

FREIE UNIVERSITÄT BERLIN
HUMBOLDT-UNIVERSITÄT ZU BERLIN



Masterarbeit

zur Erlangung des akademischen Grades
Master of Science (M.Sc.)
im Studiengang Statistik

Explorative Datenanalyse mit R Shiny

Eine Anwendungsentwicklung im Rahmen des Bildungsmonitoring Berlin-Mitte

Gutachter:

Prof. Dr. Ulrich Rendtel
Prof. Dr. Ulrike Rockmann

vorgelegt von:

Ulrike Niemann

Matrikelnummer 194238 (HU), 4877844 (FU)

ulrike.niemann@e-market-research.de

Berlin, den 5. Mai 2020

Inhaltsverzeichnis

Abbildungsverzeichnis	vi
Tabellenverzeichnis	vii
Abkürzungsverzeichnis	viii
Gender-Erklärung	x
1 Einleitung	1
2 Grundlagen und Zielsetzung	3
2.1 Bildungsmonitoring Berlin-Mitte	3
2.1.1 Allgemeine Ziele	3
2.1.2 Schwerpunkte und Arbeitsschritte	4
2.1.3 Projektpartner	5
2.2 Einschulungsuntersuchung	5
2.2.1 Rechtliche Grundlagen	6
2.2.2 Ziele	7
2.2.3 Inhalte	7
2.3 Sozialraumorientierung und räumliche Gliederungen	8
2.3.1 Lebensweltlich orientierte Räume	9
2.3.2 Einschulungsbereiche	10
2.4 Datenbasis	11
2.4.1 Erhebungsdaten	12
2.4.2 Indikatoren	12
2.4.3 Daten aus weiteren Quellen	14
2.4.4 Metadaten	14
2.5 Ziele der Anwendungsentwicklung	15
2.5.1 Allgemeine Zielsetzung	15
2.5.2 Zielpersonen	16
2.5.3 Anforderungsspezifikation	16
3 Auswahl statistischer Methoden	18
3.1 Einführung	18
3.1.1 Begriffsbildungen	18

3.1.2	Teilgebiete der Statistik und explorative Datenanalyse	19
3.1.3	Zur Anwendung induktiver Statistik bei Vollerhebungen	20
3.2	Häufigkeitsverteilungen	21
3.2.1	Eindimensionale Häufigkeiten	21
3.2.2	Kontingenztabellen	21
3.2.3	Grafische Darstellungsformen	23
3.3	Verteilungsmaßzahlen	25
3.3.1	Rangmaßzahlen	25
3.3.2	Lagemaße	26
3.3.3	Streuungsmaße	27
3.3.4	5-Zahlen-Zusammenfassung	28
3.3.5	Grafische Darstellungsformen	29
3.4	Dichteschätzung	31
3.4.1	Grundlagen	31
3.4.2	Histogramm	32
3.4.3	Kerndichteschätzer	33
3.4.4	Grafische Darstellungsformen	34
3.5	Zusammenhangmaße	35
3.5.1	Kontingenz	37
3.5.2	Rangkorrelation	39
3.5.3	Maßkorrelation	40
3.5.4	Signifikanztests	41
3.6	Thematische Karten	42
3.6.1	Choroplethenkarten	43
3.6.2	Kerndichtekarten	44
4	Anwendungsentwicklung	49
4.1	Entwurf	49
4.1.1	R	50
4.1.2	shiny	50
4.1.3	tidyverse	52
4.1.4	Programmablauf	53
4.2	Frontend: User Interface	55
4.2.1	Layout und Gestaltung	56
4.2.2	Erhöhte Benutzerfreundlichkeit mit JavaScript	57
4.2.3	Erweiterte Interaktivität mit HTMLwidgets	58
4.3	Backend: Server	59
4.3.1	Datentransformation und statistische Methoden	59
4.3.2	Datenvisualisierung	61

4.3.3	Export der Analyseergebnisse	62
4.3.4	Speichern und Laden von Analyseeinstellungen	62
4.4	Implementierung	63
4.4.1	Shiny-Module	63
4.4.2	Programmierstil	64
4.4.3	Zugriff auf die Anwendung	66
4.4.4	Qualitätssicherung	67
5	Der ESU explorer - Vorstellung der Anwendung	69
5.1	Einstieg und Navigation	69
5.2	Analysen	69
5.2.1	Einstellungen	70
5.2.2	Erweiterte Filtereinstellungen	72
5.2.3	Ergebnisse	72
5.3	Karten	76
5.3.1	Einstellungen	76
5.3.2	Ergebnisse	77
5.4	Metadaten	79
5.4.1	Variablen Informationen	80
5.4.2	Dokumentationsbögen	81
5.4.3	Hilfebereich	81
5.4.4	Über den ESU explorer	81
6	Fazit	82
6.1	Zusammenfassung	82
6.2	Ausblick	84
A	Anhang	85
A.1	Dokumentationsbogen für die Einschulungsuntersuchungen der KJGD im Land Berlin und Dokumentationsbogen für die S-ENS und SOPESS- Untertests (Schuljahr 2019)	85
A.2	Übersicht über die verwendeten R-Packages	90
A.3	Ordner- und Dateistruktur ESU explorer	91
A.4	Erweiterte Grafikeinstellungen ESU explorer	93
A.5	Hinweise zur beiliegenden CD	95
	Literaturverzeichnis	96

Abbildungsverzeichnis

2.1	Lebensweltlich orientierte Räume (LOR) Berlin-Mitte	10
2.2	Einschulungsbereiche (ESB) Berlin-Mitte	11
2.3	Indikatoren für den Sprachstand der Kinder	13
3.1	Varianten Säulen- und Balkendiagramme: Empfehlung zur schulischen Förderung nach dem Bildungsstand der Familien und der Zuwanderungs- erfahrung des Kindes, Schuljahr 2019	24
3.2	Liniendiagramm: Zuwanderungserfahrung der Kinder nach Schuljahren .	24
3.3	Kreisdiagramm: Erwerbstätigkeit der Mutter nach Zuwanderungserfahrung der Mutter, Schuljahr 2019	25
3.4	Box-Plot: Alter der Kinder bei der ESU, Schuljahr 2019	29
3.5	Streudiagramm in Kombination mit Box-Plots: Körpergröße und Körper- gewicht der Kinder bei ESU, Schuljahr 2019	30
3.6	Histogramme: Körpergröße der Kinder bei ESU, Schuljahr 2017	33
3.7	Histogramme in Kombination mit Kerndichteschätzungen: Körpergewicht der Kinder bei ESU nach Bildungsstand der Familien, Schuljahr 2019 . .	34
3.8	Violin-Plot in Kombination mit Box-Plot: Jahre in Kita bei Einschulung nach Erwerbstätigkeit der Mutter, Schuljahr 2019	35
3.9	Streudiagramm in Kombination mit Kerndichteschätzung: Körpergewicht bei ESU und Geburtsgewicht, Schuljahr 2019	36
3.10	Streudiagramme mit Regressionsgeraden: Körpergewicht nach Alter der Kinder und Zuwanderungserfahrung, Schuljahr 2019	41
3.11	Choroplethenkarten: Anteile der Kinder mit Schulischem Förderbedarf, Einschulungsbereiche, Schuljahr 2019	45
3.12	Choroplethenkarte, Kernelheaping-Karte und Hotspot-Karte: Anteil der Kinder mit eigener Zuwanderungserfahrung, Einschulungsbereiche, Schul- jahr 2019	48
4.1	Architektur einer shiny-Anwendung	52
4.2	Programmablaufplan ESU explorer	54
4.3	Auswahlmöglichkeit über selectizeInput()-Objekt	57

5.1	ESU explorer: Einstiegsseite	70
5.2	ESU explorer: Individuelle Analyse	71
5.3	ESU explorer: Erweiterte Filtereinstellungen	72
5.4	ESU explorer: Ergebnisse - Grafik	73
5.5	ESU explorer: Analyseseite Karten mit Beispieleinrichtungen	76
5.6	ESU explorer: Ergebnisse - Karte	78
5.7	ESU explorer: Metadaten - Variableninformationen	80
5.8	ESU explorer: Hilfebereich	81
A.1	Ordnerstruktur ESU explorer	91
A.2	Datei-Beziehungen ESU explorer	92
A.3	ESU explorer: Erweiterte Grafikeinstellungen für Säulen-, Balken-, Linien- und Kreisdiagramm	93
A.4	ESU explorer: Erweiterte Grafikeinstellungen für Box-Plot/Violin-Plot . . .	93
A.5	ESU explorer: Erweiterte Grafikeinstellungen für Histogramm/Dichte . . .	93
A.6	ESU explorer: Erweiterte Grafikeinstellungen für Punktediagramm	94
A.7	ESU explorer: Erweiterte Grafikeinstellungen für Choroplethen-Karten . .	94
A.8	ESU explorer: Erweiterte Grafikeinstellungen für Kernelheaping-Karten .	94

Tabellenverzeichnis

3.1	Interpretation Korrelationskoeffizient nach (Cohen, 1988)	40
-----	---	----

Abkürzungsverzeichnis

BZR	Bezirksregion
CRAN	Comprehensive R Archive Network
CSS	Cascading Style Sheets
EDA	Explorative Datenanalyse
ESB	Einschulungsbereich
ESU	Einschulungsuntersuchung
FTP	File Transfer Protocol
GBE	Gesundheitsberichterstattung
GIS	Geoinformationssystem
GsVO	Grundschulverordnung
HTML	Hypertext Markup Language
ID	Identifikator
IDE	Integrated development environment
ISBJ	Integrierte Software Berliner Jugendhilfe
ISCED	International Standard Classification of Education
ISQ	Institut für Schulqualität der Länder Berlin und Brandenburg e.V.
ISS	Integrierte Sekundarschulen
IT	Informationstechnik
KJGD	Kinder- und Jugendgesundheitsdienst
LOR	Lebensweltlich orientierte Räume
PLR	Planungsraum
PNG	Portable Network Graphics

PRG	Prognoseraum
RBS	Regionales Bezugssystem
S-ENS	Screening des Entwicklungsstandes bei Einschulungsuntersuchungen
SEM	Stochastic Expectation Maximation Algorithm
SchulG	Schulgesetz Berlin
SenGPG	Senatsverwaltung für Gesundheit, Pflege und Gleichstellung
SOPESS	Sozialpädiatrisches Entwicklungsscreening für Schuleingangsuntersuchungen
SVG	Scalable Vector Graphics
UI	User Interface

Gender-Erklärung

Aus Gründen der besseren Lesbarkeit wird in dieser Arbeit die Sprachform des generischen Maskulinums angewendet. Es wird an dieser Stelle darauf hingewiesen, dass die ausschließliche Verwendung der männlichen Form geschlechtsunabhängig verstanden werden soll und somit weibliche und andere Geschlechteridentitäten ausdrücklich mitmeint.

1 Einleitung

„A basis problem about any body of data is to make it more easily and effectively handleable by minds.“
(Tukey, 1977)

Ziel der explorativen Datenanalyse ist, ein grundlegendes Verständnis von Daten zu ermöglichen. Der Wert und Nutzen von Daten entfaltet sich erst, wenn großen Datenmengen zugänglich gemacht und zusammenfassende Informationen daraus abgeleitet werden können. Die explorative Datenanalyse ist ein systematischer Weg der Erforschung und Visualisierung von Daten durch ein iteratives Vorgehen (Wickham und Grolemund, 2016: 81). Dabei werden fortlaufend Fragen zu den Daten generiert, verfeinert und neu formuliert. Die Suche nach Antworten erfolgt vorwiegend durch einfache statistische Analysemethoden und Datenvisualisierung.

Das Ziel dieser Arbeit ist die Entwicklung einer leicht zu bedienenden computergestützten Analyseanwendung zur interaktiven explorativen Datenanalyse von spezifischen Daten im Rahmen des Projektes „Bildungsmonitoring Berlin-Mitte“. Das Projekt zielt auf eine detaillierte Betrachtung der Bildungssituation im Berliner Bezirk Mitte. Aufbauend auf empirischen Befunden und Indikatoren sollen Strategien und Handlungsempfehlungen für eine Chancengerechtigkeit im Bildungssystem entwickelt und untermauert werden (Rockmann und Leerhoff, 2018a).

Die Analyseanwendung zielt darauf ab, die Daten der in Berlin-Mitte durchgeführten Einschulungsuntersuchungen und daraus entwickelter Indikatorensatz den verschiedenen Mitarbeitern und Projektpartnern direkt zugänglich zu machen. Durch die explorative Analyse dieser Daten sollen die Verwaltungsmitarbeiter befähigt werden, ein besseres Verständnis der Bildungssituation vor Schuleintritt zu erhalten und somit eine bessere Steuerung im Bildungsbereich zu ermöglichen. Eine sozialraumorientierte Analyse der Bildungssituation in Berlin-Mitte ist dabei besonders wichtig, da der Bezirk eine sehr heterogene Bevölkerungsstruktur in Hinblick auf die soziale Situation der Kinder und Familien aufweist.

Auf Basis der Programmierumgebung R soll dafür eine shiny-Anwendung für die explorative Datenanalyse entwickelt werden. Shiny ist ein R-Package zur Erstellung von interaktiven Webanwendungen. Mithilfe von shiny und weiterer geeigneter R-Technologien

soll die Anwendung interaktive Datenanalysen ermöglichen, ohne dabei R- oder Programmierkenntnisse der Nutzer vorauszusetzen.

Der Fokus der Anwendung liegt auf einer grafischen Darstellung der Analyseergebnisse zur intuitiven Erfassung von Mustern, Trends und Zusammenhängen. Die Darstellung kleinräumiger Analysen auf einer Karte von Berlin-Mitte soll eine sozialraumorientierte Berichterstattung ermöglichen.

Zur Einführung in die Thematik werden in Kapitel 2 zunächst die theoretischen Grundlagen und Hintergründe dargelegt. Das Projekt Bildungsmonitoring Berlin-Mitte und dessen Ziele sowie die zu untersuchende Datenbasis werden vorgestellt. Aufbauend darauf werden die Anforderungen an die zu entwickelnde Analyseanwendung spezifiziert.

Anschließend folgt in Kapitel 3 die Auswahl der statistischen Methoden, welche durch die Analyseanwendung angeboten werden sollen. Der Schwerpunkt liegt hier auf Methoden der explorativen Datenanalyse und insbesondere in verschiedenen Formen der Datenvisualisierung. Dabei sollen auch kleinräumige Strukturen analysiert werden können. Dafür werden verschiedene Möglichkeiten der Darstellung thematischer Karten aufgezeigt.

Die Entwicklung der Analyseanwendung wird in Kapitel 4 beschrieben. Der Fokus liegt hier auf den programmiertechnischen Aspekten. Das Kapitel enthält detaillierte Beschreibungen des Programmentwurfs, aller verwendeter Technologien, der Implementierung und Maßnahmen zur Qualitätssicherung.

Im fünften Kapitel wird als Ergebnis die Analyseanwendung aus Nutzersicht vorgestellt. Dafür sollen die Inhalte und Funktionalitäten der entwickelten Anwendung detailliert dargestellt werden. Hinsichtlich der Namensgebung für die Analyseanwendung wurde ein Begriff gewählt, der passt und den Anwendern leicht im Gedächtnis bleibt. Als Name für die Anwendung steht bezugnehmend auf die Datenbasis und die Aufgabe der Begriff „ESU explorer“.

Dieser schriftliche Teil der Arbeit kann zudem als Dokumentation des ESU explorers verstanden werden. Die detaillierte Beschreibung der verwendeten statistischen Methoden, der programmiertechnischen Aspekte und die Vorstellung der Inhalte und Nutzungsmöglichkeiten soll den zukünftigen Nutzern ein umfassendes Verständnis des ESU explorers aus den verschiedenen Blickwinkeln ermöglichen.

2 Grundlagen und Zielsetzung

In diesem Kapitel erfolgt zunächst die Erläuterung der theoretischen Grundlagen als Einleitung in die Thematik. Es wird das Projekt Bildungsmonitoring Berlin-Mitte vorgestellt, in dessen Rahmen die Analyseanwendung zur Untersuchung von Daten genutzt werden soll. Die Einschulungsuntersuchung (ESU) als zentrale Datengrundlage wird näher vorgestellt. Es folgt eine Beschreibung der Möglichkeiten der sozialraumorientierten Berichterstattung auf Basis der zur Verfügung stehenden Daten. Unter Berücksichtigung der Ziele des Bildungsmonitoring Berlin-Mitte und der Datenbasis erfolgt eine Konkretisierung der Zielsetzung dieser Arbeit. Dabei sollen die Anforderungen an die zu entwickelnden Analyseanwendung unter Berücksichtigung der Nutzerbedürfnisse spezifiziert werden.

2.1 Bildungsmonitoring Berlin-Mitte

Das *Bildungsmonitoring Berlin-Mitte* ist ein Projekt des Bezirksamtes Berlin-Mitte in Kooperation mit dem Institut für Schulqualität der Länder Berlin und Brandenburg e.V. (ISQ). Das Projekt zum Thema „Bildungszugänge und Bildungsübergänge von im Bezirk Mitte lebenden Kindern von 0 - 18 Jahren und ihren Familien im Zeitraum 2017 - 2019“ wurde im Juni 2017 vom Bezirksamt Mitte beauftragt.

2.1.1 Allgemeine Ziele

Das Projekt Bildungsmonitoring Berlin-Mitte orientiert sich an den grundlegenden Zielsetzungen der Bildungsberichterstattung (Rockmann und Leerhoff, 2018a: 18). Bildungsberichterstattung ist problemorientiert und analytisch indem sie sich auf Indikatoren und empirische Daten bezieht. Dabei wird versucht, die Stellen und Entwicklungen im Bildungswesen aufzuzeigen, die für Politik und Öffentlichkeit von besonderem Interesse sind und auf Handlungsbedarfe im Einzelfall hinweisen. Ziel von Bildungsberichterstattung und des Projektes Bildungsmonitoring Berlin-Mitte ist dabei die Förderung von Chancengleichheit und gesellschaftlicher Teilhabe, um eine systematische Benachteiligung von Kindern aufgrund ihrer Herkunft, des Geschlechts oder anderer Merkmale entgegenzuwirken (Autorengruppe Bildungsberichterstattung, 2018: 1-2), (Rockmann und Leerhoff, 2018a: 18).

Das Projekt Bildungsmonitoring Berlin-Mitte liefert einen Beitrag dazu, das Bildungsgefüge besser zu verstehen und Gründe zu finden, die dazu führen, dass Kinder im Bildungswesen unterschiedliche Chancen haben, ihre Fähigkeiten bestmöglich zu entwickeln (Bezirksamt Mitte von Berlin, 2017).

Berlin als wachsende und multikulturelle Metropole zeigt einen hohen Anteil junger Menschen, die ohne Abschluss die Schule verlassen. So lag der Anteil der Berliner Schulabsolventen und Abgänger, die zum Ende des Schuljahres 2016/2017 keinen Schulabschluss erreichen konnten bei 9,8 %. In Berlin-Mitte erreichten sogar 11,5 % der Absolventen und Abgänger zum Ende des Schuljahres 2016/2017 keinen Schulabschluss (Amt für Statistik Berlin-Brandenburg, 2019: 29-31).

Berlin-Mitte ist dabei ein sehr heterogener Bezirk in Hinblick auf den Sozialmilieus der Familien. Ein hoher Anteil sowohl an hohen als auch niedrigen Bildungsabschlüssen der Eltern und vielen aus dem Ausland zugewanderten Familien stellt das Bildungssystem vor erhebliche Herausforderungen (Rockmann und Leerhoff, 2018b).

Eine detaillierte Betrachtung der Bildungssituation in Berlin-Mitte soll im Rahmen des Bildungsmonitoring ein besseres Verständnis für das Zusammenwirken unterschiedlicher Faktoren in Hinblick auf die Chancengerechtigkeit im Bildungssystem liefern.

2.1.2 Schwerpunkte und Arbeitsschritte

Schwerpunkt des Projektes Bildungsmonitoring Berlin-Mitte ist die Identifizierung von Risiken, die die Chancen der Kinder auf ihren Bildungserfolg verringern. Auf dieser Basis sollen regional valide Indikatorensets entwickelt werden, „die der Politik und Verwaltung eine bessere Steuerung im Bildungsbereich ermöglichen und insbesondere dabei unterstützen sollen, den hohen Anteil derer, die in Berlin-Mitte die Schule ohne Abschluss verlassen, zu reduzieren“ (Rockmann und Leerhoff, 2018a: 17-18).

Im Rahmen des Bildungsmonitoring Berlin-Mitte sollen die zu entwickelnden Indikatorensets eine kontinuierliche und langfristige Beobachtung ermöglichen, den Fokus auf die Übergänge zwischen den institutionellen Bildungseinrichtungen richten, an die nationale und regionale Bildungsberichterstattung anschlussfähig sein sowie die Berlin-spezifischen Indikatorenssysteme der Stadtentwicklung berücksichtigen (Rockmann und Leerhoff, 2018b: 8). Um eine gebietsdifferenzierte Handlungsebene zu ermöglichen, soll im Rahmen des Projektes die Betrachtung der Bildungssituation auch innerhalb des Bezirkes kleinräumig spezifisch erfolgen.

Ein erster Projektbericht des Bildungsmonitoring Berlin-Mitte zeigt Analysen in Hinblick auf den Start in das und die ersten Jahre im Bildungssystem, welche als zentrale Weichenstellung anzusehen sind (Rockmann und Leerhoff, 2018b: 7). Dafür wurden Datenbestände aus der amtlichen Statistik sowie der Berliner und bezirklichen Verwaltung analysiert und auf die Verwendbarkeit für die regional benötigten Monitoring-Indikatoren untersucht.

Schwerpunkt ist hierbei die Identifizierung von Risiken, die der Chancengerechtigkeit entgegenstehen und die Chancen der Kinder auf ihren Bildungserfolg verschlechtern (Rockmann und Leerhoff, 2018b: 8). Mittels einer Beobachtung von Indikatoren wie beispielsweise Zuwanderungserfahrung und Bildungsstand der Familien, Familiensprachen, Sprachstand der Kinder und anderer möglicher Einflussfaktoren auf den Bildungserfolg der Kinder soll Politik und Verwaltung befähigt werden, Strategien und Unterstützungsmaßnahmen zur Förderung der Chancengleichheit zu entwickeln und dabei regional spezifische Gegebenheiten berücksichtigen.

Das letztliche Ziel des Bildungsmonitoring Berlin-Mitte liegt in der genauen Definition der Indikatoren und dem Aufbau einer Analysearchitektur, welche in die eigenständige Nutzung der bezirklichen Fachverwaltungen überführt werden soll (Bezirksamt Mitte von Berlin, 2017).

2.1.3 Projektpartner

Das Bezirksamt Mitte von Berlin setzt das Bildungsmonitoring in Zusammenarbeit mit dem Institut für Schulqualität der Länder Berlin und Brandenburg e.V. (ISQ) unter der Leitung von Prof. Dr. Ulrike Rockmann um, wobei die Koordination in ressortübergreifender Zusammenarbeit von der bezirklichen Jugendhilfeplanung (Jugendamt) und dem Sprachförderzentrum Mitte stattfindet.

Auf der Internetseite zum Projekt finden sich eine aktuelle Übersicht über die beteiligten Institutionen, Projekt- und Ansprechpartner sowie bisherige Veröffentlichungen, Ziele, Quellen und Methodik (Bezirksamt Mitte von Berlin, 2019).

2.2 Einschulungsuntersuchung

Die für alle schulpflichtig werdenden Kinder verpflichtende Einschulungsuntersuchung (ESU) bietet eine Vielzahl von Merkmalen, welche die Entwicklung der im Rahmen des Bildungsmonitoring benötigten Indikatorensets unterstützen. Als seit Jahrzehnten etablierte Erhebung bietet die ESU eine Reihe von Analysemöglichkeiten, mit denen die Situation der Kinder bei ihrem Übergang zwischen Kita und Grundschule charakterisiert werden kann (Rockmann und Leerhoff, 2018a: 22). Die Erhebungsdaten der ESU liegen als Verwaltungsdaten der bezirklichen Kinder- und Jugendgesundheitsdienste (KJGD) vor und werden im Rahmen des Bildungsmonitoring Berlin-Mitte als wichtige Analysegrundlage genutzt.

Im folgenden Abschnitt werden die rechtlichen Grundlagen, Ziele und Inhalte der ESU beschrieben. Die Schuljahre werden nachfolgend nur mit dem Jahr bezeichnet in welchem das Schuljahr beginnt. So wird das Schuljahr 2019/2020 mit Schuljahr 2019 bezeichnet.

2.2.1 Rechtliche Grundlagen

Jedes Kind in Deutschland wird vor der Einschulung schulärztlich untersucht. Die Einschulungsuntersuchung (ESU) ist nach dem Schulgesetz Berlin (SchulG) verpflichtend vorgeschrieben und ergibt sich aus § 55a Abs. 6 SchulG und § 52 Abs. 2 SchulG. Die einzuschulenden Kinder und deren Erziehungsberechtigten sind demnach verpflichtet, vor der Aufnahme der Kinder in die Schule an der schulärztlichen Untersuchung teilzunehmen und die erforderlichen Angaben zu machen. Die ESU ist in Deutschland die einzige Pflichtuntersuchung im Leben eines Menschen.

Die ESU ist Voraussetzung für die Aufnahme des Kindes in der Schule. Es werden alle schulpflichtig werdenden Kinder untersucht, sowie jene, die vorzeitig in die Schule aufgenommen werden sollen und jene, die in den Vorjahren vom Schulbesuch zurückgestellt wurden. Dabei werden alle Kinder schulpflichtig, die mit Beginn eines Schuljahres (1. August) das sechste Lebensjahr vollendet haben oder bis zum folgenden 30. September vollenden werden (§ 42 Abs. 1 SchulG). Zudem können Kinder, die bis zum 31. März des folgenden Kalenderjahres das sechste Lebensjahr vollenden werden, auf Antrag der Eltern eingeschult werden (§ 42 Abs. 2 Satz 1 SchulG).

Zum Schuljahr 2017 wurde die Stichtagsregelung im SchulG insofern geändert, dass das Einschulungsalter um 3 Monate nach hinten verschoben wurde. Zuvor wurden Kinder in dem Jahr schulpflichtig, indem sie das sechste Lebensjahr vollenden werden - also bis zum 31. Dezember des Jahres, in dem das Schuljahr begann. Im Vorfeld dieser Änderung galt bereits für das Schuljahr 2016 eine Übergangsregelung, nach der eine Rückstellung der Einschulung allein aufgrund eines einfachen Antrags der Erziehungsberechtigten erfolgen konnte (§ 129 Abs. 7 SchulG). Die sogenannte „Früheinschulung“ vor dem Schuljahr 2017 führte dazu, dass viele Kinder in Berlin bereits mit fünfeinhalb Jahren eingeschult wurden. Die Früheinschulung war umstritten und führte zu gestiegenen Rückstellungsquoten, was letztlich zur Gesetzesänderung führte.

Die Eltern müssen ihre schulpflichtig werdenden Kinder (oder diejenigen die auf Antrag vorzeitig eingeschult werden sollen) im Vorjahr der Einschulung in der für sie zuständigen Grundschule anmelden. Die zuständige Grundschule wird durch das bezirkliche Schulamt basierend auf dem Einschulungsbereich des Wohnortes bestimmt (§ 55a Abs. 1 SchulG). Bei der Anmeldung erhalten die Eltern in der Regel den Termin für die ESU. Die Schule meldet alle angemeldeten Kinder dem Kinder- und Jugendgesundheitsdienst (KJGD) (§ 5 Abs. 1 Satz 1 Grundschulverordnung (GsVO)). Der KJGD der bezirklichen Gesundheitsämter führt die ESU durch. Die Untersuchungen für ein Schuljahr werden hauptsächlich von Oktober bis Juni im Jahr vor der Einschulung durchgeführt. Die Reihenfolge der Untersuchungen orientieren sich am Alter der Kinder, so dass die ältesten Kinder zuerst und jüngsten zuletzt untersucht werden sollen (§ 5 Abs. 1 Satz 2 GsVO). Das Mindestalter der zu untersuchenden Kinder beträgt 5 Jahre (§ 5 Abs. 1 Satz 3 GsVO).

2.2.2 Ziele

Eine gründliche Anamnese (griechisch, 'Erinnerung'; Erfassung der medizinischen Vorgeschichte) und körperliche Untersuchung des Kindes bietet die Gelegenheit, ein großes Spektrum an Informationen über das Kind und seine Lebensumstände sowie seine Vorgeschichte und aktuelle Symptomatiken zu erhalten. Dabei sollen der Entwicklungsstand, aber auch der gesundheitliche Zustand des Kindes durch medizinisches Fachpersonal beurteilt werden (Rosenecker und Schmidt, 2008: 4).

Die ESU zielt auf eine medizinische Beurteilung der sprachlichen, motorischen und geistigen Entwicklung des Kindes. Neben medizinischen Informationen und Beurteilungen werden soziodemografische Angaben der Familie erfasst. Für das Kind können dabei frühzeitig gesundheitliche Probleme und Gefährdungen erkannt werden. Dies bietet eine Grundlage für eine individuelle Beratung der Eltern und Förderung der Kinder im (vor-)schulischen Bereich.

Gleichzeitig sind die erhobenen Daten Grundlage für die Gesundheitsberichterstattung (GBE). Anliegen der GBE in Berlin ist, die differenzierten Lebensverhältnisse und Lebenslagen sowie die gesundheitliche Versorgungssituation im Zeitverlauf systematisch zu erfassen, darzustellen und zu bewerten (Oberwöhrmann et al., 2011: 4). In diesem Fall wird die GBE des Landes Berlin durch die Senatsverwaltung für Gesundheit, Pflege und Gleichstellung (SenGPG) realisiert, welche jährlich die „Grundauswertung zur Einschulungsuntersuchung“ veröffentlicht. Die Berichte (Beispiel: (Bettge und Oberwöhrmann, 2018)) enthalten umfassende Informationen über die Methodik der Untersuchung und tabellarische Auswertungen für das Land Berlin insgesamt sowie auf Bezirksebene.

Darüber hinaus werden einzelne Merkmale aus der ESU für die Erstellung von Bezirksregionenprofilen genutzt. Die Bezirksregionenprofile beschreiben anhand von Kernindikatoren die Berliner Bezirke zu ihrer Wohnsituation, Bevölkerungs- und Sozialstruktur und zu Entwicklungsbedingungen von Kindern und Jugendlichen. Diese sollen eine Planungs- und Entscheidungsgrundlage im Rahmen einer Sozialraumorientierung bieten (Senatsverwaltung für Stadtentwicklung, 2009: 29), (Oberwöhrmann et al., 2011: 6).

2.2.3 Inhalte

Die Bestandteile, Durchführung und Dokumentation der ESU sind standardisiert und für ganz Berlin einheitlich geregelt. Die SenGPG stimmt das Untersuchungsprogramm und den Dokumentationsbogen jährlich mit den Leitern der KJGD ab.

Inhalte der ESU sind eine allgemeine und soziale Anamnese des Kindes und der Familie, eine medizinische Anamnese des Kindes sowie ärztliche Beurteilungen und Empfehlungen. Dafür werden unter anderem erfasst:

Angaben zur ESU: eindeutiger Identifikator (ID), Untersuchungsmonat und -jahr, Untersuchernummer.

Angaben zum Kind: Wohnraum auf Planungsraum-Ebene (siehe Kapitel 2.3.1), zugewiesene Grundschule, Geburtsmonat und -jahr, Geschlecht, Zuwanderungshintergrund, Kitabesuch, durchschnittlicher täglicher Konsum elektronischer Medien, Deutschkenntnisse.

Familiärer Hintergrund: Geburtsland, Staatsangehörigkeit, Schulabschluss, berufliche Ausbildung, Erwerbstätigkeit sowie Deutschkenntnisse von Mutter und Vater, Familiensprachen, Anzahl der im Haushalt lebenden Erwachsenen und Kinder, Anzahl Raucher im Haushalt.

Medizinische Daten und Beurteilungen des Kindes: Vorsorgeuntersuchungen, Impfstatus, Geburtsgewicht, Körpergröße, Körpergewicht, Überprüfung der Hör- und Sehfähigkeiten, bisherige therapeutische Behandlungen, psychische Auffälligkeiten.

Entwicklungsdiagnostik: mittels standardisierter Testverfahren zur Erfassung des Entwicklungsstandes und von Entwicklungsstörungen (Screening des Entwicklungsstandes bei Einschulungsuntersuchungen (S-ENS) sowie ab den Untersuchungen zum Schuljahr 2012 auch zwei Subtests aus dem Sozialpädiatrischen Entwicklungsscreening für Schuleingangsuntersuchungen (SOPESS)): Körperkoordination, Visuomotorik, visuelle Wahrnehmung und Informationsverarbeitung, Sprachkenntnisse und Artikulation, Mengenvorwissen.

Empfehlungen: Empfehlung schulischer Förderungen, sozialpädagogischer Förderbedarf, Zurückstellung der Einschulung.

Die Untersuchung wird von einem Schularzt in den Räumlichkeiten des KJGD durchgeführt. Die Untersuchungsdauer beträgt ca. eine Stunde. In der Anlage A.1 finden sich als aktuelle Beispiele der Dokumentationsbogen der ESU und der Dokumentationsbogen für die S-ENS und SOPESS-Untertests für das Schuljahr 2019. Hier sind alle genauen Angaben und Antwortkategorien aufgeführt.

2.3 Sozialraumorientierung und räumliche Gliederungen

Das Bildungsmonitoring Berlin-Mitte soll vorliegende Ansätze der Bildungsberichterstattung insofern erweitern, dass spezifische regionale Gegebenheiten berücksichtigt werden können um so der Politik und der Verwaltung als Basis für lokal angepasste Handlungsempfehlungen dienen zu können (Rockmann und Leerhoff, 2018a: 18). Die Heterogenität des Bezirks Mitte in Bezug auf die Bevölkerungsstruktur und somit die Familiensituationen erfordert eine regional vertiefte Betrachtung, um handlungsrelevante Informationen zu erbringen. Die Analyse soll also nicht nur Berlin-Mitte als Bezirk insgesamt betrachten, sondern muss die sozialräumliche Gliederung des Bezirks nutzen (Rockmann und Leerhoff, 2018b: 8). Die Sozialraumorientierung ist dabei eine Strategie, die von den Bedürfnissen und Ressourcen der Bewohner eines zusammenhängenden Stadtteils ausgeht (Senatsverwaltung für Stadtentwicklung, 2009: 11).

Die Einteilung eines Stadt- oder Regionalgebietes in kleinräumige Einheiten wird räumliche Gliederung genannt. Die Berliner Variante dieser Raumgliederung wird als Regionales Bezugssystem (RBS) bezeichnet. Das RBS ist ein aktuelles zentrales Verzeichnis aller Berliner Adressen sowie der damit verbundenen Raumgliederungssystematiken (Bömermann et al., 2006: 366). Die wichtigste kleinräumige Gliederung für Berlin ist die der Lebensweltlich orientierten Räume (LOR). Daneben existieren weitere Raumgliederungen, wie beispielsweise die Einschulungsbereiche (ESB).

Innerhalb einer Sozialraumorientierung sollten insbesondere in der Jugendarbeit nicht nur Räume isoliert, sondern auch die speziellen Lebenswelten der Kinder betrachtet werden. Der Sozialraumbegriff kann also insofern erweitert werden, dass auch die Schulen und Kitas als individuelle Lebensräume der Kinder in die Betrachtung einbezogen werden und einer Situationsanalyse zu unterziehen sind (Deinet, 2009).

Bei der ESU werden Informationen zum Wohnort des Kindes, zur zuständigen Grundschule und in Berlin-Mitte zum Teil auch über die besuchte Kita erfasst, welche für sozialraumorientierte und kleinräumige Analysen genutzt werden können. Eine solche Betrachtung der Ergebnisse der ESU soll die unterschiedliche Situation in den verschiedenen Räumen aufzeigen. Dies erlaubt eine differenzierte Sozialraumentwicklung. Im Folgenden werden die im Rahmen dieser Arbeit betrachteten kleinräumigen Gliederungen vorgestellt.

2.3.1 Lebensweltlich orientierte Räume

Die Lebensweltlich orientierten Räume (LOR) stellen eine hierarchische und nach einheitlichen Kriterien gebildete regionale Gliederung für das Land Berlin dar. Dieses räumliche Bezugssystem wurde im Jahr 2006 eingeführt und soll eine sozialraumorientierte Politik ermöglichen sowie die Zusammenarbeit von verschiedenen Verwaltungsorganisationen erleichtern. Die hierarchische Gliederung umfasst drei Ebenen unterhalb der Berliner Bezirke. Die Grenzen der LOR wurden nach den Kriterien bau- und sozialstrukturell ähnlicher und zusammenhängender Wohngebiete mit ähnlicher Einwohnerzahl festgelegt (Bömermann et al., 2006: 368), (Oberwöhrmann et al., 2011: 6).

Die höchste der drei Ebenen der LOR bilden Prognoseräume (PRG), weiter untergliedert in Bezirksregionen (BZR) sowie die kleinste Einheit der Planungsräume (PLR). Im Bezirk Mitte sind dabei vier Prognoseräume, zehn Bezirksregionen und 41 Planungsräume definiert (vgl. Abbildung 2.1).

Bei der ESU wird der Wohnort der Kinder aus Datenschutzgründen nicht mit der genauen Adresse angegeben. Stattdessen wird die LOR-Ebene des Wohnortes erfasst. Dies ermöglicht eine kleinräumige Analyse auch innerhalb des Bezirkes und somit eine sozialraumorientierte Analyse innerhalb des Bildungsmonitoring Berlin-Mitte.

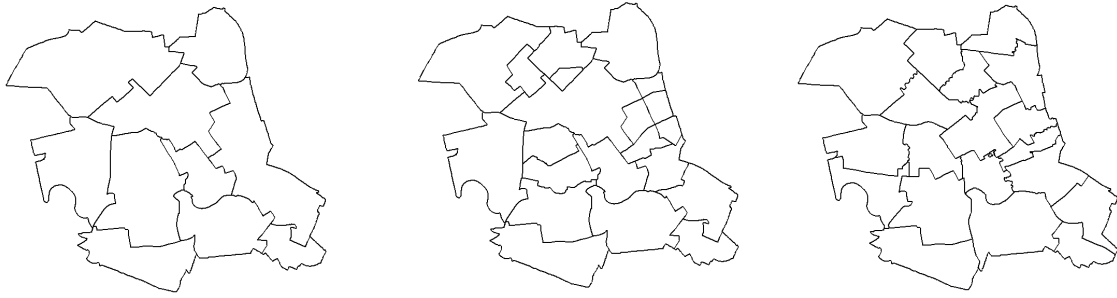


Abbildung 2.2: Einschulungsbereiche (ESB) Berlin-Mitte

Dargestellt sind die Zuschnitte der ESB für den Bezirk Berlin-Mitte für die Schuljahre 2011, 2015 und 2019 (von links nach rechts). Deutlich werden die Änderungen der ESB.

Die Geometriedaten für die ESB wurden als Shapedateien vom Bezirksamt Mitte zur Verfügung gestellt. (Eigene Darstellung)

sich für jedes Kind auch der entsprechende ESB ermitteln. Eine kleinräumige Analyse auf Basis der ESB wird somit möglich. Bei einer entsprechenden Untersuchung auf Basis der ESB im Zeitverlauf muss aber beachtet werden, dass die Ergebnisse durch die jährlichen Änderungen der ESB nicht oder nur unter Vorbehalt und Prüfung der Übereinstimmung der entsprechenden ESB über die betrachteten Schuljahre vergleichbar sein können.

Bei der Analyse nach zugewiesener Grundschule und ESB ist außerdem zu berücksichtigen, dass viele Kinder später eine andere als die zugewiesene Grundschule besuchen. Beispielsweise können die Kinder auf Elternwunsch eine andere Grundschule oder auch Privatschule besuchen. Die letztliche Schulplatzvergabe und Entscheidung erfolgt erst nach der ESU, so dass diese Information nicht in den ESU-Daten enthalten ist.

2.4 Datenbasis

Die Daten der ESU bilden im Bildungsmonitoring Berlin-Mitte eine der zentralen Analysequellen für die Situation der Kinder vor Schulbeginn. Aus den Angaben zu den Kindern und ihrer familiären Situation lassen sich viele Indikatoren ableiten, die den Bildungserfolg der Kinder beeinflussen können. Zudem bietet die ESU verschiedene Anknüpfungspunkte, um Daten aus anderen Quellen zu verbinden und somit weitergehende Analysemöglichkeiten zu schaffen. Da die ESU verpflichtend für alle einzuschulenden Kinder ist, ist gesichert, dass hieraus abgeleitete Indikatoren auch in Zukunft zur Verfügung stehen.

Die Datengrundlage für die im Rahmen dieser Arbeit zu entwickelnde Anwendung basiert auf den Daten der in Berlin-Mitte durchgeführten ESU der zu den Schuljahren 2010 - 2019 untersuchten Kinder. Die Betrachtung der Untersuchungen zu den letzten zehn Schuljahren soll die Analyse der Merkmale im Zeitverlauf ermöglichen. Des Weiteren bietet sich so die Möglichkeit, Analysen auch über mehrere Jahre hinweg durchzuführen. So können beispielsweise Subgruppenanalysen mit kleinen Fallzahlen pro Schuljahr auf der Datengrundlage mehrerer Jahre durchgeführt werden, um für verlässlichere Aussagen

eine genügend hohe Fallzahl zu erreichen. Bei der rückblickenden Betrachtung der Daten der Einschulungsuntersuchung ist zu beachten, dass das Alter der untersuchten Kinder vor der Gesetzesänderung zum Schuleintrittsalter (siehe Kapitel 2.2.1) niedriger liegt bzw. dass die Kinder ab dem Schuljahr 2017 insgesamt älter sind als in den Vorjahren. Eine Vergleichbarkeit der Ergebnisse ist somit eingeschränkt bzw. der Einfluss des Alters des Kindes ist zu berücksichtigen.

2.4.1 Erhebungsdaten

Die Daten der in Berlin-Mitte durchgeführten ESU dieser Jahre wurden dem Bezirksamt Mitte vom KJGD Berlin-Mitte zur Nutzung im Rahmen des Bildungsmonitoring Berlin-Mitte zur Verfügung gestellt. In diesen Datensätzen sind alle genauen Angaben aus den Dokumentationsbögen der ESU enthalten. Name und genaue Wohnanschrift der Kinder sind nicht bzw. nur pseudonymisiert erfasst, so dass die Anonymität gesichert ist.

