (Ja/Nein und Höhe), Bodenart und Rückstau-Beeinflussung der inspizierten Haltungen 14

Abbildung 10: Bezirk und Stadtteil der inspizierten Haltungen 14

Abbildung 11: Zustandsverteilung über alle inspizierten Haltungen (a: sechs Zustandsklassen der BWB, b: drei aggregierte Zustandsbereiche, Aggregierung nach Absprache mit BWB-AE für den Untersuchungszweck, orientiert am Sanierungsbedarf) 15

Abbildung 12: Berechnung des Korrelationskoeffizienten nach Pearson (links) und des Rangkorrelationskoeffizienten nach Spearman (rechts) am Beispiel der Exponentialfunktion 17

Abbildung 13: Berechnung von *Cramér's V* für ein einfaches Beispiel mit zwei Variablen (Alter und Material) mit je zwei Ausprägungen (Alt / Jung und Steinzeug / Stahlbeton) und insgesamt 160 Datensätzen. 19

Abbildung 14: Abhängigkeiten aller numerischen Variablen, quantifiziert über den Spearman-Koeffizienten 20

Abbildung 15: Abhängigkeiten der verbleibenden numerischen Variablen, quantifiziert über den Spearman-Koeffizienten 21

Abbildung 16: Abhängigkeiten der kategorischen Variablen, quantifiziert über *Cramér's V* 22

Abbildung 17: Datenverteilung für die neun Variablenpaare mit starker oder mittlerer Abhängigkeit (*Cramér's V* ≥ 0,3) 23

Abbildung 18: Ranking der Eingangsvariablen nach ihrem Effekt auf die Zustandsverteilung, quantifiziert über *Cramér's V* 24

Abbildung 19: Zustandsverteilung für die Variablen Alter, Profil, Länge, Grundwasserüberdeckung, Material und Bäume (Rang 1 bis 6) 25

Abbildung 20: Zustandsverteilung für die Variablen Bezirk, Abwassertyp, Breite, Bodenart, Rückstau und Überdeckung (Rang 7 bis 12) 27

Abbildung 21: Zustandsverteilung der Eingangsvariablen, die aufgrund von geringer Relevanz (oben) oder Abhängigkeiten (unten) nicht weiter untersucht wurden. 28

Abbildung 22: Schadenshäufigkeit, differenziert nach Schadenstyp, summiert über alle Inspektionsdatensätze. 31

Abbildung 23: Anzahl der schadhaften Haltungen, differenziert nach Schadenstyp 32

Abbildung 24: Einfluss der erklärenden Variablen auf das Auftreten der einzelnen Schadenstypen 33

Abbildung 25: Materialverteilung der Haltungen mit einem bestimmten Schaden (oben) und aller inspizierten Haltungen (unten) 34

Abbildung 26: Altersverteilung der Haltungen mit einem bestimmten Schaden (oben) und aller inspizierten Haltungen (unten) 34

Abbildung 27: Profilverteilung der Haltungen mit einem bestimmten Schaden (oben) und aller inspizierten Haltungen (unten) 35

Abbildung 28: Breitenverteilung der Haltungen mit einem bestimmten Schaden (oben) und aller inspizierten Haltungen (unten) 36

Abbildung 29: Bezirksverteilung der Haltungen mit einem bestimmten Schaden (oben) und aller inspizierten Haltungen (unten) 36

Abbildung 30: Längenverteilung der Haltungen mit einem bestimmten Schaden (oben) und aller inspizierten Haltungen (unten) 37

Abbildung 31: Verteilung des Abwassertyps für die Haltungen mit einem bestimmten Schaden (oben) und alle inspizierten Haltungen (unten) 37

Abbildung 32: Anteil der Inspektionen pro Inspektionsjahr (links) und pro Haltung (rechts). Alle Haltungen wurden zwischen 2001 und 2016 inspiziert. Datensatz nach Anwendung des Filterschrittes Nummer 2 (siehe Tabelle 5) 40

