Institut für Visualisierung und Interaktive Systeme
Universität Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Diplomarbeit Nr. 3346

Medical Visual Analytics

Patricius-Samuel Albu

Studiengang:	Informatik
Prüfer:	Prof. Dr. Thomas Ertl
Betreuer:	Dipl.-Inf. Michael Wörner
begonnen am:	01. Juni 2012
beendet am:	30. Januar 2013
CR-Klassifikation:	I.3.8, J.3

Kurzfassung

Visual Analytics hat in den letzten Jahren die Aufmerksamkeit vieler Forscher auf sich gezogen. Aus dem ursprünglichen Themenfeld der Katastrophen- und der Terrorbekämpfung haben sich die Anwendungen von Visual Analytics auch auf andere Bereiche erweitert. Durch die Integration von Visualisierungs- und Data Mining-Methoden können die Vorteile der menschlichen Wahrnehmung und der automatisierten Analyse verbunden und dadurch die Nachteile der jeweiligen Methode behoben werden. Basierend auf das Visual Analytics Mantra wird in dieser Arbeit ein interaktives System zur explorativen Analyse von historischen Patientendaten entwickelt. Die Daten werden basierend auf ihre geografische Zugehörigkeit auf einer Karte dargestellt und können nach verschiedenen Kriterien gruppiert, gefiltert und ausgewertet werden. In jeder Ansicht der Anwendung können weitere deskriptive Statistiken der ausgewählten Gruppen und deren zugrundeliegenden Datensätze angezeigt werden. In einem Experteninterview wurden im Anschluss an die Entwicklung eine Reihe von Anwendungsszenarien formuliert und analysiert, um die Taugligkeit des entwickelten Systems zu überprüfen.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Verzeichnis der Listings

Verzeichnis der Algorithmen

1. Einleitung

Information is not knowledge

(Albert Einstein)

1.1. Motivation

Die Technologien für das Sammeln von Daten wurden in den letzten vier Jahrzehnten stetig weiterentwickelt [DMKK12]. Neue Sensoren für die Datenaufnahme sowie schnellere und günstigere Speichermethoden haben diese Entwicklung in unterschiedlichen Branchen vorangetrieben [KMSZ06, DMKK12]. Auch in der Medizin werden umfangreiche Daten über Patienten und Krankheitsverläufe gesammelt. Bereits im Jahr 2001 wurde die Anzahl an Patienten in Nordamerika, Europa und Asien, die mindestens einen Teil ihrer medizinischen Daten in elektronischer Form abgespeichert hatten, auf etwa 750 Millionen geschätzt [CM02]. Allein der Onkologische Schwerpunkt Stuttgart[1] hat in den vergangenen 25 Jahren umfassende Informationen zum Krankheitsverlauf von über 100.000 Patienten gesammelt [OSPb].

Die Möglichkeiten Daten aufzunehmen und zu speichern sind schneller gewachsen als die Möglichkeiten diese zu analysieren, so dass man eine Informationsüberflutung befürchtet [KMSZ06]. Diese Informationen sind nur dann wertvoll, wenn sie interpretiert und neue Erkenntnisse daraus gewonnen werden können, ansonsten drohen sie zu „Datendeponien" zu werden [Kei02]. Visual Analytics hat das Ziel, die Gefahr der Informationsüberflutung in eine Gelegenheit umzuwandeln [KMSZ06].

Das Ziel dieser Arbeit ist die Erstellung eines Visual Analytics-Prototyps, der die explorative Analyse der Daten von Brustkrebspatienten ermöglichen, das Verständnis der vorliegenden Daten verbessern und das Finden neuer Informationen, die die Lebensqualität der Patienten verbessern können, fördern soll. Zwei Gründe haben dazu geführt, dass als Anwendung für die aktuelle Arbeit die Analyse von Patientendaten, die an Brustkrebs erkrankt sind, verwendet wird: die Tragweite dieser Krankheit in der heutigen Gesellschaft und die große Zahl der zur Verfügung stehenden Daten.

Der Brustkrebs (Mammakarzinom) ist eine bösartige Tumorerkrankung der Brustdrüse. Es ist die am häufigsten vorkommende Tumorerkrankung bei Frauen mit einem Anteil von 32,1% aller Tumorerkrankungen und mit geschätzten 71.660 Neuerkrankungen im Jahr 2008 [Kre12]. Die Erkrankung liegt auf Platz neun der häufigsten Todesursachen der Gesamtbevölkerung in Deutschland [Tod12a]. Sie tritt überwiegend bei Frauen auf und liegt

[1] http://www.osp-stuttgart.de

mit 17.974 Todesfällen im Jahr 2011 auf Platz vier der häufigsten Todesursachen bei Frauen [Tod12b]. Nur etwa ein Prozent der Erkrankungen treten bei Männern auf [SLJ93].

Die schon erwähnte Datensammlung des Onkologischen Schwerpunktes Stuttgart enthält eine Datenbank von etwa 20.000 Krankheitsverläufen von Mammakarzinomfällen, die durch die freundliche Unterstützung von Frau Prof. Dr. Else Heidemann vom Onkologischen Schwerpunkt Stuttgart und Dr. Peter Fritz vom Institut für Digitale Medizin Stuttgart für diese Arbeit zur Verfügung gestellt wurden. Kapitel 3 gibt einen Einblick in die Herkunft und den Aufbau dieser Datensätze.

1.2. Gliederung

Die restliche Arbeit ist in folgender Weise gegliedert:

Kapitel 2 – Grundlagen liefert die Definition des Begriffes „Visual Analytics“ und stellt die nötigen Grundlagen für das Verständnis der restlichen Kapitel vor.

Kapitel 3 – Datenquelle beschreibt die Herkunft, den Umfang und den Aufbau der zur Verfügung stehenden Daten.

Kapitel 4 – Eigenes Konzept stellt die konzeptionelle Grundlage, nach der die Arbeit strukturiert wurde, vor.

Kapitel 5 – Eigene Umsetzung beinhaltet die Beschreibung der einzelnen Programmfunktionen.

Kapitel 6 – Prototyp beschreibt die verwendeten Technologien, zeigt Implementierungsbeispiele und schildert die bei der Entwicklung aufgetretenen Probleme.

Kapitel 7 – Expertenmeinung enthält eine ärztliche Beurteilung der Programmfunktionalität aufgrund einer Reihe von Anwendungsszenarien.

Kapitel 8 – Zusammenfassung und Ausblick fasst die Ergebnisse der Arbeit zusammen und stellt Anknüpfungspunkte für mögliche Weiterentwicklungen vor.