Eine erste Herausforderung stellte die unterschiedliche Datenstruktur der einzelnen ESU-Datensätze pro Jahr dar. Durch die jährliche Abstimmung und Anpassung der Dokumentationsbögen wurden im Laufe der Jahre Antwortkategorien umformuliert oder geändert, einzelne Fragen neu hinzugenommen oder entfernt. So kann in jedem Jahr eine vom Vorjahr abweichende Datenstruktur entstehen. Im Vorfeld der hier vorliegenden Arbeit wurden die Einzeldatensätze der Einschulungsjahre 2010 - 2019 hinsichtlich der Datenstruktur bereits vereinheitlicht. Eine Zusammenführung der vorliegenden Einzeldatensätze wurde hierdurch ermöglicht. Insgesamt stehen damit 33.303 Datensätze der in Berlin-Mitte durchgeführten ESU als Datenbasis zur Verfügung.

2.4.2 Indikatoren

Neben den Erhebungsdaten aus der ESU sind daraus abgeleitete Merkmale von Interesse. Kern von Bildungsberichterstattung und des Bildungsmonitoring Berlin-Mitte ist die Identifikation geeigneter quantitativer und über einen längeren Zeitraum verfügbarer Indikatoren. Die Interpretation dieser Indikatoren soll es ermöglichen, die Entwicklungen im institutionalisierten Bildungsgeschehen zu verstehen, Stärken und Schwächen zu identifizieren und die Leistungsfähigkeit von Systemen zu vergleichen und somit politischen Handlungsbedarf verdeutlichen (Konsortium Bildungsberichterstattung, 2005: 2).

Aus den vorliegenden Erhebungsdaten aus der ESU wurden im Rahmen des Bildungsmonitoring bereits diverse neue Variablen und Indikatoren abgeleitet. Dazu gehören einfache Klassifizierungen wie beispielsweise das Zusammenfassen von Antwortkategorien als auch kompliziertere Berechnungen, bei denen die Zielindikatoren von mehreren Angaben abhängen. Als Beispiel sei hier der Bildungsstand der Familien als wichtige familiäre Ressource für die Kinder genannt (vgl. Abbildung 2.3). Der Bildungsstand von Mutter und Vater kann mittels der in der ESU erfassten Schul- und Berufsabschlüsse eingeordnet werden. Dabei

erfolgt die Einordnung des Bildungsstandes gemäß der International Standard Classification of Education (ISCED). Der Bildungsstand der Familie wird anschließend anhand des höchsten erreichten Bildungsstandes von Vater oder Mutter definiert (Rockmann und Leerhoff, 2019: 85).

Weitere Beispiele für berechnete Indikatoren betreffen diverse Zeitangaben. So werden im Dokumentationsbogen der ESU verschiedene Zeitangaben erfasst, beispielsweise Untersuchungsmonat und -jahr, Geburtsmonat und -jahr des Kindes, ggf. Monat und Jahr der Zuwanderung nach Deutschland und im Falle eines Kita-/Einrichtungsbesuches Monat und Jahr des Eintrittes in die Betreuungseinrichtung.

Aus diesen Angaben lassen sich weitere Merkmale des Kindes ableiten. So kann aus dem Untersuchungsmonat und -jahr in Verbindung mit dem Geburtsmonat und -jahr das Alter des Kindes bei der ESU abgeleitet werden (bzw. in Monaten geschätzt werden). Ebenso lässt sich für die nicht in Deutschland geborenen Kinder das Alter des Kindes bei seiner Zuwanderung ermitteln. Auch kann die Dauer eines Kitabesuches geschätzt werden, ein ebenfalls wichtiger Indikator für beispielsweise den Sprachstand der Kinder (vgl. Abbildung 2.3).

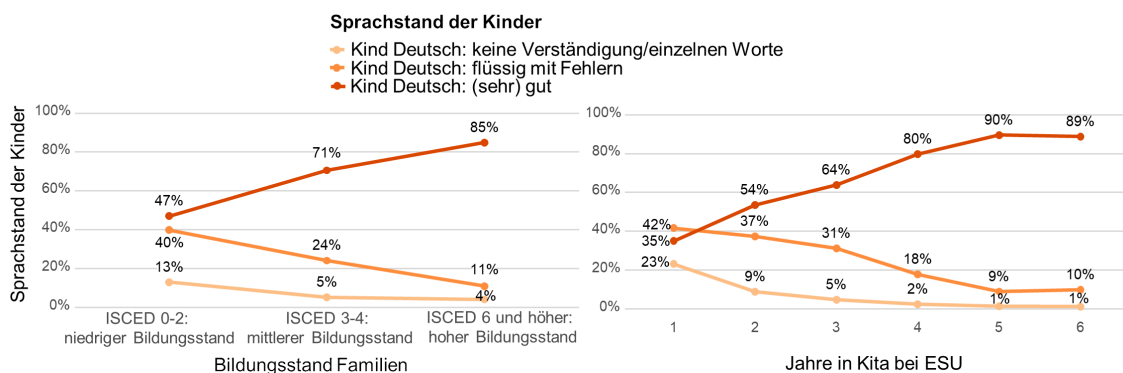


Abbildung 2.3: Indikatoren für den Sprachstand der Kinder

Die Analyse umfasst die in Berlin-Mitte untersuchten und wohnhaften Kinder der ESU der Schuljahre 2010-2019. Der starke Zusammenhang zwischen dem Sprachstand der Kinder und dem Bildungsstand seiner Familie einerseits sowie der Dauer seines Kitabesuchs andererseits tritt deutlich hervor. Die dargestellten Indikatoren sind aus Berechnungen aus den Erhebungsdaten der ESU hervorgegangen. Die Identifizierung und Ableitung von Indikatoren für den Bildungserfolg der Kinder stehen im Fokus des Bildungsmonitoring Berlin-Mitte. (Beispielgrafiken aus dem ESU explorer)

Für die vorliegende Arbeit werden somit nicht nur die Erhebungsdaten der ESU, sondern auch viele darauf basierende Indikatoren genutzt. Im Analysedatensatz befinden sich nach derzeitigem Stand ca. 300 Variablen. Etwa 180 dieser Variablen enthalten die Erhebungsdaten aus der Erfassung der Dokumentationsbögen der ESU. Weitere ca. 120 Variablen wurden berechnet und stellen Hilfsvariablen oder Indikatoren dar.

2.4.3 Daten aus weiteren Quellen

Die Daten der ESU beinhalten verschiedene Angaben, auf deren Basis weitere Datenquellen unter Berücksichtigung der datenschutzrechtlichen Verwendbarkeit verknüpft werden können. Die Vernetzung verschiedener Verwaltungsdaten und Amtlicher Statistiken kann einen Mehrwert generieren, der über den Wert einer Analyse der isolierten Daten hinaus geht. Beispielsweise ermöglicht die Angabe zum Wohnort der Kinder auf LOR-Ebene die Verknüpfung von Aggregatdaten zum entsprechenden Wohnraum aus Amtlichen Statistiken (beispielsweise der Einwohnerregisterstatistik) oder weiteren Quellen. Ohne Restriktionen sind dabei alle öffentlich verfügbaren Daten verknüpfbar.

Innerhalb der ESU wurde für die Kinder die eine Kita besuchen, in einigen Jahren die Kita-ID erfasst. Über diese ID ist es unter Berücksichtigung datenschutzrechtlicher Grundlagen theoretisch möglich, Daten zur Kita aus den Fachverfahren Integrierte Software Berliner Jugendhilfe (ISBJ) zu verknüpfen. Das ISBJ-Fachverfahren ist ein internetgestütztes zentrales IT-Verfahren auf welchem der Datenaustausch zwischen Kita-Trägern und den Jugendämtern basiert. In dieser Online-Datenbank sind Stammdaten und Merkmale der Träger und Kitas enthalten. Es lassen sich also Daten zur Kita - wie beispielsweise zum pädagogisches Personal - mit den ESU-Daten verbinden. Eine Analyse des Einflusses von Kita-Merkmalen zum Bildungsstand der Kinder kann somit ermöglicht werden.

Eine weitere Besonderheit ist die zusätzliche Verknüpfung der Angabe zur tatsächlich besuchten Grundschule der Kinder. Dies ist von besonderem Interesse, da beispielsweise im Schuljahr 2018 nur 55 % der Kinder an der zunächst zugewiesenen Schule eingeschult wurden. Wenn die Kapazität der Schule nicht ausreicht oder auf Elternwunsch können die Kinder auch eine andere als die zugewiesene Grundschule besuchen. Für das Schuljahr 2018 hat das Schulamt die aufnehmende Schule in Verbindung mit der ESU-ID für das Bildungsmonitoring Berlin-Mitte zur Verfügung gestellt. Damit wurde es ermöglicht, im ESU-Datensatz für jedes Kind die aufnehmende Grundschule anzufügen. Dies eröffnet zum einen die Möglichkeit - wie bei der Kita - Daten zu den Schulen anzufügen. Im Besonderen sollen dadurch aber auch die Sozialdaten aus der ESU bei einer Charakterisierung der Schülerschaft einer Schule innerhalb eines Schulprofils berücksichtigt werden können (Rockmann und Leerhoff, 2019: 87).

Die Auswertung von Angaben und Indikatoren auf Kita- oder Schulebene unterstützt die Sozialraumorientierung als eines der Ziele des Bildungsmonitoring Berlin-Mitte.

2.4.4 Metadaten

Als Metadaten werden übergeordnete Daten bezeichnet, die den Inhalt und die Bedeutung spezieller Daten in strukturierter Weise beschreiben. So können Metadaten als Bestandteil einer Datengrundlage verstanden werden, die Zusatzinformationen zu den eigentlichen Daten enthalten, um die eigentlichen Daten zu beschreiben und dadurch mit Informationen

anzureichern. Aus der Perspektive von Nutzern von Daten sind solche Angaben zum fachlichen Inhalt sowie Hintergrundinformationen entscheidend zur korrekten Interpretation und sachgerechten Verwendung der Daten (Lindenstruth und Claußen, 2017).

Als Metadaten der in dieser Arbeit verwendeten Datenbasis können folgende Informationen verwendet werden:

- Dokumentationsbögen für die einzelnen Schuljahre,
- Detaillierte Angaben zu den Variablen:
 - Variablenname im Datensatz,
 - Kurze Beschreibung zum Variableninhalt (Variablenlabel),
 - Ausprägungen der Variablen (Wertelabel),
 - Mess- bzw. Skalenniveau,
 - Herkunft der Variablen (Erhebungsdaten, berechnete Indikatoren, verknüpfte Daten aus weiteren Quellen),
 - Fragenummer im jährlichen Dokumentationsbogen bei Erhebungsdaten,
 - Weitere Informationen: bspw. Hinweise zur Methodik bei der Erhebung und eventuelle Änderungen im Zeitverlauf, Verwendung der Variablen, bei Indikatoren Berechnungsgrundlage, Datenqualität, Hinweise etc.

Informationen zur Datengrundlage, Methodik und Qualität sollen eine korrekte Interpretation von Analysen sicherstellen.

2.5 Ziele der Anwendungsentwicklung

Bezugnehmend auf die Ziele des Projekts Bildungsmonitoring Berlin-Mitte und die vorgestellte Datenbasis folgt in diesem Kapitel eine allgemeine und spezifizierte Zielsetzung der im Rahmen dieser Arbeit zu entwickelnden Anwendung. Dabei sollen die Ressourcen und Wünsche der Zielpersonen Berücksichtigung finden.

2.5.1 Allgemeine Zielsetzung

Zur Unterstützung der Projektbeteiligten des Bildungsmonitoring Berlin-Mitte bei der Erreichung ihrer Ziele in Hinblick auf die Bildungsberichterstattung und die Analyse der entwickelten Indikatorensets soll im Zuge dieser Arbeit eine Anwendung entwickelt werden, die es ermöglicht, Zugriff auf die vorliegenden Daten zu erhalten und auf einfache Weise differenzierte Auswertungen zu erstellen.

Das Ziel dieser Arbeit ist demnach die Entwicklung eines einfach zu bedienenden Analysewerkzeugs zur Untersuchung der Daten der ESU, daraus abgeleiteter Indikatoren und weiterer verknüpfter Daten. So sollen Zusammenhänge zwischen Merkmalen und Indikatoren untersucht werden können. Auch die Betrachtung von Langzeitentwicklungen

über die letzten Jahre hinweg ist von Interesse. Ein spezielles Ziel ist darüber hinaus die Analysemöglichkeit auf kleinräumiger Ebene um die Sozialraumorientierung als eines der Ziele des Bildungsmonitoring Berlin-Mitte zu unterstützen. Ein möglichst intuitiver, schneller und direkter Zugang zu den Ergebnissen soll dabei eine effiziente Betrachtung der Datenlage ermöglichen. Ein Schwerpunkt ist dabei die grafische Darstellung von Analyseergebnissen.

2.5.2 Zielpersonen

Zielpersonen bzw. Anwender sind zunächst die Projektbeteiligten, Mitarbeiter und Kooperationspartner des Bildungsmonitoring Berlin-Mitte. Den Anwendern soll es ermöglicht werden, innerhalb der normalen IT-Infrastruktur des Bezirks einen möglichst eingängigen Zugang zu den Daten zu erlangen. Je nach Verantwortlichkeit der Mitarbeiter sollen diese mittels der Anwendung befähigt werden, einfache und auch tiefergehende Analysen selbstständig durchzuführen, um Entscheidungsgrundlagen zu schaffen oder auch Anfragen von Bezirksmitarbeitern, Schulplanern, Politikern oder Presse zu beantworten. Es kann davon ausgegangen werden, dass die zukünftigen Anwender über grundlegende, aber nicht unbedingt vertiefte Statistikkenntnisse verfügen.

Die Spezifikation der Anforderungen an die zu entwickelnde Anwendung erfolgte in enger Abstimmung mit der Projektleiterin des Bildungsmonitoring Berlin-Mitte sowie den Projektbeteiligten und Kooperationspartnern. Im Vorfeld der Anwendungsentwicklung wurde die grundlegende Idee den folgenden Beteiligten vorgestellt und deren Hinweise und Wünsche dazu in die Zielsetzung und Anforderungsspezifikation integriert:

- Institut für Schulqualität der Länder Berlin und Brandenburg e.V.: Projektleiterin des Bildungsmonitoring Berlin-Mitte,
- Bezirksamt Mitte von Berlin: Organisationseinheit für Qualitätssicherung, Planung und Koordination des öffentlichen Gesundheitsdienstes: Aufgabenbereich Gesundheits- und Sozialberichterstattung,
- Bezirksamt Mitte von Berlin: SprachFörderZentrum Berlin-Mitte,
- Bezirksamt Mitte von Berlin: Jugendamt Mitte,
- Gesundheitsamt Mitte: Kinder- und Jugendgesundheitsdienst (KJGD).

Die Einbindung der beteiligten Institutionen und zukünftigen Nutzer der Anwendung soll eine bedarfsgerechte Zielsetzung sicherstellen.

2.5.3 Anforderungsspezifikation

Die hier spezifizierten Anforderungen an die zu entwickelnde Anwendung sollen als Grundlage für Planung, Entwicklung, Test und Beschreibung dienen. Die technischen und inhaltlichen Anforderungen an die Anwendung werden wie folgt festgelegt:

Lauffähigkeit: Ein wichtiges Kriterium an die Anwendung ist die problemfreie Integration in die IT-Infrastruktur des Bezirkes. Die Anwendung soll dafür auf grundsätzlich kostenfreien Programmen basieren (Open Source Software).

Benutzerfreundlichkeit: In Hinblick auf die verschiedenen Nutzer der Anwendung soll die Bedienung möglichst einfach und mittels intuitiver Benutzeroberfläche realisiert werden. Auch sollen sich innerhalb der Anwendung hilfreiche Informationen zur effizienten Nutzung finden.

Visualisierung der Daten: Eine grafische Darstellung der Datenlage soll durch den Einsatz von visuellen Elementen wie verschiedenen Diagrammart eine intuitive Methode zur Erkennung und zum Verständnis von Mustern, Trends und Zusammenhängen in den Daten bieten. Ziele: Erkenntnisgewinn, Zeitersparnis, Erleichterung der Kommunikation, Unterstützung datengesteuerter Entscheidungen.

Kartenerstellung: Ein besonderer Fokus liegt auf der Möglichkeit, in einfacher Weise kleinräumige Analysen auf einer Karte von Berlin-Mitte darzustellen. Anhand von interaktiven Karten sollen räumliche Muster in den Daten hinsichtlich einer effizienten Sozialraumorientierung festgestellt werden.

Tabellarische Auswertung der Daten: Die genauen Analyseergebnisse sollen darüber hinaus anhand verschiedener statistischer Maßzahlen tabellarisch dargestellt werden. Die genaue Datenlage, welche sich in ihrer Gesamtheit nicht immer durch eine Visualisierung darstellen lässt, soll dadurch umfassend beschrieben werden.

Zusammenhangsanalyse: Insbesondere in Hinblick auf die Indikatoren sollen Zusammenhänge zwischen diesen untersucht werden können. Mittels passender Zusammenhangsmaße soll ein statistischer relevanter Zusammenhang geprüft werden.

Zugang zu Metadaten: Zusätzliche Informationen zu den Analysevariablen sollen die korrekte Verwendung und Interpretation sicherstellen.

Filterführungen: Es sollen die Daten nicht nur in ihrer Gesamtheit, sondern insbesondere auch Subgruppen analysiert werden können. Eine einfach zu bedienende Möglichkeit der Filterführung einer Analyse soll bereitgestellt werden.

Export von Analyseergebnissen: Alle Ergebnisse einer Analyse sollen möglichst einfach aus der Anwendung heraus zur weiteren Verwendung exportiert werden können.

Speichern und Laden von Analysen: Auf besonderen Wunsch der zukünftigen Nutzer sollen Einstellungen für durchgeführte Analysen gespeichert und wieder geladen werden können. Eine genaue Reproduktion von durchgeführten Analysen soll so sichergestellt werden.

Basierend auf diesen Anforderungen werden im folgenden Kapitel zunächst die statistischen Methoden ausgewählt und spezifiziert. Anschließend erfolgt die genaue Beschreibung des Entwicklungsprozesses der Anwendung.

3 Auswahl statistischer Methoden

Statistik ist die Wissenschaft vom Sammeln, Beschreiben, Analysieren, Interpretieren und Präsentieren von Daten (Härdle et al., 2015: 1). In diesem Kapitel erfolgt eine Auswahl und Vorstellung der statistischen Methoden, welche durch die zu entwickelnde Analyseanwendung zur Verfügung gestellt werden sollen. Die Auswahl soll sich dabei auf wenige elementare Verfahren beschränken. Ein Fokus liegt dabei auf der Datenvisualisierung und dabei insbesondere auch auf thematischen Karten. Bei der Kartendarstellung soll zudem ein innovatives Verfahren der Darstellung nach der Methode der simulierten Geokoordinaten vorgestellt werden. Alle in diesem Kapitel gezeigten Grafiken wurden mit der entwickelten Analyseanwendung - dem ESU explorer - erzeugt und geben einen Einblick in die Situation der Kinder Berlin-Mittes vor ihrem Schuleintritt.

3.1 Einführung

Als Basis für die folgende Darstellung statistischer Methoden sollen zunächst die verwendeten Begriffe erläutert werden. Anschließend erfolgt eine Beschreibung der Teilgebiete der Statistik. Hier wird auch der Begriff *explorative Datenanalyse* erläutert und die Schwerpunktsetzung auf dieses Gebiet begründet. Zudem erfolgt eine kritische Betrachtung der Anwendung induktiver Statistik bei Vollerhebungen.

3.1.1 Begriffsbildungen

Objekte deren Eigenschaften von Interesse sind, werden *statistische Elemente* genannt. Eine Gesamtheit statistischer Elemente, die sich sachlich, zeitlich und örtlich klar abgrenzen lassen, heißt *Grundgesamtheit*. Im Falle der in dieser Arbeit verwendeten Datengrundlage umfasst die Grundgesamtheit die einzuschulenden Kinder (sachlich) bezüglich der Schuljahre 2010-2019 (zeitlich) wohnhaft in Berlin-Mitte (örtlich).

Eine beobachtbare Eigenschaft eines statistischen Elementes heißt *statistische Variable* oder *Merkmal*. Die Ausprägungen oder Messwerte der statistischen Variablen, heißen *Merkmalsausprägungen*. Die statistischen Merkmale lassen sich nach den Beziehungen zwischen den Merkmalsausprägungen - ihrem *Skalen- oder Messniveau* - klassifizieren. Hier entspricht die *Nominalskala* dem niedrigsten Skalenniveau. Dabei stellen die Merk-

malsausprägungen lediglich eine Klassifikation dar. Hier liegt keine natürliche Ordnung vor. Als Beispiel sei hier das Geschlecht einer Person genannt. Hier ist ausschließlich ein Vergleich auf Äquivalenz möglich. Lässt sich die Ordnung der Merkmalsausprägungen sinnvoll interpretieren, so liegt eine *Ordinalskala* vor. Hier lässt sich neben der Äquivalenz auch die Reihenfolge sinnvoll interpretieren, nicht jedoch der Abstand zwischen den Ausprägungen. Ein Beispiel stellt hier die Erfassung des Schulabschlusses in aufsteigender Ordnung dar. Lässt sich über Äquivalenz und Reihenfolge auch der Abstand zwischen zwei Merkmalsausprägungen sinnvoll interpretieren, so spricht man von *metrisch skalierten* Variablen. Hier sei als Beispiel die Körpergröße einer Person in Zentimetern genannt.

Nominal und ordinal skalierte Variablen sowie metrisch skalierte Variablen mit wenigen Merkmalsausprägungen werden zusammengefasst als *kategoriale Variablen* bezeichnet. Die entsprechenden Merkmalsausprägungen heißen dann *Kategorien*.

Wenn von allen statistischen Einheiten einer Grundgesamtheit die interessierenden statistischen Merkmale erhoben werden, so spricht man von einer *Vollerhebung*. Wird im Gegensatz dazu nur eine Teilmenge der Grundgesamtheit untersucht, spricht man von einer *Stichprobe* oder *Teilerhebung*.

3.1.2 Teilgebiete der Statistik und explorative Datenanalyse

Traditionell wird zwischen *deskriptiver* und *induktiver* Statistik unterschieden (Härdle et al., 2015: 1). Die deskriptive Statistik soll Daten anschaulich, übersichtlich und verständlich darstellen, indem diese zusammengefasst und visualisiert werden. Die Instrumente der deskriptiven Statistik umfassen beispielsweise Häufigkeitstabellen, Maßzahlen zur Lage und Streuung von Merkmalen sowie Zusammenhangsmaße.

Die induktive Statistik befasst sich mit den Rückschlüssen von Daten einer Stichprobe auf die Eigenschaften der dahinterliegenden Grundgesamtheit durch stochastische Modelle. Statistische Testverfahren sollen dabei helfen, Annahmen über die zugrundeliegende Grundgesamtheit zu überprüfen. Eine kritische Betrachtung der Anwendung von induktiven Verfahren bei Vollerhebungen findet sich im Kapitel 3.1.3.

Der Begriff *Explorative Datenanalyse (EDA)* umfasst einen neueren Zweig der Statistik, der auf deskriptiven Methoden basiert und damit als Teilgebiet der deskriptiven Statistik betrachtet werden kann (Polasek, 1994: 3). Der Begriff wurde durch (Tukey, 1977) geprägt. Dabei sollen zunächst einfache deskriptive und grafische Analysen effektiv helfen, Daten zu beschreiben und zu überblicken. Die EDA hat darüber hinaus zum Ziel, bisher unbekannte Strukturen und Zusammenhänge zwischen Daten aufzudecken und die Auswahl von weiteren statistischen Methoden zu unterstützen. Ein äußerst wichtiger Aspekt der EDA ist die Datenvisualisierung. Grafiken bieten einen sehr intuitiven Zugang zu Daten. Denn es ist nicht möglich, mit Worten auszudrücken, was eine Darstellung oder ein Diagramm zu zeigen vermag (Tukey, 1977: 56). Verschiedene Darstellungsformen können Verteilungen,

Maßzahlen und Zusammenhänge visualisieren. Vor allem für den Vergleich verschiedener Gruppen und Zusammenhangsanalysen bieten Grafiken einen direkten Zugang zu den Ergebnissen. Trends, Muster oder Zusammenhänge werden durch die bildliche Darstellung einfacher sichtbar.

Die folgende Auswahl statistischer Methoden wird hauptsächlich aus dem Bereich der EDA erfolgen und begründet sich in der Zielsetzung der Arbeit (siehe Kapitel 2.5). Ein Schwerpunkt liegt dabei in verschiedenen Formen der Datenvisualisierung.

3.1.3 Zur Anwendung induktiver Statistik bei Vollerhebungen

Wir können bei der Betrachtung der ESU-Daten von einer Vollerhebung sprechen, da die ESU eine verpflichtende Untersuchung und Voraussetzung für den Schulbesuch des Kindes ist (siehe Kapitel 2.2). Alle einzuschulenden Kinder werden untersucht.

Die Frage, ob induktive statistische Methoden auf Vollerhebungsdaten angewendet werden können ist umstritten und es existiert keine eindeutige Lösung für dieses Problem (Behnke, 2005). Induktive Verfahren wie die Anwendung von Signifikanztests beruhen auf der Annahme, dass die zugrunde liegenden Daten einer Zufallsstichprobe entstammen. Bei einer Vollerhebung werden jedoch alle zur Grundgesamtheit gehörenden Elemente untersucht. Die Anwendung von Testverfahren kann dennoch als sinnvoll betrachtet werden, wenn die Daten durch stochastische Komponenten beeinflusst werden (Broscheid und Gschwend, 2003). Die Unsicherheit quantitativer Daten hat verschiedene Quellen, von denen die Stichprobenziehung nur eine ist.

Berücksichtigen muss man in diesem Zusammenhang das Problem des Messfehlers (Broscheid und Gschwend, 2003), (Broscheid und Gschwend, 2005). Messfehler bedeuten, dass das was wir messen, nicht unbedingt das ist was wir messen wollen. Die Annahme, dass die Messfehler vernachlässigbar gering ausfallen, ist gerade in den Sozialwissenschaften selten gerechtfertigt. Wenn die Schwankungen eines Messwertes zumindest teilweise auf Messfehler zurückgeführt werden können, ist es sinnvoll auch bei Vollerhebungen Signifikanztests durchzuführen (Behnke, 2005).

Bei der ESU werden viele Daten erhoben, die nicht direkt beobachtbar sind, sondern erfragt werden. Dabei schleichen sich bei der Datenerhebung zwangsläufig Fehler ein - sozialwissenschaftliche Schlussfolgerungen sind daher inhärent unsicher (Broscheid und Gschwend, 2003). Bei der Erhebung der Daten der ESU können auf verschiedenen Ebenen Messfehler entstehen: beispielsweise durch unvollständige Angaben der Eltern, Verständigungsschwierigkeiten oder die Übertragung der Daten in ein Erfassungssystem. Zusammenfassend lässt sich festhalten, dass die Vollerhebungsdaten durch eine Reihe stochastischer Prozesse beeinflusst werden. Aufgrund dieser Überlegungen werden auch wenige induktive Verfahren in der folgenden Auswahl statistischer Methoden berücksichtigt.

3.2 Häufigkeitsverteilungen

Als einfachste Stufe der Analyse können zunächst einfache Auszählungen dienen. Dabei werden die Individuen zu zwei oder mehr sich einander ausschließenden Kategorien zugeordnet. Dies kann bei allen kategorialen Variablen, also unabhängig vom Skalenniveau durchgeführt werden. Durch ein Auszählen des Auftretens der Kategorien erhält man eine *Häufigkeitsverteilung* (Härdle et al., 2015: 21).

3.2.1 Eindimensionale Häufigkeiten

Sei N die Anzahl statistischer Elemente und X eine statistische Variable mit k Merkmalsausprägungen. Dann wird die Zahl von Beobachtungen, die in eine bestimmte Kategorie fallen, *Häufigkeit* genannt (Härdle et al., 2015: 17 ff.), (Schlittgen, 2008: 12).

Die Anzahl an Beobachtungen pro Kategorie ist die *absolute Häufigkeit* und wird definiert durch:

$$h(X = x_i) = h(x_i) = h_i, \quad i = 1, \dots, k$$

mit den Eigenschaften:

$$0 \leq h(x_i) \leq N \quad \text{und} \quad \sum_{i=1}^k h(x_i) = N$$

Der Anteil an Beobachtungen ist die *relative Häufigkeit*:

$$f(X = x_i) = f(x_i) = f_i = \frac{h(x_i)}{N}$$

mit den Eigenschaften:

$$0 \leq f(x_i) \leq 1 \quad \text{und} \quad \sum_{i=1}^k f(x_i) = 1$$

Relative Häufigkeiten werden oft in Prozentsätzen angegeben. Dies entspricht $100 \cdot f(x_i)$. Betrachtet man alle Ausprägungen einer einzelnen Variable, so spricht man von einer *eindimensionalen Häufigkeitsverteilung*. Für kategoriale Variablen X entspricht eine *Häufigkeitstabelle* den Häufigkeiten der k Kategorien dieser Variable.

3.2.2 Kontingenztabellen

Wenn die Kombinationen von Kategorien von zwei Merkmalen X und Y untersucht werden, ergibt sich eine *bivariate* oder *zweidimensionale Häufigkeitsverteilung*. Hier wird die Beziehung zwischen zwei Variablen untersucht. Zentral sind hier die Fragen nach dem

Zusammenhang (der *Kontingenz*) zwischen beiden Variablen. Es werden die *gemeinsamen Häufigkeiten* dargestellt. Die absoluten bzw. relativen gemeinsamen Häufigkeiten ergeben sich aus der Häufigkeit, mit der X den Wert $x_i, i = 1, \dots, k$ und Y den Wert $y_j, j = 1, \dots, m$ angenommen hat:

$$h(X = x_i, Y = y_j) = h(x_i, y_j) = h_{ij} \quad \text{bzw.} \\ f(X = x_i, Y = y_j) = f(x_i, y_j) = f_{ij} = \frac{h_{ij}}{N}$$

mit den Eigenschaften:

$$\sum_{i=1}^k \sum_{j=1}^m h_{ij} = N \quad \text{und} \quad \sum_{i=1}^k \sum_{j=1}^m f_{ij} = 1$$

Bivariate Häufigkeitstabellen werden auch *Kontingenz-* oder *Kreuztabellen* genannt. Am rechten und unteren Rand stehen hier die Zeilen- bzw. Spaltensummen - die *Randhäufigkeiten* für die einzelnen Variablen. Diese entsprechen den univariaten absoluten Häufigkeiten der einzelnen Variablen X und Y :

$$h_{i\bullet} = h_{i1} + \dots + h_{im} = \sum_{j=1}^m h_{ij} = h(X = x_i) = h(x_i), \quad i = 1, \dots, k \\ h_{\bullet j} = h_{1j} + \dots + h_{kj} = \sum_{i=1}^k h_{ij} = h(Y = y_j) = h(y_j), \quad j = 1, \dots, m$$

Dies gilt ebenso für die relativen Randhäufigkeiten:

$$f_{i\bullet} = f_{i1} + \dots + f_{im} = \sum_{j=1}^m f_{ij} = f(X = x_i) = f(x_i), \quad i = 1, \dots, k \\ f_{\bullet j} = f_{1j} + \dots + f_{kj} = \sum_{i=1}^k f_{ij} = f(Y = y_j) = f(y_j), \quad j = 1, \dots, m$$

Für die absoluten und relativen Randhäufigkeiten gelten die zu oben korrespondierenden Eigenschaften:

$$h_{\bullet\bullet} = \sum_{i=1}^k \sum_{j=1}^m h_{ij} = \sum_{i=1}^k h_{i\bullet} = \sum_{j=1}^m h_{\bullet j} = N \quad \text{und} \\ f_{\bullet\bullet} = \sum_{i=1}^k \sum_{j=1}^m f_{ij} = \sum_{i=1}^k f_{i\bullet} = \sum_{j=1}^m f_{\bullet j} = 1$$

Daraus abgeleitet werden die *bedingten relativen Häufigkeiten*. Speziell für Vergleich zweier kategorialer Merkmale interessieren dabei die Verteilungen des einen Merkmals für eine feste Ausprägung des anderen Merkmals. Dabei ist die bedingte relative Häufigkeit

mit der eine Variable X den Wert x_i angenommen hat unter der Bedingung das Variable Y den Wert y_j angenommen hat und $h(Y = y_j) > 0$ wie folgt definiert

$$f(X = x_i|Y = y_j) = \frac{h(Y = y_j, X = x_i)}{h(Y = y_j)} = \frac{h_{ij}}{h_{\bullet j}} = \frac{h_{ij}/N}{h_{\bullet j}/N} = \frac{f_{ij}}{f_{\bullet j}}$$

Über zweidimensionale Betrachtungen hinaus können auch mehr als zwei Merkmale gemeinsam in einer Häufigkeitstabelle dargestellt werden. Mit zunehmender Anzahl von betrachteten Merkmalen sinkt allerdings die Übersichtlichkeit von Kontingenztabellen.

3.2.3 Grafische Darstellungsformen

Säulen- und Balkendiagramme

Das Säulendiagramm ist eine der am häufigsten verwendeten Diagrammarten. Häufigkeiten für kategoriale Variablen lassen sich über Säulendiagramme einfach visualisieren. Jede Kategorie wird dabei durch eine Säule repräsentiert, deren Höhe proportional zur Häufigkeit ist. Die Stärke des Säulendiagramms liegt dabei im Vergleich verschiedener kategorialer Merkmale (Härdle et al., 2015: 22). Bei der Darstellung können die Säulen gestapelt oder gruppiert dargestellt werden. Die gruppierte Darstellung lässt einen Vergleich der einzelnen Kategorienhäufigkeiten zu. Alternativ sind gestapelte Säulen bei vielen Kategorien oder Gruppenvergleichen als platzsparende Variante geeignet.

Balkendiagramme sind zunächst nur gekippte Säulendiagramme. Balkendiagramme eignen sich ebenfalls gut für Vergleiche. Dabei werden die Ergebnisse mittels vertikaler Balken dargestellt, deren Längen sich bei der Betrachtung besonders einfach vergleichen lassen. Ein weiterer Vorteil liegt darin, dass sich lange Beschriftungen besser darstellen lassen. Die Abbildung 3.1 zeigt die verschiedenen Varianten von Säulen- und Balkendiagrammen.

Liniendiagramme

Um den Verlauf von Trends und zeitlichen Entwicklungen darzustellen, eignen sich Liniendiagramme besonders gut. Dabei werden auf der horizontalen Achse die Zeitangaben dargestellt, die vertikale Achse zeigt bestimmte Werte zu diesen Zeitpunkten (vgl. Abbildung 3.2). Neben Zeitreihen können Liniendiagramme natürlich auch für weitere zweidimensionale Auswertungen genutzt werden. Dabei kann ein Diagramm auch mehrere Linien enthalten. Dabei ist aber auf die Übersichtlichkeit achten.

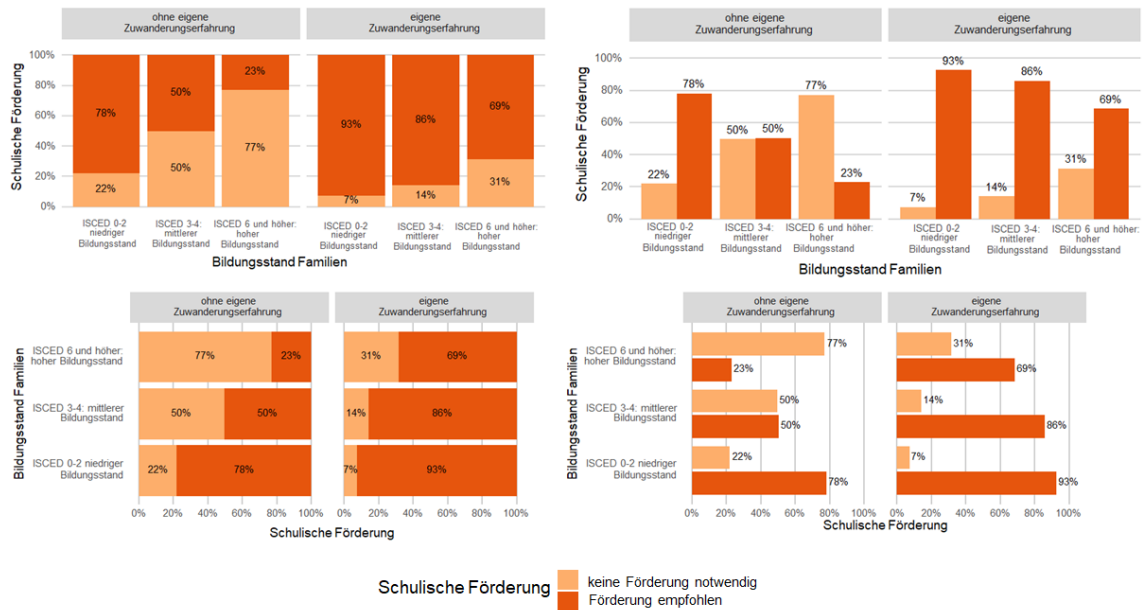


Abbildung 3.1: Varianten Säulen- und Balkendiagramme: Empfehlung zur schulischen Förderung nach dem Bildungsstand der Familien und der Zuwanderungserfahrung des Kindes, Schuljahr 2019

Die Diagramme zeigen identische Analysen bedingter relativer Häufigkeiten. Von links oben nach rechts unten: gestapelte Säulen, gruppierte Säulen, gestapelte Balken, gruppierte Balken. Sehr deutlich wird der Zusammenhang zwischen dem Bildungsstand der Familien und einer Empfehlung zur schulischen Förderung des Kindes. Dies trifft sowohl auf die Kinder ohne als auch mit Zuwanderungserfahrung zu. Über alle Bildungsstände hinweg wird Kindern mit Zuwanderungserfahrung deutlich häufiger eine schulische Förderung empfohlen. (Beispielgrafiken aus dem ESU explorer)

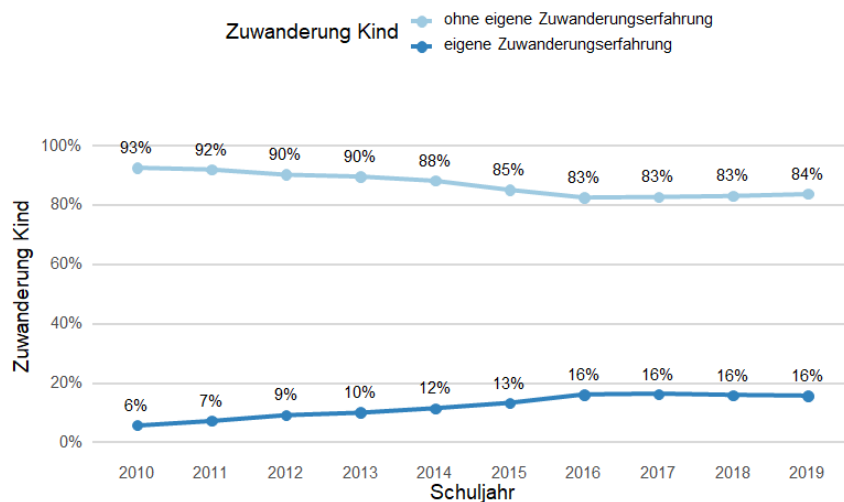


Abbildung 3.2: Liniendiagramm: Zuwanderungserfahrung der Kinder nach Schuljahren
Liniendiagramme eignen sich besonders gut für die Darstellung von zeitlichen Verläufen. Bis zum Schuljahr 2015 stieg der Anteil der Kinder mit eigener Zuwanderungserfahrung kontinuierlich an. Seit dem Schuljahr 2016 stagniert dieser Anteil bei konstanten 16 %. (Beispielgrafik aus dem ESU explorer)

Kreisdiagramme

In Kreisdiagrammen werden Häufigkeiten als Segmente eines Kreises dargestellt. Die Fläche jedes Segmentes ist proportional zur korrespondierenden relativen Häufigkeit (vgl. Abbildung 3.3). Diese Art der Informationsvermittlung kann jedoch nachteilig sein, da der Mensch besser beim Beurteilen linearer Maße als bei der Beurteilung relativer Flächen ist (Schlittgen, 2008: 14). Für den Vergleich vieler Kategorien oder Vergleichsgruppen sind daher die vorher genannten Darstellungsvarianten vorzuziehen.

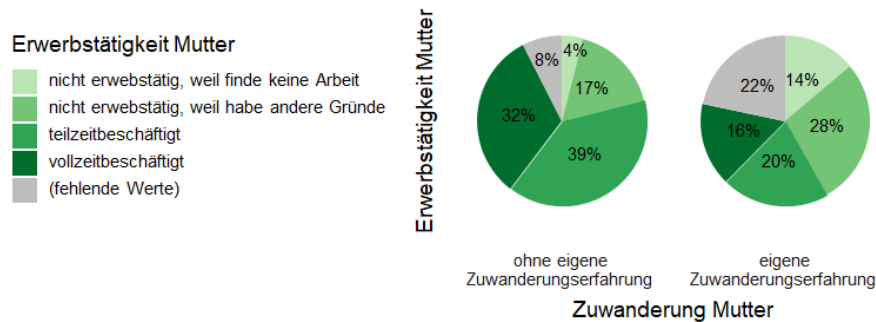


Abbildung 3.3: Kreisdiagramm: Erwerbstätigkeit der Mutter nach Zuwanderungserfahrung der Mutter, Schuljahr 2019

Für den Vergleich weniger Kategorien und Vergleichsgruppen eignen sich Kreisdiagramme. Hier deutlich wird der Zusammenhang zwischen Zuwanderungserfahrung der Mutter und ihrer Berufstätigkeit: Mütter ohne Zuwanderungserfahrung sind deutlich häufiger berufstätig als Mütter mit Zuwanderungserfahrung. Darüber hinaus ist erkennbar, dass Mütter mit Zuwanderungserfahrung deutlich häufiger keine Angabe zu dieser Frage machen wollen oder können. (Beispielgrafik aus dem ESU explorer)

3.3 Verteilungsmaßzahlen

Neben der Beschreibung einer Häufigkeitsverteilung ist es zweckmäßig, die Daten weiter zu verdichten. Die Beschreibung mittels weniger Maßzahlen erlaubt die Darstellung wichtiger Eigenschaften von Verteilungen (Schlittgen, 2008: 39). Die folgend vorgestellten Maßzahlen sollen insbesondere zur Beschreibung metrischer Merkmale dienen.

3.3.1 Rangmaßzahlen

Alle Rangmaßzahlen basieren auf dem Konzept der geordneten, also der Größe nach sortierten Urliste der Länge N eines Merkmales X , die wir *Rangliste* nennen $x_{(1)} \leq \dots \leq x_{(N)}$. Die Rangmaßzahlen werden über die Teilung der Rangliste definiert. Da Rangmaßzahlen auf der Ordnung von Daten beruhen, können diese nur für ordinale oder metrische Merkmale Anwendung finden. Rangmaßzahlen beschreiben eine Verteilung auf sehr robuste Weise.