Abbildung 33: Verteilung der Dauer zwischen der Doppelbefahrungen. Datensatz nach Anwendung des Filterschrittes Nummer 3 (siehe Tabelle 1). Das blaue Rechteck zeigt den Datensatz nach Anwendung des Filterschrittes Nummer 4 40

Abbildung 34: Verteilung des Kameratyps bzw. des Inspektionszieles pro Jahr. Datengrundlage vor Filterung (siehe oben). 41

Abbildung 35: Abweichungen zwischen erster und zweiter Inspektion. Datengrundlage nach Anwendung des Filterschrittes 4 (siehe Tabelle 5). 42

Abbildung 36: Abweichungen zwischen ersten und zweiten Inspektionen mit den drei aggregierten Zustandsbereichen. Datengrundlage nach Anwendung des Filterschrittes 4 (siehe Tabelle 5) 42

Abbildung 37. Beispiel von zwei Gompertz-Überlebenskurven für drei Zustandsklassen/-bereiche 49

Abbildung 38: Berechnungsbeispiel für den Gini-Index. Im Beispiel führt die Verzweigung zu einer Verringerung des Gini-Indexes und damit zu einer besseren Klassifizierung der Daten. 51

Abbildung 39: Darstellung eines einfachen Entscheidungsbaums zur Klassifizierung der Haltungen nach ihrer Zustandsklasse. Die Balkengrafiken zeigen die

Zustandsverteilungen für jedes Blatt (Node 3, 5, 6 und 7; Knoten hier als „Node“ bezeichnet). 52

Abbildung 40: Veranschaulichung von Hyperebene, Stützvektoren und Abstand für einen zweidimensionalen Parameterraum 53

Abbildung 41: Beispiel einer Transformation aus zwei Dimensionen in einem neuen Raum mit drei Dimensionen. Die Punkte und Kreuze zeigen den Output (z.B. Zustand) für jedes Objekt. Die Transformation erlaubt die lineare Trennung der Punkte durch die dargestellte Hyperebene. 54

Abbildung 42: Umwandlung einer kategorischen Variable in numerische „Dummy“-Variablen am Beispiel des Abwassertyps 55

Abbildung 43: Schema eines Künstlichen Neuronalen Netzes mit einem *hidden layer* 56

Abbildung 44: sigmoide Aktivierungsfunktion für Künstliche Neuronale Netze 56

Abbildung 45: Beispiel für die Berechnung der Neuronenwerte unter Anwendung der sig-Aktivierungsfunktion 57

Abbildung 46: Datenverteilungen für Trainingsdaten (*n* = 58.528) und Testdaten (*n* = 39.019) 59

Abbildung 47: Aufteilung der Daten bei der Kreuzvalidierung (Beispiel: 5-fache Kreuzvalidierung) 60

Abbildung 48: Überlebenskurven für ausgewählte Kohorten: a) Material: Beton mit hoher Tragfähigkeit; b) Bezirk: Reinickendorf, Material: Steinzeug, Profil: Kreisprofil, Bäume: Ja, Bodenart: Sand-Lehm; c) Bezirk: Tempelhof-Schöneberg, Material: Beton. Für Altersbereiche über dem beobachteten Alter der Kanäle der jeweiligen Kohorten wurden die Überlebenskurven extrapoliert. 68

Abbildung 49: Abweichungen auf Netzebene zwischen Inspektion (oben) und Prognose mit GompitZ (unten). Die drei Farben repräsentieren die drei aggregierten Zustandsbereiche „gut“ (grün; Zustandsklasse 4, 5 und 6), „mittel“ (gelb; Zustandsklasse 3) und „schlecht“ (rot; Zustandsklasse 1 und 2) 69

Abbildung 50: Einfluss der für die Kalibrierung verwendeten Datenmenge auf die Modellgüte von GompitZ auf Netzebene (Indikatoren *K1* bis *K6* und *K_Netz*). Dargestellt sind die Bewertungsindikatoren mit Mittelwert, Minimum und Maximum der 50 Wiederholungen je Stichprobenumfang. 71