2. Grundlagen

2.1. Visual Analytics

Der Begriff „Visual Analytics“ wurde im Jahr 2005 von dem Department of Homeland Security, dem Heimatschutzministrium der Vereinigten Staaten, eingeführt [MHK+10]. Mit dem Fokus auf die nationale Sicherheit und angetrieben von den Terroranschlägen vom 11. September 2001 und den Folgen des Hurrikans „Kathrina“ aus dem Jahr 2005 wurde eine Forschungs- und Entwicklungsagenda erstellt („Illuminating the Path: The Research and Development Agenda for Visual Analytics“), die dringend benötigte Forschungsgebiete und -richtungen nennt, um die Herausforderungen bei der Analyse von komplexen, heterogenen Daten in Stresssituationen zu meistern [TC05, TC06]. Das Ziel des dazu gegründeten Nationalen Visualisierungs- und Analytikzentrums (National Visualization and Analytics Center, NVAC) ist es, Analysten dabei zu unterstützen, „Erwartetes herauszufinden und Unerwartetes zu entdecken“ („detect the expected and discover the unexpected“ [TC06]). Seit dem Jahr 2005 ist das wissenschaftliche Interesse an Visual Analytics deutlich und stetig gestiegen. Abb. 2.1 zeigt die Entwicklung der Anzahl an Publikationen zum Thema Visual Analytics bei der „Association for Computing Machinery“[1] und beim „Institute of Electrical and Electronics Engineers“[2]. Ein dem NVAC ähnliches Projekt wurde 2008 in der Europäischen Union unter dem Namen „VisMaster“ gestartet. Das daraus entstandene Buch „Mastering the Information Age - Solving Problems with Visual Analytics“ beinhaltet eine Revision aller Aspekte von Visual Analytics und zeigt Entwicklungsrichtungen und -strategien für die Zukunft auf [KKEM10].

[TC05] definiert Visual Analytics als „die Wissenschaft der analytischen Beweisführung, ermöglicht durch interaktive visuelle Schnittstellen“ („the science of analytical reasoning faclilitated by interactive visual interfaces“ [TC06]). Auch weitere Definitionen von Visual Analytics sehen es als Kombination automatisierter Analyse und interaktiver Visualisierungen, wodurch die Stärken von Menschen mit denen der elektronischen Datenverarbeitung vereint werden [KKEM10, DMKK12]. [KAF+08] nennt als Herausforderung der Visual Analytics die Suche nach dem besten Algorithmus für eine bestimmte Analyse, die Identifikation dessen Schranken und die Erstellung eines flüssigen Übergangs zwischen der Analyse und der interaktiven Visualisierung. Diese zwei Aspekte von Visual Analytics sowie weitere begleitende Funktionen und Konzepte werden in den folgenden Abschnitten beschrieben.

[1] ACM, http://www.acm.org
[2] IEEE, http://www.ieee.org

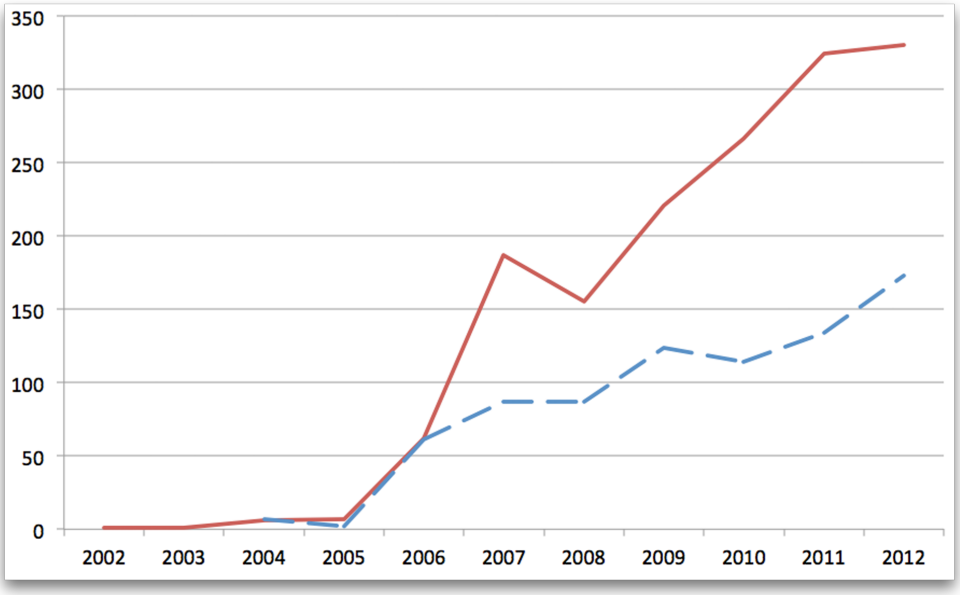


Abbildung 2.1.: Entwicklung der Anzahl an Publikationen zum Thema Visual Analytics in den digitalen Bibliotheken von ACM (rote, durchgezogene Linie) und IEEE (blaue, gestrichelte Linie) in den Jahren 2002 bis 2012.

2.2. Visualisierung

Visualisierungen als graphische Darstellungen konkreter Dinge zum Zweck der Informationsübertragung wurden von Menschen seit Jahrtausenden verwendet. [Dat] stellt eine interaktive Zeitleiste wichtiger Meilensteine in der Visualisierung zur Verfügung. Die ältesten darin vorgestellten Visualisierungen umfassten ausschliesslich Karten, Routenpläne und astrologische Visualisierungen.

Wissenschaftliche Visualisierungen als visuelle Repräsentationen gemessener oder simulierter physikalischer Daten [OL03, KAF+08] sind erst durch die Erfindung des kartesischen Koordinatensystems durch René Descartes im 17. Jahrhundert aufgekommen [Few13]. Im Gegensatz dazu beschäftigt sich die Informationsvisualisierung auch mit abstrakten Daten wie die Visualisierung von Zugriffsstatistiken von Webseiten [OL03, KAF+08].

Bei der Informationsvisualisierung von tabellarischen Daten[3] werden die Spalten „Dimensionen“ oder „Eigenschaften“, die Zeilen „Datensätze“ genannt. Die in einer Zelle der Tabelle gespeicherte Information wird „Wert“ der Eigenschaft für den Datensatz in der entsprechenden Spalte und Zeile genannt. Nach der Art der betrachteten Dimensionen werden quantitative und kategorische Eigenschaften unterschieden [OHS05]. Diese Unterscheidung

[3] Daten, die in Spalten und Zeilen strukturiert sind [OL03].

spiegelt sich in der Auswahl geeigneter Visualisierungsmethoden wieder. Auch die Anzahl der betrachteten Eigenschaften – bei der zwischen univariate und multivariate Daten unterschieden wird [Kei02] – bestimmt die geeigneten Visualisierungsmethoden.