Extremwerte

Zu den Rangmaßzahlen gehören zunächst die *Extremwerte*. Die Extremwerte von X sind das größte und kleinste Element der Rangliste: das Minimum $x_{\min} = x_{(1)}$ und das Maximum $x_{\max} = x_{(N)}$. Extremwerte können als einfache Verteilungsmaßzahlen angesehen werden. Sie stecken den Bereich ab, in den die Merkmalsausprägungen fallen.

Quartile

Als weitere Rangmaßzahlen gelten die *Quartile*, die die Rangliste in etwa vier gleich große Teile unterteilen. Dabei wird für einen Anteil p das p -Quartil x_p bestimmt. Dieser Wert unterteilt die Rangliste in die ersten $(100p\% \cdot N)$ der Beobachtungen und die letzten $((1 - 100p\%) \cdot N)$ Beobachtungen. So gibt das $x_{0,25}$ -Quartil die Zahl an, für die 25 % der Merkmalswerte kleiner oder gleich und 75 % der Merkmalswerte größer oder gleich sind. Zwischen dem ersten (unteren) Quartil $x_{0,25}$ und dem dritten (oberen) Quartil $x_{0,75}$ befinden sich 50 % aller Daten. Das zweite (mittlere) Quartil ($x_{0,5}$) teilt die Rangliste in etwa zwei gleich große Teile und spiegelt das Zentrum einer Verteilung.

3.3.2 Lagemaße

Ein Wert, um den sich Messungen gruppieren, beschreibt die *Lage* einer univariaten Verteilung als einzelne Maßzahl. Eine solche Lagemaßzahl soll möglichst zentral, also dicht bei den beobachteten Daten liegen bzw. dessen Zentrum widerspiegeln. Ein geeigneter Lageparameter soll also möglichst kleine Entfernungen zu den beobachteten Werten aufweisen (Polasek, 1994: 163), (Schlittgen, 2008: 39).

Modus

Ein einfacher Lageparameter ist der *Modus* \tilde{x} . Der Modus entspricht derjenigen Kategorie, welche die größte Häufigkeit hat und kann somit direkt aus einer vorliegenden Häufigkeitstabelle entnommen werden. Der Modus wird als Lageparameter nominal skalierten Daten empfohlen, kann aber auch bei ordinal und metrisch skalierten Daten sinnvoll sein (Schlittgen, 2008: 48). Der Modus ist gegenüber extrem abweichenden Werten (Ausreißern) unempfindlich. Ein Nachteil des Modus ist seine relative Ungenauigkeit beispielsweise bei mehreren gleich oder ähnlich häufig auftretenden Messwerten.

Median

Für mindestens ordinale Merkmale kann der *Median* angegeben werden. Der Median \tilde{x} eines Datensatzes der Länge N ist der mittlere Wert einer Rangliste und ergibt sich aus:

$$\tilde{x} = \begin{cases} x_{(\frac{N+1}{2})} & \text{bei ungeradem } N \\ \frac{1}{2}(x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}) & \text{bei geradem } N \end{cases}$$

Der Median ist identisch zum mittleren Quartil ($x_{0,5}$) und teilt die geordneten Daten in zwei Hälften. Die eine Hälfte der Daten ist nicht größer als der Median, die andere Hälfte nicht kleiner. Als wichtige Eigenschaft des Medians gilt seine Robustheit gegenüber Ausreißern.

Arithmetisches Mittel

Voraussetzung für die Berechnung des *arithmetischen Mittels* sind metrische Daten. Das einfache arithmetische Mittel \bar{x} ergibt sich aus der Summe der betrachteten Werte x_1, \dots, x_N geteilt durch ihre Anzahl N und ist somit gegeben durch:

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n$$

Die Verwendung des arithmetischen Mittels empfiehlt sich, wenn die Daten annähernd symmetrisch und eingipflig verteilt sind. Bei asymmetrischen Verteilungen ist es sinnvoll, den Median zu benutzen, da er stets zwischen Modus und arithmetischem Mittel liegt (Clauß et al., 2004: 35). Gegenüber Modus und Median reagiert das arithmetische Mittel relativ sensibel auf Ausreißer.

3.3.3 Streuungsmaße

Neben der Lage ist die *Streuung* die wichtigste Eigenschaft einer Verteilung. Streuungsmaßzahlen messen die Variabilität in den Daten und beschreiben somit die Abweichung vom Zentrum einer Verteilung. Die Erfassung der Streuung eines Merkmals basiert auf Abstandsmaßen, sie ist somit an ein metrisches Skalenniveau des entsprechenden Merkmals gebunden (Schlittgen, 2008: 50). Streuungsmaßzahlen sind auch ein Maß der Unsicherheit im Bereich der Merkmalsausprägungen. Je größer die Streuung, desto unschärfer ist die Beschreibung des Merkmals durch einen Lageparameter (Polasek, 1994: 186).

Varianz und Standardabweichung

Die Varianz ist neben Mittelwert und Median eine der wichtigsten Verteilungsmaßzahlen (Polasek, 1994: 187). Für ein Merkmal x_1, \dots, x_N mit dem arithmetischen Mittel \bar{x} ist die *Varianz* s_x^2 das arithmetische Mittel der quadrierten Abweichungen vom Mittelwert:

$$s_x^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})^2$$

Die *Standardabweichung* s_x ergibt sich aus der Wurzel der Varianz:

$$s_x = \sqrt{\frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})^2}$$

Die Varianz hat eine andere Dimension als die Daten selbst und ist somit nicht intuitiv. Mit dem Übergang zur Standardabweichung liegen die Daten wieder in derselben Dimension wie die Beobachtungen vor. Die Standardabweichung ist also anschaulicher als die Varianz.

Spannweite und Quartilsabstand

Die Varianz und Standardabweichung bilden die Abweichungen der Werte vom arithmetischen Mittel. Ein weiterer Ansatz geht von Differenzen zwischen den Beobachtungen selbst aus (Schlittgen, 2008: 57).

Die *Spannweite* R (engl. range) eines Merkmals x_1, \dots, x_N ist der Abstand der Extremwerte - vom minimalen Wert x_{min} zum maximalen Wert x_{max} - und ergibt sich somit aus:

$$R = x_{max} - x_{min}$$

Die Spannweite zeigt die gesamte Ausbreitung der Daten, reagiert jedoch stark auf einzelne extreme Werte.

Ein weiteres Streuungsmaß, welches unempfindlich gegenüber Extremwerten ist, ist der Interquartilsabstand, kurz *Quartilsabstand*. Der Quartilsabstand s_Q kennzeichnet den Abstand zwischen dem unteren Quartil ($x_{0,25}$) und oberem Quartil ($x_{0,75}$) :

$$s_Q = x_{0,75} - x_{0,25}$$

Innerhalb des Quartilsabstandes liegen 50 % der Werte, er kennzeichnet somit die Ausbreitung des zentralen Bereiches der Daten. Der Quartilsabstand wird durch Ausreißer nicht beeinflusst.

3.3.4 5-Zahlen-Zusammenfassung

Die 5-Zahlen-Zusammenfassung bietet eine komprimierte Darstellung der Lage und Streuung einer Verteilung und dient dazu, diese kurz und einfach zu beschreiben. Sie besteht aus den beiden Extremwerten sowie unterem Quartil, Median und oberem Quartil:

$$x_{min}, x_{0,25}, x_{0,5} = \tilde{x}, x_{0,75}, x_{max}$$

Die Angabe dieser fünf Werte geben einen informativen Überblick über die Verteilung eines metrischen Merkmals. Tukey beschrieb erstmals diese Form der Beschreibung eines Merkmals und benannte sie 5-Zahlen-Zusammenfassung (Tukey, 1977: 33).

3.3.5 Grafische Darstellungsformen

Box-Plots

Box-Plots bieten eine anschauliche Darstellung der 5-Zahlen-Zusammenfassung. Durch die Darstellung lässt sich inhaltlich mehr erfassen als durch eine einfache Auflistung der entsprechenden Werte (Tukey, 1977: 40). Ein schneller und intuitiver Eindruck über die Verteilung eines metrischen Merkmals wird dadurch ermöglicht. Box-Plots gehören zu den Basiswerkzeugen der explorativen Datenanalyse. Es existieren verschiedene Darstellungsformen (vgl. (Polasek, 1994: 52 ff.)), hier vorgestellt sei die Form von (Tukey, 1977).

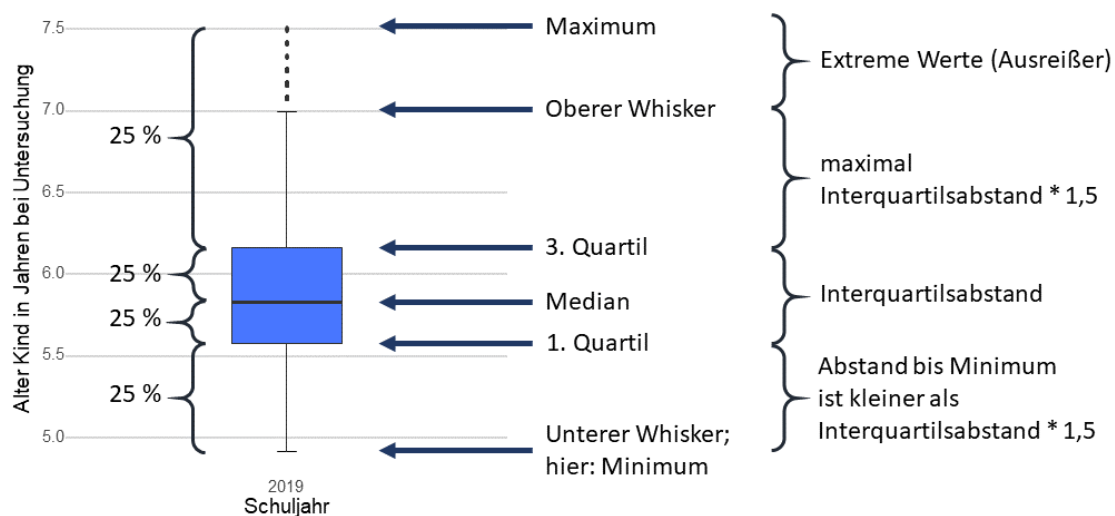


Abbildung 3.4: Box-Plot: Alter der Kinder bei der ESU, Schuljahr 2019

Am Beispiel der Verteilung des Alters der Kinder zum Untersuchungszeitpunkt (Schuljahr 2019) ist hier ein Boxplot mit Interpretationshilfen dargestellt. Das Minimum liegt hier innerhalb der Grenze vom 1,5-fachen des Interquartilsabstandes. Damit endet der untere Whisker beim Minimum. Dabei liegt das Minimum unter 5 Jahre. Dies ist ein Beleg dafür, dass die Vorgabe zum Mindestalter der zu untersuchenden Kinder - die Kinder sollen demnach zum Zeitpunkt der ESU mindestens 5 Jahre alt sein (vgl. Kapitel 2.2.1) - in der Praxis in einigen Fällen nicht eingehalten wird. (Box-Plot erstellt mit dem ESU explorer)

Ein Box-Plot besteht dabei aus folgenden Elementen (vgl. Abbildung 3.4):

Einer Skala parallel zur Hauptachse des Box-Plots: Die Skala kennzeichnet den Wertebereich des dargestellten Merkmals.

Einem Rechteck (der Box) vom unteren Quartil $x_{0,25}$ zum oberen Quartil $x_{0,75}$:

Die Länge der Box entspricht somit dem Quartilsabstand und kennzeichnet den Bereich, in dem die mittleren 50 % der Daten liegen.

Einem Querstrich innerhalb der Box auf der Höhe des Medians: Der Median teilt die Daten in zwei Bereiche, in denen jeweils 50 % der Daten liegen. Über die Lage des Median innerhalb der Box lässt sich die Schiefe einer Verteilung beurteilen: liegt der Median unterhalb der Mitte der Box, ist die Verteilung rechtsschief und umgekehrt.

Den oberen und unteren Antennen (Whisker): Diese kennzeichnen den Wertebereich der außerhalb der Box liegenden Werte. Die Definition nach (Tukey, 1977) besagt, die Länge der Whisker auf das maximal 1,5-fache des Quartilsabstandes zu beschränken. Der Whisker endet dabei aber bei dem Wert, der noch innerhalb dieser Grenze liegt. Daraus resultiert, dass die Whisker nicht unbedingt gleich lang sein müssen.

Die extremen Werte der Verteilung: Fallen Werte außerhalb der Grenze des 1,5-fachen des Quartilsabstandes oberhalb und unterhalb der Box, so werden diese als Ausreißer bezeichnet. Sie werden als einzelne Punkte außerhalb der Whisker dargestellt.

Box-Plots eignen sich besonders gut für einen schnellen und explorativen Überblick über eine Verteilung, deren Schiefe und zur Identifikation von Ausreißern. Ein großer Vorteil besteht im übersichtlichen Vergleich von Verteilungen verschiedener Untergruppen, da sie wenig Platz in der Darstellung nebeneinander benötigen.

Streudiagramme

Streudiagramme sind die einfachste Möglichkeit die Verteilung von zwei metrischen Variablen darzustellen. Erkennen lassen sich die Extremwerte und Spannweiten der Merkmale sowie ein visueller Eindruck ihrer Lage und Streuung. Es können sich eventuelle Abhängigkeitsmuster der betrachteten Merkmale aufzeigen. Streudiagramme lassen sich mit der Darstellung von Box-Plots beider Merkmale kombinieren (vgl. Abbildung 3.5).

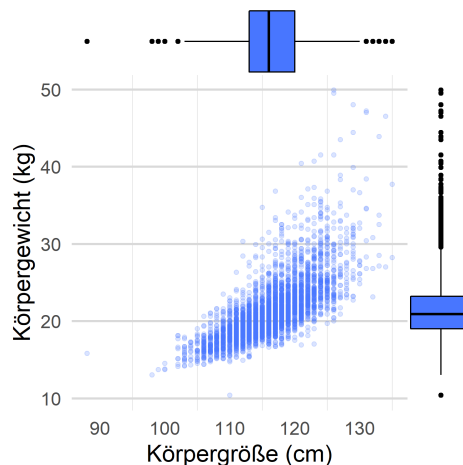


Abbildung 3.5: Streudiagramm in Kombination mit Box-Plots: Körpergröße und Körpergewicht der Kinder bei ESU, Schuljahr 2019

An der Form und Ausbreitung der Punktwolke lässt sich der starke Zusammenhang zwischen Körpergröße und Gewicht erkennen. Die gleichzeitige Darstellung von Box-Plots spezifiziert zusätzlich die Lage und Streuung der beiden Merkmale. (Beispielgrafik aus dem ESU explorer)

3.4 Dichteschätzung

Die traditionellen Verteilungsmaßzahlen geben immer nur Teilaspekte der Eigenschaften von Verteilungen metrischer Merkmale wieder. Um detailliertere Informationen über eine solche Verteilung zu erhalten, werden Verfahren der parameterfreien Dichteschätzung verwendet. Diese vermeiden strikte Annahmen über die zugrundeliegende Datenstruktur. Die Dichteschätzungen werden insbesondere zur grafischen Darstellung verwendet.

3.4.1 Grundlagen

Bei metrischen Merkmalen mit einer hohen Anzahl an Ausprägungen wird eine Häufigkeitstabelle oft lang und entsprechende grafische Darstellungen unübersichtlich. Für die weitere Analyse und Darstellung werden solche Merkmale oft in überlappungsfreie und angrenzende Klassen eingeteilt (*Klassierung*). Es wird davon ausgehend betrachtet, mit welcher relativen Häufigkeit ein Wert in ein bestimmtes Intervall - also in eine der Klassen - fällt. Für ein metrisches Merkmal X seien dabei für alle Klassen k mit x_k^u die unteren Klassengrenzen und x_k^o die oberen Klassengrenzen angegeben. Die Klassenbreiten Δx_k ergeben sich damit aus:

$$\Delta x_k = x_k^u - x_k^o, \quad k = 1, \dots, s$$

Die Anzahl bzw. der Anteil der betrachteten statistischen Elemente, deren Ausprägung in eine bestimmte Klasse fallen, heißt absolute bzw. relative Klassenhäufigkeit, symbolisiert durch:

$$h(x_k) = h_k = h(x_k^u \leq X < x_k^o) \quad \text{bzw.} \quad f(x_k) = f_k = f(x_k^u \leq X < x_k^o)$$

Die absolute bzw. relative *Häufigkeitsdichte* ergibt sich aus dem Quotienten der Häufigkeit und der Klassenbreite:

$$\hat{h}(x_k) = \frac{h(x_k)}{\Delta x_k} \quad \text{bzw.} \quad \hat{f}(x_k) = \frac{f(x_k)}{\Delta x_k}$$

Die Häufigkeitsdichte standardisiert die Klassenhäufigkeiten nach den jeweiligen Klassenbreiten. Häufigkeiten für unterschiedliche Klassenbreiten werden somit vergleichbar gemacht (Härdle et al., 2015: 18). Sind die Klassen alle gleich breit, so sind Häufigkeitsdichte und absolute bzw. relative Häufigkeiten proportional zueinander.

Fasst man ein stetiges Merkmal X als Zufallsvariable auf, so existieren theoretisch unendlich viele Ausprägungen des Merkmals. Die Dichte der Verteilung von X innerhalb eines bestimmten Intervalls wird dabei durch die *Dichtefunktion* angegeben. Für ein stetiges

Merkmal existiert eine Dichtefunktion $f(x)$, für die gilt:

$$P(a < X \leq b) = \int_a^b f(x)dx, \quad a < b$$
$$f(x) \geq 0, \quad \int_{-\infty}^{+\infty} f(x)dx = 1$$

Dabei gibt die Dichtefunktion die Wahrscheinlichkeit an, mit der ein mögliches Ergebnis innerhalb eines bestimmten Intervalls $[a, b]$ liegt. Das Integral der Dichtefunktion ergibt - wie die Summe einer kategorialen relativen Häufigkeitsverteilung - gleich 1. Die Dichtefunktion hat vor allem die Aufgabe, einen visuellen Eindruck der Verteilung von X zu vermitteln. Dabei können sich Hinweise auf die Charakteristika einer Verteilung bieten. Es gibt verschiedene Möglichkeiten, die zunächst unbekannte Verteilung zu schätzen.

3.4.2 Histogramm

Das Histogramm ist die bekannteste Form der grafischen Darstellung klassierter Daten. Gleichzeitig lässt es sich als einfachste Methode zur visuellen Einschätzung einer Dichtefunktion auffassen. Dabei werden für jede Klasse ein Block bzw. ein Rechteck gezeichnet. Die Klassenhäufigkeiten werden durch die Flächen der aneinandergrenzenden Rechtecke repräsentiert. Die Breiten der Rechtecke entsprechen den Klassenbreiten und die Höhen der Rechtecke entsprechen den Häufigkeitsdichten. Damit ist der Flächeninhalt der Rechtecke proportional zur relativen Häufigkeit (*Prinzip der Flächentreue*, (Schlittgen, 2008: 20)). Ein Histogramm zeigt den Kurvenverlauf der dargestellten Verteilung und bietet auch Hinweise zur Lage und Streuung der Verteilung.

Für die Darstellung eines Histogramms muss die Entscheidung nach der Klassenanzahl k bzw. Klassenbreite h getroffen werden. Ein einfaches Histogramm besitzt gleiche Klassenbreiten (Polasek, 1994: 29). Die Klassenbreite ergibt sich in diesem Fall aus $x_{\max} - x_{\min}/k$. Je nach Klassenanzahl können dabei deutliche Unterschiede zwischen Histogrammen von ein- und demselben Merkmal entstehen. Je mehr Klassen gewählt werden, umso schmäler sind die resultierenden Klassen. Werden zu viele Klassen gewählt, so bleibt die Verteilung auch nach der Gruppierung noch zu unübersichtlich. Werden zu wenige, also sehr breite Klassen gewählt, besteht die Gefahr Charakteristiken der Verteilung zu verwischen und damit nicht zu verdeutlichen (vgl. Abbildung 3.6).

Es gibt keine allgemeingültige Regel, welche Klassenanzahl am günstigsten ist. Diese kann beispielsweise durch verschiedene Daumenregeln bestimmt werden (Polasek, 1994: 27f.). Das Problem der Klassenzahl ist in der EDA jedoch von untergeordneter Bedeutung. Hier sollte die optimale Kommunikationsform immer Vorrang vor anderen Prinzipien haben (Polasek, 1994: 27). Es muss also im individuellen Fall entschieden werden, welche Wahl der Klassen bei einer Darstellung am sinnvollsten ist.

Das Histogramm ist die Darstellung der Häufigkeitsverteilung klassierter metrischer Merk-

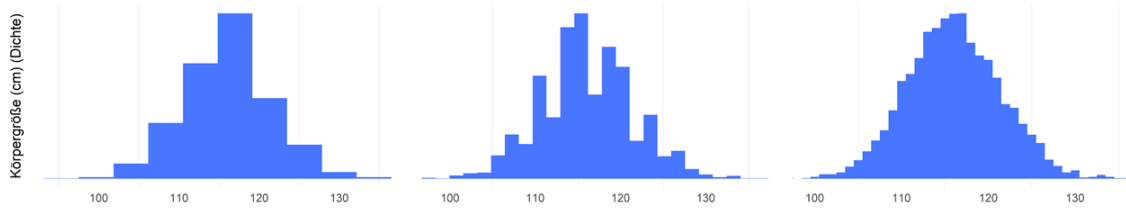


Abbildung 3.6: Histogramme: Körpergröße der Kinder bei ESU, Schuljahr 2017

Am Beispiel der Verteilung der Körpergröße der Kinder zum Untersuchungszeitpunkt sind hier Histogramme der Häufigkeitsverteilung dargestellt. Die Darstellungen zeigen, inwiefern die Wahl der Klassenanzahl Einfluss auf den visuellen Eindruck des Histogramms hat. Dargestellt sind von links nach rechts: 10, 25 und 40 Klassen. Bei 10 Klassen ist die Darstellung noch zu grob. Bei 25 Klassen ergeben sich mehrere „Spitzen“. Bei 40 Klassen ist die Darstellung am glattesten und lässt eine annähernde Normalverteilung vermuten. (Beispielgrafiken aus dem ESU explorer)

male. Die Darstellung ist lokal konstant, unstetig und keine glatte Funktion. Das Histogramm ist somit nur ein sehr einfacher Schätzer der zugrundeliegenden Dichtefunktion.

3.4.3 Kerndichteschätzer

Ein Nachteil des Histogramms ist die unstetige Darstellung der Verteilung eines metrischen Merkmales. Es ist anzunehmen, dass die zugrunde liegende Verteilung eine stetige Dichtefunktion hat. Die Schätzung einer stetigen Dichte ist mittels Kerndichteschätzungen möglich. Die Kerndichteschätzung $\hat{f}_h(x)$ an der Stelle x ist dabei gegeben durch:

$$\hat{f}_h(x) = \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x - X_i}{h}\right), \quad h > 0$$

Dabei symbolisiert $K(\cdot)$ eine Kernfunktion, N die Anzahl Beobachtungen und h die Band- oder Intervallbreite. Die Kernfunktion $K(\cdot)$ muss die Anforderungen an eine Dichtefunktion erfüllen (vgl. Kapitel 3.4.1). Dabei ist die Wahl der Kernfunktion hinsichtlich der Effizienz des Schätzers von untergeordneter Bedeutung (Silverman, 1986: 43). Oft wird als Kernfunktion die Dichte der Standardnormalverteilung gewählt. Diese wird als Gaußkern bezeichnet und ist gegeben durch:

$$K(u) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}u^2\right), \quad u = (x - X_i)/h$$

Die Kernfunktion $K(\cdot)$ wird dabei jedem Punkt X_i zugeordnet. Der Kerndichteschätzer $\hat{f}_h(x)$ entsteht durch das Aufsummieren und Mitteln der einzelnen Kernfunktionen. Der Grad der Glättung der Dichteschätzung kann dabei durch die Bandbreite h kontrolliert werden. Dies ist gleichzusetzen mit dem Problem der Wahl der Klassenanzahl bei der Darstellung eines Histogramms. Eine zu kleine Bandbreite kann zu einer eher nicht glatten, unruhigen Darstellung der Dichte führen. Eine zu große Bandbreite führt hingegen zu einer zu glatten Funktion, die mit einem Informationsverlust einhergeht. Für einen Gaußkern

lässt sich die optimale Bandbreite beispielsweise durch die Daumenregel von (Silverman, 1986: 47f.) ermitteln:

$$h = 0,9 \cdot \text{Min}(s_x^2; s_Q/1,34) \cdot N^{-1/5}$$

Dabei symbolisieren s_x^2 die Standardabweichung und s_Q den Interquartilsabstand des Merkmals X .

3.4.4 Grafische Darstellungsformen

Histogramme und Kerndichteschätzungen

Die Kerndichteschätzung kann als geglättete Variante des Histogramms aufgefasst werden. Oft werden beide Darstellungsvarianten der Dichteschätzung gemeinsam dargestellt (vgl. Abbildung 3.12).

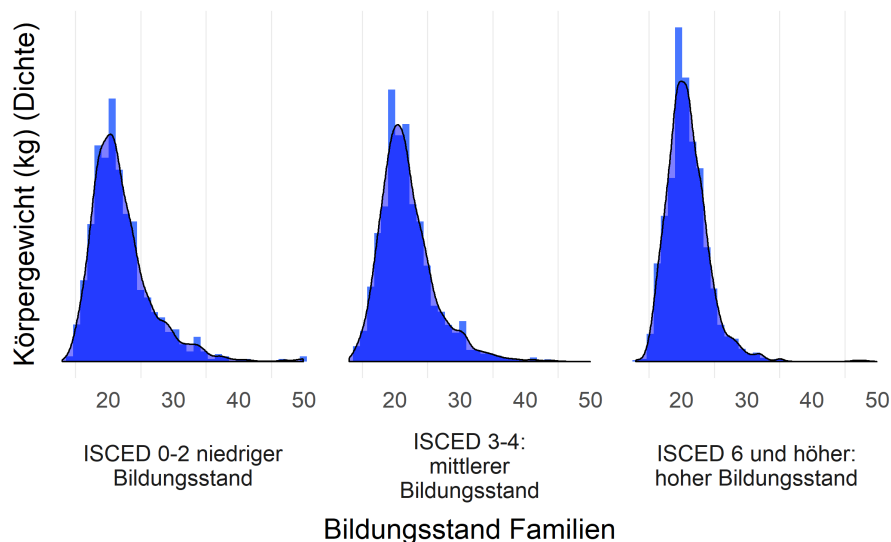


Abbildung 3.7: Histogramme in Kombination mit Kerndichteschätzungen: Körpergewicht der Kinder bei ESU nach Bildungsstand der Familien, Schuljahr 2019

Am Beispiel der Verteilung des Gewichtes der Kinder zum Untersuchungszeitpunkt sind hier Histogramme und Kerndichteschätzungen dargestellt. Die Kerndichteschätzungen wurden mit dem Gaußkern und der Daumenregel nach (Silverman, 1986: 47f.) ermittelt. Bei niedrigerem Bildungsstand der Familien schwankt das Körpergewicht der Kinder deutlich stärker und liegt im Mittel höher als bei Kindern aus Familien mit hohem Bildungsstand. (Beispielgrafik aus dem ESU explorer)

Violin-Plots

Der Violin-Plot ist eine Weiterentwicklung des Box-Plots und soll die Verteilung des Merkmals besser visualisieren. Der Violin-Plot greift die ursprüngliche Idee von (Benjamini, 1988) auf, den Box-Plot um weitere Informationen zur Verteilung zu erweitern.

Der Violin-Plot zeigt dabei zusätzlich die Dichte der Daten. Dabei wird eine rotierte

Kerndichteschätzung zu beiden Seiten gezeigt. Durch die gespiegelte Darstellung sollen die Charakteristika der Verteilung besser hervortreten. Die Kerndichteschätzung zeigt den Verlauf und damit die Höhen, Spitzen und Tiefen einer Verteilung. Diese Informationen fehlen beim Box-Plot. In seiner ursprünglichen Variante wird ein Violin-Plot gemeinsam mit einem Standard-Box-Plot oder einzelner Kennzahlen daraus dargestellt (Hintze und Nelson, 1998). Die Kombination beider Darstellungsvarianten soll die Vorteile beider Methoden in einer Grafik ermöglichen. Vor allem ein Vergleich verschiedener Verteilung und ihrer Verläufe wird so vereinfacht (vgl. Abbildung 3.8).

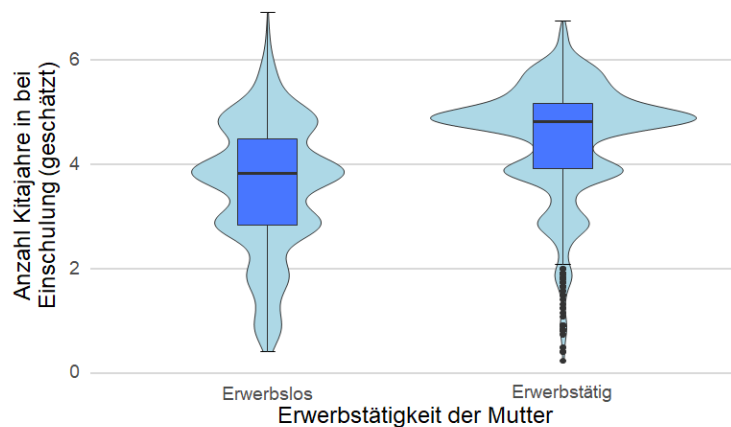


Abbildung 3.8: Violin-Plot in Kombination mit Box-Plot: Jahre in Kita bei Einschulung nach Erwerbstätigkeit der Mutter, Schuljahr 2019

Die kombinierte Darstellung von Box-Plot und Violin-Plot soll die Vorteile beider Darstellungsvarianten nutzen. Durch den Violin-Plot werden die Spitzen der Verteilungen deutlich, diese Information fehlt beim Box-Plot. Die Verteilungen zeigen hier mehrere Spitzen. Dies resultiert aus der Tatsache, dass die Kitas zu Beginn eines Schuljahres - wenn die schulpflichtigen Kinder die Kita verlassen - vermehrt jüngere Kinder aufnehmen. So haben viele Kinder zum Zeitpunkt ihrer Einschulung etwa volle 3, 4 oder 5 Jahre in der Kita verbracht. Dabei besuchen die Kinder erwerbstätiger Mütter deutlich länger die Kita als Kinder mit erwerbslosen Müttern. Die Kerndichteschätzungen wurden mit dem Gaußkern und der Daumenregel nach (Silverman, 1986: 47f.) ermittelt. (Beispielgrafik aus dem ESU explorer)

Streudiagramme und Kerndichteschätzungen

Auch die bereits vorgestellten Streudiagramme von zwei metrischen Merkmalen lassen sich mit der Darstellung von Kerndichteschätzungen beider Merkmale kombinieren. Die Vorteile beider Darstellungsvarianten können so genutzt werden (vgl. Abbildung 3.9).

3.5 Zusammenhangmaße

Bei der Untersuchung von zwei Merkmalen besteht oft Interesse daran, ob die beiden Merkmale zusammenhängen bzw. korrelieren. Mit den bereits aufgeführten Darstellungsvarianten lassen sich eventuelle Zusammenhänge visuell entdecken. Im Folgenden geht es um die Frage inwiefern sich ein solcher Zusammenhang quantifizieren lässt. Einfach

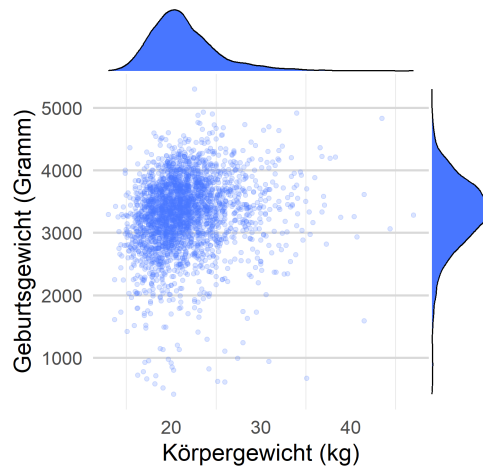


Abbildung 3.9: Streudiagramm in Kombination mit Kerndichteschätzung: Körpergewicht bei ESU und Geburtsgewicht, Schuljahr 2019

Hier werden mit Streudiagramm und Kerndichteschätzungen beider Merkmale ebenfalls zwei Darstellungsvarianten gemeinsam visualisiert. Die Kerndichteschätzungen wurden mit dem Gaußkern und der Daumenregel nach (Silverman, 1986: 47f.) ermittelt. (Beispielgrafik aus dem ESU explorer)

ausgedrückt, hängen zwei Variablen voneinander ab, wenn sie nicht unabhängig sind. Dabei spricht man von Unabhängigkeit, wenn die Änderung eines Wertes der einen Variable keinen Einfluss auf den Wert einer anderen Variable hat (Liebetrau, 1983: 6).

Zwei Variablen X und Y sind somit *unabhängig* wenn (Härdle et al., 2015: 436)

1. die bedingten relativen Verteilungen von X einander gleich sind und gleich zur entsprechenden relativen Randverteilung. Das heißt für die bedingte Verteilung von X und Y :

$$f(x_i|y_j) = f(x_i|y_h) = f_{i\bullet} = f(x_i), \text{ für alle } j, h = 1, \dots, m \text{ und für alle } i = 1, \dots, k$$

$$f(y_j|x_i) = f(y_j|x_h) = f_{\bullet j} = f(y_j), \text{ für alle } i, h = 1, \dots, k \text{ und für alle } j = 1, \dots, m$$

2. die gemeinsame relative Häufigkeitsverteilung gleich dem Produkt der entsprechenden Randverteilungen ist:

$$f(x_i, y_j) = f_{ij} = f_{i\bullet} \cdot f_{\bullet j} = f(x_i) \cdot f(y_j)$$

Dabei können *kontingente* Zusammenhänge zweier nominaler Merkmale, *monotone* Zusammenhänge zweier ordinal skalierten Merkmale oder *lineare* Zusammenhänge zweier metrisch skalierten Merkmale untersucht werden. Für Zusammenhangsanalysen zwischen verschiedenen skalierten Merkmalen gilt, dass das Maß für das jeweils niedrigere Skalenniveau anwendbar ist.

Zusammenhangsmaße quantifizieren die Beziehung zwischen Variablen und messen somit den in konkreten Daten vorliegenden Zusammenhang (Liebetrau, 1983: 6). Die Stärke und

Richtung eines Zusammenhanges sind hier von Interesse. Üblicherweise sind Zusammenhangsmaße beschränkt auf einen Wertebereich von 0 bis 1 (standardisierte Maße), wobei 0 keinen Zusammenhang und 1 einen perfekten Zusammenhang repräsentiert. Zusammenhangsmaße in einem Wertebereich von -1 bis 1 beschreiben zusätzlich die Richtung des Zusammenhanges.

Ein möglicher Zusammenhang impliziert dabei keine Kausalität, also Ursache-Wirkungs-Beziehung. Die Richtung, in welcher die Merkmale aufeinander wirken und ob zwei Variablen ursächlich zusammenhängen, können durch ein Zusammenhangsmaß nicht beschrieben werden.

Im Folgenden wird eine Auswahl gebräuchlicher Zusammenhangsmaße in Abhängigkeit vom Skalenniveau der betrachteten Merkmale beschrieben. Es existieren weitere Gruppen von Maßen, die beispielsweise auf der Minimierung von Fehlern in Vorhersagen basieren (Fehlerreduktionsmaße) (Liebetrau, 1983). Auf diese wird im Rahmen dieser Arbeit nicht eingegangen. Zu den ausgewählten und darüber hinausgehenden Zusammenhangsmaßen sei auf (Liebetrau, 1983) und (Bortz et al., 2008: 325 ff.) verwiesen.

3.5.1 Kontingenz

Der Ausgangspunkt für die Zusammenhanganalyse zweier kategorialer Merkmale ist die gemeinsame Häufigkeitsverteilung - die Kontingenztabelle (siehe Kapitel 3.2.2) (Bortz et al., 2008: 326). Wie einführend erläutert, entspricht die relative Häufigkeit im Falle von Unabhängigkeit zweier Variablen dem Produkt der relativen Häufigkeiten der Randverteilungen beider Merkmale (Härdle et al., 2015: 450):

$$f_{ij} = f_{i\bullet} \cdot f_{\bullet j} \quad \text{und} \quad \frac{h_{i\bullet} \cdot h_{\bullet j}}{N} = N \cdot f_{i\bullet} \cdot f_{\bullet j}$$

Die erwartete Häufigkeit bei Vorliegen von Unabhängigkeit ergibt sich damit:

$$e_{ij} = \frac{h_{i\bullet} \cdot h_{\bullet j}}{N}$$

Auf diesen Überlegungen aufbauend lässt sich die *quadrierte Kontingenz*, repräsentiert durch χ^2 (sprich: Chi-Quadrat), berechnen. Der χ^2 -Wert entspricht der Summe der quadrierten standardisierten Abweichungen aller Tabellenwerte h_{ij} von den entsprechenden erwarteten Häufigkeiten bei Unabhängigkeit e_{ij} :

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^m \left(\frac{h_{ij} - e_{ij}}{e_{ij}} \right)^2 \quad \text{mit} \quad 0 \leq \chi^2 \leq N \cdot (\min(k, m) - 1)$$

Der Chi-Quadrat-Wert wurde erstmals von (Pearson, 1900) als Maßzahl für die Abweichung von Unabhängigkeit beschrieben.

Für die quadrierte Kontingenz gilt: $\chi^2 = 0$ genau dann, wenn die Merkmale unabhängig sind, ansonsten $\chi^2 > 0$ mit einem Maximalwert von $\chi^2 = N \cdot (\min(k, m) - 1)$ bei einem höchstmöglichen Zusammenhang, wobei k Anzahl der Zeilen und m Anzahl der Spalten in der Kontingenztabelle entspricht. Je größer χ^2 , desto größer sind die Abweichungen und desto größer ist der Unterschied zwischen beobachteten und bei Unabhängigkeit erwarteten Häufigkeiten. Der Chi-Quadrat-Wert gibt somit die Stärke eines Zusammenhanges wieder. Der χ^2 -Wert gibt keine Auskunft darüber, wie genau eine Veränderung in einer Variablen eine andere Variable beeinflusst. Außerdem sind χ^2 -Werte nicht standardisiert und somit schwierig zu vergleichen. Für vergleichbare Maße sind weitere Modifikationen der Größe notwendig (Liebetrau, 1983: 13).

Die folgenden auf χ^2 basierten standardisierten Zusammenhangsmaße sind unabhängig von Fallzahl und Dimensionen und sind deshalb untereinander und mit anderen standardisierten Zusammenhangsmaßen vergleichbar.

Phi-Koeffizient

Der Φ - (*Phi*-)Koeffizient wird als gebräuchlichstes Zusammenhangsmaß für den Spezialfall der Analyse von zwei nominalen Merkmalen mit je nur zwei Ausprägungen genutzt (Bortz et al., 2008: 327). Er berücksichtigt die Fallzahl N und ist definiert als:

$$\Phi = \sqrt{\frac{\chi^2}{N}}$$

Ist der χ^2 Wert 0, so ist auch $\Phi = 0$, es liegt kein Zusammenhang vor. Als maximalen Wert kann $\Phi = 1$ annehmen, in diesem Fall liegt ein perfekter Zusammenhang vor.

Cramér's V

Ebenfalls auf χ^2 basiert *Cramér's V*, auch als *Cramér's Φ -Koeffizient* bezeichnet (Bortz et al., 2008: 355). V kann für alle Kontingenztafeln unabhängig von deren Dimensionen verwendet werden und ist definiert als:

$$V = \sqrt{\frac{\chi^2}{N \cdot (h - 1)}} \quad \text{mit} \quad h = \min(k, m)$$

Wie vorher, steht 0 für keinen Zusammenhang und 1 für maximalen Zusammenhang. Dieses Zusammenhangsmaß wurde im speziellen für Kontingenztabellen mit verschiedenen Dimensionen entwickelt. Bei einer Anwendung auf eine 2x2 Tabelle entspricht V dem absoluten Betrag des Φ -Koeffizienten.

Kontingenzkoeffizienten nach Pearson

Ein oft genutztes Maß zur Darstellung der Stärke des Zusammenhangs zwischen nominal skalierten Variablen ist der *Kontingenzkoeffizient* nach Pearson (Pearson, 1904). Basierend auf χ^2 kann der Kontingenzkoeffizient nach Pearson C wie folgt berechnet werden:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

Dabei ist C stets größer als Cramérs V (Bortz et al., 2008: 358). Ein Kontingenzkoeffizient $C = 0$ zeigt statistische Unabhängigkeit. Der Kontingenzkoeffizient erreicht niemals 1, selbst nicht im Falle eines perfekten Zusammenhanges. Für C gilt:

$$0 \leq C < 1 \quad \text{bzw.} \quad 0 \leq C \leq \sqrt{\frac{h-1}{h}} \quad \text{mit} \quad h = \min(k, m)$$

Um den Einfluss der Dimensionen der Kontingenztabelle auf die Obergrenze von C zu umgehen und somit die Vergleichbarkeit von Ergebnissen sicherzustellen, kann der korrigierte (normierte) Kontingenzkoeffizient $C_{\text{kor}}r$ berechnet werden (Härdle et al., 2015: 451):

$$C_{\text{kor}}r = C \cdot \sqrt{\frac{h}{h-1}}$$

dabei gilt $0 \leq C_{\text{kor}}r \leq 1$.

3.5.2 Rangkorrelation

Ausgangspunkt für die Analyse von monotonen Zusammenhängen zwischen mindestens ordinalen Merkmalen sind die Rangwerte $R(x_n), R(y_n), n = 1, \dots, N$. Die beobachteten Werte x_i und y_i werden mit den jeweiligen Rängen $R(x_i)$ und $R(y_i)$ ersetzt. Kommen gleiche Werte (auch Bindungen genannt) vor, so werden mittlere Ränge vergeben.

Der *Rangkorrelationskoeffizient* r_s von *Spearman* ist der für die getrennt bestimmten Rangwerte $R(x_n), R(y_n)$ sowie den entsprechenden arithmetischen Mitteln $\overline{R(x)}, \overline{R(y)}$ ermittelte Korrelationskoeffizient:

$$r_s = \frac{\sum_{n=1}^N (R(x_n) - \overline{R(x)})(R(y_n) - \overline{R(y)})}{\sqrt{\sum_{n=1}^N (R(x_n) - \overline{R(x)})^2} \sqrt{\sum_{n=1}^N (R(y_n) - \overline{R(y)})^2}} \quad \text{dabei gilt } -1 \leq r_s \leq 1$$

Eine Alternative zu Spearmans r_s bietet Kendalls Tau τ . Er beruht auf einem paarweisen Vergleich der Ränge der beiden Merkmale. Da bei einem vollständigen Paarvergleich $\frac{N(N-1)}{2}$ Vergleiche entstehen (Härdle et al., 2015: 449), ist τ für hohe Fallzahlen sehr berechnungsintensiv und damit eher für kleine Datensätze geeignet. Aus diesem Grund wird τ im Rahmen dieser Arbeit nicht verwendet.