Abbildung 51: Einfluss der für die Kalibrierung verwendeten Datenmenge auf die Modellgüte von GompitZ auf Haltungsebene (Indikatoren *K7* bis *K12* und *K_Haltung*). Dargestellt sind die Bewertungsindikatoren mit Mittelwert, Minimum und Maximum der 50 Wiederholungen je Stichprobenumfang. 72

Abbildung 52: Einfluss des Gewichtungsfaktors *w1* (oben) und *w2* (unten) auf ausgewählte Güteindikatoren für die Vorhersage mit Random Forest auf Netzebene (*K_Netz*, links) und Haltungsebene (*K_Haltung* und *K9*, Mitte und rechts). Optimale Parameterbereiche sind farblich hervorgehoben. 73

Abbildung 53: Einfluss der Parameter *nodesize* und *mtry* auf die Modellgüte von Random Forest auf Netzebene (Indikatoren *K1* bis *K6* und *K_Netz*). Fixierte Parameter: *w1* = 2,0; *w2* = 1,0; ntrees = 100. 74

Abbildung 54: Einfluss der Parameter *nodesize* und *mtry* auf die Modellgüte von Random Forest auf Haltungsebene (Indikatoren *K7* bis *K12* und *K_Haltung*) Fixierte Parameter: *w1* = 0,8; *w2* = 1,0; *ntrees* = 100. 74

Abbildung 55: Abweichungen auf Netzebene zwischen Inspektion (oben) und Prognose mit Random Forest (Modell A, unten). Die drei Farben repräsentieren die drei aggregierten Zustandsbereiche „gut“ (grün; Zustandsklasse 4, 5 und 6), „mittel“ (gelb; Zustandsklasse 3) und „schlecht“ (rot; Zustandsklasse 1 und 2) 76

Abbildung 56: Einfluss der für das Training verwendeten Datenmenge auf die Modellgüte von Random Forest (Modell A) auf Netzebene (Indikatoren *K1* bis *K6* und *K_Netz*). Dargestellt sind die Bewertungsindikatoren mit Mittelwert, Minimum und Maximum der 50 Wiederholungen je Stichprobenumfang. 78

Abbildung 57: Einfluss der für das Training verwendeten Datenmenge auf die Modellgüte von Random Forest (Modell B) auf Haltungsebene (Indikatoren *K7* bis *K12* und *K_Haltung*). Dargestellt sind die Bewertungsindikatoren mit Mittelwert, Minimum und Maximum der 50 Wiederholungen je Stichprobenumfang. 79

Abbildung 58: Einfluss des Gewichtungsfaktors *w2* (oben) und *w3* (unten) auf ausgewählte Güteindikatoren für die Vorhersage auf Netz- und Haltungsebene (*K_Netz, K_Haltung* und *K9*). Optimale Parameterbereiche sind farblich hervorgehoben. 80

Abbildung 59: Einfluss der Parameter *C* und *sigma* auf die Modellgüte von Support Vector Machine auf Netzebene (Indikatoren *K1* bis *K6* und *K_Netz*) . Fixierte Parameter: für *w2* = 1,4; *w3* = 1,8. 81

Abbildung 60: Einfluss der Parameter *C* und *sigma* auf die Modellgüte von Support Vector Machine auf Haltungsebene (Indikatoren *K7* bis *K12* und *K_Haltung*). Fixierte Parameter: für *w2* = 1,4; *w3* = 1,8. 81

Abbildung 61: Abweichungen auf Netzebene zwischen Inspektion (oben) und Prognose mit Support Vector Machine (Modell A, unten). Die drei Farben repräsentieren die drei aggregierten Zustandsbereiche „gut“ (grün; Zustandsklasse 4, 5 und 6), „mittel“ (gelb; Zustandsklasse 3) und „schlecht“ (rot; Zustandsklasse 1 und 2) 83