2.2.1. Quantitative und kategorische Eigenschaften

Als quantitative Dimensionen werden die Eigenschaften bezeichnet, die als Wert eine Zahl annehmen können [OHS05]. Quantitative Werte verschiedener Datensätze können miteinander verglichen werden und die Werte mehrerer Datensätze können mit Hilfe verschiedener Funktionen aggregiert weden. Kategorische Dimensionen dagegen erlauben die Einteilung der Datensätze in einzelne Gruppen, die aber in der Regel keine Wertung haben und nicht miteinander verglichen werden können [OHS05]. Geschlecht ist eine solche kategorische Eigenschaft. Es bezeichnet keine Charakteristik, von der eine Gruppe mehr besitzt, als die andere. Qualitative Informationen von Datensätzen, die als „schlecht“, „gut“ oder „sehr gut“ angegeben werden können, sind zwar vergleichbar, ein Computersystem kann diese ohne weiteres Wissen allerdings nicht interpretieren. Ist die qualitative Bedeutung dieser Werte wichtig, so werden diese in der Regel mit Hilfe von Zahlen kodiert. Mit der Kodierung 1 = „schlecht“, 2 = „gut“ und 3 = „sehr gut“ können die angegebenen Werte miteinander verglichen werden.

2.2.2. Univariate und multivariate Visualisierungen

Bei univariaten Visualisierungen wird die Ausprägung einer einzigen Dimension grafisch dargestellt. Balkendiagramme, die das Alter unterschiedlicher Patienten anzeigen, aber auch Kaplan-Meier-Schätzer (2.3.3) für die Visualisierung von Ereigniszeitanalysen gehören zu den univariaten Visualisierungsmethoden. Multivariate Verfahren dagegen zeigen den Zusammenhang zwischen den Werten mehrerer Eigenschaften in einer Visualisierung an. Zu diesen Verfahren gehören beispielsweise die Parallelen Koordinaten [ID90]. Zu den multivariaten Verfahren gehört auch die bivariate Visualisierung, die genau zwei Dimensionen beinhaltet.

Eine Sonderform der multivariaten Visualisierungsmethoden stellen die Geovisualisierungen dar. Hier sind in der Regel zwei Dimensionen mit der Anzeige der geografischen Länge und Breite belegt. Die dargestellten Objekte müssen anhand weiterer Eigenschaften wie Größe, Form, Ausrichtung oder Farbe oder durch separate Grafiken, die mit der angezeigten Karte synchronisiert werden, die Ausprägung der analysierten Werte visualisieren [SMG02]. Manche Visualisierungen integrieren auch die Zeit, meist in der dritten Dimension [KKEM10, AAD+10, TSWS05]. Eine der bekanntesten und als „vermutlich beste statistische Grafik aller Zeiten“ [Tuf83, S. 40] bezeichnete Visualisierung, ist die 1861 entstandene Zeichnung von Charles Joseph Minard, die den Feldzug Napoleons in Russland darstellt (Abb. 2.2). Neben den zwei geografischen Dimensionen (Länge und Breite) und der Zeitdimension werden die Größe und die Marschrichtung der Armee sowie die Temperaturen zu unterschiedlichen Zeitpunkten angezeigt [Tuf83].

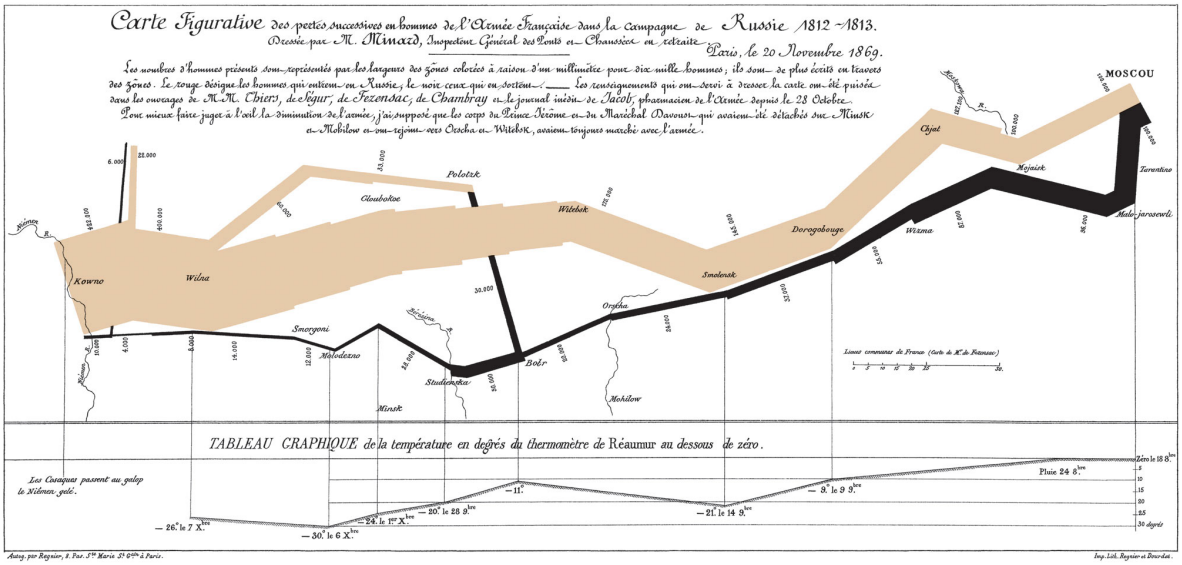


Abbildung 2.2.: Feldzug Napoleons gegen Frankreich. Grafik von Charles Joseph Minard (1861) mit Anzeige von sechs Dimensionen auf einer zweidimensionalen Fläche [Min].

Während Minards Visualisierung eine retrospektive Grafik darstellte, wurde die Zeichnung des Arztes John Snow bei der Suche nach der Ursache für die Choleraepidemie vom September 1854 in London prospektiv verwendet. Snow zeichnete alle Fälle von Choleratote und die Position der Wasserpumpen auf einer Londoner Karte ein (Abb. 2.3). Die Pumpe in der Broad Street wurde durch die Häufung der Krankheitsfälle in dieser Gegend als Ursache der Erkrankungen vermutet. Daraufhin ließ Snow den Schwengel der Pumpe abbauen, was zur Beendung der Epidemie führte [Tuf83, S. 24].