3.5.3 Maßkorrelation

Die Basis für die Betrachtung der Zusammenhänge zwischen metrisch skalierten Merkmalen bildet die *Kovarianz* zweier Merkmale X und Y (Schlittgen, 2008: 93). Die Kovarianz s_{XY} ergibt sich aus den Daten $(x_n, y_n), n = 1, \dots, N$ und basiert auf den Abweichungen der Merkmalswerte von ihrem Mittelwert:

$$s_{XY} = \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})(y_n - \bar{y})$$

Die Kovarianz bestimmt die Stärke eines Zusammenhangs zweier Merkmale und misst deren linearen Zusammenhang. Die Kovarianz ist wie die Varianz von den Skalen, in der die Merkmale erfasst werden, stark abhängig. Um diese Abhängigkeit zu beseitigen, kann die Kovarianz standardisiert werden, indem sie auf die Standardabweichung bezogen wird. Es ergibt sich der *Korrelationskoeffizient* r_{XY} von *Bravais-Pearson*:

$$r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{\frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})(y_n - \bar{y})}{\sqrt{\frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})^2} \sqrt{\frac{1}{N} \sum_{n=1}^N (y_n - \bar{y})^2}}$$

mit den Eigenschaften $-1 \leq r_{XY} \leq 1$, und $r_{XY} = \pm 1$ bei Vorliegen eines perfekten linearen Zusammenhangs und $r_{XY} = 0$ bei Unabhängigkeit von X und Y . Weiterhin ist der Korrelationskoeffizient symmetrisch: $r_{XY} = r_{YX}$ (Härdle et al., 2015: 439f.).

Zur Interpretation der Höhe des Korrelationskoeffizienten und anderer standardisierter Zusammenhangsmaße werden in der Literatur verschiedene Richtlinien genannt (vgl. beispielsweise (Schlittgen, 2008: 97), (Härdle et al., 2015: 440)). In Tabelle 3.1 dargestellt werden die Einteilungen der Höhe des Betrages des Koeffizienten nach (Cohen, 1988: 79f.). Die Anwendung dieser Interpretationsregeln ist in den Sozialwissenschaften üblich.

Tabelle 3.1: Interpretation Korrelationskoeffizient nach (Cohen, 1988)

Korrelationskoeffizient $ r $	Interpretation
0	keine Korrelation
0,1	schwache Korrelation
0,3	mittlere Korrelation
0,5	starke Korrelation
1	perfekte Korrelation

Ein linearer Zusammenhang kann darüber hinaus auf Streudiagrammen visualisiert werden. Dargestellt wird dabei eine Gerade, die den Zusammenhang möglichst gut aufzeigt. Die Abstände zwischen der Geraden und den Merkmalsausprägungen sollen dafür minimiert werden. Die gesuchte Gerade $y = a + bx$ soll eine Abhängigkeit des Merkmals Y von X beschreiben. Dabei gibt a den Achsenabschnitt auf der y -Achse an, b die Steigung der Geraden an. Es gilt die Beziehung $y_i = a + bx_i + e_i$, dabei sind e_i die Abweichungen zu

den tatsächlichen y -Werten. Dafür wird eine Gleichung $\hat{y} = a + bx$ geschätzt, für die gilt:

$$\sum_{i=1}^N e_i^2 = \sum_{i=1}^N (y_i - \hat{y}_i)^2 = \text{Minimum}$$

Eine solche Gerade heißt Regressionsgerade nach der Methode der kleinsten Quadrate. Zu den Grundideen der einfachen Regressionsrechnung sei auf (Schlittgen, 2008: 101ff.) verwiesen. Eine Darstellung von Regressionsgeraden findet sich in Abbildung 3.10.

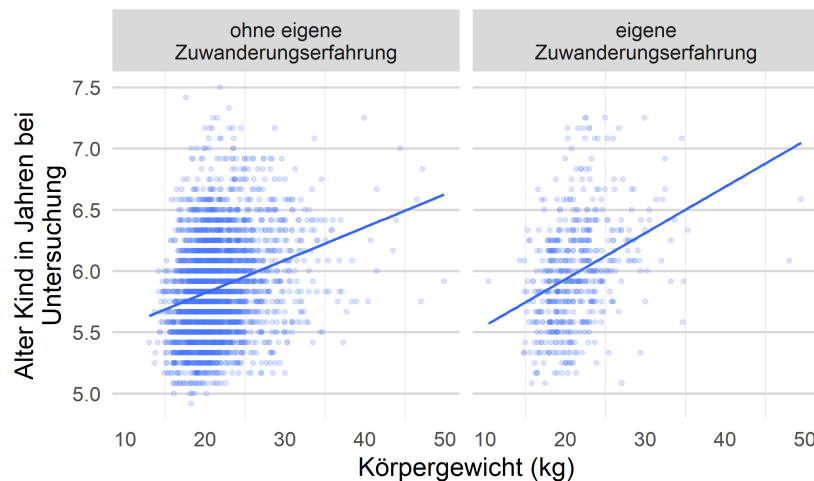


Abbildung 3.10: Streudiagramme mit Regressionsgeraden: Körpergewicht nach Alter der Kinder und Zuwanderungserfahrung, Schuljahr 2019

In Streudiagrammen lassen sich lineare Zusammenhänge mittels einer Regressionsgeraden visualisieren. Erkennbar ist ein positiver Zusammenhang zwischen Alter und Gewicht der Kinder. Der Zusammenhang ist stärker ausgeprägt bei Kindern mit Zuwanderungserfahrung. Der Korrelationskoeffizient r_{XY} für die Kinder ohne Zuwanderungserfahrung liegt bei 0,26 (schwacher Zusammenhang), für die Kinder mit Zuwanderungserfahrung bei 0,35 (mittlerer Zusammenhang).

Das Alter der Kinder wurde aus dem Geburtsmonat und -jahr sowie dem Untersuchungsmonat und -jahr abgeleitet. Aus Datenschutzgründen werden diese Daten nicht tagesgenau erfasst, so dass die daraus abgeleiteten Zeiträume nur auf Monaten basieren. Erkennbar ist dies in der Darstellung an der Diskretisierung des Alters der Kinder (horizontale Linienbildung). (Beispielgrafik aus dem ESU explorer)

3.5.4 Signifikanztests

Über die reine Angabe der Zusammenhangsmaße hinaus sind *Signifikanztests* bzw. Hypothesentests möglich. Traditionell wird die Nullhypothese H_0 „ X und Y sind statistisch unabhängig“ vs. Alternativhypothese H_1 „ X und Y sind nicht statistisch unabhängig“ geprüft. Zu den Grundkonzepten statistischer Tests sei auf (Härdle et al., 2015: 311ff.) und (Bortz et al., 2008: 28ff.) verwiesen.

Tests basieren auf *Prüfgrößen*, welche daraufhin überprüft werden, wie weit sie von einer unter der Nullhypothese angenommenen Verteilung entfernt liegen. Signifikanztests ermitteln dabei die Wahrscheinlichkeit p , mit der das vorliegende Ergebnis auftreten kann, wenn die Populationsverhältnisse der Nullhypothese entsprechen. Je kleiner diese Wahrschein-

lichkeit ist, um so unwahrscheinlicher ist die Nullhypothese. Als *Signifikanzniveau* oder *Irrtumswahrscheinlichkeit* bezeichnet man dabei die festgelegte maximale Wahrscheinlichkeit dafür, dass die Nullhypothese abgelehnt wird, obwohl sie in Wirklichkeit wahr ist. Die Irrtumswahrscheinlichkeit wird oft auf maximal 5 % festgelegt - dann werden p -Werte kleiner als 0,05 als *statistisch signifikant* bezeichnet.

Ein Zusammenhang zwischen kategorialen Merkmalen kann mit dem *Chi-Quadrat-Unabhängigkeits-Test* untersucht werden. Der Test ist auf alle Skalenniveaus anwendbar und nichtparametrisch - das heißt er basiert auf keinen Verteilungsannahmen. Der Chi-Quadrat-Test wurde erstmals von (Pearson, 1900) beschrieben. Die Prüfgröße ist die quadrierte Kontingenz χ^2 (siehe Kapitel 3.5.1). Unter der Nullhypothese ist χ^2 annähernd Chi-Quadrat-verteilt mit $(k - 1) \cdot (m - 1)$ Freiheitsgraden. Die Approximation ist hinreichend genau, wenn für alle i, j die erwarteten Häufigkeiten $e_{ij} \geq 5$ entsprechen (Schlittgen, 2008: 410). Für alle auf dem χ^2 -Wert basierenden Zusammenhangsmaße (Phi Φ , Cramers V , Kontingenzkoeffizient C und C_{corr}) gilt: sie gelten als statistisch signifikant, wenn der dazugehörige χ^2 -Wert signifikant ist.

Bei der Überprüfung eines monotonen Zusammenhanges, repräsentiert durch den Rangkorrelationskoeffizient r_s von Spearman, wird die Prüfgröße durch $r_s \cdot \sqrt{N - 1}$ ermittelt. Diese ist unter der Nullhypothese und einer Fallzahl $N \geq 10$ annähernd standardnormalverteilt (Schlittgen, 2008: 413).

Ein linearer Zusammenhang kann über den Korrelationskoeffizient r_{XY} nach Pearson getestet werden. Hier wird die Prüfgröße wie folgt berechnet:

$$T = \frac{\sqrt{N - 2} \cdot r_{XY}}{\sqrt{1 - r_{XY}^2}}$$

Der Wert T folgt bei Gültigkeit der Nullhypothese einer t -Verteilung mit $N - 2$ Freiheitsgraden. Die Voraussetzung dieser Annahme ist eine bivariate Normalverteilung der beiden untersuchten Merkmale. Dieser Test gehört damit zu den parametrischen Tests. Eine Verletzung dieser Voraussetzung hat jedoch vor allem bei großen Fallzahlen in der Regel keinen nennenswerten Einfluss auf die Validität des Signifikanztests (Bortz et al., 2008: 447).

3.6 Thematische Karten

Karten zeigen ein maßstabsgetreues und verallgemeinertes Abbild räumlicher Strukturen. Thematische Karten visualisieren bestimmte Merkmale oder Themen. Die Darstellung von Themen anhand von Karten bietet viele Vorteile. Neben dem Datenwert ergeben sich zusätzliche Informationen zur räumlichen Umgebung und wo sich dieser Wert in der räumlichen Struktur befindet. Zudem lassen sich die Werte durch eine farbliche Darstellung leichter und schneller erfassen. Informationen, die in Tabellenform schlecht zu erfassen

sind, werden somit mittels einer thematischen Karte auf einen Blick sichtbar. Der wichtigste Effekt ist somit der Informationsgewinn (Olbrich et al., 2002: 4).

Die digitale Erfassung und Analyse raumbezogener Daten erfolgt mittels Geoinformationssystemen (GIS). Die Darstellung thematischer Karten durch Geoinformationssysteme bildet ein wichtiges Werkzeug explorativer Datenanalyse in kleinräumigen Strukturen.

Bei der ESU wird der Wohnort der Kinder auf LOR-Ebene und die ihnen zugewiesene Grundschule erfasst, welche als Grundlage für die Erstellung thematischer Karten genutzt werden können. Die Sozialraumorientierung als eines der Ziele des Bildungsmonitoring Berlin-Mitte soll durch entsprechende Karten unterstützt werden (vgl. Kapitel 2.3).

3.6.1 Choroplethenkarten

Die klassische Darstellungsform kleinräumiger Verteilungen sind Choroplethenkarten, auch Kartogramme genannt. Choroplethenkarten sind thematische Karten, die flächenbezogene Daten darstellen, indem die Flächen unterschiedlich eingefärbt werden. Zumeist werden in Choroplethenkarten Verhältniszahlen wie Anteile dargestellt. Die Verteilungen dieser Anteile sind als Dichte zu interpretieren. Bei der Darstellung von Absolutwerten kann aufgrund unterschiedlicher Flächengrößen ein stark fehlleitender optischer Eindruck entstehen (Olbrich et al., 2002: 39).

Auf der Karte werden zunächst kleinräumige Einheiten wie Lebensweltlich orientierte Räume (LOR) oder Einschulungsbereiche (ESB) in ihrer Form und geografischen Anordnung dargestellt. Die Einfärbung dieser Flächen mittels Farbschematas soll dann hohe und niedrige Werte visualisieren. Die jeweiligen Angaben gelten jeweils für die gesamte Fläche, ohne interne Variationen zu beachten.

Ein Hauptmerkmal der Choroplethenkarte besteht darin, dass sie Daten oftmals klassifiziert darstellt. Dabei nimmt die Wahl der Klassenbildung einen besonders hohen Stellenwert bei der Erstellung der Karten ein. Die Anzahl und Aufteilung der Klassen beeinflussen den visuellen Eindruck und damit die Interpretationsmöglichkeiten stark. Eine allgemeingültige Lösung gibt es hierfür nicht. Die Methode der Klassenbildung hängt stark von den vorliegenden Daten, vom Thema und von der Zielgruppe ab (Olbrich et al., 2002: 39). So gibt es beispielsweise folgende Möglichkeiten:

Gleiche Intervalle: Alle Klassen haben dieselbe Breite. Der Vorteil liegt hier in der Vergleichbarkeit verschiedener Karten mit derselben Klasseneinteilung. Ein Nachteil besteht darin, dass diese Methode nur für Daten geeignet ist, die eine relativ gleichmäßige Verteilung aufweisen.

Quantile: Die Klassen werden so gewählt, dass alle Klassen gleich häufig besetzt sind. Hier gibt es keine gleichen Klassengrenzen. Eine Vergleichbarkeit zwischen verschiedenen Karten ist schwierig, da sich die Klassen nur auf eine bestimmte Karte anwenden lassen.

Bei einer Klassenbildung ist die Klassenanzahl zu bestimmen. Je mehr Klassen genutzt werden, umso deutlicher lassen sich auch unterschiedliche Ergebnisse erkennen. Andererseits haben zu viele Klassen den Nachteil, zu unübersichtlichen Karten und Legenden zu führen. Je nach Anzahl und Einteilung der Klassen können sehr unterschiedliche Karten entstehen und die Wahrnehmung erheblich beeinflussen. Nach Datenlage ist individuell und visuell zu entscheiden, welche Art der Klassenbildung und welche Klassenanzahl das Thema am besten wiedergeben. Es besteht aber auch die Möglichkeit, auf eine Klassenbildung zu verzichten. Dann werden die Werte in einer stetigen Farbgebung dargestellt, indem beispielsweise dem kleinsten Wert eine helle Farbe und dem größten Wert eine dunkle Farbe zugeordnet wird. Alle weiteren Werte erhalten dann eine Abstufung der Farben die zwischen den beiden Extremen liegen. In Abbildung 3.11 werden verschiedene Möglichkeiten der Klassenbildung und Farbgebung einer Choroplethenkarte dargestellt. Der Nachteil von Choroplethenkarten besteht darin, dass die einzelnen Flächen farblich einheitlich dargestellt und wahrgenommen werden, auch wenn in der Realität die Ergebnisse innerhalb des Raums schwanken. Die Unterschiede innerhalb der Räume werden dabei unterschlagen. Durch eine Klassifizierung gehen außerdem Informationen verloren.

3.6.2 Kerndichtekarten

Die genannten Nachteile der Choroplethenkarten wie die einheitliche variationsfreie und somit unrealistische Färbung der einzelnen räumlichen Einheiten mit Farbsprüngen an den Raumgrenzen können durch ein innovatives Verfahren der Darstellung umgangen werden. So sollen Karten, welche nicht oder weniger stark auf den Raumgrenzen basieren, eine realistischere Perspektive auf die vorliegende Verteilung bieten.

Eine realistischere, stetige Dichteverteilung über Flächen kann nur dargestellt werden, wenn für die einzelnen Beobachtungen eine genaue Lageposition im Raum vorliegt. Räumliche Positionen werden mittels Geokoordinaten (also Breiten- und Längengrad) erfasst. Die Erfassung eines Gebietsraumes statt der genauen Geokoordinaten erfolgt meist aus Datenschutzgründen und kann als eine Art Rundung der bivariaten Angabe der Geokoordinaten aufgefasst werden. Die gerundeten Geokoordinaten werden dabei als Punkt innerhalb des Gebietsraumes - in der Regel dem geografischen Mittelpunkt des Raumes - festgelegt. Man spricht hierbei auch von lokal aggregierten Daten. Diese ungenaue Erfassung von Geokoordinaten wiederum kann als Messfehler angesehen werden (Groß et al., 2017).

Um aus den lokal aggregierten Daten eine stetige Dichte über die Gebietsräume abzuleiten, muss eine Rückverteilung der ungenauen Geokoordinaten über die darzustellenden Flächen erfolgen. Dafür soll ein nichtparametrisches Verfahren der multivariaten Kerndichteschätzung genutzt werden. Da die Anwendung einer klassischen Kerndichteschätzung auf gerundete Daten teilweise zu schlechten Ergebnissen führt, berücksichtigt dass von (Groß et al., 2017) entwickelte *Kernelheaping-Verfahren* das Vorliegen von Messfehlern -

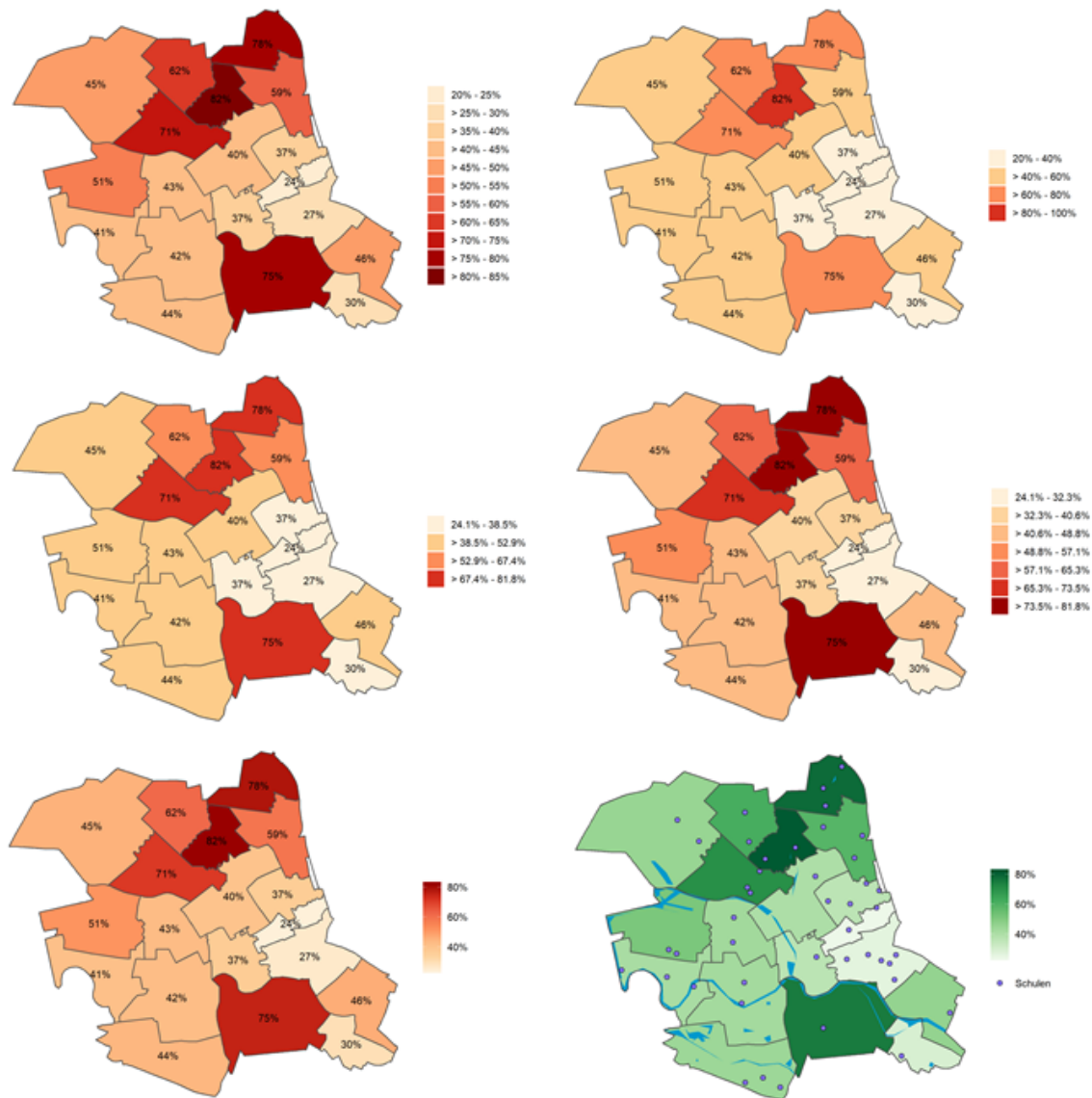


Abbildung 3.11: Choroplethenkarten: Anteile der Kinder mit Schulischem Förderbedarf, Einschulungsbereiche, Schuljahr 2019

Dargestellt sind die Einschulungsbereiche zum Schuljahr 2019 auf einer Karte von Berlin-Mitte. Das Thema der Karten ist der Anteil der zum Schuljahr 2019 untersuchten Kinder mit schulischem Förderbedarf. Alle Karten basieren auf denselben Daten, sie unterscheiden sich in der Wahl der Klassenbildung und Klassenanzahl.

Von links oben nach rechts unten: Klassenbildung mit gleichen Intervallen, Klassenbreite 5 %, Klassenanzahl 11 (links oben); Klassenbildung mit gleichen Intervallen, Klassenbreite 20 %, Klassenanzahl 4 (rechts oben); Klassenbildung mit Quantilen, Klassenanzahl 4 (links Mitte); Klassenbildung mit Quantilen, Klassenanzahl 7 (rechts Mitte); kontinuierliche Farbgebung ohne Klassenbildung (links unten); kontinuierliche Farbgebung ohne Klassenbildung in einer anderen Farbskala - außerdem sind auf dieser letzten Karte zusätzlich die Gewässer und Standorte der Schulen dargestellt (rechts unten). Je nach Darstellungsvariante unterscheidet sich der visuelle Eindruck der Karten. Mehr oder weniger stark zu erkennen ist die starke Variation der Anteile der Kinder mit Förderbedarf zwischen den einzelnen Einschulungsbereichen. Die höchsten Anteile mit 82 % und 78 % zeigen sich im mittleren nördlichen Bereich des Bezirkes Mitte (Gesundbrunnen/Wedding), ein weiterer Bereich im Süden von Mitte (Regierungsviertel) mit 75 %. Der niedrigste Anteil findet sich etwa im Bereich Zentrum/Brunnenviertel (24 %). Die Grundschulen der einzelnen Einschulungsbereiche stehen also vor grundsätzlich verschiedenen Herausforderungen was die Förderung der Kinder betrifft.

Geometriedaten zu Gewässern, Grünflächen etc. sind unter (OpenStreetMap, 2019) frei verfügbar. Die Standorte der Schulen wurden vom Bezirksamt Mitte zur Verfügung gestellt.

(Beispielgrafiken aus dem ESU explorer)

im speziellen auch nur gerundet erfassten Geokoordinaten - und kann deshalb als Fehlerkorrekturmodell bezeichnet werden.

Die Idee basiert auf einer Kartendarstellung von Kerndichteschätzungen der genauen Geokoordinaten, welche punkt- statt flächenbezogene Daten darstellen. Die räumlichen Inhomogenitäten werden dabei durch eine Fehlerkorrektur geschätzt. Der neuartige Ansatz integriert dabei das Problem der Bandbreitenwahl (siehe auch im eindimensionalen Fall Kapitel 3.4.3) in den Schätzungsprozess.

Multivariate Kerndichteschätzung

Wenn das Merkmal X die zweidimensionale Erfassung von Geokoordinaten durch Längen- und Breitengrad darstellt, dann ist $X_i, i = 1, \dots, N$ gegeben durch (X_{i1}, X_{i2}) . Eine multivariate Kerndichteschätzung ist am Punkt x gegeben durch:

$$\hat{f}_H(x) = \frac{1}{n|H|^{\frac{1}{2}}} \sum_{i=1}^n K(H^{-\frac{1}{2}}(x - X_i))$$

Dabei symbolisiert $K(\cdot)$ eine multivariate Kernfunktion und H eine Bandbreitenmatrix. Oft und so auch in der vorliegenden Methode wird eine multivariate Standardnormalverteilung als Kernfunktion verwendet. Wie im eindimensionalen Fall ist die Wahl der Bandbreiten entscheidend für die Qualität der Kerndichteschätzung. In der hier angewendeten Methode wird ein Plug-in-Schätzer nach (Wand und Jones, 1994) zur Bandbreitenwahl verwendet.

Kerndichteschätzung für lokal aggregierte Geokoordinaten

Das Kernelheaping-Verfahren - auch als Methode der simulierten Geokoordinaten bezeichnet - berücksichtigt insbesondere das Vorliegen von Messfehlern durch die lokal aggregierten Daten.

Der Ansatz des Kernelheaping-Verfahrens simuliert die zweidimensionale Dichte mittels eines iterativen Prozesses. Dabei wird eine Variante des Stochastischen-Erwartungs-Maximierungs-Algorithmus (Stochastic Expectation Maximization Algorithm (SEM)) angewandt. Die Kernidee ist, mit einem Modell zu starten und sich durch einen iterativen Prozess von wiederholten Kerndichteschätzungen und Stichprobenziehungen der wahren Dichte anzunähern. Das ursprüngliche Modell soll also soweit angepasst und optimiert werden, dass die wahre Dichte möglichst gut rekonstruiert werden kann.

Die grundsätzliche Idee basiert auf folgenden Schritten:

1. Initialisierung: Als Startwerte der lokal aggregierten Daten dienen die Mittelpunkte der einzelnen kleinräumigen Einheiten. Darüber hinaus wird eine Rastergröße vorgegeben, auf welcher die zu schätzende Dichte basieren soll. Das Raster deckt die Flächen vollständig ab. Bestimme eine einfache zweidimensionale Kerndichteschätzung \hat{f}_0 auf diesem vorgegebenen Raster.

2. Ziehe für jeden Raum zufällige Stichproben mit Zurücklegen aus dem vorgegebenen Raster. Es werden jeweils so viele Elemente gewählt, wie dem Raum ursprünglich zugeordnet waren. Die einzelnen Rasterpunkte werden mit einer Wahrscheinlichkeit proportional zur aktuellen Dichteschätzung gewählt (geschichtete Stichprobe).
3. Schätze die Bandbreitenmatrix mit dem multivariaten Plugin-Schätzer von (Wand und Jones, 1994) und schätze die Kerndichte $\hat{f}_H(x)$ aus den simulierten Geokoordinaten erneut.
4. Wiederhole die Schritte 2 und 3. Die Schritte werden $B + N$ mal wiederholt, dabei steht B für die Anzahl an Burnin-Iterationen und N für die Anzahl weiterer Iterationen.
5. Die ersten B Iterationen hängen noch zu stark von den Startwerten des ersten Modells ab, deshalb werden diese Ergebnisse verworfen (Burnin-Phase).
6. Verwende den Mittelwert aus den verbleibenden N Dichteschätzungen $\hat{f}_H(x)$ auf dem vorgegebenen Raster als Ergebnis.

Zu den Details des Verfahrens, deren Grundlagen und dem dahinterliegenden Messfehlermodell wird auf (Groß et al., 2017) verwiesen. Hier finden sich auch Vergleiche zur Effizienz des Schätzers.

Qualität des Kernelheaping-Verfahrens

(Erfurth, 2018) zeigte, dass das Verfahren bereits bei einer geringen Anzahl an Iterationen ein Optimum bezüglich der Qualität des Schätzers zeigt. Darüber wurde gezeigt, dass die Wahl des Aggregationslevels - also die Zahl bzw. Größen der Einzelflächen im Verhältnis zur Gesamtgröße - einen hohen Einfluss auf die Qualität des Ergebnisses hat. Zwar bietet das Kernelheaping-Verfahren für alle Aggregationslevel die besten Schätzergebnisse im Vergleich zu einfachen Kerndichteschätzungen und Choroplethenkarten, jedoch nimmt die Qualität mit zunehmender Anzahl von Aggregationsstufen zu. Das bedeutet etwa, wenn Daten auf Planungsraum (PLR)-Ebene in die Schätzung eingehen, ist mit besseren Ergebnissen zu rechnen, als wenn die Schätzung auf Daten lediglich auf Prognoseraum (PRG)-Ebene basiert (zu den räumlichen Gliederungen vgl. Kapitel 2.3.1).

Eine mit dem Kernelheaping-Verfahren dargestellte Dichte ist realistischer, da in Wirklichkeit keine Sprünge an Raumgrenzen vorliegen. Der Verlauf der Dichte ist glatt und lässt Konzentrationsgebiete leicht erkennen. Ein Informationsverlust durch klassierte Farbdarstellung wird vermieden. Die Kernelheaping-Karten sind also nicht durch Gebietsgrenzen und Klassierung eingeschränkt. Die klassischen Nachteile der Choroplethenkarten werden somit vermieden. Auch ist es möglich, spezifische Konzentrationsgebiete (Hotspots) aus

den Kernelheaping-Karten abzuleiten und darzustellen. Ein Vergleich einer Choroplethenkarte, der entsprechenden Karte nach dem Kernelheaping-Verfahren und einer daraus abgeleiteten Hotspot-Karte findet sich in Abbildung 3.12.

Als Nachteile des Verfahrens sind eine vergleichsweise hohe Rechenintensität und ein gewisser Erklärungsbedarf zu nennen.

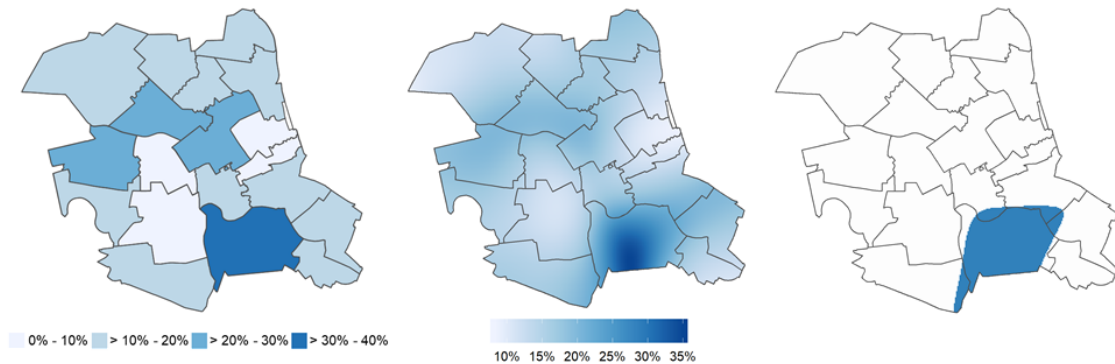


Abbildung 3.12: Choroplethenkarte, Kernelheaping-Karte und Hotspot-Karte: Anteil der Kinder mit eigener Zuwanderungserfahrung, Einschulungsbereiche, Schuljahr 2019

Am Beispiel der Verteilung des Anteils der Kinder mit Zuwanderungserfahrung sind hier verschiedene Formen von Karten dargestellt. Die klassische Choroplethenkarte (links) hat Nachteile wie harte Farbsprünge an den Raumgrenzen und Informationsverlust durch Klassenbildung. Diese Nachteile umgeht die Karte nach dem Kernelheaping-Verfahren (Mitte): eine stetige und somit realistischere Darstellung der Anteilsverteilung wird ermöglicht. Aus der Kernelheaping-Karte lassen sich Hotspots (Konzentrationsgebiete) ableiten und darstellen. Rechts dargestellt sind die 10 % der Fläche Berlin-Mittes mit den höchsten Anteilen an Kindern mit eigener Zuwanderungserfahrung. (Beispielgrafiken aus dem ESU explorer)

Die in diesem Kapitel vorgestellten statistischen Methoden und grafischen Darstellungsformen sollen durch die im Rahmen dieser Arbeit zu entwickelnde Analyseanwendung zur Verfügung gestellt werden. Mithilfe der Maßzahlen und Visualisierungen wird es den Nutzern ermöglicht, einen Überblick über die komplexen Daten zu erhalten und die Datenlage explorativ zu erkunden.

Während in diesem Kapitel die Methodik aus statistischer Sicht spezifiziert wurde, sind die methodischen Aspekte der technischen Umsetzung und Implementierung der Analyseanwendung Thema des nächsten Kapitels.

4 Anwendungsentwicklung

In diesem Kapitel wird die Entwicklung der Analyseanwendung beschrieben. Ziel ist es, den gesamten Prozess der Realisierung der Anwendung zu beleuchten. Zunächst erfolgt eine Beschreibung des Programmentwurfes. Ausgehend von den Anforderungen werden hier die wichtigen grundlegenden Technologien und der allgemeine Programmablauf vorgestellt. Anschließend folgt eine genauere Vorstellung und Begründung der verwendeten Techniken und Hilfsprogramme. Am Ende des Kapitels wird die konkrete Umsetzung hinsichtlich der realisierten Programmmodule, des Programmierstils, des Zugriffs auf die Anwendung und einzelner Aspekte von Qualitätssicherung beschrieben.

Dieses Kapitel beleuchtet also den programmiertechnischen Aspekt der Anwendungsentwicklung. Es hat dabei nicht den Anspruch, eine detaillierte Einführung in die Programmierung zu geben. Jedoch sollen die Konzepte, alle verwendeten Technologien und beispielhafte Funktionalitäten referenziert und vorgestellt werden, soweit sie zum Verständnis des Programmaufbaus beitragen. Zum tiefer gehenden Verständnis sei auf die Dokumentationen der verwendeten Technologien und die weiteren genannten Quellen verwiesen.

Innerhalb von Beschreibungen zum Thema Programmierung und Entwicklung ist es häufig schwierig, für gebräuchliche englische Fachausdrücke deutsche Synonyme zu finden und verständlich zu verwenden. Im Rahmen dieses Kapitels wird deshalb bei einigen Ausdrücken der englische gebräuchliche Begriff verwendet und auf eine durchgehende Übersetzung verzichtet. Die Namen von Programmen und Funktionen werden durch Textformatierung gekennzeichnet (Bsp. `shiny` oder `function()`).

4.1 Entwurf

Ziel des Entwurfes ist die Festlegung der Basistechnologien und des grundlegenden Verhaltens der Analyseanwendung. Den Ausgangspunkt für den Entwurf bilden die im Kapitel 2.5.3 spezifizierten Anforderungen. Dafür sollen die im Kapitel 3 ausgewählten und vorgestellten statistischen Methoden und grafischen Darstellungsformen in die Anwendung integriert werden. Daneben sollen die weiteren Anforderungen wie Lauffähigkeit, Benutzerfreundlichkeit, Zugang zu Metadaten, Möglichkeiten von Filterführungen, Export von Analyseergebnissen sowie Speichern und Laden von Analyseinstellungen erfüllt werden.

In Hinblick auf diese Anforderungen werden im Folgenden R und shiny als grundlegend verwendete Technologien vorgestellt. Der anschließend dargestellte Programmablauf zeigt die elementaren Programmschritte zur Lösung der Aufgabenstellung.

4.1.1 R

Die Implementierung der Anwendung erfolgt in der Statistiksoftware und Programmierumgebung R. Dabei stellt R eine Programmiersprache und Umgebung für statistische Berechnungen und Grafiken zur Verfügung. Die frei erhältliche (Open Source) Software R wird von einer internationalen Entwicklergemeinschaft ständig weiterentwickelt und bietet ein breites Spektrum statistischer und grafischer Techniken (Hornik, 2020). Damit erfüllt R die Anforderung, die zu entwickelnde Analyseanwendung auf einer kostenfreien Statistiksoftware zu basieren.

R enthält für grundlegende Funktionalitäten einige Basispakete (engl. core-packages - im Folgenden wird der englische Begriff Packages weiterverwendet), und kann mit weiteren R-Packages in individuellem Umfang erweitert werden. Ein R-Package ist eine Sammlung von Funktionen, Daten und Dokumentationen, welche die Leistungsfähigkeit von R erweitert (Wickham und Grolemund, 2016: xvi). Das Comprehensive R Archive Network (CRAN) (<https://cran.r-project.org/>) bietet ein Netzwerk von File Transfer Protocol (FTP)- und Web-Servern, über welche die Software, Packages, Dokumentationen sowie Programmcode frei erhältlich sind (R Core Team, 2019). Allein hier finden sich im Moment über 15.000 Packages zur freien Nutzung. Dabei lässt sich R über das ebenfalls frei erhältliche RStudio bedienen. RStudio bietet eine integrierte Entwicklungsumgebung (Integrated development environment (IDE)) und komfortable grafische Benutzeroberfläche für R (RStudio Team, 2020a).

Neben zahlreichen statistischen Methoden und Möglichkeiten der Datenvisualisierung bietet R auch die Möglichkeit, diverse Packages zur Geodatenanalyse und Kartenerstellung zu nutzen. R kann somit auch eine vollständige Geoinformationssystem (GIS)-Lösung bieten (Bivand et al., 2013).

4.1.2 shiny

Das R-Package shiny (Chang et al., 2018) bietet ein leistungsstarkes Rahmenwerk zum Erstellen von Webanwendungen mit R. Mithilfe von shiny können statistische Analysen in interaktiven Webanwendungen durchgeführt und dargestellt werden. Eine shiny-Anwendung (engl. App) resultiert in eine Internetseite mit Benutzeroberfläche (engl. User Interface (UI)). Im Hintergrund läuft dabei eine interaktive R-Session. Der Zugriff auf sämtliche R-Funktionalitäten ist damit möglich. Der Endanwender einer shiny-App benötigt keinerlei Programmierkenntnisse und interagiert ausschließlich über die grafische Benutzeroberfläche. Somit sind shiny-Apps besonders benutzerfreundlich.

Eine umfangreiche und empfehlenswerte Einführung in shiny mit vielen Artikeln und Tutorials findet sich unter <https://shiny.rstudio.com/>. Empfehlenswert für den Einstieg ist auch der shiny-Schummelzettel unter <https://rstudio.com/wp-content/uploads/2015/08/shiny-german.pdf>.

Reaktivität

Grundsätzlich basiert shiny auf reaktiver Programmierung. *Reaktivität* bedeutet, dass ständig automatisch die Abhängigkeiten der einzelnen Objekte geprüft und verfolgt werden. Damit kann die Anwendung unverzüglich auf Benutzereingaben reagieren und davon abhängige Objekte - wie beispielsweise tabellarische und grafische Ausgaben - automatisch erzeugen und aktualisieren (Wickham, 2020a: Ch. 4). Reaktivität ermöglicht damit Interaktivität, die shiny-Anwendung reagiert also auf Benutzereingaben.

Dabei werden reaktive Objekte in reaktive Werte und reaktive Ausdrücke unterschieden (Beeley und Sukhdeve, 2018: 43). Reaktive Werte (beispielsweise Benutzereingaben) nehmen Einfluss auf reaktive Ausdrücke (beispielsweise Berechnungsergebnisse, tabellarische oder grafische Ausgaben). Reaktive Ausdrücke werden dabei nur dann neu berechnet, wenn sich einer der Eingabe-Werte, von dem der Ausdruck abhängt, ändert. Reaktive Ausdrücke können ebenfalls voneinander abhängen. Reaktive Programmierung ist also ein Programmierstil, der mit reaktiven Werten startet und aufbauend die davon abhängigen Ausdrücke automatisch aktualisiert.

Architektur

Shiny-Anwendungen basieren grundsätzlich auf zwei Komponenten. Dies ist zum einen die Definition und Gestaltung der für den Endnutzer sichtbaren grafischen Benutzeroberfläche (User Interface (UI): Frontend) und zum anderen die Definition der dahinterliegenden Prozesse und Funktionen (Server-Komponenten: Backend). Das Backend ist also der Teil der Anwendung, der für die Benutzer nicht sichtbar ist, aber jegliche Programmierungen für die Datenverarbeitung, Berechnungen und Reaktionen enthält.

Die Beschreibung der grafischen Benutzeroberfläche einer shiny-App erfolgt mit einem *UI-Objekt*. Die Bereitstellung der nötigen Programme und Funktionalitäten erfolgt in der sogenannten *Server-Funktion*. Bei größeren Anwendungen mit viel Programmcode empfiehlt sich der Übersichtlichkeit halber die Anlage zweier getrennter R-Skripte namens `ui.R` und `server.R` (Beeley und Sukhdeve, 2018: 38). Liegen diese beiden R-Skripte im selben Arbeitsverzeichnis, kann die Anwendung hieraus mit dem Befehl `runApp()` gestartet werden. In Abbildung 4.1 ist die grundlegende Architektur einer shiny-Anwendung visualisiert.

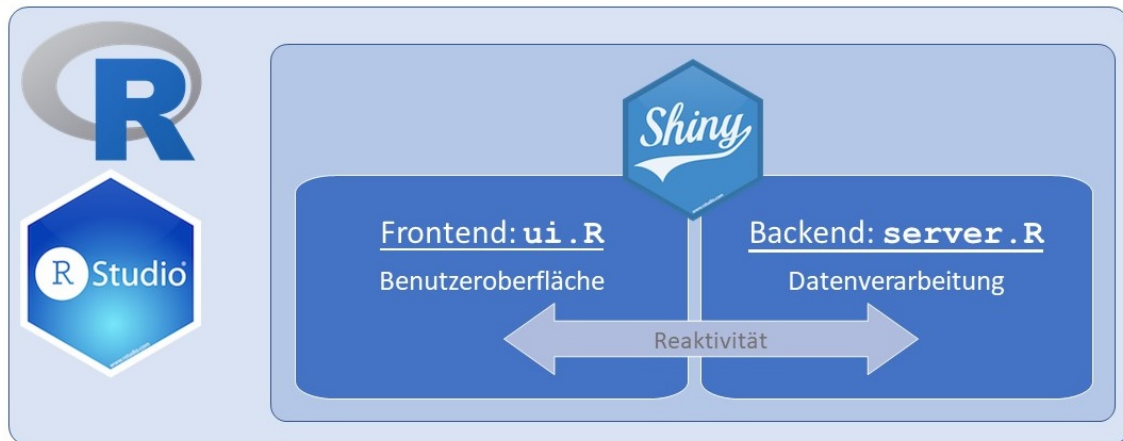


Abbildung 4.1: Architektur einer shiny-Anwendung

Die Basistechnologie für eine shiny-Anwendung bildet die Statistiksoftware und die Programmiersprache R. Über die IDE RStudio lässt sich R komfortabel bedienen. Das R-Package shiny bietet ein Rahmenwerk für die Erstellung benutzerfreundlicher Anwendungen auf Basis von Webtechnologien. Der Programmaufbau einer shiny-Anwendung unterscheidet grundsätzlich zwischen den Komponenten Frontend - der Definition der grafischen Benutzeroberfläche - und Backend - welches alle Schritte der Datenverarbeitung festlegt. Beide Komponenten interagieren miteinander und ermöglichen Reaktivität.