Abbildung 62: Einfluss der Parameter *nhid* und *actfun* auf die Modellgüte des Künstlichen Neuronalen Netzes auf Netzebene (Indikatoren *K1* bis *K6* und *K_Netz*) 85

Abbildung 63: Einfluss der Parameter *nhid* und *actfun* auf die Modellgüte des Künstlichen Neuronalen Netzes auf Haltungsebene (Indikatoren *K7* bis *K12* und *K_Haltung*) 86

Abbildung 64: Abweichungen auf Netzebene zwischen Inspektion (oben) und Prognose mit einem Künstlichen Neuronalen Netz (unten). Die drei Farben repräsentieren die drei aggregierten Zustandsbereiche „gut“ (grün; Zustandsklasse 4, 5 und 6), „mittel“ (gelb; Zustandsklasse 3) und „schlecht“ (rot; Zustandsklasse 1 und 2) 87

Abbildung 65: Modellgüte auf Netzebene für die vier untersuchten Modellansätze. Erläuterung der Bewertungsindikatoren in Kap. 4.2.5. 89

Abbildung 66: Modellgüte auf Haltungseebene für die vier untersuchten Modellansätze. Erläuterung der Bewertungsindikatoren in Kap. 4.2.5. 90

Abbildung 67: Prognose der Zustandsentwicklung von 2017 bis 2066 mit GompitZ 92

Abbildung 68: Prognose der Zustandsentwicklung von 2017 bis 2066 mit Random Forest (Modell A) 92

Abbildung 69: Datenverteilungen der Variablen untereinander 103

Abbildung 70: Datenverteilung für Grundwasserüberdeckung (a), Bäume (b), Bodentyp (c), Rückstau (d) und Überdeckung (e) für die Haltungen mit einem bestimmten Schaden (jeweils oben) und alle inspizierten Haltungen (jeweils unten) 106

Abbildung 71. Darstellung der Überlebenskurven für alle Kohorten (nur für 57 von 59 Kohorten). Für zwei Kohorten („Mauerwerk“; „StegZeh_Steinzeug_Kreis_Ja_SandLehm“) war die Kalibrierung nicht erfolgreich (keine Konvergenz bei der Maximum-Likelihood-Methode). Für diese Kohorten wurden die Überlebenskurven der nächsten konvergierenden Kohorte übernommen. 110

Abbildung 72: Random Forest - Einfluss des Parameters *mtry* auf die Güteindikatoren auf Netz- und Haltungsebene (Parameterwerte für *mtry*, *nodesize*, *ntrees*, *w1* und *w2* zufällig variiert). 111

Abbildung 73: Random Forest - Einfluss des Parameters *nodesize* auf die Güteindikatoren auf Netz- und Haltungsebene (Parameterwerte für *mtry*, *nodesize*, *ntrees*, *w1* und *w2* zufällig variiert). 112

Abbildung 74: Random Forest - Einfluss des Parameters *ntrees* auf die Güteindikatoren auf Netz- und Haltungsebene (Parameterwerte für *mtry*, *nodesize*, *ntrees*, *w1* und *w2* zufällig variiert). 113

Abbildung 75: Random Forest - Einfluss des Parameters *w1* auf die Güteindikatoren auf Netz- und Haltungsebene (Parameterwerte für *mtry*, *nodesize*, *ntrees*, *w1* und *w2* zufällig variiert). 114

Abbildung 76: Random Forest - Einfluss des Parameters *w2* auf die Güteindikatoren auf Netz- und Haltungsebene (Parameterwerte für *mtry*, *nodesize*, *ntrees*, *w1* und *w2* zufällig variiert). 115

Abbildung 77: Support Vector Machine - Einfluss des Parameters *C* auf die Güteindikatoren auf Netz- und Haltungsebene (Parameterwerte für *C*, *sigma*, *w2* und *w3* zufällig variiert). 116

Abbildung 78: Support Vector Machine - Einfluss des Parameters *sigma* auf die Güteindikatoren auf Netz- und Haltungsebene (Parameterwerte für *C*, *sigma*, *w2* und *w3* zufällig variiert). 117