2.2.3. Aggregationsfunktionen

Mit Hilfe von Aggregationsfunktionen wird die Menge der übermittelten Informationen für eine Gruppe von Datensätzen auf einen einzigen Wert reduziert [Agg]. Anstelle eines Balkendiagramms mit den Altersangaben aller Patienten wird beispielsweise nur der Mittelwert des Alters angezeigt. Durch die gewonnene Platzeinsparung können mehrere Gruppen von Patienten übersichtlich miteinander verglichen werden.

Die folgende Aufzählung beinhaltet einige Beispiele von Aggregationsfunktionen, die auf quantitative Eigenschaften angewandt werden können:

Minimum Das Minimum gibt den kleinsten Wert einer quantitativen Eigenschaft zurück.

Maximum Das Maximum gibt den größten Wert einer quantitativen Eigenschaft zurück.

Abbildung 2.3.: John Snows Karte zur Cholera-Epidemie von 1854. Die Todesfälle in Folge von Cholera sowie die Positionen der Wasserpumpen sind auf der Karte eingezeichnet [Sno].

Mittelwert Der Mittelwert (arithmetisches Mittel) wird als das Verhältnis der Summe zur Anzahl aller Werte einer quantitativen Dimension definiert.

$$\mu = \frac{\sum x_i}{N} \tag{2.1}$$

Median Der Median ist der mittlere aus einer Reihe von sortierten Werten. Die betrachtete Eigenschaft besitzt ebensoviele kleinere, wie größere Werte, als der Median. Falls die Anzahl der Werte in der Reihe gerade ist, stellt der Median das arithmetische Mittel der mittleren zwei Werte dar. Der Median ist gegen Ausreißer resistenter als der Mittelwert, da er nicht von den Werten der anderen Zahlen abhängt.

$$x = \begin{cases} x_{\frac{n+1}{2}} & n \text{ ungerade} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n+1}{2}}\right) & n \text{ gerade} \end{cases} \tag{2.2}$$

Varianz Die Varianz ist ein Maß für die Verteilung der Werte einer Dimension um ihren Mittelwert. Sie wird verwendet, um eine genauere Beschreibung der analysierten

Datensätze zu erhalten. Zwei Reihen von Datensätzen können bei unterschiedlich verteilten Elementen den gleichen Mittelwert und Median besitzen:

- $4, 5, 5, 5, 6$ und
- $1, 2, 5, 8, 9$

haben beide den Mittelwert und den Median 5. Durch die Angabe der zwei Varianzen ($0,4$ und 10) wird die unterschiedliche Verteilung der Werte verdeutlicht. Eine beispielhafte Anwendung ist die Analyse der sogenannten „Sozialen Schere“ [Soz], dem wachsenden Vermögensunterschied zwischen Armen und Reichen in einer Gesellschaft im Verlauf der Zeit. Diese wird durch die Analyse des Mittelwertes oder des Medians nicht aufgespürt, da sich diese nicht signifikant ändern. Erst die Analyse der Verteilung der Vermögen deckt dieses Problem auf.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2 \tag{2.3}$$

Standardabweichung Auch die Standardabweichung ist ein Maß für die Verteilung der Werte um ihren Mittelwert. Sie wird als Quadratwurzel der Varianz berechnet und wird in der Maßeinheit der analysierten Daten angegeben. Dadurch wird die Standardabweichung der Varianz oft bevorzugt.

$$\sigma = \sqrt{\sigma^2} \tag{2.4}$$

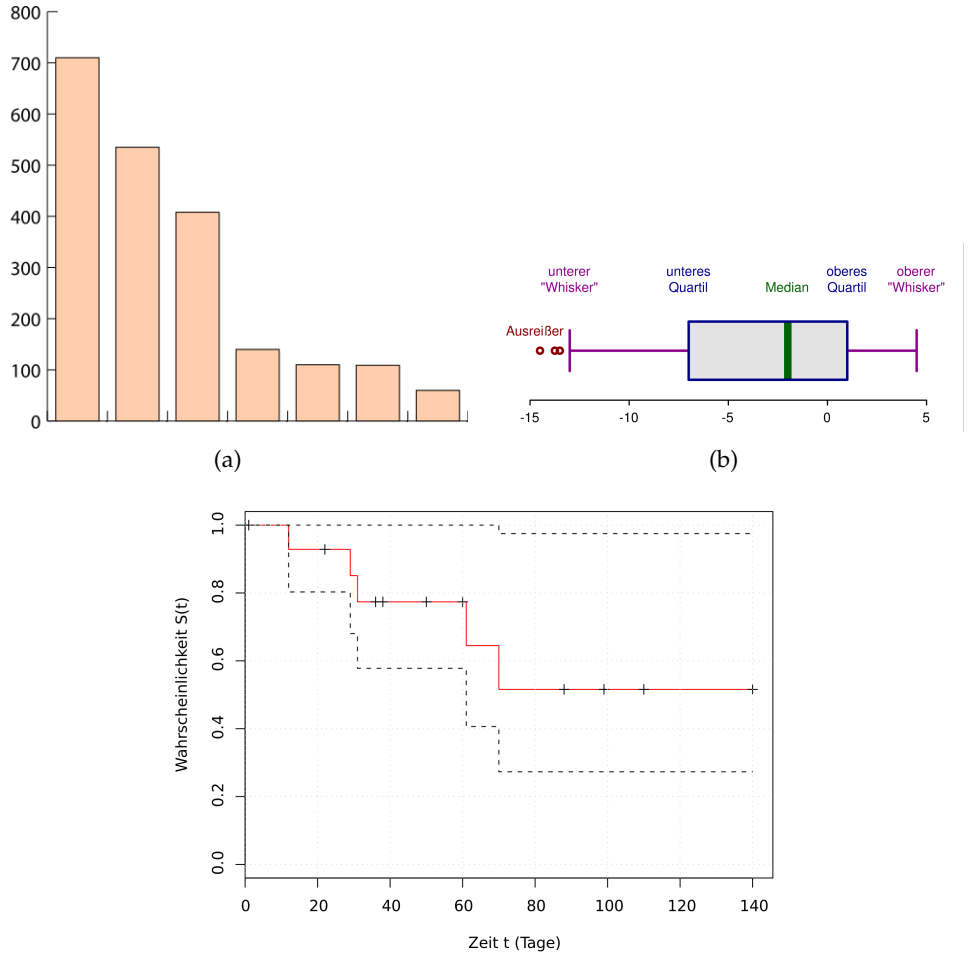
2.3. Visualisierungsbeispiele

Im Folgenden werden einige grundlegende Visualisierungsbeispiele dargestellt. Im Abschnitt 2.6 werden aktuelle Visualisierungsentwicklungen im Bereich der Medizin und der Visual Analytics vorgestellt.