(Eigene Darstellung. R-Logo: (R Core Team, 2019), Hexagon Sticker: (RStudio Team, 2020b))

4.1.3 tidyverse

Das tidyverse (Wickham, 2017) ist eine Sammlung verschiedener R-Packages, welche optimiert zusammenarbeiten und auf einer gemeinsamen Philosophie von Daten- und Programmstruktur aufbauen (Wickham und Grolemund, 2016: xvi). Das tidyverse eignet sich insbesondere für die explorative Datenanalyse (Wickham und Grolemund, 2016: 81ff.).

tidyverse Packages

Mit dem tidyverse werden R-Packages für Datenwissenschaftler bereitgestellt: für Datenimport, Datenaufbereitung, Datentransformation, Visualisierung und Modellierung. Für die vorliegende Arbeit grundlegend genutzte Packages sind dies:

- dplyr: zur Datenmanipulation (Wickham et al., 2019) (siehe Kapitel 4.3.1)
- ggplot2: zur Datenvisualisierung (Wickham et al., 2018) (siehe Kapitel 4.3.2)

Daneben werden durch das tidyverse eine Reihe weitere R-Packages genutzt. Das tidyverse wird laufend aktualisiert und weiterentwickelt. Der jeweils aktuelle Stand wird unter <https://www.tidyverse.org/> und <https://tidyverse.tidyverse.org/> dokumentiert. Eine hervorragende Einführung in die tidyverse-Packages und deren Funktionalitäten bieten (Wickham und Grolemund, 2016).

tidyverse Philosophie

Das tidyverse bietet dokumentierte Prinzipien und Richtlinien für einen konsistenten Programmaufbau. Durch die Anwendungen solcher Richtlinien und eines entsprechenden Programmierstils soll die Qualität von Programmtexten und deren Struktur hinsichtlich der Lesbarkeit und Wartbarkeit sichergestellt werden. Die *tidyverse Prinzipien* (Wickham, 2020b) und der *tidyverse style guide* (Wickham, 2020c) bieten entsprechenden Vorgaben und werden im Rahmen dieser Arbeit berücksichtigt und angewandt (siehe Kapitel 4.4).

Tidy Evaluation

Der Begriff *tidy evaluation* bezeichnet ein Rahmenwerk für Metaprogrammierung in R. Bei der Metaprogrammierung wird ein Programm genutzt, um eigenen Programmcode zu erzeugen, zu bearbeiten oder zu ändern (Wickham und Henry, 2020). Bei der Arbeit mit dem tidyverse in Verbindung mit shiny ist bei komplexen Aufgaben die Anwendung und Implementierung von tidy evaluation unerlässlich. Dies gilt insbesondere für die Erstellung von shiny-Anwendungen zur interaktiven explorativen Datenanalyse (Wickham, 2020a: Ch. 13). Eine ausführliche Einführung in die Themen Metaprogrammierung und tidy evaluation findet sich in (Wickham, 2019: 371ff.) und unter <https://tidyeval.tidyverse.org/>. Funktionalitäten rund um das Thema tidy evaluation bietet das R-Package rlang (Henry und Wickham, 2019).

Die Anwendungsentwicklung im Rahmen dieser Arbeit soll grundlegend auf tidyverse-Packages basieren, sich an der tidyverse-Philosophie ausrichten sowie unter Verwendung von tidy evaluation umgesetzt werden.

4.1.4 Programmablauf

Der Programmablauf beschreibt die logische Abfolge einzelner Programmschritte zur Lösung der Aufgabenstellung. Der in Abbildung 4.2 dargestellte Programmablaufplan des ESU explorer visualisiert die grundlegende Logik und Reihenfolge der Programmschritte für die Durchführung einer Analyse und dient dem Verständnis und der übersichtlichen Darstellung der Programmlogik.

Der Ablauf der Operationen im ESU explorer ergibt sich aus folgenden Schritten (vgl. Abbildung 4.2):

1. Auswahl der Analysevariablen (erforderlich) und Filtereinstellungen (optional) durch den Nutzer.
2. Ermittlung der Datenbasis, bedingt durch die unter 1. erfolgte Auswahl.
3. Berechnung von Häufigkeiten, Verteilungsmaßzahlen, Zusammenhangsmaßen und

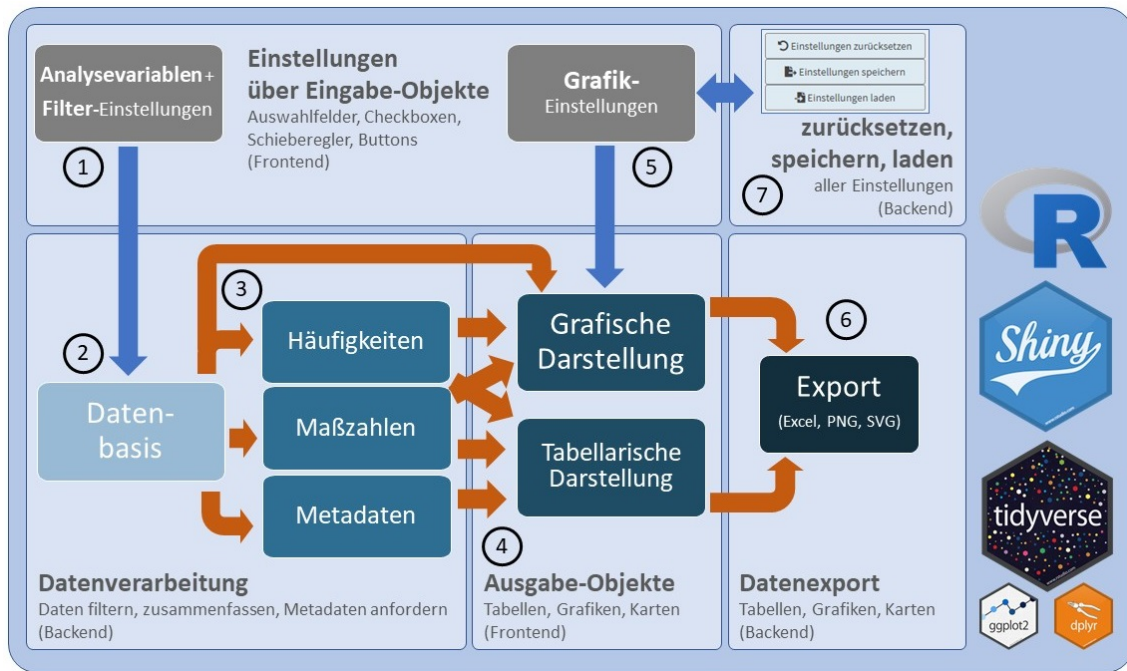


Abbildung 4.2: Programmablaufplan ESU explorer

Der Programmablaufplan visualisiert die einzelnen Programmschritte für die Erstellung einer Analyse mit dem ESU explorer. Die Rahmenwerke für das Programm stellen R, shiny, tidyverse und alle weiteren genannten R-Packages. Während die Eingabe- und Ausgabeobjekte im Frontend definiert werden, ergeben sich sämtliche Schritte der Datenverarbeitung aus den Funktionalitäten des Backends.

(Eigene Darstellung. R-Logo: (R Core Team, 2019), Hexagon Sticker: (RStudio Team, 2020b))

Anforderung der Metadaten, bedingt durch die unter 1. erfolgte Auswahl der Analysevariablen und die unter 2. ermittelte Datenbasis.

4. Grafische Darstellung von Häufigkeiten, Verteilungsmaßzahlen und Dichteschätzungen; Tabellarische Darstellung von Häufigkeiten, Verteilungsmaßzahlen, Zusammenhangsmaßen und Metadaten, bedingt durch die unter 3. erfolgten Berechnungen.
5. Spezifikation von Grafikeinstellungen durch den Nutzer (optional): Auswahl verschiedener Grafiktypen und vom Grafiktyp abhängiger Grafikparameter wie beispielsweise Diagrammelemente, Farbpaletten, Beschriftungs-, Legenden- und Skalierungsoptionen.
6. Export der Analyseergebnisse in verschiedenen Dateiformaten (optional).
7. Speichern der unter 1. und 5. ausgewählten aktuellen Einstellungen (optional). Die Einstellungsparameter werden als CSV-Datei im Programmordner gespeichert und können zu einem späteren Zeitpunkt zur Reproduktion der Analyse wieder geladen werden. Darüber hinaus die Möglichkeit, alle Einstellungen auf die Ausgangswerte zurückzusetzen (optional).

Die grundlegenden Rahmenwerke für diese Anwendung bilden die Programmiersprache R, das Rahmenwerk shiny zum Erstellen einer benutzerfreundlichen Webanwendung sowie die tidyverse-Familie mit den R-Packages zur Datenanalyse und den grundlegenden

Programmierprinzipien. Im Folgenden werden die speziellen Programmelemente und die Beschreibung der Umsetzung unter Verwendung von R-Packages und weiterer Technologien getrennt nach den beiden Komponenten Frontend und Backend dargestellt. Dabei ist zu beachten, dass die Techniken teilweise fließend ineinander übergreifen, da im Backend Prozesse ausgelöst werden, die entsprechend Einfluss auf die Darstellung im Frontend haben. Dabei wird ein Bezug zu den einzelnen Schritten des Programmablaufs hergestellt. Eine tabellarische Übersicht über alle im Rahmen dieser Arbeit verwendeten R-Packages findet sich in Anhang A.2.

4.2 Frontend: User Interface

Das Frontend - also die grafische Benutzeroberfläche (User Interface (UI)-Objekt) enthält sämtliche für den Endanwender sichtbaren Objekte der shiny-Anwendung. Hierzu gehören Texte, Eingabe- und Ausgabeobjekte sowie das Layout, also die Struktur, Anordnung und Darstellung dieser Elemente und Objekte.

Als Eingabeobjekte bietet shiny verschiedene Kontrollelemente. Damit können Eingaben und Angaben des Endanwenders erfasst werden. Zu diesen Elementen zählen beispielsweise Auswahlfelder wie Select-Boxen, Checkboxes und Schieberegler aber auch Buttons, die bestimmte Funktionalitäten auslösen. Die Auswahl der Analysevariablen, die Filtereinstellungen sowie die Grafik-Einstellungen (Programmschritte 1 und 5, siehe Kapitel 4.1.4) erfolgen im ESU explorer über solche Eingabeobjekte. Die Werte von Eingabeobjekten werden als reaktive Werte bezeichnet (Beeley und Sukhdeve, 2018: 43). Andere Objekte können von ihnen abhängen, also auf diese Eingaben reagieren (siehe Kapitel 4.1.2).

Weiterhin werden innerhalb der Benutzeroberfläche auch die Bereiche für Ausgaben definiert. Hierzu gehören beispielsweise die Ausgabe von Texten, Tabellen oder Grafiken (Programmschritt 4, siehe Kapitel 4.1.4). Solche Ausgaben werden als reaktiv bezeichnet, wenn sie automatisch auf Eingaben vom Endanwender oder auch andere Objekte reagieren. Innerhalb des Frontends erfolgt jedoch zunächst nur die Festlegung der Bereiche, in welchen Ausgaben erfolgen sollen. Die Berechnung und Erstellung solcher Ausgaben wird an anderer Stelle - im Backend (der Server-Funktion) - ausgelöst (siehe Kapitel 4.3). Die dort durchgeführten Berechnungen werden zurück zum UI-Objekt gesendet (Beeley und Sukhdeve, 2018: 41).

Das Frontend der resultierenden shiny-Anwendung basiert auf den bekannten Webtechnologien Hypertext Markup Language (HTML) (Auszeichnungssprache für Dokumentenstruktur und Vernetzung von Internetseiten), Cascading Style Sheets (CSS) (Formatierungssprache für die grafische Gestaltung von HTML-Dokumenten) und JavaScript (Integration von Interaktivität in HTML-Seiten). Mit dem shiny-Package lassen sich dabei Anwendungen zunächst gänzlich ohne Kenntnisse dieser Web-Sprachen erstellen. Eine eigene Integration entsprechender Sprach-Elemente ist jedoch machbar, erweitert die

Möglichkeiten und kann die Qualität einer shiny-App stark steigern (Fay et al., 2020).

4.2.1 Layout und Gestaltung

Die Gestaltung und Anordnung von Informationen werden als Layout bezeichnet. Shiny-Anwendungen basieren standardmäßig auf dem Bootstrap Layout (Beeley und Sukhdeve, 2018). Bootstrap ist ein Rahmenwerk für HTML, CSS und JavaScript und bietet thematische Grafikvorlagen für nutzerfreundliche Internetseiten. Das Layout einer Anwendung bestimmt maßgeblich deren Benutzerfreundlichkeit. Unter diesem Aspekt wurde das shiny-Standardlayout im Rahmen der Anwendungsentwicklung erweitert und angepasst durch die Verwendung des R-Packages shinydashboard sowie der Nutzung von HTML- und CSS-Elementen.

Layout mit shinydashboard

Das Package shinydashboard bietet ein Rahmenwerk für die grafische Gestaltung von shiny-Anwendungen und ermöglicht die einfache Darstellung ansprechender Oberflächen für Dashboards (Chang und Borges Ribeiro, 2018). Dashboard (zu Deutsch Armaturenbrett) ist ein Oberbegriff für grafische Benutzeroberflächen auf denen Informationen übersichtlich dargestellt werden.

Eine shiny-Benutzeroberfläche (UI) wird wie eine Webseite aus HTML-Bausteinen aufgebaut. Das shinydashboard Package bietet eine Reihe von Funktionen zur Erzeugung solcher HTML-Bausteine zum Aufbau benutzerfreundlicher Dashboards. Die Grundbausteine sind dabei eine Kopfzeile, ein Seitenmenü und der eigentliche Seiteninhalt. Innerhalb der Anwendung lassen sich Inhalte und alle in shiny bereits enthaltenen Eingabe- und Ausgabeobjekte auf Unterseiten, mit gestaltbaren Boxen, auf Wunsch mit weiteren Unterteilungen (Tabs) anordnen. Die Unterteilung einer Anwendung auf mehrere Einzelseiten und -inhalte hat einen weiteren Vorteil: die Berechnung der Ausgaben erfolgt jeweils nur für die jeweils ausgewählte Seite bzw. den ausgewählten Inhalt (Beeley und Sukhdeve, 2018: 49). Die verschiedenen Inhalte des ESUexplorer werden so thematisch gegliedert auf verschiedenen Seiten dargestellt, welche über das Seitenmenü schnell zugänglich sind. Damit bietet shinydashboard eine besonders ansprechende und benutzerfreundliche Oberfläche für die shiny-Anwendung.

HTML und CSS

HTML ist die Auszeichnungssprache, mit der Internetseiten erzeugt und dargestellt werden. Shiny erstellt diesen HTML-Code im Hintergrund automatisch. Shiny bietet darüber hinaus die Möglichkeit, eigenen HTML-Code einzubinden (Beeley und Sukhdeve, 2018: 59). Damit ist es möglich, die Darstellung von Inhalten zu optimieren. In der vorliegenden Anwendung wurden HTML-Elemente beispielsweise für Textformatierungen, Erzeugung

von Links, Darstellung von Listen oder Zuweisung von CSS-Klassen integriert.

Mit Cascading Style Sheets (CSS) steht ein mächtiges Werkzeug für die Gestaltung von HTML-Seiten zur Verfügung. Dabei lassen sich die einzelnen Elemente der HTML-Seite gestalten und dynamisch anpassen (Bühler et al., 2017: 44f.). Für die Gestaltung des ESU explorers wurden Anpassungen des Layouts durch CSS vorgenommen. Dafür wurde eine externe CSS-Datei in das UI-Objekt eingebunden. Zur Anwendung gekommen sind verschiedene Gestaltungsmöglichkeiten wie Schriftformate und Formatierungen von Buttons und weiterer Objekte.

4.2.2 Erhöhte Benutzerfreundlichkeit mit JavaScript

Neben HTML und CSS stellt JavaScript die dritte Säule für die Erstellung von Internetseiten dar und sorgt für Dynamik und Interaktivität eines Dokuments. Shiny-Anwendungen selbst basieren auf interaktiven JavaScript-Komponenten. Darüber hinaus lassen sich für eine Optimierung der Benutzerfreundlichkeit R-Packages einbinden, die weitere JavaScript-Funktionalitäten bieten. Damit können die einzelnen Eingabe- und Ausgabeobjekte der shiny-Anwendung interaktiv gestaltet werden.

Auswahlfelder mit `selectize.js`

Die JavaScript-Bibliothek `selectize.js` bietet sehr flexible Eingabeobjekte vom Typ Select-Box. Mittels eines `selectizeInput()`-Objektes ist es möglich, nach den angebotenen Optionen mittels Texteingabe zu suchen, die Optionen nach Bereichen zu gliedern sowie die Anzahl der darzustellenden Optionen festzulegen (Xie, 2017). Die Benutzerfreundlichkeit kann dadurch vor allem bei langen Auswahllisten entscheidend gesteigert werden. In Abbildung 4.3 ist eine Auswahlmöglichkeit mittels `selectizeInput()`-Eingabeobjekt dargestellt.

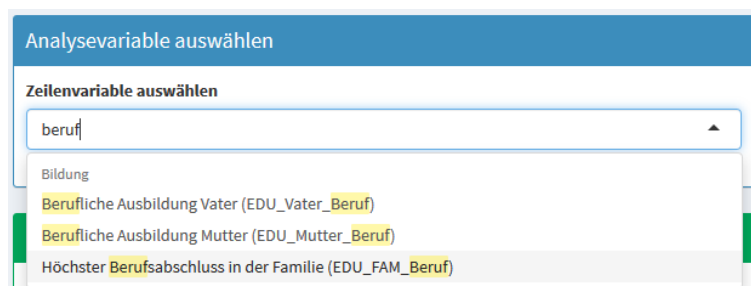


Abbildung 4.3: Auswahlmöglichkeit über `selectizeInput()`-Objekt

Das dargestellte Eingabeobjekt vom Typ `selectizeInput()` basiert auf der `selectize.js`-Bibliothek. Die Auswahl ist dabei über ein Dropdown-Menü möglich. Die Auswahlmöglichkeiten lassen sich in Bereiche gliedern (hier sichtbar: Bereich *Bildung*). Insbesondere die Möglichkeit der Suche nach Textinhalten verbessert die Benutzerfreundlichkeit speziell bei vielen Auswahlmöglichkeiten entscheidend. Eine partielle Übereinstimmung des zu suchenden Textes mit den Optionen ist ausreichend. (Screenshot ESU explorer)

Ladeanzeigen mit shinycssloaders

Mit dem R-Package shinycssloaders (Sali und Hass, 2017) lassen sich Ladeanzeigen in die Anwendung integrieren, um dem Endnutzer zu signalisieren, dass Berechnungen stattfinden. Diese Ladeanzeigen erscheinen beim ESU explorer während der Berechnung der grafischen und tabellarischen Darstellungen (Programmschritt 4, siehe Kapitel 4.1.4).

Bedingte Anzeigen mit shinyjs

Das R-Package shinyjs ermöglicht es, die Ein- oder Ausgabeobjekte im Frontend per JavaScript variabel ein- und auszublenden. So werden beispielsweise bei der Spezifikation der Grafikeinstellungen im ESU explorer durch diese Funktionalitäten nur die Auswahlobjekte angezeigt, welche für die ausgewählte Grafikform relevant sind (Programmschritt 5, siehe Kapitel 4.1.4).

4.2.3 Erweiterte Interaktivität mit HTMLwidgets

Weiterhin lassen sich spezielle JavaScript-Bibliotheken zur Erhöhung von Interaktivität von Tabellen und Grafiken in shiny-Apps einbinden. Sogenannte *HTMLwidgets* bieten Rahmenwerke zur Verknüpfung von JavaScript-Bibliotheken und R (Vaidyanathan et al., 2020). Jedes HTMLwidget wird in einem eigenen R-Package bereitgestellt.

Interaktive Tabellen mit DT

Mit dem HTMLwidget und R-Package DT (Xie et al., 2019) lassen sich einfache HTML-Tabellen vielfältig individualisieren und interaktiv gestalten. Das Package realisiert die Einbindung der JavaScript-Bibliothek DataTables. Das erklärte Ziel von DataTables ist die verbesserte Zugänglichkeit von Daten in HTML-Tabellen (SpryMedia, 2020). Damit lassen sich bei den Tabellen im ESU explorer beispielsweise Funktionalitäten zur Suche nach Tabelleninhalten oder zum automatischen Download der Tabellen bereitstellen, lange Tabellen unterteilen sowie individuelle JavaScripte zur Erweiterung der Interaktivität integrieren.

Interaktive Grafiken mit ggiraph

Das HTMLwidget und R-Package ggiraph (Gohel und Skintzos, 2019) bindet die JavaScript Bibliothek d3 ein. Diese ist auf eine interaktive Visualisierung von Daten innerhalb von HTML-Dokumenten ausgerichtet. Mithilfe von ggiraph lassen sich Grafiken mit Tooltips ergänzen. Tooltips sind kleine Popups, welche weitere Informationen enthalten. Diese erscheinen, wenn der Benutzer den Mauszeiger auf eine bestimmte Fläche der Grafik richtet. Diese Funktionalität wurde im ESU explorer bei der Darstellung thematischer Karten genutzt.

4.3 Backend: Server

Das Backend definiert die Datenverarbeitung im Hintergrund der Anwendung. Während das Frontend den für den Benutzer sichtbaren Teil der Anwendung darstellt, enthält das Backend alle nicht-sichtbaren Prozesse, Programmabläufe und Funktionen. Über das Frontend interagiert der Nutzer mit dem Backend. Das Backend einer shiny-Anwendung wird auch als Server bezeichnet.

Im `server.R`-Skript werden dabei alle Objekte, die im `ui.R`-Skript zunächst nur angelegt wurden, beschrieben und mit Inhalten gefüllt. Die im `server.R`-Skript enthaltene Server-Funktion enthält also die Anweisungen, wie die Objekte erstellt und aktualisiert werden. Die Objekte gelten dann als reaktiv, wenn sie auf Eingabeobjekte oder berechnete Werte reagieren (siehe Kapitel 4.1.2). Im `server.R`-Skript werden dabei auch Abhängigkeiten der Objekte untereinander festgelegt. Shiny definiert diese Abhängigkeiten automatisch. Sobald ein reaktiver Ausdruck auf ein Eingabeobjekt oder einen weiteren reaktiven Wert zugreift, werden beide Objekte miteinander vernetzt. Damit wird ein Ausgabeobjekt in dem Moment automatisch neu berechnet, wenn sich ein Wert ändert, von welchem das Ausgabeobjekt abhängt (Beeley und Sukhdeve, 2018: 43).

Innerhalb des Backends werden Technologien für Datenmanipulation, Datentransformation, Datenvisualisierung und spezielle statistische Methoden genutzt. Im Folgenden werden alle für das Backend genutzte R-Packages genannt und die entsprechende Verwendung dokumentiert.

4.3.1 Datentransformation und statistische Methoden

Die Basis für die Datenanalyse und die Berechnung der ausgewählten statistischen Methoden bilden R-Packages für Datenaufbereitung, Datenmanipulation, Datentransformation und spezielle statistische Methoden (Programmschritte 2 und 3, siehe Kapitel 4.1.4). Ein Großteil dieser Transformationen und Berechnung sind über die R-Packages der `tidyverse`-Familie realisiert worden (siehe Kapitel 4.1.3). Daneben werden weitere R-Packages für spezielle Aufgabenstellungen genutzt.

Datenmanipulation mit `dplyr`

Bei einer statistischen Analyse sind Aufbereitung, Transformation und Zusammenfassen von Daten unerlässlich. Das `dplyr`-Package aus der `tidyverse`-Familie bietet alle grundlegenden Funktionalitäten für entsprechende Aufgaben der Datenmanipulation (Wickham und Golemund, 2016: 45). Mit `dplyr` lassen sich Zeilen und Spalten eines Datensatzes filtern, Daten gruppieren, neue Variablen berechnen und zusammenfassende Statistiken berechnen. Die Ermittlung der Datenbasis aufgrund der vom Anwender spezifizierten Filtereinstellungen erfolgt durch die Funktionalitäten des `dplyr`-Packages. Anschließend

werden die in dieser Arbeit gewünschten Häufigkeitsverteilungen (siehe Kapitel 3.2) und Verteilungsmaßzahlen (siehe Kapitel 3.3) mittels `dplyr` berechnet.

Kerndichteschätzungen mit Kernelheaping

Das R-Package `Kernelheaping` (Groß, 2018) implementiert den in Kapitel 3.6.2 beschriebenen Algorithmus zur Kerndichteschätzung nach der Methode der simulierten Geokoordinaten. Das Package bietet auch die `dshapebivrrProp()`-Funktion, welche eine bivariate Kerndichteschätzung von prozentualen Anteilen für lokal aggregierte Daten ermöglicht. Das Verfahren erlaubt darüber hinaus auch eine Grenzkorrektur. Da die an den Rändern liegenden Punkte weniger Nachbarpunkte aufweisen, die auf diese Punkte Einfluss nehmen, kann die Dichte an diesen Punkten unterschätzt werden. Zu der Schätzung von prozentualen Anteilen und der Grenzkorrektur sei auch auf die Arbeiten von (Groß et al., 2018) und (Erfurth, 2018) verwiesen. Der ESU explorer nutzt das `Kernelheaping`-R-Package für die Berechnung der Dichteschätzung für die Darstellung von Kerndichtekarten.

Die Funktionen des `Kernelheaping`-Packages wurden für die Anwendung innerhalb der shiny-App modifiziert. Die Berechnungen für den Schätzungsprozess nehmen je nach Parameterwahl einige Zeit in Anspruch. In Hinblick auf die Benutzerfreundlichkeit war es hier Ziel, den Endnutzer über den Fortschritt dieser Berechnungen auf dem Laufenden zu halten. Damit ein regelmäßiger Fortschritt aus den Berechnungsschritten abgeleitet werden kann, wird dafür an mehreren Stellen der aktuelle Berechnungsstand an die shiny-Anwendung übergeben.

Weitere Funktionalitäten

Weitere R-Packages zur Datentransformation wurden für einzelne Funktionalitäten des shiny-Backends genutzt. Dabei kamen zum Einsatz:

- Weitere Packages der `tidyverse`-Familie (Wickham, 2017):
 - `tidyr` und `tibble` für eine konsistente Datenstruktur (zum Konzept einer guten Datenstruktur (*tidy data*) siehe (Wickham, 2014) und (Wickham und Grolemund, 2016: 149)),
 - `stringr` zum Arbeiten mit Textvariablen,
 - `purrr` für iterative Programmierung,
- `DescTools` (Signorell, 2019): zur Berechnung spezifischer Zusammenhangsmaße,
- `janitor` (Firke, 2019) zur Ergänzung von Summenzeilen in Häufigkeitstabellen,
- `Hmisc` (Harrell, 2019) für die Nutzung des Label-Attributes einer Variable,
- `rgdal` (Bivand et al., 2019), `maptools` (Bivand und Lewin-Koh, 2019) und `rgeos` (Bivand und Rundel, 2019) für die vorbereitende Datenaufbereitung der Geodaten.

4.3.2 Datenvisualisierung

Die Datenvisualisierung nimmt einen hohen Stellenwert bei der Darstellung der Analyseergebnisse ein (Programmschritt 4, siehe Kapitel 4.1.4). Alle mit dem ESU explorer erstellten Diagrammen und Karten werden durch das `ggplot2`-Package realisiert.

Datenvisualisierung mit `ggplot2`

Das R-Package `ggplot2` (Wickham et al., 2018) gehört zur *tidyverse*-Familie und bietet vielfältige Funktionalitäten zur Datenvisualisierung. Dabei gehört `ggplot2` zu den am meisten heruntergeladenen R-Packages überhaupt (Wickham, 2016).

`ggplot2` basiert auf einer eigenen Grammatik: *A Layered Grammar of Graphics*, entworfen von (Wickham, 2010). Das *gg* in `ggplot2` verweist darauf. Diese speziell für R entwickelte Grammatik basiert wiederum auf den grundlegenden Konzepten und Sprachregeln von (Wilkinson, 2005), der *Grammar of Graphics*. Dabei werden die Grafiken aus verschiedenen Schichten aufgebaut. Die grundlegenden Elemente jeder Schicht bilden die Spezifikation der Daten und ihre Zuordnung zu den visuellen Merkmalen der Grafik sowie die Spezifikation der geometrischen Objekte (Grafiktypen wie Punkte, Linien, Säulen, Flächen etc.). Weiterhin können statistische Transformationen, Skalen, das Koordinatensystem, eventuelle Datenunterteilungen sowie das Layout (Farben, Schriften usw.) spezifiziert werden (Wickham, 2010: 8), (Wickham, 2016: 4-5) .

Ein großer Vorteil von `ggplot2` liegt in der Kombinationsmöglichkeit verschiedener Diagrammtypen. Dafür werden mehrere Grafiksichten kombiniert dargestellt, um Vorteile verschiedener Darstellungsformen in einer einzigen Grafik zu nutzen. Durch den geschichteten Aufbau eignet sich `ggplot2` auch hervorragend für die Darstellung von thematischen Karten.

Erweiterung der `ggplot2`-Grafiken

Darüber hinaus werden weitere R-Packages zur Erweiterung der Funktionalitäten von `ggplot2` genutzt:

- `mapproj` (McIlroy, 2019): Zur Kartenprojektion bei der Darstellung thematischer Karten.
- `RColorBrewer` (Neuwirth, 2014): Zur Erzeugung von Farbpaletten die optimiert sind für die Darstellung geordneten Daten, wie sie beispielsweise bei thematischen Karten vorliegen.
- `scales` (Wickham und Seidel, 2019): Funktionalitäten zur Darstellung von Anteilswerten im Prozent-Format.
- `ggExtra` (Attali und Baker, 2019): Funktionalität zur Kombination von `ggplot2`-Streudiagrammen mit Randgrafiken (vgl. Abbildungen 3.5 und 3.9).

4.3.3 Export der Analyseergebnisse

Die tabellarischen und grafischen Analyseergebnisse sollen aus der shiny-Anwendung heraus zur weiteren Verwendung exportiert werden können (Programmschritt 6, siehe Kapitel 4.1.4). Einzelne tabellarische Ergebnisse lassen sich über die Funktionalitäten des DT-Packages als Excel- oder Druckformat exportieren (siehe Kapitel 4.2.3). Einzelne Grafiken lassen sich über shiny- und ggplot2-Funktionalitäten als Grafikformat (Portable Network Graphics (PNG)) exportieren. Im ESU explorer finden sich Export-Buttons bei den entsprechenden Ausgabeobjekten.

Aufbau von Excel-Dateien mit openxlsx

Für noch benutzerfreundlichere Analyseexporte sollen die verschiedenen tabellarischen und grafischen Ergebnisse auch gemeinsam in einer einzigen Datei verfügbar sein. Für die Erzeugung und Bearbeitung von Excel-Dateien wurde das R-Package openxlsx verwendet (Schauberger und Walker, 2019). Damit lassen sich alle Ergebnisse einer mit dem ESU explorer durchgeführten Analyse - Tabellen, Maßzahlen, Metadaten und Grafiken - zusammen in einer einzigen Excel-Datei exportieren.

Editierbare Vektorgrafiken mit rvg und officer

Die Möglichkeit der individuellen Nachbearbeitung von Grafiken kann hilfreich sein, wenn die grafischen Analyseergebnisse wie Diagramme und Karten für Kommunikation oder Präsentationen genutzt werden sollen. Dafür müssen die Grafiken als editierbare Vektorgrafiken im Scalable Vector Graphics (SVG)-Format vorliegen. Excel unterstützt die Bearbeitung von SVG-Grafiken. Eine Funktionalität zur Erstellung von SVG-Grafiken bietet das R-Package rvg (Gohel, 2020b). In Verbindung mit dem R-Package officer (Gohel, 2020a) lassen sich diese editierbaren Grafiken in Excel-Dateien einfügen. Somit lassen sich die exportierten Grafiken nachbearbeiten, beispielsweise durch ein Hinzufügen oder Entfernen von Bildelementen, das Ändern von Farben oder das Anpassen einzelner Beschriftungselemente.

4.3.4 Speichern und Laden von Analyseeinstellungen

Auf den Analyseseiten des ESU explorer finden sich Buttons zum Zurücksetzen, Speichern und Laden von Analyseeinstellungen (Programmschritt 7, siehe Kapitel 4.1.4).

Der Button „Einstellungen zurücksetzen“ löst ein Zurücksetzen aller Eingabeobjekte aus. Alle durch den Benutzer ausgewählten Einstellungen zur Auswahl der Analysevariablen, Filter- und Grafikeinstellungen werden hier auf ihre Ursprungswerte zurückgesetzt.

Durch die Betätigung des Buttons „Einstellungen speichern“ wird ein Prozess ausgelöst, der die Werte aller Eingabeobjekte erfasst und in einem einfachen CSV-Format im Programmordner abspeichert. Der Anwender kann einen Namen für diese Datei angeben.

Dabei wird automatisch geprüft, ob unter diesem Namen bereits etwas abgespeichert ist. Mit dem Button „Einstellungen laden“ wird dem Nutzer eine Auswahlliste angezeigt, welche die abgespeicherten AnalyseEinstellungen unter den Dateinamen anbietet. Der Nutzer kann hier eine gespeicherte Analyse auswählen. Anschließend werden automatisch alle Eingabeobjekte durchlaufen und auf die entsprechend spezifizierten Werte aktualisiert. So kann eine gespeicherte Analyse vollständig reproduziert werden.

Diese Funktionalitäten konnten mit den Grundfunktionen von shiny und R realisiert werden. Hier werden für eine benutzerfreundliche Anwendung des ESU explorers Modal Dialoge (kleine Fenster mit den entsprechenden Auswahlobjekten und Buttons) genutzt.

4.4 Implementierung

Die Implementierung beschreibt die konkrete Umsetzung der Aufgabenstellung auf Basis der Programmiersprache R und den vorgestellten verwendeten Technologien. Im Folgenden werden dafür die wichtigsten Module, der verwendete Programmierstil, der realisierte Zugriff auf die Anwendung sowie Maßnahmen zur Qualitätssicherung vorgestellt.

4.4.1 Shiny-Module

Durch Module bleiben das `server.R` und `ui.R`-Skript übersichtlich, indem einzelne Komponenten mit einer bestimmten Aufgabe ausgelagert werden. Für shiny-Module werden sowohl der Frontend- als auch der Backend-Teil einer Aufgabenstellung durch zwei zusammengehörige Funktionen definiert. Shiny-Module bieten dabei eine Abstraktionsebene, die über die von Funktionen hinausgeht. Module bieten einen eigenen Namensraum und können von einer shiny-App auch mehrfach verwendet werden (Cheng, 2019). Damit können Duplikationen von Programmcode vermieden und Programmtexte übersichtlich strukturiert werden. Für die Unterseiten mit den verschiedenen Funktionalitäten des ESU explorer wurden verschiedene Module implementiert.

Modul analysis

Das wichtigste Modul realisiert den in Kapitel 4.1.4 dargestellten Programmablauf für bivariate statistische Analysen (R-Skript `analysis.R`). Auf der Benutzeroberfläche kann der Anwender hier die einzelnen Analysevariablen auswählen und Filtereinstellungen spezifizieren. Als Ausgabe erscheint eine zum Skalenniveau passende grafische Darstellung, welche der Anwender anschließend ändern und durch verschiedene Einstellungen spezifizieren kann. Als tabellarische Ausgaben werden absolute und relative Häufigkeiten immer, Verteilungsmaßzahlen und Zusammenhangsmaße jedoch abhängig vom Skalenniveau dargestellt. Die Zusammenhangsmaße werden in der Ausgabe mit Interpretationshilfen ergänzt. Weiterhin werden die Metadaten der Analysevariablen sowie die Filtereinstellungen

gen tabellarisch aufgeführt. Alle Analyseergebnisse lassen sich einzeln oder gemeinsam exportieren.

Modul maps

Ein zweites Programmmodul wurde ausschließlich für eine sozialraumorientierte Analyse mit Kartendarstellungen erstellt (R-Script `maps.R`). Damit können Anteilswerte auf einer Karte von Berlin-Mitte dargestellt werden. Dafür können die verschiedenen räumlichen Gliederungen von Berlin-Mitte und eine kategoriale Analysevariable ausgewählt werden. Dargestellt werden die relativen Häufigkeiten für eine Ausprägung dieser Analysevariable. Eine Choroplethen-Karte wird automatisch erzeugt. Der Anwender hat hier die Möglichkeit, verschiedene Methoden der Klassenbildung anzuwenden, um die Darstellung zu optimieren (siehe Kapitel 3.6.1). Darüber hinaus kann der Anwender die Erstellung von Kerndichtekarten nach dem Kernelheaping-Verfahren anfordern (siehe Kapitel 3.6.2). Als tabellarische Ausgaben werden hier die Häufigkeiten, Metadaten und Filtereinstellungen gezeigt. Auch hier lassen sich die Ergebnisse einzeln oder gemeinsam exportieren.

Beide Programmmodule nutzen dieselben Programmcode-Bausteine zu den Filtereinstellungen, zur Erzeugung der Datenbasis und zum Speichern und Laden von Analyseeinstellungen. Für die einzelnen Programmschritte wurden Funktionen definiert. Der Aufbau von Funktionen für Datenanalyse und Datenvisualisierung erfolgte unter Anwendung von `tidy evaluation` (siehe Kapitel 4.1.3).

Module für Metadaten

Darüber hinaus wurden Module für die Inhalte zur Darstellung der Metadaten implementiert. Hier werden dem Anwender Unterseiten mit einer tabellarischen Darstellung aller Metadaten, zum Download der ESU-Dokumentationsbögen und ein Hilfebereich zur Verfügung gestellt (R-Skripte `metaTab.R`, `downloadDoku.R`, `help.R`). Die verschiedenen Inhalte und Unterseiten des ESU explorer werden in Kapitel 5 dargestellt.

4.4.2 Programmierstil

Programmcode soll möglichst einfach zu lesen, zu bearbeiten und zu prüfen sein. Das Ziel der Anwendung entsprechender Gestaltungsrichtlinien (engl. *style guides*) ist ein konsistenter und strukturierter Aufbau von Programmcode. Die Anwendung entsprechender Richtlinien bei der Programmcode-Erstellung wird auch als Programmierstil bezeichnet. Das in dieser Arbeit stark genutzte `tidyverse` bietet mit den `tidyverse` Prinzipien und dem `tidyverse style guide` entsprechende Richtlinien.

tidyverse Prinzipien

Für das tidyverse wurden konsistente Prinzipien entworfen, um eine einheitliche Basis zu schaffen und Design und Zusammenarbeit von Programmen und Packages zu optimieren. (Wickham, 2020b) beschreibt dafür im Artikel *The tidy tools manifesto* vier Basis-Prinzipien:

1. Vorhandene Datenstrukturen wiederverwenden.
2. Zusammenstellen einfacher Funktionen unter Verwendung des Pipe-Operators (`%>%`).
3. Ausrichtung an funktionaler Programmierung.
4. Entworfen für Menschen.

Auf diese tidyverse Prinzipien ist auch die vorliegende Programmentwicklung ausgerichtet. Datenstrukturen wurden soweit wie möglich wiederverwendet und die einzelnen Schritte des Programmablaufs in Funktionen unter Verwendung des Pipe-Operators ausgelagert. Der Pipe-Operator `%>%` sorgt dabei für gute Lesbarkeit und Verständlichkeit (zur Anwendung des Pipe-Operators siehe (Wickham und Grolemund, 2016: 261ff.)). Ein gut lesbarer und wartbarer Programmcode soll dabei durch die Verwendung des tidyverse style guide sichergestellt werden.

tidyverse style guide

Der *tidyverse style guide* bietet Richtlinien für die Erstellung von gut lesbaren Programmtexten (Wickham, 2020c). Wichtige Kriterien bei der Anwendung des tidyverse style guide sind der Aufbau der Syntax sowie der Aufbau der Dateistruktur.

Unter Syntax versteht man ein Regelsystem für den Aufbau der Programmtexte. Der tidyverse style guide empfiehlt dafür beispielsweise eine konsistente und aussagekräftige Namensgebung von Objekten, Funktionen und Argumenten sowie eine Strukturierung und Dokumentation der Programmtexte durch kurze beschreibende Kommentare. Darüber hinaus soll der Code übersichtlich und lesbar bleiben durch Formatierungsregeln zu Zeilenumbrüchen, Einsatz von Leerzeichen und Einrückungen (Wickham, 2020c). Diese Formatierungsregeln konnten mit Hilfe des R-Packages `styler` (Müller und Walthert, 2019) automatisch sichergestellt werden.

Der tidyverse style guide empfiehlt darüber hinaus eine sinnvolle Dateistruktur und eine aussagekräftige und konsistente Namensgebung einzelner R-Skripte (Wickham, 2020c). Die interne Aufspaltung eines Programms in einzelne R-Skripte erhöht die Übersichtlichkeit und Wartbarkeit des Programmcodes. Insbesondere bei komplexen shiny-Anwendungen empfiehlt sich eine entsprechende Dateistrukturierung. Bei der Programmentwicklung wurden dafür folgende Strukturmaßnahmen realisiert:

- Einleitende Programmschritte und Objektdefinitionen wurden in ein `global.R`-Skript ausgelagert,

- Anwendung von shiny-Modulen für die einzelnen Unterseiten des ESU explorer,
- Auslagern von Funktionsdefinitionen für die einzelnen Programmschritte,
- Aufteilung der für die Anwendung notwendigen R-Skripte, Daten und Dateien innerhalb einer sinnvollen Ordnerstruktur.

Die Ordnerstruktur und die Beziehungen zwischen den Daten- und Programmdateien des ESU explorer sind im Anhang A.3, Abbildungen A.1 und A.2 visualisiert.

4.4.3 Zugriff auf die Anwendung

Für die Bereitstellung einer shiny-Anwendung bestehen grundsätzlich zwei verschiedene Möglichkeiten (RStudio Team, 2020c):

- Teilen der shiny-Anwendung als R-Skripte oder
- Teilen der shiny-Anwendung als Internetseite.