Abbildung 79: Support Vector Machine - Einfluss des Parameters *w2* auf die Güteindikatoren auf Netz- und Haltungsebene (Parameterwerte für *C*, *sigma*, *w2* und *w3* zufällig variiert). 118

Abbildung 80: Support Vector Machine - Einfluss des Parameters *w3* auf die Güteindikatoren auf Netz- und Haltungsebene (Parameterwerte für *C*, *sigma*, *w2* und *w3* zufällig variiert). 119

Literaturverzeichnis

Agresti, A. (2007): An Introduction to Categorical Data Analysis, John Wiley & Sons Hoboken, N.J.

Barandela, R., Valdovinos, R.M., Salvador Sánchez, J. and Ferri, F.J. (2004): The imbalanced training sample problem: under or over sampling?, pp. 806-814.

Bärthel, H. (2003): Geklärt! 125 Jahre Berliner Stadtentwässerung, HUSS-Medien GmbH, Verlag Bauwesen, pp.10-23

Berger, C. und Falk, C. (2009): Zustand der Kanalisation in Deutschland, Ergebnisse der DWA Umfrage 2009, Korrespondenz Abwasser 2011, pp. 24-39

Berger, C., Falk, C., Hetzel, F., Pinnekamp, J., Roder, S., Ruppelt, J. (2015): Zustand der Kanalisation in Deutschland, Ergebnisse der DWA Umfrage 2015, Korrespondenz Abwasser 2016, p. 6

Biegel, M. (2016): Der Schatz im Untergrund, Artikel aus WLB Wasser, Luft und Boden, p. 10.

Breiman, L., J. H. Friedman, C. J. Stone and R. A. Ohlsen (1984): Classification and Regression Trees. United Kingdom, Taylor & Francis Ltd.

Broggi, A., Cerri, P., Medici, P., Porta, P.P. and Ghisio, G. (2007) Real time road signs recognition, pp. 981-986.

BWB (2001): Schadenskatalog für Abwasserkanäle, Version 11. Berliner Wasserbetriebe, Grundlagenplanung und Investitionssteuerung.

Caradot, N., Rouault, P., Clemens, F. and Cherqui, F. (2017): Evaluation of uncertainties in sewer condition assessment. Structure and Infrastructure Engineering, 1-10.

Cohen, J. (1988): Statistical power and analysis for the behavioral sciences (2nd ed.), Hillsdale, N.J., Lawrence Erlbaum Associates, Inc.

Cramér, H. (1946): Mathematical Methods of Statistics. Princeton: Princeton University Press, page 282 (Chapter 21. The two-dimensional case). ISBN 0-691-08004-6

Davies, J.P., Clarke, B.A., Whiter, J.T. & Cunningham, R.J. (2001): Factors influencing the structural deterioration and collapse of rigid sewer pipes. Urban Water 3(1-2), 73-89.

Díaz-Uriarte, R. and S. Alvarez de Andrés (2006): Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7.

Elandt-Johnson, R.C. & Johnson, N.L. (1980): Survival Models and Data Analysis, Wiley.

Fenner, R.A. & Sweeting, L. (1999): A decision support model for the rehabilitation of 'non-critical' sewers, pp. 193-200.

Fisher, R.A. (1936): The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, 179-188.

Harvey, R. R. and E. A. McBean (2014): Predicting the structural condition of individual sanitary sewer pipes with random forests. Canadian Journal of Civil Engineering 41(4): 294-303.

Hastie, T., R. Tibshirani and J. H. Friedman (2008): The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer.

Hebb, D. O. (1949): The Organization of Behavior: A Neuropsychological Theory. New York: Wiley and Sons

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. and Kingsbury, B. (2012) Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine 29(6), 82-97.