2.3.1. Das Säulendiagramm

Das Säulendiagramm, manchmal auch „Stabdiagramm“ oder bei horizontaler Ausrichtung „Balkendiagramm“ genannt, ist eine univariate Visualisierungsmethode, die sich für die Anzeige von quantitativen Werten eignet. Dabei wird für jeden beobachteten Datensatz eine Säule auf der horizontalen Achse angeordnet (Abb. 2.4(a)). Die Längen der Säulen entsprechen den Werten der betrachteten Dimension der einzelnen Datensätze. Sonderformen, wie das gestapelte Säulendiagramm [Kum05, S. 255], erweitern das Säulendiagramm zu einer bi- oder multivariaten Visualisierungsmethode. [KHDH02] stellt eine Methode vor, die die multivariate Visualisierung großer Datenmengen ermöglicht.

Säulen- und Balkendiagramme eignen sich nicht für die Darstellung kategorischer Eigenschaften, da ihre Länge eine wertmäßige Ordnung der Kategorien suggeriert, die zu falschen Schlussfolgerungen führen kann (siehe Abschnitt 2.2.1).



(a) (b)

(c)

(d) (e)

Abbildung 2.4.: Beispiel unterschiedlicher Visualisierungsmethoden: (a) Säulendiagramm (verändert übernommen aus [Bar]); (b) Boxplot-Diagramm [Box]; (c) Kaplan-Meier-Schätzer [Kap]; (d) Kreisdiagramm (verändert übernommen aus [Pie]); (e) Ringdiagramm (verändert übernommen aus [Pie]).

2.3.2. Das Boxplot-Diagramm

Das Boxplot-Diagramm ist eine weitere univariate Visualisierungsmethode, das die Verteilung der Werte einer Eigenschaft durch die „Fünf-Zahlen-Zusammenfassung“ visuell darstellt [Poto6]. Dabei werden alle vorkommenden Werte einer Eigenschaft in vier gleich große Mengen – „Quartile“ genannt – aufgeteilt. Die zwei inneren Quartile, die die mittleren 25% bis 75% der Werte beinhalten, bestimmen den „Interquartilsabstand“ – die mittleren 50% der Daten – und werden durch ein Rechteck (engl: „Box“) dargestellt. Eine Trennlinie innerhalb dieses Rechteckes verdeutlicht die Position des Medians. Die zwei Antennen (engl: „Whiskers“) reichen bis zu dem minimalen bzw. dem maximalen Wert der Daten. Bei der Analyse der Verteilung mehrerer Gruppen eignet sich das Boxplot-Diagramm durch seine „visuelle Eleganz“ besser, als ähnliche Visualisierungen [WPK89].

Eine Sonderform des Boxplot-Diagramms beschränkt die Länge der Antennen auf das 1,5-fache des Interquartilabstandes. Werte außerhalb dieses Bereiches werden als Ausreißer bezeichnet und gesondert dargestellt (Abb. 2.4(b)). [Poto6] und [Kamo8] stellen weitere Sonderformen des Boxplot-Diagramms vor, die mehr Informationen über die betrachtete Eigenschaft beinhalten oder multivariate Visualisierungen ermöglichen.

2.3.3. Der Kaplan-Meier-Schätzer

Die Ereigniszeitanalyse (auch bekannt als „Überlebenszeitanalyse“) untersucht die Zeit bis zum Auftreten eines bestimmten Ereignisses in einer Menge von Daten [Zhao5]. In medizinischen Studien ist das untersuchte Ereignis häufig der Tod der Patienten, es kann aber auch als Heilung oder Rückfall definiert werden. Besonders in medizinischen Studien kann die Beobachtung der untersuchten Patienten häufig nicht bis zum Eintreten des Ereignisses bei allen Patienten verfolgt werden. Ist bis zum Ende der Beobachtung das Ereignis noch nicht eingetreten, spricht man von „rechts zensierten Datensätzen“. Als „links zensierte Datensätze“ werden die Datensätze bezeichnet, bei denen das Ereignis zwar eingetreten ist, der Zeitpunkt des Eintretens aber vor der Beobachtung liegt.

Der Kaplan-Meier-Schätzer ist eine Methode der Ereigniszeitanalyse, die die Wahrscheinlichkeit, dass das beobachtete Ereignis innerhalb eines Zeitintervalls nicht eintritt, visualisiert. In Form einer Treppenfunktion wird dabei die Entwicklung des Anteils der beobachteten Datensätze, bei denen das Ereignis nicht eingetreten ist, im Verlauf der Zeit visualisiert (Abb. 2.4(c)). Für die Erstellung der Grafik wird die Zeitdauer bis zum Eintreten des Ereignisses für jeden Datensatz sowie die Angabe, ob der Datensatz zensiert ist, benötigt. Da der Kaplan-Meier-Schätzer das Eintreten eines einzigen Ereignisses visualisiert, gehört er zu den univariaten Visualisierungen. Im Unterschied zu naiven Methoden, die die zensierten Daten von der Beobachtung ausschliessen, betrachtet der Kaplan-Meier-Schätzer diese vor Eintritt der Zensierung als Teil der Risikomenge.

2.3.4. Das Kreisdiagramm

Das Kreisdiagramm (manchmal auch „Kuchen-“ oder „Tortendiagramm“ genannt) wurde erstmals im Jahr 1801 von William Playfair verwendet. Es hat die Form eines Kreises, der in mehrere Kreissektoren aufgeteilt ist. Die einzelnen Kreissektoren und ihre Größen stellen einzelne Teilmengen und deren Anteil dar (Abb. 2.4(d)). Das Kreisdiagramm stellt somit eine univariate Visualisierung dar. Eine Sonderform von Kreisdiagrammen sind Ringdiagramme, deren Fläche nicht bis zur Mitte des Kreises ausgefüllt ist (Abb. 2.4(e)). Kreisdiagramme können durch Verwendung des Kreisradiuses als quantitatives Maß zu einer bivariaten Visualisierung erweitert werden.

2.4. Data Mining

Data Mining ist der Prozess der Entdeckung auffälliger Muster in Datenmengen. Es grenzt sich von der Statistik insofern ab, als dass nur neue, interessante Informationen, die im Voraus nicht vermutet werden konnten, gesucht werden [Milo8]. Somit steht bei Data Mining die Generierung anstatt der Bestätigung von Hypothesen im Mittelpunkt. Data Mining-Algorithmen können eine große Menge an Muster generieren, so dass ein Maß für die Interessantheit der entdeckten Muster entwickelt werden muss, um den Benutzer bei der Analyse zu unterstützen [Milo8].