Jeder Rechner der die Programme R und RStudio sowie einen beliebigen Internetbrowser installiert hat, kann die vorliegende shiny-Anwendung ausführen. Die Anwender benötigen dafür nur eine Kopie des Programmordners mit den notwendigen R-Skripten und Daten. Der Programmordner kann per Datenträger oder auch als verschlüsselte zip-Datei auf weitere Rechner übertragen werden. Der Start der Anwendung kann über RStudio erfolgen. Ein Nachteil bei dieser Methode ist, dass Änderungen in der Anwendung nicht zentral durchgeführt werden können, sondern an jedem Rechner einzeln erfolgen müssen. Weit häufiger werden shiny-Anwendungen als Internetseite an zentraler Stelle bereitgestellt. Der Endnutzer ruft die Anwendung in diesem Fall ausschließlich genau wie eine Internetseite auf und benötigt auf dem eigenen Rechner kein R (RStudio Team, 2020c). Dabei bestehen verschiedene Möglichkeiten die Anwendung zu hosten. RStudio bietet beispielsweise einen eigenen Hostingservice (<https://www.shinyapps.io/>) sowie ein Programm zum Aufbau eines eigenen Internet-Servers speziell für shiny-Anwendungen (<https://github.com/rstudio/shiny-server>).

Für den Endanwender ist ein Zugriff auf eine shiny-Anwendung über eine zentral gehostete Internetseite natürlich am einfachsten. Diese Möglichkeit wurde für die Anwendung im Rahmen des Bildungsmonitoring Berlin-Mitte jedoch ausgeschlossen. Zunächst bestehen hohe Datenschutzanforderungen, so dass das Übertragen der Daten auf einen Rechner außerhalb des Bezirksamtes ausgeschlossen werden muss. Die Nutzung eines eigenen Shiny-Servers im internen Netzwerk wäre mit hohen Formalitäten und Bereitstellung entsprechender Hard- und Software verbunden.

Aus diesen Gründen wird der ESU explorer als Programmordner an die Mitarbeiter verteilt. Auf der höchsten Ebene des Programmordners findet sich eine README.txt-Datei mit hilfreichen Informationen zur Installation und zum Start der Anwendung. Für den

Programmstart wird ein R-Skript namens `ESUexplorer.R` bereitgestellt. Hier finden sich nur wenige Zeilen Programmcode, um dem Endanwender den Start besonders einfach zu ermöglichen. Die zukünftigen Anwender erhalten Unterstützung bei der Erstinstallation und eine persönliche Einweisung in die Arbeit mit dem `ESUexplorer`.

4.4.4 Qualitätssicherung

Maßnahmen zur Qualitätssicherung sollen sicherstellen, dass die Anforderungen an die Anwendung erfüllt werden. Gängige Kriterien für Softwarequalität finden sich in der Norm (ISO/IEC 9126, 2001). Als Qualitätsmerkmale gelten demnach Funktionalität, Zuverlässigkeit, Benutzbarkeit, Effizienz, Änderbarkeit und Übertragbarkeit.

Funktionalität: Die Anwendung erfüllt die im Vorfeld bestimmten Funktionalitäten zur explorativen Datenanalyse der spezifizierten Datenbasis. Die in Kapitel 3 ausgewählten statistischen Methoden und Darstellungsformen wurden implementiert. Die Korrektheit der Funktionalitäten und Analyseergebnisse wurde während des Entwicklungsprozesses laufend kontrolliert. Dafür wurden die Programmfunktionen der einzelnen Schritte des Programmablaufs individuell und in verschiedenen Kombinationen getestet. Dabei wurde die Richtigkeit der Berechnungen sowie die Benutzerfreundlichkeit geprüft.

Zuverlässigkeit: Unter Zuverlässigkeit versteht man die Reife und Fehlertoleranz einer Anwendung. Bezüglich dieser Aspekte wurden diverse Prüfungen im Programmcode implementiert, um mögliches Fehlverhalten zu vermeiden. So werden die Berechnung und Ausgabe von Analyseergebnissen erst dann ausgelöst, wenn alle notwendigen Angaben vorliegen. Bei diversen zwingend notwendigen Auswahlfeldern wird automatisch sichergestellt, dass eine sinnvolle Auswahl vorliegt. Darüber hinaus wird bei der Anwendung statistischer Methoden das festgelegte Skalenniveau der ausgewählten Merkmale berücksichtigt.

Benutzbarkeit: Ziel bei der Benutzbarkeit ist eine möglichst einfache, verständliche und intuitive Bedienung der Anwendung. In Hinblick auf die Benutzbarkeit wurden umfangreiche Maßnahmen getroffen. Das Frontend wurde insbesondere in Hinblick auf die Benutzerfreundlichkeit gestaltet (siehe Kapitel 4.2). Ein Test einer ersten Programmversion durch Mitarbeiter des Bildungsmonitoring Berlin-Mitte erbrachte wertvolle Hinweise darüber, an welchen Stellen Unklarheiten bezüglich der Bedienung des `ESUexplorer` herrschten. Aufgrund dieses Feedbacks wurde das Programm optimiert. So werden dem Benutzer optisch hervorgehobene Hinweise angezeigt, beispielsweise zu den für eine Analyse notwendigen Angaben oder auch wenn aus anderen Gründen keine Ergebnisse dargestellt werden. Bezüglich der Erlernbarkeit der Bedienung wurde zusätzlich ein Hilfebereich integriert, in dem sich nützliche Informationen für den Erstanwender finden.

Effizienz: Die Effizienz einer Anwendung ergibt sich aus dem Zeit- und Verbrauchsverhalten. Die Analyseberechnungen benötigen beim ESU explorer nur wenige Sekunden, so dass der Anwender ohne lange Wartezeiten die Ergebnisse betrachten kann. Eine Ausnahme bildet die Kerndichteschätzung für die Darstellung von Karten nach dem Kernelheaping-Verfahren. Die Parameter zur Dichteschätzung wurden so gewählt, dass eine Balance zwischen Qualität und Berechnungszeit besteht (Rastergröße 300 Pixel, 5 Burnin-Iterationen + 7 weitere Iterationen) (siehe Kapitel 3.6.2). Die Dichteschätzungen benötigen somit etwa eine Minute Rechenzeit. Auf die Möglichkeit der Grenzkorrektur wurde aus Gründen der Effizienz verzichtet, da die Dichteschätzung mit dieser Option mindestens dreimal so viel Zeit in Anspruch nimmt und sich die Ergebnisse rein optisch kaum voneinander unterscheiden. Dem Anwender wird der Berechnungsstand mit einer Fortschrittsanzeige visualisiert.

Änderbarkeit/Wartbarkeit: Mögliche Änderungen betreffen einerseits die Datengrundlage und andererseits die Funktionalitäten der Anwendung. Bezüglich der Datengrundlage wird ein eigenes R-Skript bereitgestellt, mit dem bei möglichen Änderungen oder Erweiterungen des ESU-Datensatzes dieser aus dem vorliegenden SPSS-Format direkt in ein aufbereitetes R-Datenformat überführt und somit einfach ersetzt werden kann. Die Metadaten zu den einzelnen Merkmalen liegen im Programmordner als Excel-Datei vor und werden in dieser Form in die Anwendung eingelesen. Somit wird es besonders einfach möglich sein, Änderungen an den Metadaten vorzunehmen, indem diese Excel-Datei editiert wird. Hinsichtlich von Änderungen oder Wartungen der Programmfunktionalitäten wurden verschiedene Richtlinien beim Programmierstil angewandt (siehe Kapitel 4.4.2).


Übertragbarkeit: Die Übertragbarkeit beinhaltet die Installierbarkeit und die Kompatibilität zu verschiedenen Betriebssystemen. Die Möglichkeiten der Bereitstellung bzw. Installation der shiny-Anwendung wurden im Kapitel 4.4.3 beleuchtet. Der ESU explorer wird als Programmordner direkt auf den Rechner der Anwender kopiert. Durch das Programm wird automatisch sichergestellt, dass alle notwendigen R-Packages auf dem Anwenderrechner installiert werden. Der ESU explorer wurde problemlos auf Betriebssystemen der zukünftigen Nutzer (Windows und Mac OS) unter Verwendung der gängigen Internetbrowser getestet.

Die Berücksichtigung dieser Qualitätsmerkmale soll eine möglichst zuverlässige, fehlerfreie und benutzerfreundliche Programmnutzung sicherstellen. Dadurch sollen die gestellten Anforderungen und die Bedürfnisse und Erwartungen der Nutzer möglichst gut erfüllt werden. Im folgenden Kapitel werden die grafischen Benutzeroberflächen des ESU explorers und deren Funktionsumfang aus der Perspektive der Nutzer dargestellt.

5 Der ESU explorer - Vorstellung der Anwendung

Dieses Kapitel präsentiert als Ergebnis den ESU explorer aus Anwendersicht. Ziel ist die Vorstellung der einzelnen Oberflächen und Funktionalitäten. Dafür werden die Einstellungsmöglichkeiten für die Durchführung von Analysen sowie die verschiedenen Analyseergebnisse beschrieben. Daneben sollen die zusätzlichen Inhalte des ESU explorer gezeigt werden. Über die Analysemöglichkeiten hinaus finden sich hier Metadaten und weitere Informationen zur Unterstützung der Anwender.

5.1 Einstieg und Navigation

Die Anwendung startet im Internetbrowser mit einer Einstiegsseite. Auf dieser und allen Unterseiten werden die Kopfzeile sowie das Seitenmenü angezeigt. Der Titel der Anwendung **ESU explorer** ist links in der Kopfzeile aufgeführt. Die Navigation ist innerhalb der Anwendung immer über das Seitenmenü möglich. Im Seitenmenü ist die jeweils aktive Seite farblich gekennzeichnet. Die Anwendung kann über das Symbol  beendet werden. Auf der Einstiegsseite sind zur einfachen Navigation die verschiedenen Unterseiten optisch ansprechend aufgeführt und verlinkt. Jede Unterseite ist zudem mit einem passenden Symbol verknüpft - dieses findet sich sowohl auf der Einstiegsseite als auch im Seitenmenü. Die Unterseiten sind dabei unterteilt in den Bereich *Analysen* und den Bereich *Metadaten*. Die jeweils aktive Unterseite ist im Seitenmenü optisch hervorgehoben. Die Einstiegsseite ist in Abbildung 5.1 dargestellt.

5.2 Analysen

Die statistischen Analysen können auf verschiedenen Unterseiten durchgeführt werden. Dabei wird zunächst unterteilt in *Zeitreihen Analyse*, *Jährliche Analyse* und *Individuelle Analyse*. Die Analyseseite *Karten* hat eine spezielle Aufgabe und wird erst im folgenden Abschnitt vorgestellt.

Für viele Merkmale ist der Zeitverlauf besonders interessant. Oft stehen auch die Ergebnisse des letzten oder eines speziellen Schuljahres im Fokus. Damit stellen *Zeitreihen* und

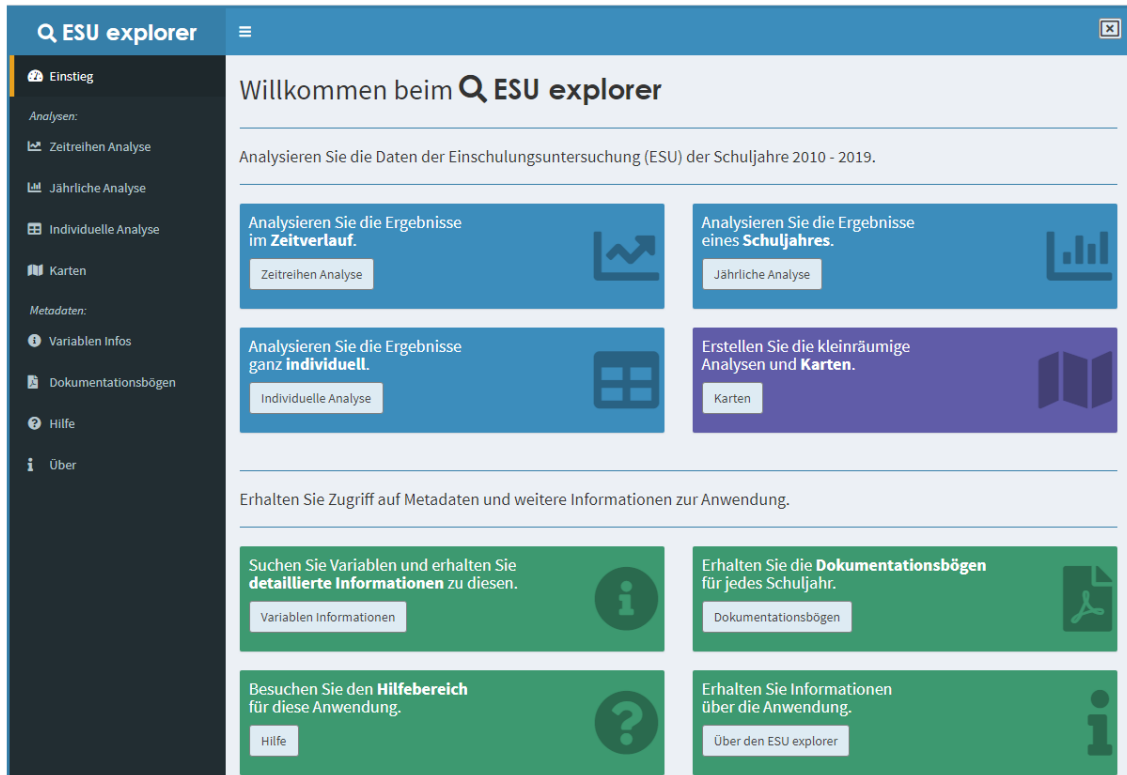


Abbildung 5.1: ESU explorer: Einstiegsseite

Eine übersichtliche Startseite soll einen einfachen Einstieg in die Anwendung gewährleisten. Die Kopfzeile sowie das Seitenmenü sind auf allen Unterseiten identisch.

Jährliche Analysen die häufigsten Anwendungsfälle dar. Über *Individuelle Analysen* werden die Möglichkeiten auf weitere Analysefälle erweitert. Diese Analysearten sind ähnlich aufgebaut, aber je nach Ziel werden bestimmte Einstellungen automatisch voreingestellt:

Zeitreihen Analyse: Hier lässt sich die zeitliche Entwicklung der Analysemerkmale betrachten. Als Spaltenvariable ist hier automatisch das Schuljahr festgelegt.

Jährliche Analyse: Diese Seite ist speziell für die Analyse der Daten eines Schuljahres vorgesehen. Hier kann ein Schuljahr ausgewählt werden, das letzte verfügbare Schuljahr ist dabei voreingestellt.

Individuelle Analyse: Diese Seite ist ohne Beschränkung auf eine Zeitreihe oder ein spezielles Schuljahr. Hier können etwa auch Analysen über alle Schuljahre hinweg durchgeführt werden, wenn beispielsweise bei Subgruppenanalysen pro Schuljahr zu kleine Fallzahlen vorliegen.

5.2.1 Einstellungen

Auf jeder dieser Seiten ist eine Analysevariable vom Anwender auszuwählen. Für eine zweidimensionale Auswertung ist die Angabe einer Spaltenvariable notwendig. Optional kann die Auswertung nach einer dritten Variablen unterteilt werden. Dabei stehen verschiede-

dene Kontrollelemente zur Festlegung der Analysevariablen und übergreifender Datenfilter bereit. Für jede der ausgewählten Variablen kann über eine Checkbox angegeben werden, ob fehlende Werte aus der Analyse ausgeschlossen werden sollen. Als globaler Filter kann ausgewählt werden, ob die Auswertung nur auf Kinder mit Wohnort in Berlin-Mitte beschränkt werden soll (aus verschiedenen Gründen werden auch wenige Kinder in Berlin-Mitte untersucht, die aber nicht ihren Wohnsitz dort haben.) Als weiterer globaler Filter kann das Schuljahr der Untersuchung ausgewählt werden. Diese Option wird nicht unter *Zeitreihen Analyse* angeboten. Bei *Jährliche Analyse* ist hier aus den zehn verfügbaren Schuljahren auszuwählen. Unter *Individuellen Analyse* steht hier zusätzlich die Option „Alle Jahre“ zur Verfügung. Alle Eingaben zu den Analysevariablen und globalen Filtern sind reaktiv, das heißt, die Analyseergebnisse werden bei Änderungen direkt erstellt und laufend aktualisiert. Als Beispiel ist in Abbildung 5.2 die Seite *Individuelle Analyse* in ihren Starteinstellungen dargestellt.

Abbildung 5.2: ESU explorer: Individuelle Analyse

In dieser Form wird die Seite *Individuellen Analyse* zum Programmstart angezeigt. Unten findet sich der Hinweis an den Nutzer, dass eine Analysevariable auszuwählen ist. Ohne diese Auswahl sind im unteren Ergebnis-Bereich auch keine Ausgaben enthalten. Die Box „Erweitere Filtereinstellungen“ ist zum Start geschlossen und kann bei Bedarf mit dem Plus-Symbol geöffnet werden.

Auf den drei Analyseseiten finden sich die Buttons „Einstellungen zurücksetzen“ und „Einstellungen speichern“. Über „Einstellungen zurücksetzen“ wird die jeweilige Analyse-seite auf ihre Starteinstellungen zurückgesetzt. Dabei werden alle Eingaben des Nutzers gelöscht. Über „Einstellungen speichern“ werden sämtliche Einstellungen ausgelesen und im CSV-Format im Programmordner gespeichert. Der Anwender kann einen Namen für diese Datei angeben. Aus Gründen der Fehlertoleranz ist der Button „Einstellungen laden“

nur auf der Seite *Individuellen Analyse* zu finden. Hier können aber alle, auch über die beiden anderen Analyseseiten gespeicherten Analyseeinstellungen geladen werden. Der Nutzer wählt dafür aus einer Liste der Namen der gespeicherten Analyseeinstellungen aus. Alle durchgeführten und gespeicherten Analysen sind damit vollständig reproduzierbar. Einige interessante gespeicherte Analysen stehen den Anwendern bereits zur Verfügung.

5.2.2 Erweiterte Filtereinstellungen

Auf jeder Analyseseite wird die Box „Erweitere Filtereinstellungen“ dargestellt. Diese kann bei Bedarf über das Plus-Symbol geöffnet werden. Hier können zusätzlich für bis zu drei Variablen individuelle Filter gesetzt werden. An dieser Stelle wurde auf Reaktivität aufgrund von Fehlervermeidung und Effizienz verzichtet. Das heißt, die Anwendung reagiert zunächst nicht auf die Eingaben in dieser Box. Die Ausführung der erweiterten Datenfilterung erfolgt erst beim Klick auf den Button „Filter anwenden“. Ebenso werden diese zusätzlichen Filter erst wieder entfernt über den Button „Filter zurücksetzen“. Unten in dieser Box ist eingblendet, welche Filter zum aktuellen Zeitpunkt aktiv sind. Der Vollständigkeit halber sind hier auch die globalen Filter aufgeführt (wohnhaft in Mitte, fehlende Werte der Analysevariablen). Die Box für die erweiterten Filtereinstellungen ist mit einer Beispielauswahl in Abbildung 5.3 dargestellt.

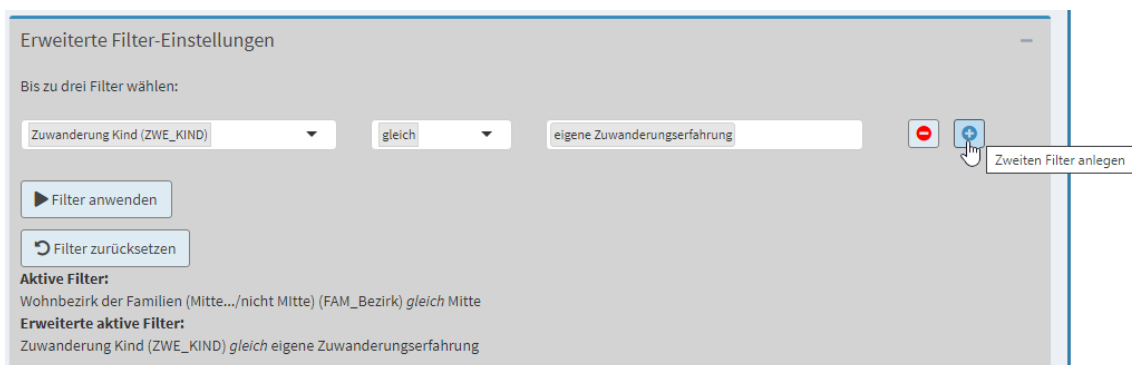


Abbildung 5.3: ESU explorer: Erweiterte Filtereinstellungen

In der Box „Erweitere Filtereinstellungen“ können bis zu drei individuelle Filter gewählt werden. Der Nutzer muss erst den Button „Filter anwenden“ betätigen, damit die ausgewählte Filterführung angewandt wird. Für den Nutzer werden am unteren Ende der Box die derzeit aktiven Filter textlich angezeigt. Enthalten sind hier auch die Filterangaben aus den vorherigen Einstellungen: Hier ist zusätzlich der globale Filter „nur Wohnort Berlin-Mitte“ gesetzt (vgl. Abbildung 5.2).

5.2.3 Ergebnisse

Der Ergebnis-Bereich enthält jeweils drei verschiedene Unterseiten, die über Reiter (Tabs) angesteuert werden können. Als Analyseergebnisse werden eine Grafik, tabellarische Auswertungen sowie entsprechende Metadaten aufgeführt.

Alle erzeugten Ergebnisse können über einen Button gemeinsam in eine Excel-Datei exportiert werden. Die Inhalte werden dabei auf einzelnen Tabellenblättern ausgegeben.

Die Grafik wird sowohl als Bildformat als auch als editierbare Vektorgrafik exportiert. Eine editierbare Vektorgrafik bietet den Vorteil, dass Inhalte, Farben, Schriftgrößen usw. einzeln angepasst werden können. Somit lassen sich die erstellten die Grafiken noch stärker individualisieren. Ein Name für die Export-Datei wird automatisch gebildet, kann aber vom Nutzer modifiziert werden.

Grafik

Die erste und somit vorausgewählte Ergebnisseite enthält eine grafische Darstellung der Analyseergebnisse. Die Grafik stellt das wichtigste Werkzeug der explorativen Datenanalyse dar und soll einen direkten und intuitiven Zugang zu den Ergebnissen sicherstellen. Die Ergebnis-Box mit einer automatisch erzeugten Grafik ist in Abbildung 5.4 dargestellt.

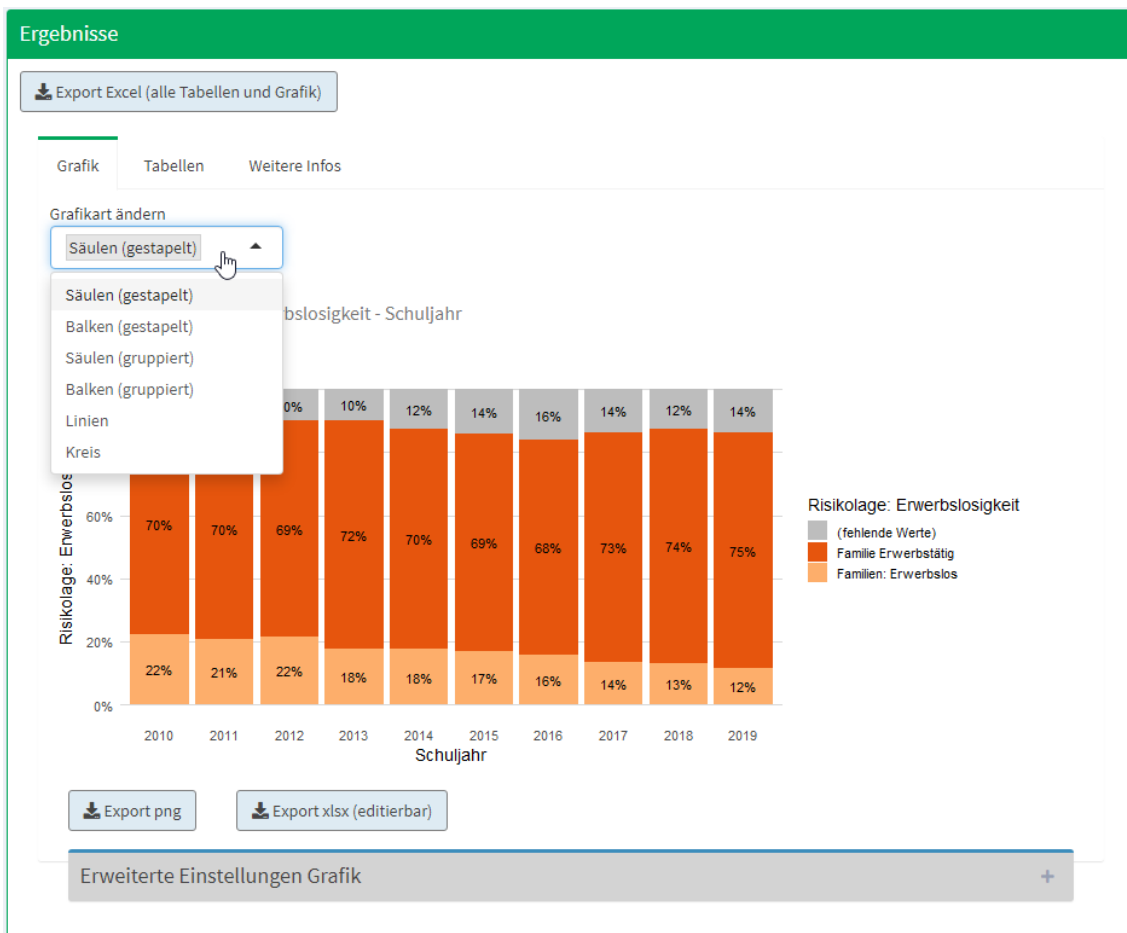


Abbildung 5.4: ESU explorer: Ergebnisse - Grafik

Dargestellt ist eine automatisch erzeugte Grafik in der Ergebnis-Box. Über die oberhalb und unterhalb der Grafik vorhandenen Buttons können die Analyseergebnisse entweder komplett oder aber nur die Grafik exportiert werden. Über die Auswahl „Grafikart ändern“ lässt sich ein Diagrammtyp wählen. Zur Auswahl stehen dabei nur die zum Skalenniveau der Analysevariablen passenden Grafikarten. Über die untere Box „Erweiterte Einstellungen Grafik“ lässt sich die Grafik noch stärker an die individuellen Bedürfnisse anpassen.

Für kategoriale Analysevariablen ist die Grafikart „Säulen (gestapelt)“ voreingestellt. Abhängig vom Skalenniveau der Analysevariablen können folgende Grafikarten angefordert

werden (siehe auch alle Abbildungen in Kapitel 3):

- Säulen (gestapelt)
- Balken (gestapelt)
- Säulen (gruppiert)
- Balken (gruppiert)
- Linien
- Kreis
- Box-Plot/Violin-Plot (nur bei metrischer Analysevariable)
- Histogramm/Dichte (nur bei metrischer Analysevariable)
- Punkte (nur wenn Analysevariable und Spaltenvariable metrisch)

Unter der Grafik ist zusätzlich eine zunächst geschlossene Box „Erweitere Einstellungen Grafik“ zu finden. Die Einstellungsmöglichkeiten sind dabei abhängig vom ausgewählten Grafiktyp. Hier lassen sich beispielsweise Werte-Beschriftungen ganz oder teilweise ausblenden und die Position der Legende, Textgrößen, oder das Seitenverhältnis der Grafik ändern. Die farbliche Gestaltung kann durch drei verschiedene Farbskalen geändert werden. Dabei werden fehlende Werte immer in grau dargestellt. Außerdem können für die kategorialen Analysevariablen Ausprägungen ausgewählt werden, welche in der Grafik nicht dargestellt werden sollen.

Beim Grafiktyp Box-Plot/Violin-Plot lassen sich die Grafikinhalt variabel wählen. Dabei können Box-Plots, Violin-Plots und Mittelwerte sowohl gemeinsam, also auch in verschiedenen Kombinationen dargestellt werden.

Beim Grafiktyp Histogramm lässt sich in den erweiterten Grafikeinstellungen die Klassenzahl variabel per Schieberegler zwischen 10 und 50 Klassen wählen. Somit ist es ausgesprochen schnell möglich, den Eindruck verschiedener Klassenzahlen visuell zu überprüfen. Zusätzlich lässt sich eine einfache Kerndichteschätzung zu diesem Histogramm ein- oder ausblenden.

Bei Punktediagrammen kann gewählt werden, ob ein linearer Trend und ob Box-Plots oder Kerndichteschätzungen als Randgrafiken angezeigt werden sollen.

Mit den Einstellungsmöglichkeiten lassen sich die Grafiken vielfältig modifizieren und an die individuellen Bedürfnisse anpassen. Die Einstellungsmöglichkeiten für die verschiedenen Grafiktypen sind in Anhang A.4, Abbildungen A.3 bis A.6 dargestellt.

Die erstellte Grafik lässt sich über Buttons auch einzeln als Bildformat (PNG) oder editierbare Vektorgrafik in Excel exportieren.

Tabellen

Die tabellarischen Auswertungen auf dem zweiten Reiter der Ergebnis-Box enthalten die statistischen Verteilungen und Maßzahlen. Hauptsächlich werden kategoriale Merkmale

analysiert, deshalb werden Häufigkeitstabellen und kontingente Zusammenhangsmaße immer ausgegeben. Verteilungsmaßzahlen und weitere Zusammenhangsmaße werden nur dann berechnet, wenn die Analysevariablen das entsprechenden Skalenniveau vorweisen. Die Ergebnistabellen enthalten entsprechend:

- Kontingenztabelle relative und absolute bedingte Häufigkeiten (vgl. Kapitel 3.2),
- Tabelle mit Verteilungsmaßzahlen (nur bei metrischer Analysevariable): 5-Zahlen-Zusammenfassung (Minimum, 1. Quartil, Median, 3. Quartil, Maximum), Mittelwert, Standardabweichung und gültige Fallzahl (vgl. Kapitel 3.3),
- Tabelle mit Zusammenhangsmaßen (abhängig vom Skalenniveau der Analysevariablen) (vgl. Kapitel 3.5).

Bei den Zusammenhangsmaßen werden immer ein Chi-Quadrat-Test und darauf basierende Zusammenhangsmaße angefordert. Das Ergebnis enthält die Werte Chi-Quadrat χ^2 , Freiheitsgrade und p -Wert sowie Cramers V ¹ und Kontingenzkoeffizienten nach Pearson C und C_{corr} . Vor der Anforderung des Tests und dieser Maßzahlen wird automatisch geprüft, ob die zugrundeliegende Kontingenztabelle entsprechende minimale Dimensionen von mindestens zwei Ausprägungen pro Merkmal aufweist. Ist eine der erwarteten Häufigkeiten kleiner 5, so wird eine entsprechende Warnmeldung ergänzt.

Abhängig von den Skalenniveaus der Analysevariablen werden weitere Zusammenhangsmaße berechnet. Bei einer Analyse zweier mindestens ordinaler Merkmale wird Rangkorrelationskoeffizient r_s von Spearman berechnet. Beim Vergleich zweier metrischer Merkmale wird zusätzlich der Korrelationskoeffizient r_{XY} von Bravais-Pearson berechnet.

Bei allen Zusammenhangsmaßen wird in einer weiteren Spalte eine Interpretationshilfe abhängig von der Höhe der Maßzahl angegeben. Die Tabelle 3.1 (Interpretation Korrelationskoeffizient nach (Cohen, 1988)) wurde für eine Einordnung zugrunde gelegt.

Es ist möglich, die Analysetabellen einzeln im Excel- oder Druckformat zu exportieren.

Weitere Informationen

Auf dem Reiter „Weitere Infos“ sind zusätzliche Informationen zur angeforderten Analyse aufgeführt. Dazu gehören

- Meta-Informationen zur Analysevariable,
- Meta-Informationen zur Spaltenvariable,
- Meta-Informationen zur Unterteilungs-Variable (nur wenn vorhanden),
- Auflistung aller gesetzten Filter.

¹da Cramers V in einer 4-Felder-Tafel Φ entspricht (siehe Kapitel 3.5.1), wurde auf die explizite Angabe von Φ verzichtet.

Bei den Variableninformationen werden der Variablen-Name, das Label, das Mess- bzw. Skalenniveau sowie weitere Informationen ausgegeben. Die weiteren Informationen können Angaben zur Erhebungsmethodik, zur Berechnung oder zur Verwendung des entsprechenden Merkmals enthalten. Diese Informationen in Verbindung mit der Angabe aller aktiven Filter sollen eine korrekte Interpretation der Analyse unterstützen.

5.3 Karten

Für die Erstellung von Karten für sozialraumorientierte und kleinräumige Analysen steht im ESU explorer eine eigene Analyseseite bereit. Hier können Choroplethenkarten und Karten nach dem Kernelheaping-Verfahren erstellt und modifiziert werden. Die Basis bildet eine Karte von Berlin-Mitte.

5.3.1 Einstellungen

Für die Erstellung einer Karte ist zunächst eine räumliche Gliederungsebene zu wählen. Der Nutzer kann hier die Darstellung für die Lebensweltlich orientierten Räume (Berlin-Mitte gesamt, Prognoseräume, Bezirksregionen und Planungsräume) sowie für die Einschulungsbereiche anfordern. Als globaler Filter ist ein Schuljahr zu wählen, das letzte verfügbare Schuljahr ist hierbei voreinstellt. Der Anwender kann aus einer Liste der kategorialen Variablen eine Analysevariable auswählen. Anschließend ist eine Ausprägung dieser Variable auszuwählen, dessen Anteilswert auf der Karte dargestellt werden soll. In Abbildung 5.5 sind die Einstellungen an einem Beispiel dargestellt.

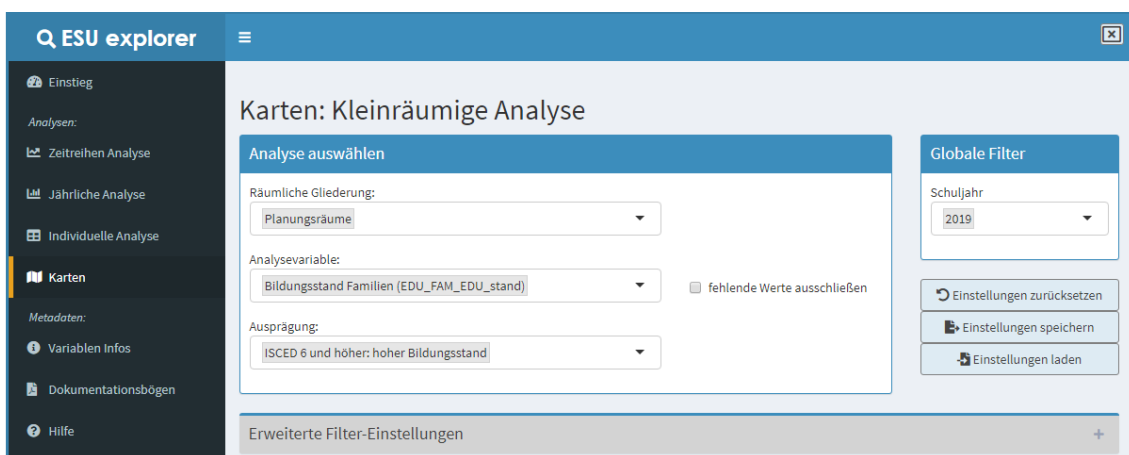


Abbildung 5.5: ESU explorer: Analyseseite Karten mit Beispieleinstellungen

Auf der Analyseseite Karten können für die verschiedenen räumlichen Gliederungen Karten von Berlin-Mitte erzeugt werden. Dafür ist eine räumliche Gliederungsebene auszuwählen sowie eine kategoriale Analysevariable und eine Ausprägung dieser Variable. Voreingestellt ist die Auswertung für das letzte verfügbare Schuljahr.

Daneben ist wie bei den zuvor vorgestellten Analyseseiten die Box „Erweiterte Filtereinstellungen“ enthalten. Hier kann die Datenbasis mit Filtern für bis zu drei Variablen

spezifiziert werden. Die Darstellung und Funktionalität dieser Filtereinstellungen sind identisch zu den anderen Analyseseiten (vgl. Kapitel 5.2.2). Ebenso finden sich die Buttons „Einstellungen zurücksetzen“, „Einstellungen speichern“ und „Einstellungen laden“ für eine einfache Anwendung und Reproduktion von Karten und Analyseergebnissen.

5.3.2 Ergebnisse

Analog zu den anderen Analyseseiten enthält die Box „Ergebnisse“ verschiedene Reiter. Als erstes Ergebnis wird automatisch eine Choroplethenkarte von Berlin-Mitte dargestellt. Auf dem zweiten Reiter können für die gewählte Analyse Karten nach dem Kernelheaping-Verfahren angefordert werden. Weiterhin sind tabellarische Auswertungen sowie Metadaten enthalten. Auch hier können per Button alle erzeugten Ergebnisse gemeinsam in eine Excel-Datei exportiert werden.

Choroplethenkarte

Auf dem ersten voreingestellten Reiter ist eine Choroplethenkarte von Berlin-Mitte dargestellt. Diese Karte ist zentral für die schnelle Darstellung und intuitive Erfassung von kleinräumigen Unterschieden. Der Grafik-Ergebnisbereich mit einer automatisch erzeugten Choroplethenkarte ist in Abbildung 5.6 dargestellt.

Unter „Erweiterte Einstellungen Grafik“ können auch hier einige Änderungen an der Darstellung durchgeführt werden. Die Einstellungsmöglichkeiten für die Kartendarstellung sind in Anhang A.4, Abbildung A.7 dargestellt. Hier lassen sich die Beschriftungen ein- oder ausblenden sowie in Format und Größe modifizieren. Die Position der Legende kann geändert und die farbliche Darstellung durch drei verschiedene Farbskalen variiert werden. Über Checkboxes lassen sich die verschiedenen Schichten der Karte spezifizieren. So können die Raumgrenzen, Schulstandorte, Gewässer, Grünflächen oder andere unbewohnte Flächen auf der Karte dargestellt werden.

Als wichtigste Einstellung kann hier die Art der Klassenbildung individuell spezifiziert werden (vgl. Kapitel 3.6.1 und Abbildung 3.11). Dafür stehen vier Möglichkeiten zur Auswahl:

Automatische Klassen: Automatische Klassenbildung mit gleichlangen durch 1, 2, 5 oder 10 teilbaren Intervallen; ca. 5 Intervalle je nach Datenlage.

Gleiche Intervalle: Klassenbildung mit gleichen Intervalllängen, die Länge der Intervalle kann mittels Schieberegler festgelegt werden, die Anzahl der Intervalle ergibt sich dann aus der Spannweite der Analysedaten und der gewählten Intervalllänge.

Quantile: die darzustellenden Analysedaten werden durch Quantile dargestellt, die Anzahl der Quantile kann mittels Schieberegler zwischen 2 und 10 variiert werden.

Kontinuierliche Skala: die Daten werden nicht in Klassen unterteilt, die Farbgebung erfolgt mittels kontinuierlichem Farbverlauf.

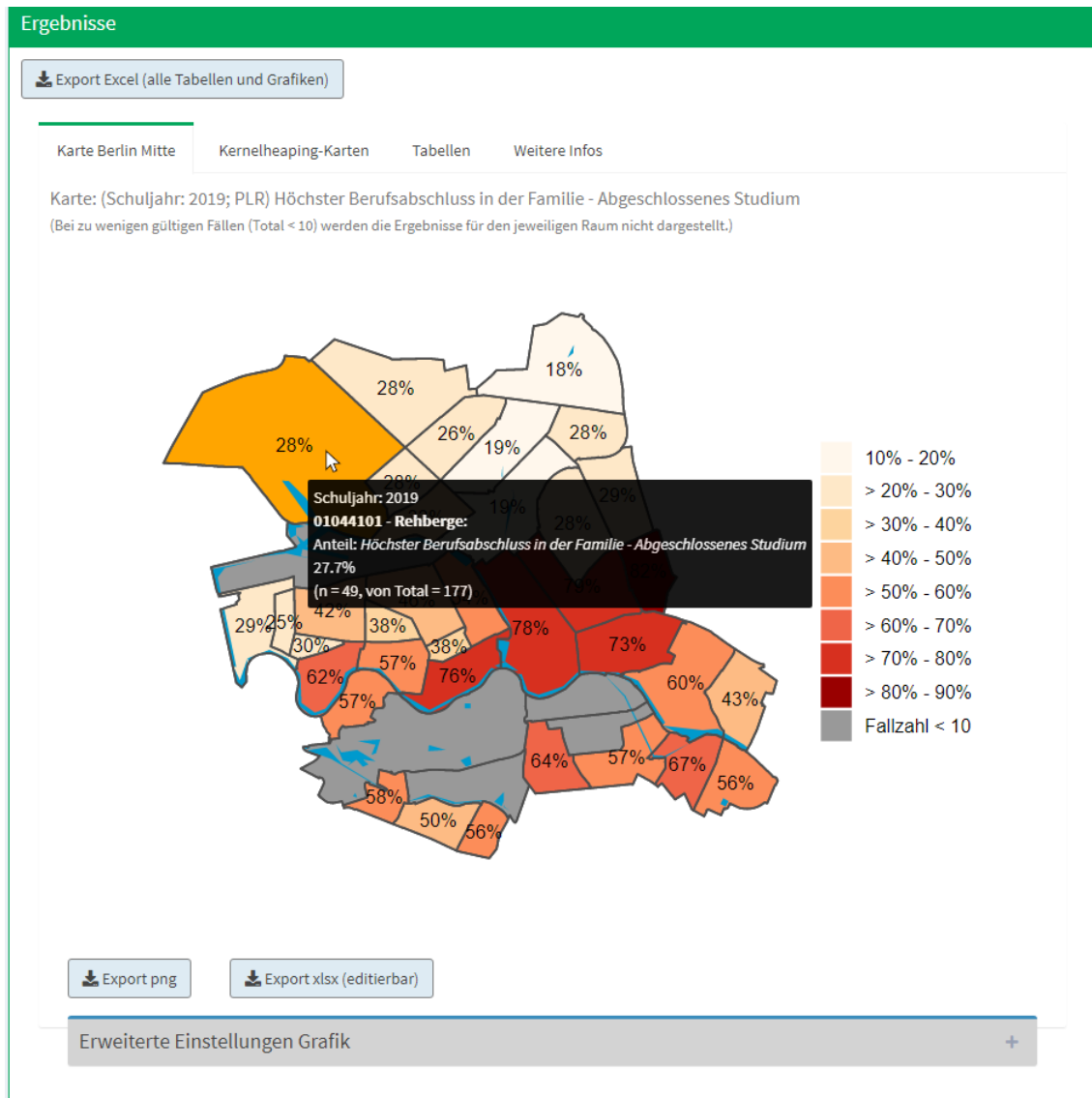


Abbildung 5.6: ESU explorer: Ergebnisse - Karte

Dargestellt ist eine automatisch erzeugte Choroplethenkarte im Ergebnisbereich auf der Analyseseite „Karten“. Die Karten werden interaktiv dargestellt. Dabei werden beim Überfahren eines Raumes mit dem Mauszeiger zusätzliche Informationen angezeigt. Die Darstellung und insbesondere die Methode der Klassenbildung kann über die Box „Erweiterte Einstellungen Grafik“ modifiziert werden.