Hoffmann, M. (2006): Support Vector Machines - Kernels and the Kernel Trick, An elaboration for the Hauptseminar “Reading Club: Support Vector Machines". http://www.cogsys.wiai.uni-bamberg.de/teaching/ss06/hs_svm/slides/SVM_Seminarbericht_Hofmann.pdf

Huang, G.B., Zhu, Q.Y. and Siew, C.K. (2006): Extreme learning machine: Theory and applications. Neurocomputing 70(1-3), 489-501.

Japkowicz, N. and Stephen, S. (2002): The class imbalance problem: A systematic study. Intelligent Data Analysis 6(5), 429-449.

Jones, G. M. A. (1984): The structural deterioration of sewers. In international conference on the planning, construction, maintenance & operation of sewerage systems, Reading, UK, September 1984.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012): ImageNet classification with deep convolutional neural networks, pp. 1097-1105.

Le Gat, Y. (2008); Modeling the deterioration process of drainage pipelines, Urban Water Journal 5 97-106.

Liaw, A. and Wiener, M. (2002): Classification and Regression by Random Forest. R News 2, 18-22.

Lin, M., Lucas, H.C. and Shmueli, G. (2013): Too big to fail: Large samples and the p-value problem. Information Systems Research 24(4), 906-917.

Maalouf, M. and Trafalis, T.B. (2011): Rare events and imbalanced datasets: An overview. International Journal of Data Mining, Modelling and Management 3(4), 375-388.

Markow, A.A. (2006) Classical text in translation: An example of statistical investigation of the text “Eugene Onegin” concerning the connection of samples in chains - Lecture at the physical-mathematical faculty, Royal Academy of Sciences, St. Petersburg, 23 January 1913. Science in Context 19(4), 591-600.

McHugh, Mary L. (2013): The Chi-square test of independence, Biochem Med (Zagreb). 2013 Jun; 23(2): 143–149. doi: 10.11613/BM.2013.018; PMCID: PMC3900058; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900058/

O'Reilly M. P., Prosbrook R. B., Cox G. C., McCloskey A. (1989): Analysis of defects in 180 km of pipe sewers in southern water authority, TRRL Research Report 172.

Pados, D.A. and Papantoni-Kazakos, P. (1994): Note on the estimation of the generalization error and the prevention of overfitting. IEEE International Conference on Neural Networks - Conference Proceedings 1, 321-326.

Pearson, K. (1900): On the criterion that a given system of derivations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 50(5), 157–175.

Philip Runarsson, T. and Yao, X. (2005) Search biases in constrained evolutionary optimization, IEEE Trans. on Systems, Man, and Cybernetics Part C: Applications and Reviews, vol. 35 (no. 2), pp. 233-243.

Rokach, L. and Maimon, O. (2008): Data Mining with Decision Trees: Theory and Applications, World Scientific Publishing Co., Inc., Singapore.

Rosenblatt, F. (1958): The Perceptron, a Probabilistic Model for Information Storage and Organisation in the Brain. Psycological Review 62(386).

Schmidhuber, Jürgen (2015): Deep learning in neural networks: An overview. Neural Networks. 61: 85–117

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: A software Environment for integrated models of biomolecular interaction networks. Genome Research 13(11), 2498-2504.

Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D. and Summers, R.M. (2016): Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. IEEE Transactions on Medical Imaging 35(5), 1285-1298.

Smith, B. T. (2004): Lagrange Multipliers Tutorial in the Context of Support Vector Machines, Memorial University of Newfoundland St. John's, Newfoundland, Canada.

Vapnik, V. und Chervonenkis, A. (1974): Theory of Pattern Recognition (Deutsche Übersetzung: Theorie der Zeichenerkennung), Akademie Verlag, Berlin.

Zhang, W., Li, C., Peng, G., Chen, Y. and Zhang, Z. (2018): A deep convolutional neural network with new training methods for bearing fault diagnosis under noisy environment and different working load. Mechanical Systems and Signal Processing 100, 439-453.

Zhu, X. and Wu, X. (2004): Class Noise vs. Attribute Noise: A Quantitative Study of Their Impacts. Artificial Intelligence Review 22, 177-210.