Zu den Data Mining-Methoden gehören die Klassifikation, die Assoziationsanalyse, die Klassifikation und Prognose, die Clusteranalyse und die Ausreißererkennung [Milo8]. Die letzten beiden Methoden werden in den folgenden Abschnitten erläutert.

2.4.1. Ausreißererkennung

[AY01] definiert Ausreißer (engl: „outlier detection“) als „Datenpunkte, die von dem Rest der Daten, basierend auf einem bestimmten Maß, sehr unterschiedlich sind“. Obwohl Ausreißer auf fehlerhafte Beobachtungen zurückführbar sein können, können sie auch interessante Erkenntnisse über seltene beobachtete Fälle vermitteln. Somit sollte nicht unbedingt eine der zwei Extremfälle – das Ein- oder das Ausschließen der Ausreißer bei der Analyse – verfolgt werden. Mögliche Alternativen sind Methoden, die Ausreißer zwar einschließen, aber ihren Einfluss auf die Analyse minimieren [BL94, S. 3]. Durch die Kombination von Data Mining und Visualisierung in Visual Analytics wird der Unsicherheit bei dem Umgang mit Ausreißern durch die Möglichkeit des Benutzers, die Ausreißer zu analysieren und zu bewerten, entgegengewirkt.

Ein Spezialfall von Ausreißer sind räumliche Ausreißer, die zwar in den „nicht-räumlichen“ Eigenschaften konsistent zu den restlichen Datensätzen sein können, sich aber von den Datensätzen in ihrer räumlichen Umgebung bedeutsam unterscheiden [SLZ03, Milo8].

Methoden für die Ausreißererkennung großer Datenmengen und für die multivariate Ausreißererkennung werden in [SLZ03] , [AY01], [CSM02] und [BG05] vorgestellt.

2.4.2. Clusteringanalyse

Die Clusteranalyse befasst sich mit der Einteilung der Datensätze in Ähnlichkeitsgruppen. Diese Ähnlichkeitsgruppen sind nicht vordefiniert, sondern werden von dem Analysealgorithmus mitentwickelt. Algorithmen für räumliches Clustering verwenden bei der Bildung der Cluster räumliche Beziehungen zwischen den Datensätzen [Mil08].

2.5. Das Visual Analytics Mantra

Die Zusammenarbeit der zwei vorgestellten Bestandteile, Visualisierung und Data Mining, bestimmen den Arbeitsablauf in Visual Analytics-Anwendungen. Aufbauend auf das Visual Information Seeking Mantra [Shn96], das den zyklischen Ablauf bei der visuellen Datenanalyse beschreibt, wird in [KMS+08] das „Visual Analytics Mantra“ vorgeschlagen.

Das Visual Information Seeking Mantra basiert auf den folgenden drei Schritten [Shn96]:

Overview First: Der Benutzer soll eine Übersicht der vorhandenen Daten bekommen. Zu den Strategien bei der Implementierung dieser Funktion gehören der Fischaugeneffekt und das „Focus+Context“-Prinzip [Fco02], das neben der Detailsansicht eine Übersichtsansicht aller Daten zeigt, die der Orientierung des Benutzers dient. Die Übersicht beinhaltet ein bewegliches Fenster, das den sichtbaren Ausschnitt der Daten darstellt. Außerdem soll die Übersicht das Schwenken innerhalb der Daten erlauben.

Filter and Zoom: Das Herausfiltern unerwünschter Elemente durch dynamische Abfragen erlaubt die Konzentration auf die Interessensschwerpunkte. Diese Konzentration soll auch durch die Möglichkeit, Daten zu vergrößern oder zu verkleinern, unterstützt werden.

Details on Demand: Nach der Fokussierung bestimmter Daten sollen weitere Informationen abgerufen werden können. Das gewöhnliche Vorgehen dabei ist das Anklicken der jeweiligen Datenelemente.

Durch die Integration der Analysefunktionen wurde das Visual Information Seeking Mantra zum Visual Analytics Mantra erweitert, das aus folgenden Schritten besteht:

- Analyse First
- Show the Important
- Zoom, Filter and Analyse Further
- Details on Demand

Aufgrund der umfangreichen Daten, in denen der Benutzer möglicherweise keinen Überblick hat, soll der erste Schritt des Arbeitsablaufes eine automatische Analyse der geladenen Daten sein. Anders als in dem Visual Information Seeking Mantra, werden dem Benutzer nicht die kompletten Daten in einer Übersicht angezeigt, sondern er wird auf das Wichtige geleitet, das von den Analysen als solches aufgedeckt wird. Ab diesem Punkt beginnt der

normale Arbeitsablauf der Visualisierung mit Zoom, Filter und Details on Demand. Der Benutzer hat weiterhin die Möglichkeit, Analysen selbst anzustoßen, die Parameter dieser Analysen anzupassen und dadurch den Kreislauf des Visual Analytics Mantras zu schließen. Um die Flexibilität des Arbeitsablaufes zu erhöhen soll der erste Schritt („Analyse first“) deaktivierbar sein und der Benutzer soll dadurch selbst die Richtung der Datenexploration bestimmen können.

2.6. Verwandte Arbeiten

Aufgrund der großen Menge an Daten, die im medizinischen Bereich anfallen und aufgrund der Heterogenität der gespeicherten Daten, wurden schon vor der Etablierung des Begriffes Visual Analytics, Systeme zur Visualisierung von historischen medizinischen Daten entwickelt. LifeLines [PMS+98] ermöglichte 1998 die Visualisierung von Ereignissen in der Krankengeschichte eines Patienten. Das System ermöglichte die Suche nach bestimmten Ereignissen und die Anzeige dazugehörender Details aus den Patientenakten. Der Nachfolger LifeLines2 [WPQ+08] ermöglicht den Vergleich der Krankengeschichte mehrerer Patienten miteinander. Durch die Ausrichtung der Zeitachse mehrerer Patienten nach bestimmten Ereignissen können Muster in dem weiteren Verlauf der Krankheit analysiert werden. CareCruiser [GAK+11] ermöglicht die Anzeige der Behandlungen und deren Folgen in der Krankengeschichte eines Patienten. Interessante Werteentwicklungen werden farblich gekennzeichnet. Der Vergleich mehrerer Patienten ist ebenfalls möglich.