Der schnelle Wechsel zwischen unterschiedlichen Arten der Klassenbildung erlaubt es, Alternativen zu betrachten und in ihrer Wirkung zu überprüfen, um das bestmögliche Ergebnis zu erhalten.

Die Darstellung der Choroplethenkarte ist zudem interaktiv. Damit werden beim Überfahren eines Raumes mit dem Mauszeiger alle für diesen Raum relevanten Informationen in einer Infobox angezeigt. Dazu gehören die Angaben zum Schuljahr, Angaben zum Raum (Nummer, wenn vorhanden Name), Analysevariable und Ausprägung, Anteil, Fallzahl und Total-Fallzahl. Werden die Schulen auf der Karte dargestellt, so wird beim Zeigen mit der Maus auf einen Schulstandort der Name der entsprechenden Schule angezeigt.

Räume, die auf einer Fallzahl < 10 basieren, werden auf diesen Choroplethenkarten ausgegraut und ohne Wertebeschriftung dargestellt. Anteile, die auf einer solch kleinen Fallzahl beruhen, können zu hohe Schwankungen aufweisen, so dass auf eine Interpretation verzichtet werden sollte. Die tatsächlichen Anteile und Fallzahlen sind trotzdem mit dem Mauszeiger oder auch über die Häufigkeitstabelle zugänglich.

Kernelheaping-Karten

Auf einem weiteren Reiter kann die Erstellung von Kerndichtekarten nach dem Kernelheaping-Verfahren angefordert werden (vgl. Kapitel 3.6.2). Um möglichst gute Schätzergebnisse zu erhalten, ist diese Option auf Analysen auf Basis der PLR und ESB beschränkt. Da die Kerndichteschätzung eine erhöhte Rechenzeit erfordert, wird diese erst durch einen Klick auf den Button „Kerndichtekarten anfordern“ ausgelöst. Während der Berechnung wird dem Nutzer der Berechnungsfortschritt mit einer fortschreitenden Ladeanzeige angezeigt. Als Ergebnisse werden eine Kernelheaping-Karte mit einer kontinuierlichen Dichteschätzung und eine daraus abgeleitete Hotspot-Karte dargestellt. Die Hotspot-Karte zeigt die 10% der Fläche von Berlin-Mitte mit den höchsten Anteilswerten der Dichteschätzung (vgl. Abbildung 3.12). Unter „Erweiterte Grafikeinstellungen“ können die Position der Legende, die Farbpalette und die verschiedenen Kartenschichten wie Raumgrenzen, Schulen oder Gewässer geändert werden (vgl. Anhang A.4, Abbildung A.8).

Tabellen

Der Reiter „Tabellen“ enthält Häufigkeitstabellen für die angeforderte Analyse. Dargestellt werden die relativen und absoluten bedingten Häufigkeiten für die Analysevariable nach der räumlichen Gliederungsebene. Der Vollständigkeit und Übersichtlichkeit halber sind hier die Häufigkeiten für alle Ausprägungen der Analysevariable aufgeführt. Die Darstellung auf den Karten basiert jedoch nur auf den Häufigkeiten der ausgewählten Ausprägung.

Weitere Informationen

Unter dem Reiter „Weitere Infos“ sind wie bei den anderen Analyseseiten Informationen zur erstellten Analyse aufgeführt. Es werden die Meta-Informationen zur Analysevariable sowie zur räumlichen Gliederungsebene ausgegeben. Ebenso ist eine tabellarische Übersicht aller aktiven Filter enthalten.

5.4 Metadaten

Der Bereich *Metadaten* im ESU explorer enthält verschiedene Unterseiten mit weiterführenden Informationen. Hier kann der Nutzer auf Informationen zu allen Variablen, die Dokumentationsbögen, einen Hilfebereich und allgemeine Angaben zum ESU explorer zugreifen.

5.4.1 Variablen Informationen

Auf dieser Seite ist die Suche nach und der Zugriff auf die Metadaten der einzelnen Variablen möglich. Übergeordnete Informationen zu allen Variablen werden tabellarisch aufgeführt. Dabei sind zunächst nur eine laufende Nummer, Variablenname und Label sichtbar. Am Beginn der Zeile ist jeweils ein grünes Plus-Symbol enthalten²: beim Klick auf dieses Zeichen öffnet sich ein weiterer Bereich, in dem weitere Informationen zur Variablen aufgeführt sind. Hier finden sich Angaben zur Herkunft oder Berechnung, Fragebogennummern pro Schuljahr, Ausprägungen, Messniveau sowie Zusatzinformationen zur Methodik oder Verwendung. Mit dem roten Minus-Symbol lässt sich dieser Bereich wieder schließen.

Die Tabelle bietet vielfältige Elemente zur Erhöhung der Benutzerfreundlichkeit. Am wichtigsten ist dabei eine Suchfunktion, um schnell an bestimmte Inhalte zu gelangen. Die Tabelle wird unterteilt mit einer übersichtlichen Anzahl Zeilen dargestellt. Auch lassen sich die angezeigten Spalten auf- oder absteigend sortieren. Die Tabelle ist in Abbildung 5.7 dargestellt.

Nr	Variable	Label
1	Gesamt	Gesamt
2	A2q_Schuljahr	Schuljahr
3	EDU_FAM_Beruf	Höchster Berufsabschluss in der Familie
<p>Herkunft: berechnet: EDU_Mutter_Beruf, EDU_Vater_Beruf Ausprägungen: kein Beruf in Ausbildung / Studium Abgeschlossene Berufsausbildung Abgeschlossenes Studium (fehlende Werte) Messniveau: ordinal Zusatzinformationen: Höchster Berufsabschluss von Mutter und Vater (der höchste Abschluss wird übernommen). Angabe dient zur Einschätzung der sozialen Statusgruppe, der die Familie des Kindes zugeordnet wird.</p>		
4	EDU_FAM_EDU_stand	Bildungsstand Familien
5	EDU_FAM_ISCED	ISCED der Familie
6	EDU_FAM_Schule	Höchster Schulabschluss in der Familie
7	EDU_Mutter_Beruf	Berufliche Ausbildung Mutter
8	EDU_Mutter_EDU_Stand	Bildungsstand Mutter
9	EDU_Mutter_ISCED	ISCED der Mutter
10	EDU_Mutter_Schule	Schulabschluss Mutter

Abbildung 5.7: ESU explorer: Metadaten - Variableninformationen

Die weiterführenden Informationen und Metadaten der einzelnen Variablen sind in einer übersichtlichen und benutzerfreundlichen Tabelle dargestellt. Hier kann der Nutzer besonders einfach nach Inhalten suchen und auf alle Informationen zugreifen.

²Aussehen und Funktionalität der Tabelle wurde der Beispieltabelle auf der Startseite von (SpryMedia, 2020) nachempfunden.

5.4.2 Dokumentationsbögen

Auf dieser Seite ist der Download aller Dokumentationsbögen der Schuljahre 2010 - 2019 möglich. Der Zugriff auf die genauen Frage- und Antwortgestaltungen soll eine korrekte Interpretation sicherstellen. Die Dokumentationsbögen werden jährlich angepasst, so dass für alle Jahre die entsprechende Version enthalten ist. In den Metadaten zu den einzelnen Variablen ist unter anderem auch der Verweis auf die Fragennummer pro Schuljahr enthalten, um eine fehlerfreie Zuordnung sicherzustellen. Die Dokumentationsbögen stehen dem Nutzer als PDF-Datei zur Verfügung.

5.4.3 Hilfebereich

Der Hilfebereich soll insbesondere Erstanwender dabei unterstützen, den ESU explorer effizient zu nutzen. Zu verschiedenen Inhalten und Nutzungsmöglichkeiten sind hier Informationen aufgeführt. Hier wurden Fragen und Hinweise berücksichtigt, welche nach einem Test einer ersten Programmversion durch Mitarbeitende des Bildungsmonitoring Berlin-Mitte erbracht wurden.

Dabei sollte eine Ausgewogenheit bezüglich der Länge dieses Hilfebereichs und dessen Übersichtlichkeit erreicht werden. So sind hier Hinweise zur grundsätzlichen Nutzung der Anwendung enthalten, nicht aber theoretische Hintergründe zu den verwendeten statistischen Methoden. Ein Teil des Hilfebereichs ist in Abbildung 5.8 dargestellt.



Abbildung 5.8: ESU explorer: Hilfebereich

Der Hilfebereich bietet grundsätzliche Informationen zu den Inhalten und zur Nutzung des ESU explorers. Dargestellt ist ein Teil des Hilfebereichs. In der Inhaltsliste oben sind die einzelnen Themen verlinkt.

5.4.4 Über den ESU explorer

Auf dieser Seite sind Name und E-Mail-Adresse der Autorin enthalten sowie ein Hinweis auf die vorliegende Arbeit und den Nutzungsbereich. Diese Angaben stellen damit eine Art Impressum dar und runden den Inhalt des ESU explorer ab.

6 Fazit

Ziel dieser Arbeit war es, eine Anwendung zur interaktiven explorativen Analyse der Daten der Einschulungsuntersuchungen in Berlin-Mitte und daraus abgeleiteter Indikatoren zu entwickeln. Damit sollen die Mitarbeiter und Projektpartner des Bildungsmonitoring Berlin-Mitte zur eigenständigen explorativen Analyse und Visualisierung der Datenlage befähigt werden. Dafür wurde die Analyseanwendung ESU explorer konzipiert und entwickelt.

6.1 Zusammenfassung

Die Ausgangslage bildeten die Ziele des Bildungsmonitoring Berlin-Mitte in Bezug auf eine analytische Bildungsberichterstattung zur Förderung von Chancengleichheit im Bildungswesen. Die Einschulungsuntersuchungen bilden dafür eine solide Basis zur Ableitung von Indikatoren für eine Analyse der Situation der Kinder und ihrer Familien vor Schuleintritt. Mit den vielfältigen Informationen zum Wohnort, zur sozialen sowie zur gesundheitlichen Lage der Kinder lassen sich auch sozialraumorientierte Analysen durchführen. So soll eine Betrachtung der heterogenen Sozialmilieus in Berlin-Mitte lokal angepasste Handlungsempfehlungen ermöglichen. Unter Berücksichtigung der Ziele des Bildungsmonitoring und der Wünsche der zukünftigen Nutzer wurden die Anforderungen an die zu entwickelnde Analyseanwendung konkretisiert.

Die Planung und Entwicklung der Analyseanwendung wurden methodisch aus zwei Blickwinkeln beleuchtet. Die erste Sicht bildete die Auswahl und Vorstellung statistischer Methoden, welche mit dieser Analyseanwendung durchgeführt werden sollen. Im Fokus stehen hier klassische Methoden zur Beschreibung von Verteilungen und Zusammenhängen sowie passende visuelle Darstellungsformen. Dabei eignen sich für die explorative Analyse Säulen-, Balken-, Linien- und Kreisdiagramme für kategoriale sowie Box-Plots, Violin-Plots, Histogramme oder Streudiagramme für metrische Merkmale. Eine kleinräumige Analyse wird mittels Choroplethenkarten visualisiert. Dabei nimmt die Wahl der Klassenbildung eine bedeutende Rolle für eine optimale Darstellung ein. Eine Darstellung von Kerndichtekarten nach dem Kernelheaping-Verfahren zeigt dagegen eine stetige und somit realistischere Darstellung von Verteilungen über räumliche Einheiten.

Den zweiten methodischen Blickwinkel bildeten die programmiertechnischen Aspekte der Anwendungsentwicklung. Die technologische Basis für die Analyseanwendung setzt sich

aus der Programmierumgebung R sowie dem R-Package shiny zur Erstellung von interaktiven Webanwendungen zusammen. Darüber hinaus tat sich die tidyverse-Familie als solide Grundlage für die Umsetzung der Programmentwicklung hervor. Die tidyverse-R-Packages sowie die tidyverse-Philosophie bieten einen durchdachten Rahmen für die Durchführung explorativer Datenanalysen mit R. Der federführende Autor des tidyverse und Verfasser vieler Schriften zum Thema ist Hadley Wickham. Seine Werke stellten sich als exzellente Quellen für die vorliegende Arbeit heraus. Mithilfe von shiny, tidyverse und weiterer R-Packages konnten die gewünschten Funktionalitäten, Methoden und Inhalte durch verschiedene Programmmodule realisiert werden. Bei der Gestaltung der Analyseanwendung wurde dabei insbesondere eine hohe Benutzerfreundlichkeit fokussiert.

Als Ergebnis der methodischen Überlegungen und Schritte wurde die Analyseanwendung namens ESU explorer aus Anwendersicht vorgestellt. Der ESU explorer bietet verschiedene Inhalte und Ansätze zur Analyse der ESU- und darauf aufbauender Daten. Der ESU explorer ermöglicht dabei etwa die Betrachtung von zeitlichen Entwicklungen der letzten 10 Jahre, eines speziellen Schuljahres oder übergreifender Ergebnisse. Ein spezieller Bereich des ESU explorers ist auf die Erstellung von thematischen Karten von Berlin-Mitte ausgerichtet. Auf klassischen Chroplethenkarten lassen sich relative Häufigkeiten von Ausprägungen kategorialer Analysevariablen darstellen. Das Problem der Klassenwahl wurde berücksichtigt. Der Anwender kann hier auf einfachem Wege aus verschiedenen Methoden der Klassenbildung auswählen und die Klassenzahl spezifizieren. Zusätzlich können Kerndichtekarten nach dem Kernelheaping-Verfahren angefordert werden.

Die mit dem ESU explorer erzeugten Grafiken können zur schnellen und wiederholten explorativen Datenanalyse einen wertvollen Beitrag leisten. Darüber hinaus lassen sich die verschiedenen Grafiktypen mittels spezifischer Einstellungsmöglichkeiten vielfältig individualisieren. Die Analyseergebnisse enthalten neben der grafischen Darstellung auch Tabellen mit Häufigkeitsverteilungen, Verteilungsmaßzahlen, Zusammenhangsmaßen und Metadaten. Alle Analyseergebnisse können in eine einzige Excel-Datei exportiert werden. Dabei werden die Grafiken auch als editierbare Vektorgrafiken exportiert. Für eine Verwendung zu Kommunikationszwecken können die Grafiken somit noch individueller angepasst und optimiert werden.

Für alle Analysen lässt sich die Datenbasis durch individuelle Filter spezifizieren. Dadurch sind auch weiterführende Subgruppenanalysen einfach durchführbar. Durch Funktionalitäten zum Speichern und Laden von Analyseinstellungen können alle mit dem ESU explorer durchgeführten Analysen vollständig reproduziert werden. Die Einbettung verschiedener Metadaten zu den Analysevariablen und ein Hilfebereich für Erstanwender runden den Inhalt des ESU explorer ab.

Der ESU explorer ermöglicht es den Mitarbeitern des Bezirksamts Mitte, schnell und interaktiv Zugriff auf detaillierte Ergebnisse der ESU zu erhalten. Damit wird den Anwendern die explorative Datenanalyse auf benutzerfreundliche Weise ermöglicht. So können lo-

kal spezifische Handlungsempfehlungen und Entscheidungen abgeleitet und untermauert werden.

6.2 Ausblick

Der ESU explorer wurde speziell für die Analyse der ESU-Daten für Berlin-Mitte konzipiert. Dadurch bestehen wenige Anforderungen an die Datenbasis. Die Variablen für das Schuljahr und die verschiedenen räumlichen Gliederungsebenen müssen im Datensatz vorhanden sein. Darüber hinaus ist der ESU explorer vollkommen flexibel im Hinblick auf die verwendeten Daten. Das heißt, dass zukünftig beliebig viele Variablen ergänzt oder geändert werden können, ohne dass die Funktionalität des ESU explorers beeinträchtigt wird.

Der ESU explorer wird als Programmordner mit den notwendigen R-Skripten und Daten an die Anwender verteilt. Ein Nachteil dabei ist, dass Programm und Daten nicht zentral geändert werden können. Sofern das Bezirksamt Mitte die Voraussetzungen in der IT-Infrastruktur für das interne Hosting über einen shiny-Server schafft, ließen sich ein zentraler Zugriff und eine zentrale Daten- und Programmpflege realisieren. Auch eine Nutzung durch externe Interessenten wäre damit möglich. Denkbar wäre hier eine Erweiterung des ESU explorers durch die Nutzung einer geschützten Datenbank und die Implementierung einer regelbasierten Datenschutzroutine, welche die Anforderungen an die Geheimhaltung von Einzeldaten sicherstellt.

A Anhang

A.1 Dokumentationsbogen für die Einschulungsuntersuchungen der KJGD im Land Berlin und Dokumentationsbogen für die S-ENS und SOPESS-Untertests (Schuljahr 2019)

Name, Vorname des Kindes:

Geb.-Datum:

Datum:

Dokumentationsbogen für die Einschulungsuntersuchungen der KJGD im Land Berlin

KJGD-Stelle:

Schuljahr: **2019**

Laufende Nummer

--	--	--	--	--

1. Allgemeine und soziale Anamnese1. Planungsraum LOR

--	--	--	--	--	--	--	--

2. Untersuchungsmonat und -jahr

--	--	--	--	--	--

3. Nummer der Schule

--	--	--	--	--	--

(Achtung! Schulnr. von SenBJF)4. Nummer Untersucher(in)

--	--

5. Anmeldung zur Untersuchung
☐ 1 Schulpflichtig (bis 30.09.2013)
☐ 2 Antragsweise (von 1.10.2013 bis 31.03.2014)
☐ 3 Nach Zurückstellung im Vorjahr6. Geburtsmonat und -jahr

--	--	--	--	--	--

7. Geschlecht ☐ 1 männlich ☐ 2 weiblich8. Kind ist in Deutschland geboren
☐ 1 ja
☐ 0 nein, dann bitte letzte Zuwanderung nach D
Zeitpunkt (Monat/Jahr)

--	--	--	--	--	--

☒ 99 keine Angabe9. Geburtsland der Mutter
☐ 1 Deutschland

--

☒ 99 keine Angabe10. Staatsangehörigkeit der Mutter
erste

--

weitere

--

11. Geburtsland des Vaters
☐ 1 Deutschland

--

☒ 99 keine Angabe12. Staatsangehörigkeit des Vaters
erste

--

weitere

--

13. Familiensprache(n)
1.

--

2.

--

3.

--

14. Kita-/Einrichtungsbesuch
seit (Monat/Jahr)

--	--	--	--	--	--

☐ 0 Kind hat keine Kita/Einrichtung besucht
☒ 99 keine Angabe
15. Kind lebt überwiegend bei
☐ 1 den Eltern
☐ 2 allein erziehendem Elternteil
☐ 3 anderswo
☒ 99 keine Angabe

16. Schulabschluss (ggf. den höchsten angeben)

Mutter	Vater
<input type="radio"/> 0	<input type="radio"/> 0 ohne Abschluss
<input type="radio"/> 1	<input type="radio"/> 1 Hauptschulabschluss
<input type="radio"/> 2	<input type="radio"/> 2 mittlere Reife / MSA / 10. Klasse
<input type="radio"/> 3	<input type="radio"/> 3 Abitur/Fachabitur
<input checked="" type="radio"/> 99	<input checked="" type="radio"/> 99 keine Angabe

17. Berufliche Ausbildung (ggf. die höchste angegeb.)

Mutter	Vater
<input type="radio"/> 0	<input type="radio"/> 0 ohne bzw. ohne abgeschlossene Berufsausbildung
<input type="radio"/> 1	<input type="radio"/> 1 in Ausbildung/Studium
<input type="radio"/> 2	<input type="radio"/> 2 abgeschlossene Berufsausbildung/Fachschulabschluss
<input type="radio"/> 3	<input type="radio"/> 3 abgeschlossenes Studium (Uni, Fachhochschule)
<input checked="" type="radio"/> 99	<input checked="" type="radio"/> 99 keine Angabe

18. Erwerbstätigkeit der Eltern

Mutter	Vater
<input type="radio"/> 0	<input type="radio"/> 0 nicht erwerbstätig, weil
<input type="radio"/> 1	<input type="radio"/> 1 finde keine Arbeit
<input type="radio"/> 2	<input type="radio"/> 2 habe andere Gründe
<input type="radio"/> 3	<input type="radio"/> 3 teilzeitbeschäftigt
<input checked="" type="radio"/> 99	<input checked="" type="radio"/> 99 vollzeitbeschäftigt
	<input checked="" type="radio"/> 99 keine Angabe

19. Anzahl aller im Haushalt lebenden Personen

Erwachsene (älter 18 J.)	<table border="1"><tr><td></td><td></td></tr></table>		
Kinder (bis 18 J.)	<table border="1"><tr><td></td><td></td></tr></table>		
keine Angabe	<input checked="" type="radio"/> 99		

20. Anzahl der Raucher im Haushalt

keine Angabe	<table border="1"><tr><td></td><td></td></tr></table>		
	<input checked="" type="radio"/> 99		

Dokumentationsbogen für die Einschulungsuntersuchungen der KJGD im Land Berlin																																																																
KJGD-Stelle:	Schuljahr: 2019																																																															
<div style="display: flex; justify-content: space-between; align-items: center;"> <div style="border: 1px solid black; padding: 2px;"> Laufende Nummer <div style="display: flex; gap: 5px;"> <div style="border: 1px solid black; width: 20px; height: 20px;"></div> <div style="border: 1px solid black; width: 20px; height: 20px;"></div> <div style="border: 1px solid black; width: 20px; height: 20px;"></div> <div style="border: 1px solid black; width: 20px; height: 20px;"></div> </div> </div> <div style="text-align: center; font-weight: bold;">2. Soziale und medizinische Anamnese</div> </div>																																																																
<p>21. Durchschnittl. tägl. Konsum elektron. Medien</p> <p> <input type="radio"/> gar nicht <input type="radio"/> max. 1 Stunde <input type="radio"/> max. 2 Stunden <input type="radio"/> max. 3 Stunden <input type="radio"/> über 3 Stunden <input checked="" type="radio"/> keine Angabe </p> <p>eigenes elektronisches Gerät des Kindes</p> <p> <input type="radio"/> TV <input type="radio"/> andere <input type="radio"/> kein Gerät <input checked="" type="radio"/> k. A. </p> <p>22. Vorsorgestatus</p> <p> <input type="radio"/> Heft fehlt <input type="radio"/> Heft vorhanden </p> <p>Die folgenden Untersuchungen fehlen:</p> <p> <input type="radio"/> U1 <input type="radio"/> U2 <input type="radio"/> U3 <input type="radio"/> U4 <input type="radio"/> U5 <input type="radio"/> U6 <input type="radio"/> U7 <input type="radio"/> U7a <input type="radio"/> U8 <input type="radio"/> U9 </p> <p>23. Impfstatus</p> <p> <input type="radio"/> Heft fehlt <input type="radio"/> Heft fehlt, Kind hat keinerlei Impfungen <input type="radio"/> Heft vorhanden </p> <div style="display: flex; justify-content: space-between; margin-top: 10px;"> <div style="width: 45%;"> <p>Anzahl der Impfdosen (keine Impfungen = 0)</p> </div> <div style="width: 45%;"> <p>nur bei 3 dok. Impf. Abstand zw. 2. und 3. Impfung ≥ 6 Monate</p> </div> </div> <p>Diphtherie <input type="checkbox"/> <input type="radio"/> ja <input type="radio"/> nein</p> <p>Pertussis <input type="checkbox"/> <input type="radio"/> ja <input type="radio"/> nein</p> <p>4. Pertussisimpfung (M u. J) <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/></p> <p>Tetanus <input type="checkbox"/> <input type="radio"/> ja <input type="radio"/> nein</p> <p>Polio <input type="checkbox"/> <input type="radio"/> ja <input type="radio"/> nein</p> <p>2 Dosen Virelon® erhalten <input type="checkbox"/> <input type="radio"/> ja <input type="radio"/> nein</p> <p>Hib <input type="checkbox"/> <input type="radio"/> ja <input type="radio"/> nein</p> <p>Hepatitis B <input type="checkbox"/> <input type="radio"/> ja <input type="radio"/> nein</p> <p>Pneumokokken <input type="checkbox"/> <input type="radio"/> ja <input type="radio"/> nein</p> <p>1. Pneumokokkenimpf. (M u. J) <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/></p> <p>Masern <input type="checkbox"/> <input type="radio"/> ja <input type="radio"/> nein</p> <p>2. Masernimpfung (M u. J) <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/></p> <p>Mumps <input type="checkbox"/> <input type="radio"/> ja <input type="radio"/> nein</p> <p>Röteln <input type="checkbox"/> <input type="radio"/> ja <input type="radio"/> nein</p> <p>Varizellen <input type="checkbox"/> <input type="radio"/> ja <input type="radio"/> nein</p> <p>Meningokokken C <input type="checkbox"/> <input type="radio"/> ja <input type="radio"/> nein</p> <p>Rotavirus <input type="checkbox"/> <input type="radio"/> ja <input type="radio"/> nein</p>	<p>24. Geburtsgewicht (g) <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/></p> <p>keine Angabe <input checked="" type="radio"/></p> <p>25. Körpergröße (cm) <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/></p> <p>keine Angabe <input checked="" type="radio"/></p> <p>26. Körpergewicht (kg) <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/></p> <p>keine Angabe <input checked="" type="radio"/></p> <p>27. Sehen</p> <p>Brille <input type="radio"/> nein <input type="radio"/> ja <input checked="" type="radio"/> keine Angabe</p> <p>Visus <input type="radio"/> ohne Brille <input type="radio"/> mit Brille <input checked="" type="radio"/> k.A. / nicht mögl.</p> <table border="1" style="width: 100%; border-collapse: collapse; margin-top: 10px;"> <thead> <tr> <th colspan="2">Visus</th> <th colspan="2">Vorschaltlinse</th> </tr> <tr> <th>rechts</th> <th>links</th> <th>rechts</th> <th>links</th> </tr> </thead> <tbody> <tr> <td colspan="2" style="text-align: center;">Rodenstock</td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> </tr> <tr> <td colspan="2" style="text-align: center;">Sehtafel</td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> </tr> <tr> <td></td> <td></td> <td><input type="radio"/> 1</td> <td><input type="radio"/> 1 besser</td> </tr> <tr> <td></td> <td></td> <td><input type="radio"/> 2</td> <td><input type="radio"/> 2 gleich</td> </tr> <tr> <td></td> <td></td> <td><input type="radio"/> 3</td> <td><input type="radio"/> 3 schlechter</td> </tr> <tr> <td></td> <td></td> <td><input checked="" type="radio"/></td> <td><input checked="" type="radio"/> k. Angabe</td> </tr> </tbody> </table> <p>Stereosehen <input type="checkbox"/> Anzahl erkannter Stereobilder <input type="checkbox"/></p> <p>keine Angabe / verweigert <input checked="" type="radio"/></p> <p>Farbsehen <input type="radio"/> unauffällig <input type="radio"/> auffällig <input checked="" type="radio"/> keine Angabe / verweigert</p> <p>28. Hören</p> <table border="1" style="width: 100%; border-collapse: collapse; margin-top: 10px;"> <thead> <tr> <th colspan="2">Audiogramm</th> <th colspan="5">Frequenz [Hz]</th> </tr> <tr> <th colspan="2">dB</th> <th>500</th> <th>1.000</th> <th>2.000</th> <th>4.000</th> <th>6.000</th> <th>k.A.</th> </tr> </thead> <tbody> <tr> <td>rechts</td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input checked="" type="radio"/></td> </tr> <tr> <td>links</td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input type="checkbox"/></td> <td><input checked="" type="radio"/></td> </tr> </tbody> </table> <p>Auswertung verwendbar <input type="radio"/> ja <input type="radio"/> nein</p> <p>29. Sinnesorgane o.B. auff. k.A. N B Ü kK</p> <p>Sehen <input type="radio"/> 0 <input type="radio"/> 1 <input checked="" type="radio"/> 99 <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4</p> <p>Hören <input type="radio"/> 0 <input type="radio"/> 1 <input checked="" type="radio"/> 99 <input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4</p> <p>30. Zustand des Gebisses</p> <p> <input type="radio"/> naturgesund / versorgt <input type="radio"/> sanierungsbedürftig (Karies) <input type="radio"/> Stümpfe / Extraktion(en) wg. Karies <input checked="" type="radio"/> keine Angabe </p> <p>Zähneputzen in der Kita</p> <p> <input type="radio"/> ja <input type="radio"/> nein <input type="radio"/> nicht sicher <input checked="" type="radio"/> k.A. </p>	Visus		Vorschaltlinse		rechts	links	rechts	links	Rodenstock		<input type="checkbox"/>	<input type="checkbox"/>	Sehtafel		<input type="checkbox"/>	<input type="checkbox"/>			<input type="radio"/> 1	<input type="radio"/> 1 besser			<input type="radio"/> 2	<input type="radio"/> 2 gleich			<input type="radio"/> 3	<input type="radio"/> 3 schlechter			<input checked="" type="radio"/>	<input checked="" type="radio"/> k. Angabe	Audiogramm		Frequenz [Hz]					dB		500	1.000	2.000	4.000	6.000	k.A.	rechts	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="radio"/>	links	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="radio"/>
Visus		Vorschaltlinse																																																														
rechts	links	rechts	links																																																													
Rodenstock		<input type="checkbox"/>	<input type="checkbox"/>																																																													
Sehtafel		<input type="checkbox"/>	<input type="checkbox"/>																																																													
		<input type="radio"/> 1	<input type="radio"/> 1 besser																																																													
		<input type="radio"/> 2	<input type="radio"/> 2 gleich																																																													
		<input type="radio"/> 3	<input type="radio"/> 3 schlechter																																																													
		<input checked="" type="radio"/>	<input checked="" type="radio"/> k. Angabe																																																													
Audiogramm		Frequenz [Hz]																																																														
dB		500	1.000	2.000	4.000	6.000	k.A.																																																									
rechts	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="radio"/>																																																									
links	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="radio"/>																																																									

Dokumentationsbogen für die Einschulungsuntersuchungen der KJGD im Land Berlin																																																																																															
KJGD-Stelle:			Schuljahr: 2019																																																																																												
<div style="display: flex; justify-content: space-between; align-items: center;"> <div style="text-align: center;"> laufende Nummer <input style="width: 30px; height: 20px; border: 1px solid black;" type="text"/> <input style="width: 30px; height: 20px; border: 1px solid black;" type="text"/> <input style="width: 30px; height: 20px; border: 1px solid black;" type="text"/> <input style="width: 30px; height: 20px; border: 1px solid black;" type="text"/> </div> <div style="text-align: center; font-weight: bold;">3. Ärztliche Beurteilung / Empfehlungen</div> </div>																																																																																															
31. Deutschkenntnisse <table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="width: 15%;">Kind</th> <th style="width: 15%;">Mutter</th> <th style="width: 15%;">Vater</th> <th></th> </tr> </thead> <tbody> <tr> <td style="text-align: center;">①</td> <td style="text-align: center;">①</td> <td style="text-align: center;">①</td> <td>nicht</td> </tr> <tr> <td style="text-align: center;">②</td> <td style="text-align: center;">②</td> <td style="text-align: center;">②</td> <td>einzelne Worte</td> </tr> <tr> <td style="text-align: center;">③</td> <td style="text-align: center;">③</td> <td style="text-align: center;">③</td> <td>flüssig mit erh. Fehlern</td> </tr> <tr> <td style="text-align: center;">④</td> <td style="text-align: center;">④</td> <td style="text-align: center;">④</td> <td>(sehr) gut</td> </tr> <tr> <td></td> <td style="text-align: center;">9</td> <td style="text-align: center;">9</td> <td>hat nicht begleitet</td> </tr> <tr> <td style="text-align: center;">99</td> <td style="text-align: center;">99</td> <td style="text-align: center;">99</td> <td>keine Angabe</td> </tr> </tbody> </table>		Kind	Mutter	Vater		①	①	①	nicht	②	②	②	einzelne Worte	③	③	③	flüssig mit erh. Fehlern	④	④	④	(sehr) gut		9	9	hat nicht begleitet	99	99	99	keine Angabe	35. Ärztliche Beurteilung der Entwicklung <table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="width: 10%;">o.B.</th> <th style="width: 10%;">auff.</th> <th style="width: 10%;">k.A.</th> <th style="width: 10%;">N</th> <th style="width: 10%;">B</th> <th style="width: 10%;">Ü</th> <th style="width: 10%;">kK</th> </tr> </thead> <tbody> <tr> <td>Körperkoord.</td> <td style="text-align: center;">0</td> <td style="text-align: center;">1</td> <td style="text-align: center;">99</td> <td style="text-align: center;">①</td> <td style="text-align: center;">②</td> <td style="text-align: center;">③</td> <td style="text-align: center;">④</td> </tr> <tr> <td>Visuomotorik</td> <td style="text-align: center;">0</td> <td style="text-align: center;">1</td> <td style="text-align: center;">99</td> <td style="text-align: center;">①</td> <td style="text-align: center;">②</td> <td style="text-align: center;">③</td> <td style="text-align: center;">④</td> </tr> <tr> <td>vis. Wahrn.</td> <td style="text-align: center;">0</td> <td style="text-align: center;">1</td> <td style="text-align: center;">99</td> <td style="text-align: center;">①</td> <td style="text-align: center;">②</td> <td style="text-align: center;">③</td> <td style="text-align: center;">④</td> </tr> <tr> <td>Sprache</td> <td style="text-align: center;">0</td> <td style="text-align: center;">1</td> <td style="text-align: center;">99</td> <td style="text-align: center;">①</td> <td style="text-align: center;">②</td> <td style="text-align: center;">③</td> <td style="text-align: center;">④</td> </tr> <tr> <td>Mengenvorw.</td> <td style="text-align: center;">0</td> <td style="text-align: center;">1</td> <td style="text-align: center;">99</td> <td style="text-align: center;">①</td> <td style="text-align: center;">②</td> <td style="text-align: center;">③</td> <td style="text-align: center;">④</td> </tr> <tr> <td>em.-soz. Entw.</td> <td style="text-align: center;">0</td> <td style="text-align: center;">1</td> <td style="text-align: center;">99</td> <td style="text-align: center;">①</td> <td style="text-align: center;">②</td> <td style="text-align: center;">③</td> <td style="text-align: center;">④</td> </tr> <tr> <td>kognitive Entw.</td> <td style="text-align: center;">0</td> <td style="text-align: center;">1</td> <td style="text-align: center;">99</td> <td style="text-align: center;">①</td> <td style="text-align: center;">②</td> <td style="text-align: center;">③</td> <td style="text-align: center;">④</td> </tr> </tbody> </table>			o.B.	auff.	k.A.	N	B	Ü	kK	Körperkoord.	0	1	99	①	②	③	④	Visuomotorik	0	1	99	①	②	③	④	vis. Wahrn.	0	1	99	①	②	③	④	Sprache	0	1	99	①	②	③	④	Mengenvorw.	0	1	99	①	②	③	④	em.-soz. Entw.	0	1	99	①	②	③	④	kognitive Entw.	0	1	99	①	②	③	④
Kind	Mutter	Vater																																																																																													
①	①	①	nicht																																																																																												
②	②	②	einzelne Worte																																																																																												
③	③	③	flüssig mit erh. Fehlern																																																																																												
④	④	④	(sehr) gut																																																																																												
	9	9	hat nicht begleitet																																																																																												
99	99	99	keine Angabe																																																																																												
	o.B.	auff.	k.A.	N	B	Ü	kK																																																																																								
Körperkoord.	0	1	99	①	②	③	④																																																																																								
Visuomotorik	0	1	99	①	②	③	④																																																																																								
vis. Wahrn.	0	1	99	①	②	③	④																																																																																								
Sprache	0	1	99	①	②	③	④																																																																																								
Mengenvorw.	0	1	99	①	②	③	④																																																																																								
em.-soz. Entw.	0	1	99	①	②	③	④																																																																																								
kognitive Entw.	0	1	99	①	②	③	④																																																																																								
32. Bisherige Behandlungen des Kindes <table style="width: 100%; border-collapse: collapse;"> <tbody> <tr> <td style="width: 30%;">Physiotherapie</td> <td style="width: 10%; text-align: center;">① ja</td> <td style="width: 10%; text-align: center;">② nein</td> <td style="width: 10%; text-align: center;">99 k. A.</td> </tr> <tr> <td>Ergotherapie</td> <td style="text-align: center;">① ja</td> <td style="text-align: center;">② nein</td> <td style="text-align: center;">99 k. A.</td> </tr> <tr> <td>Logopädie</td> <td style="text-align: center;">① ja</td> <td style="text-align: center;">② nein</td> <td style="text-align: center;">99 k. A.</td> </tr> <tr> <td>Psychotherapie</td> <td style="text-align: center;">① ja</td> <td style="text-align: center;">② nein</td> <td style="text-align: center;">99 k. A.</td> </tr> </tbody> </table>		Physiotherapie	① ja	② nein	99 k. A.	Ergotherapie	① ja	② nein	99 k. A.	Logopädie	① ja	② nein	99 k. A.	Psychotherapie	① ja	② nein	99 k. A.	36. Schulische Förderung empfohlen <input type="radio"/> keine Förderung notwendig <input type="radio"/> Sprache <input type="radio"/> Visuomotorik <input type="radio"/> visuelle Wahrnehmung <input type="radio"/> körperliche und motorische Entwicklung <input type="radio"/> emotionale/soziale Entwicklung <input type="radio"/> Lernen																																																																													
Physiotherapie	① ja	② nein	99 k. A.																																																																																												
Ergotherapie	① ja	② nein	99 k. A.																																																																																												
Logopädie	① ja	② nein	99 k. A.																																																																																												
Psychotherapie	① ja	② nein	99 k. A.																																																																																												
33. Psychische Auffälligkeiten (SDQ) - optional - <table style="width: 100%; border-collapse: collapse;"> <tbody> <tr> <td style="width: 30%;">emot. Probl.</td> <td style="width: 10%; text-align: center;"><input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/></td> <td style="width: 10%; text-align: center;">Peer-Probleme</td> <td style="width: 10%; text-align: center;"><input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/></td> </tr> <tr> <td>Verhalt.-Probl.</td> <td style="text-align: center;"><input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/></td> <td style="text-align: center;">prosoz. Verh.</td> <td style="text-align: center;"><input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/></td> </tr> <tr> <td>Hyperaktivität</td> <td style="text-align: center;"><input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/></td> <td></td> <td></td> </tr> <tr> <td>psy. Auffälligkeiten</td> <td style="text-align: center;">② nein</td> <td style="text-align: center;">① ja</td> <td style="text-align: center;">99 k. A.</td> </tr> <tr> <td>in Behandlung oder Diagnostik</td> <td style="text-align: center;">② nein</td> <td style="text-align: center;">① ja</td> <td style="text-align: center;">99 k. A.</td> </tr> <tr> <td>zur Diagnostik/Behandlung überwiesen</td> <td style="text-align: center;">② nein</td> <td style="text-align: center;">① ja</td> <td style="text-align: center;">99 k. A.</td> </tr> <tr> <td>Rückmeldung nach Überweisung</td> <td colspan="3"> <input type="radio"/> psy. Auffälligkeiten n. bestätigt <input type="radio"/> psy. Auffälligkeiten bestätigt <input checked="" type="radio"/> keine Angabe </td> </tr> </tbody> </table>		emot. Probl.	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	Peer-Probleme	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	Verhalt.-Probl.	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	prosoz. Verh.	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	Hyperaktivität	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>			psy. Auffälligkeiten	② nein	① ja	99 k. A.	in Behandlung oder Diagnostik	② nein	① ja	99 k. A.	zur Diagnostik/Behandlung überwiesen	② nein	① ja	99 k. A.	Rückmeldung nach Überweisung	<input type="radio"/> psy. Auffälligkeiten n. bestätigt <input type="radio"/> psy. Auffälligkeiten bestätigt <input checked="" type="radio"/> keine Angabe			37. Sonderpädagogischer Förderbedarf <input type="radio"/> kein Antrag empfohlen <input type="radio"/> Sehen <input type="radio"/> Hören <input type="radio"/> Sprache <input type="radio"/> körperliche und motorische Entwicklung <input type="radio"/> geistige Entwicklung <input type="radio"/> autistische Behinderung <input type="radio"/> emotionale/soziale Entwicklung <input type="radio"/> Lernen																																																																	
emot. Probl.	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	Peer-Probleme	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>																																																																																												
Verhalt.-Probl.	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	prosoz. Verh.	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>																																																																																												
Hyperaktivität	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>																																																																																														
psy. Auffälligkeiten	② nein	① ja	99 k. A.																																																																																												
in Behandlung oder Diagnostik	② nein	① ja	99 k. A.																																																																																												
zur Diagnostik/Behandlung überwiesen	② nein	① ja	99 k. A.																																																																																												
Rückmeldung nach Überweisung	<input type="radio"/> psy. Auffälligkeiten n. bestätigt <input type="radio"/> psy. Auffälligkeiten bestätigt <input checked="" type="radio"/> keine Angabe																																																																																														
34. Entwicklungsdiagnostik / S-ENS + SOPESS <table style="width: 100%; border-collapse: collapse;"> <tbody> <tr> <td style="width: 30%;">Körperkoordination</td> <td style="width: 10%; text-align: center;"><input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/></td> <td style="width: 10%; text-align: center;">99 k. A.</td> </tr> <tr> <td>Visuomotorik</td> <td style="text-align: center;"><input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/></td> <td style="text-align: center;">99 k. A.</td> </tr> <tr> <td>Visuelle Wahrnehmung</td> <td style="text-align: center;"><input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/></td> <td style="text-align: center;">99 k. A.</td> </tr> <tr> <td>Pseudowörter</td> <td style="text-align: center;"><input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/></td> <td style="text-align: center;">99 k. A.</td> </tr> <tr> <td>Wörter ergänzen</td> <td style="text-align: center;"><input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/></td> <td style="text-align: center;">99 k. A.</td> </tr> <tr> <td>Sätze nachsprechen</td> <td style="text-align: center;"><input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/></td> <td style="text-align: center;">99 k. A.</td> </tr> <tr> <td>Pluralbildung</td> <td style="text-align: center;"><input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/></td> <td style="text-align: center;">99 k. A.</td> </tr> <tr> <td>Artikulation</td> <td style="text-align: center;"><input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/></td> <td style="text-align: center;">99 k. A.</td> </tr> <tr> <td>Mengenvorwissen</td> <td style="text-align: center;"><input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/></td> <td style="text-align: center;">99 k. A.</td> </tr> </tbody> </table>		Körperkoordination	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	99 k. A.	Visuomotorik	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	99 k. A.	Visuelle Wahrnehmung	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	99 k. A.	Pseudowörter	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	99 k. A.	Wörter ergänzen	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	99 k. A.	Sätze nachsprechen	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	99 k. A.	Pluralbildung	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	99 k. A.	Artikulation	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	99 k. A.	Mengenvorwissen	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	99 k. A.	38. Antrag auf Zurückstellung <input type="radio"/> nein <input type="radio"/> ja <input type="radio"/> wird erwogen <input type="radio"/> k. A.																																																																		
Körperkoordination	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	99 k. A.																																																																																													
Visuomotorik	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	99 k. A.																																																																																													
Visuelle Wahrnehmung	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	99 k. A.																																																																																													
Pseudowörter	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	99 k. A.																																																																																													
Wörter ergänzen	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	99 k. A.																																																																																													
Sätze nachsprechen	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	99 k. A.																																																																																													
Pluralbildung	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	99 k. A.																																																																																													
Artikulation	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	99 k. A.																																																																																													
Mengenvorwissen	<input style="width: 20px; height: 20px; border: 1px solid black;" type="text"/>	99 k. A.																																																																																													
		39. Einschulung von KJGD befürwortet <input checked="" type="radio"/> ja <input type="radio"/> nein <input type="radio"/> keine Angabe																																																																																													
		40. Ggf. 2. ESU erforderlich <input type="radio"/> nein <input type="radio"/> ja																																																																																													
		41. Zurückstellung v. Schulaufsicht erfolgt <input checked="" type="radio"/> ja																																																																																													
		42. Zusatzangabe (Senat) <input style="width: 100%; height: 20px; border: 1px solid black;" type="text"/>																																																																																													
		43. Zusatzangabe (Bezirk) <input style="width: 100%; height: 20px; border: 1px solid black;" type="text"/>																																																																																													