Durch den Aufbruch von Visual Analytics in den letzten Jahren sind auch im medizinischen Bereich und speziell in der Krebsforschung viele Visual Analytics-Systeme entstanden. [PMZ+08] nutzt Visual Analytics bei der Klassifizierung von Mammakarzinomen und verwendet dabei Informationen aus unterschiedlichen Quellen wie Pathologieberichte, zwei- und dreidimensionale Bilder und prognostische Informationen aus einer Tumordatenbank. VisCareTrails [LHFS11] unterstützt die Analyse von Zusammenhängen zwischen unterschiedlichen Eriegnissen und bietet zu den dargestellten Visualisierungen eine Reihe von Statistiken. [MDR+10] implementiert eine geografische Darstellung von Statistiken zu Krebserkrankungen. Dabei konzentriert sich die Arbeit auf die Lösung der statistischen Probleme, die durch eine zu kleine Anzahl an Datensätzen in einer geografischen Region entstehen. Außerdem werden nur demografische Daten der Patienten bei der Analyse der Verteilung berücksichtigt.

Der in dieser Arbeit entwickelte Prototyp soll sich durch die geografische und zeitabhängige Darstellung der Visualisierung von den hier beschriebenen Systemen abgrenzen. Dadurch soll die Entdeckung von geografischen und zeitlichen Zusammenhängen in allen erfassten Eigenschaften der Krankheitsverläufe gefördert werden. Desweiteren soll durch die Implementierung des Visual Analytics Mantras eine nahtlose Kombination von Visualisierungs- und Data Mining-Methoden ermöglicht werden.

3. Datenquelle

Der Onkologische Schwerpunkt Stuttgart (OSP) ist ein Kooperationsforum von dreizehn Stuttgarter Krankenhäusern, das sich zur Aufgabe gemacht hat, „die Situation von Tumorkranken ständig weiter zu verbessern." [OSPa]. Zu den Maßnahmen, die dieses Ziel ermöglichen sollen, gehört ein Krebsregister, das 1986 geplant und seit 1988 geführt und weiterentwickelt wird [OSPb].

Das Krebsregister des OSP Stuttgart umfasste im Jahr 2010 106.500 Patientendaten zu 84 Tumorerkrankungen. Die Daten werden von 15 in Vollzeit angestellten Dokumentaren/innen und Dokumentationsassistenten/innen gepflegt und stammen ausschließlich aus den 13 Stuttgarter Mitgliedskrankenhäusern.

Die am häufigsten in den Datensätzen vorkommende Tumorerkrankung ist der Brustkrebs. 1992 wurde entschieden, dass Brustkrebsdatensätze mit erster Priorität zu dokumentieren sind. Die Brustkrebsdatenbank des OSP Stuttgart enthält Daten über 19.946 Patienten, deren Diagnosezeitpunkt zwischen den Jahren 1988 und 2010 liegt [OSPb].

Der vorliegenden Arbeit steht ein Großteil der Datensätze der Brustkrebsdatenbank als csv-Datei[1] (im Folgenden „Brustkrebsdatei" genannt) zur Verfügung. Die enthaltenen Datensätze sind anonymisiert, so dass keine Rückschlüsse auf betroffene Patienten gemacht werden können. Nähere Informationen zu den Inhalten dieser Brustkrebsdatei werden in den folgenden Kapiteln beschrieben.

3.1. Datenumfang

Die zur Verfügung stehende Brustkrebsdatei beinhaltet anonymisierte Informationen zu 16.341 Patienten, die an Brustkrebs erkrankt sind. Das Diagnosedatum der Patienten liegt zwischen dem 02.01.1989 und dem 02.01.2009. Aus der Brustkrebsdatenbank wurden Erkrankungsfälle bei Männern und nichtinvasive Fälle[2] entfernt.

Die Herkunft der erfassten Patienten verteilt sich auf ganz Deutschland und sogar bis ins Ausland (14 Datensätze von Patienten aus Belgien, den Niederlanden, der Schweiz, Liechtenstein, Italien, Schweden, Irland, Türkei, Ägypten, Russland, Kasachstan und den USA). Da alle Patienten in den Stuttgarter Krankenhäusern behandelt wurden, nimmt die Anzahl der erfassten Datensätze mit der Entfernung zur Stadt Stuttgart deutlich ab. Außerdem ist

[1] „Comma-separated values"
[2] Fälle im Frühstadium, bei denen es noch nicht zu einer Diagnose gekommen ist.

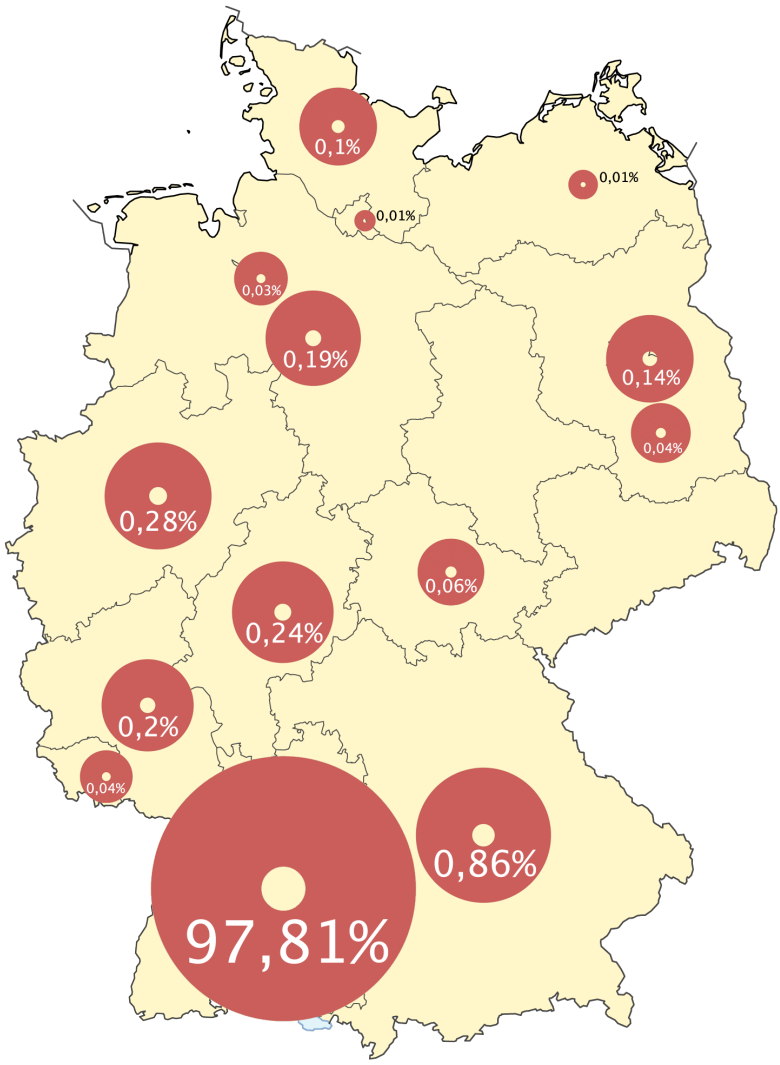


Abbildung 3.1.: Verteilung der Brustkrebsdatensätze innerhalb Deutschlands (logarithmische Skala).

für Patienten aus entfernteren Gebieten eine schlechtere Qualität der gesammelten Daten erkennbar (siehe Kapitel 7). Abb. 3.1 zeigt die Verteilung der Patienten nach Wohnort innerhalb Deutschlands. Innerhalb von Baden-Württemberg stammen die meisten Patienten aus Stuttgart (41,3%), gefolgt von dem Rems-Murr-Kreis (18,7%) und den Landkreisen Ludwigsburg (13,8%) und Esslingen (8,4%).