**Dokumentationsbogen für die S-ENS und SOPESS-Untertests bei den
Einschulungsuntersuchungen in Berlin 2019**

Name des Kindes: _____		auffällig	grenzw.	unauffäll.	keine Angabe
Körperkoordination Seitliches Hin- und Herspringen auf der Hüpfmatte. Nach den ersten fünf Sprüngen wird mit der Zeitmessung begonnen. Gezählt wird die Anzahl von Sprüngen innerhalb von 10 Sekunden.		0 - 6	7	≥ 8	
KöKo Anzahl der Sprünge: <input type="text"/> <input type="text"/>					
Visuomotorik Drachen (0 - 7 Punkte möglich) <input type="text"/> <div style="display: flex; align-items: center; justify-content: space-around;"> <div> Tisch <input type="text"/> + Kreuz <input type="text"/> + Baum <input type="text"/> = Summe <input type="text"/> </div> <div> Faktor <input type="text"/> 3 <input type="text"/> = Summe TKB <input type="text"/> <input type="text"/> </div> </div> <div style="display: flex; justify-content: space-between; font-size: small;"> (jeweils 0 - 2 Punkte möglich) (max. 6 Punkte) (max. 18 Punkte) </div>		0 - 13	14 - 16	17 - 25	k.A.
Visuo Drachen + Summe TKB <input type="text"/> <input type="text"/>					
Visuelle Wahrnehmung und Informationsverarbeitung Was passt dazu? Es wird auch die Einführungsaufgabe bewertet. 1 <input type="checkbox"/> 2 <input type="checkbox"/> 3 <input type="checkbox"/> 4 <input type="checkbox"/> 5 <input type="checkbox"/> 6 <input type="checkbox"/> 7 <input type="checkbox"/> 8 <input type="checkbox"/> 9 <input type="checkbox"/> 10 <input type="checkbox"/>		0 - 5	6	7 - 10	k.A.
Was sieht genauso aus? Es wird auch die Einführungsaufgabe bewertet.					
ViWa Gesamtsumme <input type="text"/> <input type="text"/>					
Pseudowörter nachsprechen Die beim Vorsprechen zu betonenden Buchstaben sind fett gedruckt. Die beiden Einführungsaufgaben werden auch bewertet. <input type="checkbox"/> Zippelzack <input type="checkbox"/> Fangofänger <input type="checkbox"/> Kimiklri <input type="checkbox"/> Risolamu <input type="checkbox"/> Maramula <input type="checkbox"/> Sangatima		0 - 3	4	5 - 6	k.A.
Pseu Summe "Pseudowörter" <input type="text"/>					
Wörter ergänzen Es wird auch die Einführungsaufgabe bewertet. <input type="checkbox"/> Scho?olade <input type="checkbox"/> Flugzeu? <input type="checkbox"/> Spa?etti <input type="checkbox"/> Kro?o?il <input type="checkbox"/> Sonnenschei? <input type="checkbox"/> Tee?öfifel <input type="checkbox"/> Finger?agel <input type="checkbox"/> Ele?ant		0 - 5	6	7 - 8	k.A.
Wort Summe "Wörter ergänzen" <input type="text"/>					
Sätze nachsprechen Es wird auch die Einführungsaufgabe bewertet. <input type="checkbox"/> Das grüne Pferd kann schnell rennen. <input type="checkbox"/> Da gehen drei Kinder zur Schule. <input type="checkbox"/> Der Teppich wird von dem Vater ausgeklopft. <input type="checkbox"/> Die kleine Maus wird von der Schildkröte gejagt. <input type="checkbox"/> Die Katze schnuppert an dem Blumenstrauß.		0 - 2	3	4 - 5	k.A.
Sätze Summe "Sätze nachsprechen" <input type="text"/>					
Pluralbildung Die Beispielaufgabe "Bananen" wird nicht bewertet. <input type="checkbox"/> Bäume <input type="checkbox"/> Gespenster <input type="checkbox"/> Häuser <input type="checkbox"/> Kirschen <input type="checkbox"/> Vögel <input type="checkbox"/> Fotos <input type="checkbox"/> Sterne		0 - 3	4 - 5	6 - 7	k.A.
Plural Summe "Pluralbildung" <input type="text"/>					
Artikulation Es wird dokumentiert, in welchen Stammelfehlergruppen die Artikulation auffällig ist. <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> S, Z SCH T, D CH (2) G, K L, N R F, PF B CH (1)		Summe "Artikulation" <input type="text"/> <input type="text"/>		k.A.	
Artiku					
Zahlen- und Mengenvorwissen Simultanerfassung Es wird auch die erste Aufgabe bewertet. <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> A(2) B(4) C(3) D(4) E(4) F(3) G(4) H(3)		0 - 10	11 - 13	14 - 16	k.A.
Mengenvergleich Es wird auch die erste Aufgabe bewertet. <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> A (blau) B (gelb) C (gelb) D (blau) E (gelb) F (gelb) G (blau) H (blau)					
Zahl Summe "Zahlen- und Mengenvorwissen" <input type="text"/> <input type="text"/>					
Händigkeit <input type="checkbox"/> rechtshändig <input type="checkbox"/> linkshändig <input type="checkbox"/> beidhändig <input type="checkbox"/> keine Angabe					

A.2 Übersicht über die verwendeten R-Packages

R-Package (Quelle)	Version	Anwendung
DescTools (Signorell, 2019)	0.99.28	Zusammenhangsmaße
DT (Xie et al., 2019)	0.8	Interaktive Tabellen
ggExtra (Attali und Baker, 2019)	0.9	Streudiagramme mit Randgrafiken
ggiraph (Gohel und Skintzos, 2019)	0.7.0	Interaktive Grafiken (Karten)
Hmisc (Harrell, 2019)	4.2.0	Datenaufbereitung (SPSS-Format)
janitor (Firke, 2019)	1.2.0	Summenzeilen.
Kernelheaping (Groß, 2018)	2.2.1	Kernelheaping-Verfahren
mapproj (McIlroy, 2019)	1.2.6	Kartenprojektion.
maptools (Bivand und Lewin-Koh, 2019)	0.9-5	[Datenaufbereitung (Shapefiles)]
scales (Wickham und Seidel, 2019)	1.1.0	Zahlenformate im Prozent-Format
shiny (Chang et al., 2018)	1.3.2	Webanwendungen mit R
shinycssloaders (Sali und Hass, 2017)	0.2.0	Animierte Ladeanzeigen
shinydashboard (Chang und Borges Ribeiro, 2018)	0.7.1	Layout Frontend
shinyjs (Attali, 2018)	1.0	JavaScript-Funktionalitäten
styler (Müller und Walthert, 2019)	1.1.1	Anwendung tidyverse style guide
tidyverse (R-Packages ggplot2, dplyr, tidyr, readr, purr, tibble, stringr, forcats) (Wickham, 2017)	1.2.1	Datenaufbereitung, Datenmanipulation, Statistiken, Datenvisualisierung
officer (Gohel, 2020a)	0.3.8	Manipulation von Excel-Dateien
openxlsx (Schauberger und Walker, 2019)	4.1.0.1	Datenexport in Excel-Dateien
RColorBrewer (Neuwirth, 2014)	1.1.2	Farbskalen für Grafiken
rgdal (Bivand et al., 2019)	1.4-4	[Datenaufbereitung (Shapefiles)]
rgeos (Bivand und Rundel, 2019)	0.5-1	[Datenaufbereitung (Shapefiles)]
rlang (Henry und Wickham, 2019)	0.4.0	tidy evaluation Rahmenwerk
rvg (Gohel, 2020b)	0.2.4	Export von SVG-Grafiken

A.3 Ordner- und Dateistruktur ESU explorer

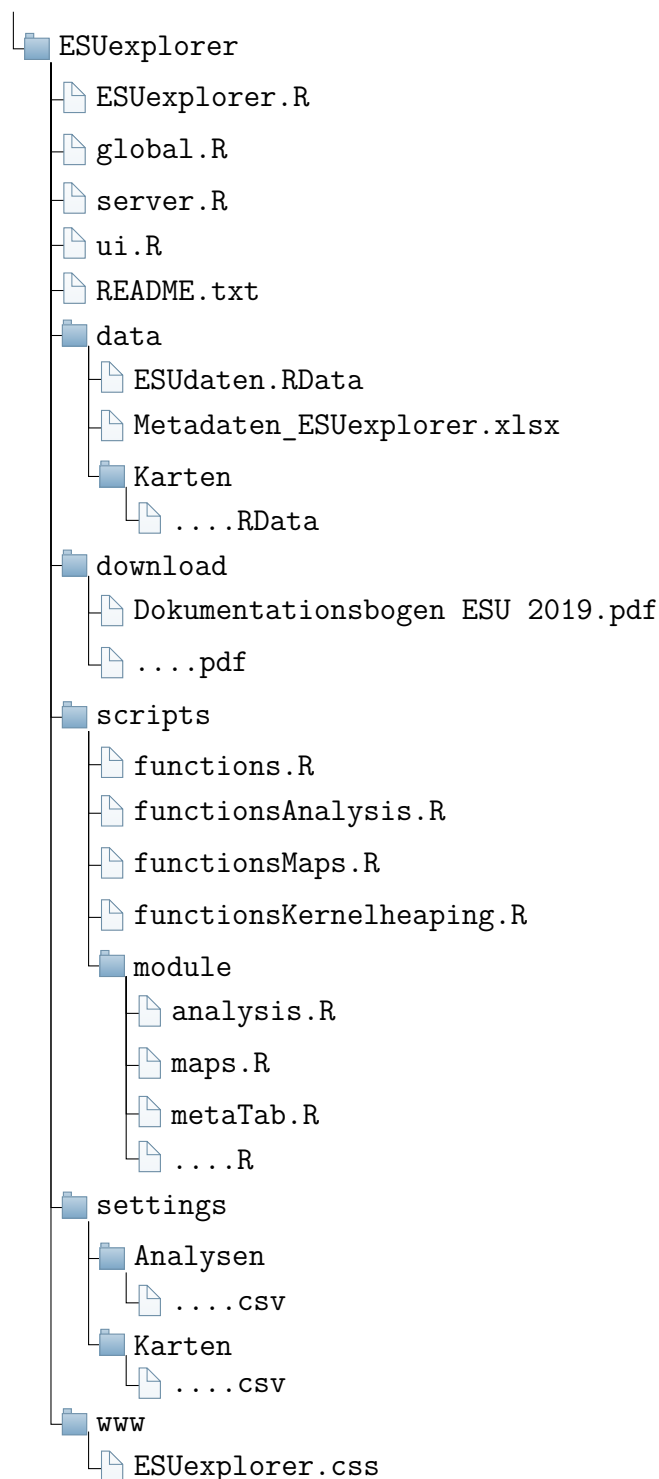


Abbildung A.1: Ordnerstruktur ESU explorer

Die sinnvolle Strukturierung und Ordnung der für den ESU explorer notwendigen Dateien soll die Übersichtlichkeit und Wartbarkeit sicherstellen.

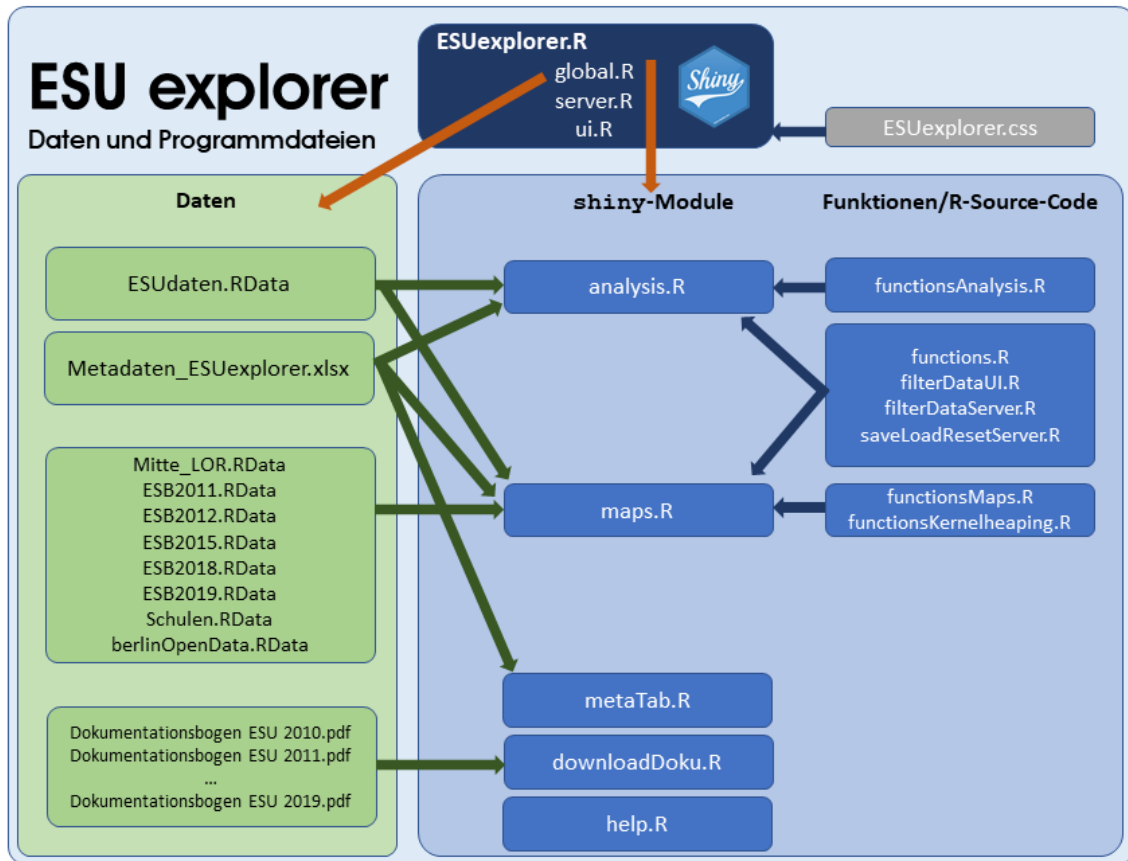


Abbildung A.2: Datei-Beziehungen ESU explorer

In dieser Darstellung sind alle für den ESU explorer notwendigen Dateien aufgeführt. Das Programm wird durch die Ausführung der Datei ESUexplorer.R gestartet. Damit werden die Haupt-Skripte global.R, ui.R und server.R automatisch eingebunden. Diese binden in Folge die notwendigen Datendateien, shiny-Module und weitere Skript-Dateien ein. Die shiny-Module nutzen verschiedene Daten, Funktionsdefinitionen und Programmbausteine.

A.4 Erweiterte Grafikeinstellungen ESU explorer

Erweiterte Einstellungen Grafik

☒ Beschriftungen ausblenden ($\leq \dots$ %) 0% 100% ☐ Beschriftung ohne Prozentzeichen % Textgröße Werte:

Textgröße Achsen/Legende: Position Legende: Farbpalette: Seitenverhältnis Grafik:

Kategorien der Analysevariable ausblenden (mind. eine Kategorie muss dargestellt werden):

Kategorien der Spaltenvariable ausblenden (mind. eine Kategorie muss dargestellt werden):

Abbildung A.3: ESU explorer: Erweiterte Grafikeinstellungen für Säulen-, Balken-, Linien- und Kreisdiagramm

Erweiterte Einstellungen Grafik

Textgröße Achsen/Legende: Seitenverhältnis Grafik:

Kategorien der Spaltenvariable ausblenden (mind. eine Kategorie muss dargestellt werden):

☒ Box-Plot anzeigen
☒ Violin-Plot anzeigen
☐ Mittelwert anzeigen

Abbildung A.4: ESU explorer: Erweiterte Grafikeinstellungen für Box-Plot/Violin-Plot

Erweiterte Einstellungen Grafik

Textgröße Achsen/Legende: Seitenverhältnis Grafik:

Kategorien der Spaltenvariable ausblenden (mind. eine Kategorie muss dargestellt werden):

Anzahl Histogramm-Klassen 5 50

☒ Kern-Dichteschätzung anzeigen

Abbildung A.5: ESU explorer: Erweiterte Grafikeinstellungen für Histogramm/Dichte

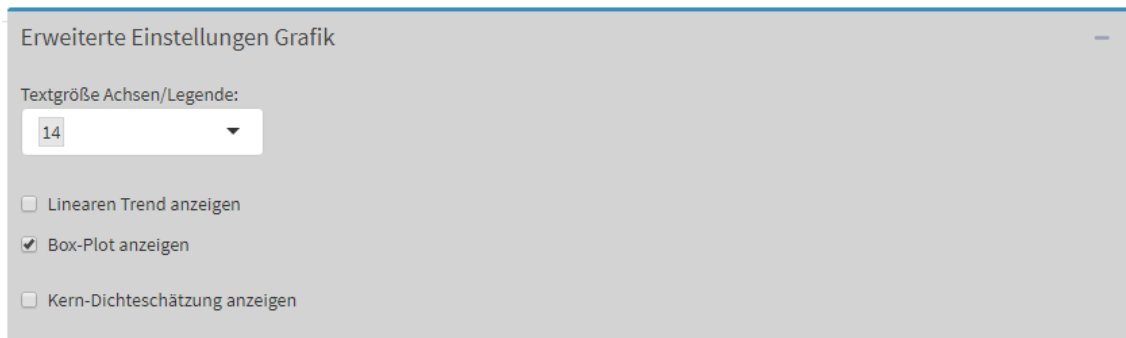


Abbildung A.6: ESU explorer: Erweiterte Grafikeinstellungen für Punktediagramm

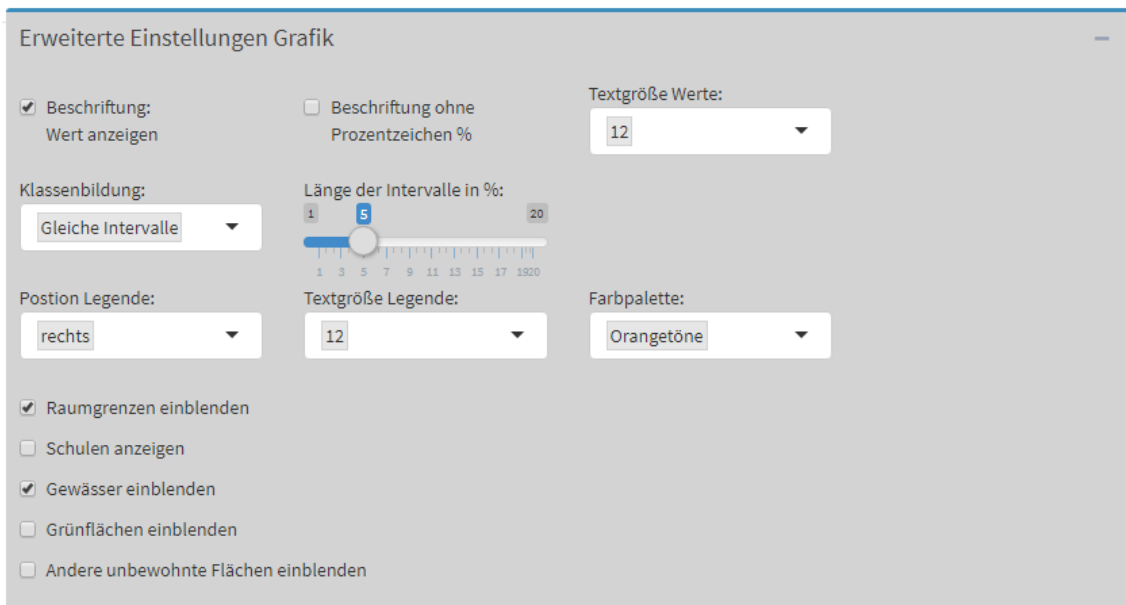


Abbildung A.7: ESU explorer: Erweiterte Grafikeinstellungen für Choroplethen-Karten

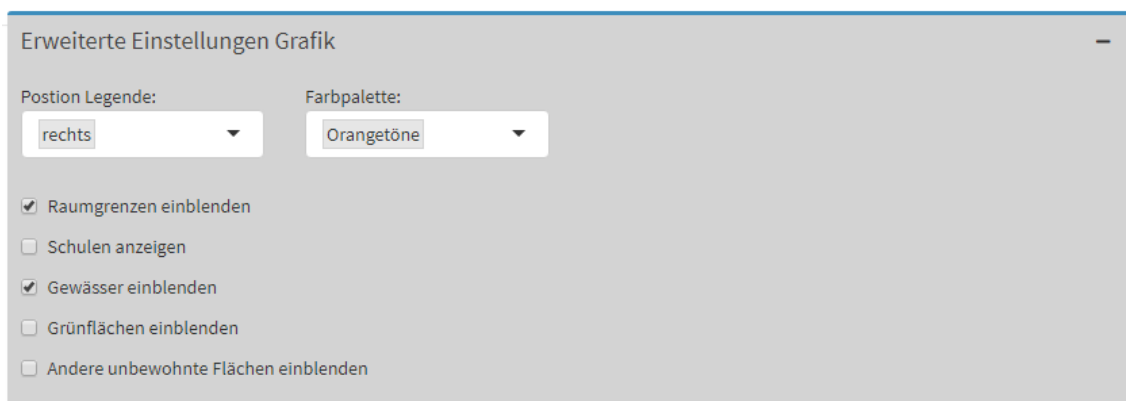


Abbildung A.8: ESU explorer: Erweiterte Grafikeinstellungen für Kernelheaping-Karten

A.5 Hinweise zur beiliegenden CD

Dieser Arbeit ist als Anlage eine CD mit folgendem Inhalt beigelegt:

- **UlrikeNiemann_ExplorativeDatenanalyseMitRShiny.PDF**

Version dieser Arbeit im PDF-Format.

- **Programmordner ESU explorer**

Zum Start der Anwendung ist die Datei `ESUexplorer.R` auszuführen. Im Programmordner ist eine `README.txt`-Datei mit weiteren Informationen enthalten.

Die originale im `ESU explorer` verwendete Datenbasis aus rund 300 Variablen darf aus datenschutzrechtlichen Gründen hier nicht verwendet und gezeigt werden. Aus diesem Grund enthält die beiliegende Version des `ESU explorers` nur wenige und vollständig simulierte Daten zum Zwecke des Tests der Funktionalitäten. Die mit dieser Demo-Version erstellten Analysen zeigen also keine echten oder realistischen Ergebnisse. Auch die bereits enthaltenen gespeicherten Analyseinstellungen zeigen nur beispielhaft die verschiedenen Grafiktypen und keine realistischen Zahlen.

- **Ordner Datenaufbereitung**

Aufbereitung der einzelnen Geodaten in das `RData`-Format, Aufbereitung der `ESU`-Daten aus dem `SPSS`- in das `RData`-Format (hier: Demo-Daten).

Literaturverzeichnis

- Attali, D. (2018): *shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds*. R Package Version 1.0. <https://cran.r-project.org/web/packages/shinyjs/>
- Attali, D. und C. Baker (2019): *ggExtra: Add Marginal Histograms to 'ggplot2', and More 'ggplot2' Enhancements*. R Package Version 0.9. <https://cran.r-project.org/web/packages/ggExtra/>
- Amt für Statistik Berlin-Brandenburg (2015): RBS-LOR, Lebensweltlich orientierte Räume, Dezember 2015. (Geometrien/Shapedateien) https://www.statistik-berlin-brandenburg.de/opendata/RBS_OD_LOR_2015_12.zip (Zugriff am 01.08.2019).
- Amt für Statistik Berlin-Brandenburg (2019): *Statistischer Bericht B I 1 - j / 17. Allgemeinbildende Schulen im Land Berlin Schuljahr 2017/18*. Potsdam.
- Autorengruppe Bildungsberichterstattung (2018): *Bildung in Deutschland 2018. Ein indikatorengestützter Bericht mit einer Analyse zu Wirkungen und Erträgen von Bildung*. Bielefeld: wbv Media.
- Beeley C. und S. R. Sukhdeve (2018): *Web Application Development with R Using Shiny*. 3. Auflage, Birmingham: Packt Publishing.
- Behnke, J. (2005): Lassen sich Signifikanzen auf Vollerhebungen anwenden? Einige essayistische Anmerkungen. in: *Politische Vierteljahresschrift*, 46. Jg. (2005), Heft 1, S. O-1–O-15.
- Benjamini, Y. (1988): Opening the Box of a Boxplot. in: *The American Statistician*, 42:4, 257-262.
- Bettge, S. und S. Oberwöhrmann; Senatsverwaltung für Gesundheit, Pflege und Gleichstellung. Referat Gesundheitsberichterstattung, Epidemiologie, Gesundheitsinformationssysteme, Statistikstelle (2018): *Grundauswertung der Einschulungsdaten in Berlin 2017*. Berlin.

- Bezirksamt Mitte von Berlin (2017): Bezirksamtsvorlage Nr. 114/2017 zur Beschlussfassung zum Bildungsmonitoring Berlin-Mitte. <https://www.berlin.de/ba-mitte/politik-und-verwaltung/bezirksamt/beschluesse-des-bezirksamts/2017/artikel.606838.php> (Zugriff am 11.12.2019).
- Bezirksamt Mitte von Berlin (2019): Bildungsmonitoring Berlin Mitte. <https://www.berlin.de/ba-mitte/politik-und-verwaltung/beauftragte/integration/bildungsmonitoring/> (Zugriff am 05.01.2020).
- Bivand, R.S., E. Pebesma und V. Gómez-Rubio (2013): *Applied Spatial Data Analysis with R*. 2. Auflage, New York, Heidelberg, Dordrecht, London: Springer.
- Bivand, R. und N. Lewin-Koh (2019): *maptools: Tools for Handling Spatial Objects*. R Package Version 0.9-5. <https://cran.r-project.org/web/packages/maptools/>
- Bivand, R. und C. Rundel (2019): *rgeos: Interface to Geometry Engine*. R Package Version 0.5-1. <https://cran.r-project.org/web/packages/rgeos/>
- Bivand, R., T. Keitt und B. Rowlingson (2019): *rgdal: Bindings for the 'Geospatial' Data Abstraction Library*. R Package Version 1.4-4. <https://cran.r-project.org/web/packages/rgdal/>
- Bömermann, H., S. Jahn und K. Nelius (2006): Lebensweltlich orientierte Räume im Regionalen Bezugssystem. Werkstattbericht zum Projekt „Vereinheitlichung von Planungsräumen“ in: *Berliner Statistik*, Monatsschrift 8/06, 366-371.
- Bortz, J., G. A. Lienert und K. Boehnke (2008): *Verteilungsfreie Methoden in der Biostatistik*. 3. Auflage, Heidelberg: Springer Medizin Verlag.
- Broscheid, A. und T. Gschwend (2003): Augäpfel, Murmeltiere und Bayes: Zur Auswertung stochastischer Daten aus Vollerhebungen. in: *MPIfG Working Paper*, 46. Jg. (2005), Heft 1, S. O-16–O-26.
- Broscheid, A. und T. Gschwend (2005): Zur statistischen Analyse von Vollerhebungen. in: *Politische Vierteljahresschrift*, 03/7.
- Bühler, P., P. Schlaich und D. Sinner (2017): *HTML5 und CSS3. Semantik – Design – Responsive Layouts*. Berlin: Springer-Verlag.
- Chang, W., J. Cheng, J.J. Allaire, Y. Xie und J. McPherson (2018): *shiny: Application Framework for R*. R Package Version 1.3.2. <https://cran.r-project.org/web/packages/shiny/>
- Chang, W. und B. Borges Riberio (2018): *shinydashboard: Create Dashboards with 'Shiny'*. R Package Version 0.7.1. <https://cran.r-project.org/web/packages/shinydashboard/>

- Cheng, J. (2019): Modularizing Shiny app code. <https://shiny.rstudio.com/articles/modules.html> (Zugriff am 06.04.2020).
- Clauß, G., F.-R. Finze und L. Parzsch (2004): *Statistik. Für Soziologen, Pädagogen, Psychologen und Mediziner*. 5. Auflage, Frankfurt am Main: Wissenschaftlicher Verlag Harri Deutsch GmbH.
- Cohen, J. (1988): *Statistical power analysis for the behavioral sciences*. 2. Auflage, Hillsdale, NJ: Lawrence Erlbaum.
- Deinet, U. (2009): Grundlagen und Schritte sozialräumlicher Konzeptentwicklung. in: Deinet, U. (Hrsg.), *Sozialräumliche Jugendarbeit*, 3. Auflage, Ort: VS Verlag für Sozialwissenschaften, 13-26.
- Erfurth, Kerstin (2018): Gütebeurteilung und Einsatz simulierter Geokoordinaten bei der regionalen Analyse zur Bundestagswahl 2017. Master Thesis, Freie Universität Berlin.
- Fay, C., S. Rochette, V. Guyader, C. Girard und (2020): Engineering Production-Grade Shiny Apps. <https://engineering-shiny.org/> (Buch in Entwicklungsphase, Veröffentlichung geplant in 2020 innerhalb der Buch-Reihe R Series: Chapman & Hall. Zugriff am 04.03.2020).
- Firke, S. (2019): *janitor: Simple Tools for Examining and Cleaning Dirty Data*. R Package Version 1.2.0. <https://cran.r-project.org/web/packages/janitor/>
- Gohel, D. (2020a): *officer: Manipulation of Microsoft Word and PowerPoint Documents*. R Package Version 0.3.8. <https://cran.r-project.org/web/packages/officer/>
- Gohel, D. (2020b): *rvgl: R Graphics Devices for Vector Graphics Output*. R Package Version 0.2.4. <https://cran.r-project.org/web/packages/rvgl/>
- Gohel, D. und P. Skintzos (2019): *ggiraph: Make 'ggplot2' Graphics Interactive*. R Package Version 0.7.0. <https://cran.r-project.org/web/packages/ggiraph/>
- Groß, M.; U. Rendtel, T. Schmid, S. Schmon und N. Tzavidis (2017): Estimating the density of ethnic minorities and aged people in Berlin: multivariate kernel density estimation applied to sensitive georeferenced administrative data protected via measurement error, in: *Journal of the Royal Statistical Society Series A* 180 (1), 161-183.
- Groß, M.; U. Rendtel, T. Schmid, H. Bömermann und K. Erfurth (2018): Simulated geo-coordinates as a tool for map-based regional analysis , in: *Diskussionsbeiträge des Fachbereichs Wirtschaftswissenschaft der Freien Universität Berlin* 2018/3.
- Groß, M. (2018): *Kernelheaping: Kernel Density Estimation for Heaped and Rounded Data*. R Package Version 2.2.0. <https://cran.r-project.org/web/packages/Kernelheaping/>

- Härdle, W; M. Müller; S. Sperlich und A. Werwatz. (2004): *Nonparametric and Semiparametric Models*. Berlin, Heidelberg: Springer-Verlag.
- Härdle, W; S. Klinkle und B. Rönz. (2015): *Introduction to Statistics*. Using Interactive MM*Stat Elements. Cham, Heidelberg, New York, Dordrecht, London: Springer.
- Harrell , F.E. (2019): *Hmisc: Harrell Miscellaneous*. R Package Version 4.2.0. <https://cran.r-project.org/web/packages/Hmisc/>
- Henry, L. und H. Wickham (2019): *rlang: Functions for Base Types and Core R and 'Tidyverse' Features*. R Package Version 0.4.0. <https://cran.r-project.org/web/packages/rlang/>
- Hintze, J. L. und R. D. Nelson (1998): Violin Plots: A Box Plot-Density Trace Synergism. in: *The American Statistician*, 52:2, 181-184.
- Hornik, K. (2020): R FAQ. Frequently Asked Questions on R. <https://CRAN.R-project.org/doc/FAQ/R-FAQ.html> (Zugriff am 20.02.2020).
- ISO/IEC 9126-1:2001 (2001): *Software engineering — Product quality — Part 1: Quality model*. Genf.
- Konsortium Bildungsberichterstattung (2005): Gesamtkonzeption der Bildungsberichterstattung. <https://www.bildungsbericht.de/de/forschungsdesign/pdf-grundlagen/gesamtkonzeption.pdf> (Zugriff am 15.12.2019).
- Liebetrau, A. M. (1983): *Measures of Association*. Thousand Oaks: SAGE Publications.
- Lindenstruth, T. und S. Claußen (2017): Metadatenmanagement als neue Integrationsarchitektur. in: *WISTA – Wirtschaft und Statistik*, 5 (2017), 76-86.
- McIlroy, D. (2019): *mapproj: Map Projections*. R Package Version 1.2.6. <https://cran.r-project.org/web/packages/mapproj/>
- Müller, K. und L. Walthert (2019): *styler: Non-Invasive Pretty Printing of R Code*. R Package Version 1.1.1. <https://cran.r-project.org/web/packages/styler/>
- Neuwirth, E. (2014): *RColorBrewer: ColorBrewer Palettes*. R Package Version 1.1-2. <https://cran.r-project.org/web/packages/RColorBrewer/>
- Oberwöhrmann, S., S. Bettge und S. Hermann; Senatsverwaltung für Gesundheit, Umwelt und Verbraucherschutz Berlin, Referat Gesundheitsberichterstattung, Epidemiologie, Gemeinsames Krebsregister, Sozialstatistisches Berichtswesen, Gesundheits- und Sozialinformationssysteme (2011): *Kernindikatoren für Bezirksregionenprofile aus den Einschulungsdaten in Berlin*. Berlin.
- Olbrich, G., M. Quick und J. Schweikart (2002): *Desktop Mapping. Grundlagen und Praxis in Kartographie und GIS*. 3. Auflage, Berlin, Heidelberg: Springer-Verlag.

- Openstreetmap.org (2019): OpenStreetMap Daten für Berlin. (Geometrien/Shapedateien) <http://download.geofabrik.de/europe/germany/berlin-latest-free.shp.zip> (Zugriff am 01.08.2019).
- Pearson, K. (1900): On the criterion that a given system of deviations 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. in: *Philosophical Magazine Series 5*, 50:302, 157 — 175.
- Pearson, K. (1904): On the theory of contingency and its relation to association and normal correlation. in: *Draper's Company 'Memoirs*, Biometric Series 1.
- Polasek, W. (1994): *EDA Explorative Datenanalyse. Einführung in die deskriptive Statistik*. 2. Auflage, Berlin, Heidelberg: Springer-Verlag.
- R Core Team: R Foundation for Statistical Computing (2019): R: A Language and Environment for Statistical Computing. <https://www.r-project.org/> (Zugriff am 19.12.2019).
- Rockmann, U. und H. Leerhoff (2018a): Pilotprojekt Bildungsmonitoring in Berlin-Mitte. in: *Stadtforschung und Statistik: Zeitschrift des Verbandes Deutscher Städtestatistiker*, 31(1), 17-22.
- Rockmann, U. und H. Leerhoff (2018b): *Bildungszugänge und Bildungsübergänge von Kindern im Alter von 0 bis 18 Jahren im Bezirk Berlin-Mitte - 1. Projektbericht: Eine Charakterisierung des Bezirks und erste Befunde*. Berlin.
- Rockmann, U. und H. Leerhoff (2019): Pilotprojekt Bildungsmonitoring in Berlin-Mitte: Schulpflichtig werdende Kinder mit eigener Zuwanderungserfahrung. in: *Stadtforschung und Statistik : Zeitschrift des Verbandes Deutscher Städtestatistiker*, 32 (2), 81-88.
- Rosenecker, J. und H. Schmidt (Hrsg.) (2008): *Pädiatrische Anamnese, Untersuchung, Diagnose*. Heidelberg: Springer Medizin Verlag.
- RStudio Team: RStudio, Inc. (2020a): RStudio: Integrated Development Environment for R. <http://www.rstudio.com/> (Zugriff am 04.03.2020).
- RStudio Team: RStudio, Inc. (2020b): R hex stickers. <https://github.com/rstudio/hex-stickers/tree/master/PNG> (Zugriff am 29.03.2020).
- RStudio Team: RStudio, Inc. (2020c): Share your apps. <https://shiny.rstudio.com/tutorial/written-tutorial/lesson7/> (Zugriff am 09.04.2020).
- Sali, A. und L. Hass (2017): *shinycssloaders: Add CSS Loading Animations to 'shiny' Outputs*. R Package Version 0.2.0. <https://cran.r-project.org/web/packages/shinycssloaders/>

- Schauberger, P. und A. Walker (2019): *openxlsx: Read, Write and Edit xlsx Files*. R Package Version 4.1.0.1. <https://cran.r-project.org/web/packages/openxlsx/>
- Schlittgen, R. (2008): *Einführung in die Statistik*. Analyse und Modellierung von Daten. 11. Auflage, München: Oldenbourg Wissenschaftsverlag.
- Senatsverwaltung für Stadtentwicklung (Hrsg.) (2009): Handbuch zur Sozialraumorientierung. Grundlage der integrierten Stadt(teil)entwicklung Berlin. https://www.stadtentwicklung.berlin.de/soziale_stadt/sozialraumorientierung/download/SFS_Handbuch_RZ_screen.pdf (Zugriff am 16.12.2019).
- Signorell, A. (2019): *DescTools: Tools for Descriptive Statistics*. R Package Version 0.99.28. <https://cran.r-project.org/web/packages/DescTools/>
- Silverman, B.W. (1986): *Density Estimation for Statistics and Data Analysis*. London, New York: Chapman and Hall.
- SpryMedia (2020): DataTables. Add advanced interaction controls to your HTML tables the free & easy way. <https://datatables.net/> (Zugriff am 19.01.2020).
- Tukey, J. W. (1977): *Exploratory Data Analysis*. Reading, Massachusetts: Addison-Wesley Publishing Company.
- Vaidyanathan, R., K. Russell und RStudio, Inc. (2020): htmlwidgets for R. <https://www.htmlwidgets.org/> (Zugriff am 16.04.2020).
- Wand, M. und M. Jones (1994): Multivariate plug-in bandwidth selection. in: *Computational Statistics*. 9:2, 97-116.
- Werner O., Spezialkanzlei für Schulrecht in Berlin (2018): Gemeinsame Einzugsbereiche. <https://www.schulrecht-rechtsanwalt.de/berlin/grundschule/einschulung/einzugsgebiete/gemeinsame-einzugsbereiche.php> (Zugriff am 18.12.2019).
- Wickham, H. (2010): A Layered Grammar of Graphics. in: *Journal of Computational and Graphical Statistics*, 19:1, 3-28.
- Wickham, H. (2014): Tidy Data. in: *Journal of Statistical Software*, Vol. 59 (2014), Issue 10.
- Wickham, H. (2016): *ggplot2: Elegant Graphics for Data Analysis*. 2. Auflage, New York: Springer Verlag.
- Wickham, H. (2017): *tidyverse: Easily Install and Load the 'Tidyverse'*. R Package Version 1.2.1. <https://cran.r-project.org/web/packages/tidyverse/>
- Wickham, H. (2019): *Advanced R, Second edition*. 2. Auflage, Boca Raton: CRC Press. Online verfügbar unter <https://adv-r.hadley.nz/> (Zugriff am 12.03.2020).

- Wickham, H. (2020a): Mastering Shiny. <https://mastering-shiny.org/> (Buch in Entwicklungsphase, Veröffentlichung geplant in 2020 von O'Reilly Media. Zugriff am 04.03.2020).
- Wickham, H. (2020b): The tidy tools manifesto. <https://tidyverse.tidyverse.org/articles/manifesto.html> (Zugriff am 04.03.2020).
- Wickham, H. (2020c): The tidyverse style guide. <https://style.tidyverse.org/> (Zugriff am 04.04.2020).
- Wickham, H., W. Chang, L. Henry, T. Lin Pedersen, K. Takahashi, C. Wilke und K. Woo (2018): *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. R Package Version 3.2.0. <https://cran.r-project.org/web/packages/ggplot2/>
- Wickham, H., R. François, L. Henry und K. Müller (2019): *dplyr: A Grammar of Data Manipulation*. R Package Version 0.8.3. <https://cran.r-project.org/web/packages/dplyr/>
- Wickham, H. und G. Grolemund (2016): *R for Data Science*. Sebastopol: O'Reilly.
- Wickham, H. und L. Henry (2018): *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. R Package Version 0.8.3. <https://cran.r-project.org/web/packages/tidyr/>
- Wickham, H. und L. Henry (2020): Tidy evaluation. <https://tidyeval.tidyverse.org/> (Zugriff am 12.03.2020).
- Wickham, H. und D. Seidel (2019): *scales: Scale Functions for Visualization*. R Package Version 1.1.0. <https://cran.r-project.org/web/packages/scales/>
- Wilkinson, L. (2005): *The Grammar of Graphics*. 2. Auflage, New York: Springer.
- Xie, Y. (2017): Using selectize input. <https://shiny.rstudio.com/articles/selectize.html> (Zugriff am 15.01.2020).
- Xie, Y., J. Cheng und X. Tan (2019): *DT: A Wrapper of the JavaScript Library 'DataTables'*. R Package Version 0.8. <https://cran.r-project.org/web/packages/DT/>

Ich erkläre hiermit, dass ich die vorliegende Arbeit mit dem Titel *Explorative Datenanalyse mit R Shiny* selbständig angefertigt, keine anderen Hilfsmittel als die im Literaturverzeichnis genannten benutzt und alle aus den Quellen und der Literatur wörtlich oder sinngemäß übernommenen Stellen als solche gekennzeichnet habe. Ich erkläre weiterhin, dass die vorliegende Arbeit noch nicht im Rahmen eines anderen Prüfungsverfahrens eingereicht wurde.

Berlin, den 5. Mai 2020