3.2. Datenqualität

Die primäre Bezugsquelle für Patientendaten sind die Krankenhausinformationssysteme (KIS) der beteiligten Krankenhäuser. Hinzu kommen Informationen aus dem Entlassbrief, aus zusätzlichen Behandlungsakten, von niedergelassenen Ärzten und aus Patientenfragebögen. Die Daten werden dem OSP von den Krankenhäusern im Behandlungszusammenhang übertragen. Dadurch wird sichergestellt, dass neue Informationen zu den richtigen Datensätzen hinzugefügt und Doubletten vermieden werden [OSPb].

Seit 2003 können in Baden-Württemberg Vitaldaten der Patienten mit den Melderegistern der Einwohnermeldeämter abgeglichen werden. Somit sind für 98% der Datensätze Aussagen zum Vitalstatus des Patienten möglich. Das erhöht die Qualität der Datensätze, da die Überlebenszeit als das wichtigste Qualitätskriterium bei der Behandlung von Tumorerkrankungen angesehen wird [OSPb].

3.3. Fehlende und fehlerhafte Daten

Trotz der priorisierten Dokumentation beinhaltet die Brustkrebsdatenbank sowohl fehlende, als auch fehlerhafte Daten. Folgende Daten sind beispielsweise davon betroffen:

- die Spalte „RFS“ (rezidivfreies Überleben) beinhaltet den Zeitraum zwischen der Diagnose und dem erneuten Auftreten der Krankheit nach einer Therapie. Dieser Wert kann definitionsbedingt nicht kleiner als Null und nicht größer als die Überlebenszeit des Patienten (Spalte „OVS“) sein. In 39 Fällen in der Brustkrebstabelle ist der RFS-Wert kleiner als Null, in 20 Fällen ist er größer als der OVS-Wert.
- zu drei der 4639 verstorbenen Patienten fehlt das Sterbedatum.
- in 637 Datensätzen fehlt der T-Wert, in 989 der N-Wert und in 892 der M-Wert der TNM-Klassifikation[3].

Einige der fehlenden Daten sind nicht als Fehler zu interpretieren. Bei den cerbB2-Werten in der Spalte „ihc“ ist die regelmäßige Erfassung in den Datensätzen erst bei Patienten mit Diagnosedatum ab den Jahren 2000-2002 zu beobachten. Das lässt auf die wissenschaftliche Entwicklung schließen, die erst zu diesem Zeitpunkt die Erfassung dieses Wertes als therapierelevant betrachtet hat. Dieser Wert ist bei Datensätzen mit einem Diagnosedatum vor dem Jahr 2000 nur in seltenen Fällen vorhanden.

Manche Spalten der Brustkrebsdatei eignen sich nicht für eine Analyse, da die eingetragenen Daten nicht standardisiert sind und somit eine zu große Anzahl an unterschiedlichen Werten beinhalten. Die Spalte „ikbefall_1“ beispielsweise beinhaltet 247 unterschiedliche Werte, von denen die meisten nur einmalig vorkommen. [CM02] führt dieses Problem auf die

[3] Die TNM-Klassifikation ist eine gängige Einteilung von Tumoren [Bur04, FKG+10] und spielt eine entscheidende Rolle bei der Auswahl der Therapie.

fehlende kanonische Form medizinischer Begriffe und die daraus entstehenden alternativen Schreibweisen zurück. Solche Spalten sollten von dem Import der Daten ausgeschlossen werden, da sie sonst zu verfälschten Ergebnissen führen können.

3.4. Aufbau der Brustkrebsdatei

Die Brustkrebsdatei ist eine Tabelle, in der jede Zeile einen Erkrankungsfall darstellt. Die Spalten beinhalten die einzelnen Informationen zu den Erkrankungsfällen. Die Tabelle ist nicht normalisiert, somit werden mehrfache, gleiche Informationen über einen Patienten – wie Therapiedaten – in separaten Spalten aufgeführt.

Die Spaltennamen der Brustkrebsdatei sind in den meisten Fällen abgekürzt und erlauben keine Rückschlüsse auf den Inhalt der jeweiligen Spalte. Während der Inhalt der Spalten „geschl“ (Geschlecht), „stand“ (Datum der letzten Beobachtung des Patienten) und „brusterh“ (Brusterhaltung) beim Betrachten der gespeicherten Daten erahnt werden kann, muss die Bedeutung der Spalten „abgru“ (Abbruchgrund) oder „ihc“ (cerbB2-Wert) in der mitgelieferten Spaltennamendatei nachgeschlagen werden. Diese Datei beinhaltet allerdings nicht die Namen aller Spalten, so dass hier die Hilfe eines Experten benötigt wird.

3.5. Datenaufbereitung

Die Brustkrebsdatei muss vor dem Importieren in die Anwendung bearbeitet werden, da sie zusätzlich zu den inhaltlichen Fehlern auch syntaktische Fehler enthält, die das korrekte Einlesen in der Anwendung verhindern.

Für die benötigten Änderungen wurde die Software TextWrangler[4] unter dem Betriebssystem Mac OS X verwendet. TextWrangler ist eine kostenlose Textverarbeitungssoftware, deren Such- und Ersetzenfunktionen reguläre Ausdrücke unterstützen.

Im Folgenden werden die syntaktischen Unregelmäßigkeiten der Brustkrebsdatei vorgestellt, die angepasst werden müssen. Tabelle 3.1 zeigt die regulären Ausdrücke, die für die Suche und das Ersetzen benötigt werden.

1. Die Zeilen enden mit vier Semikolons. Wenige Hundert davon enden sogar mit drei oder weniger Semikolons. Bei insgesamt mehr als 16.000 Zeilen kann man davon ausgehen, dass es sich dabei um Fehler handelt. Beim Einlesen der Datei reicht der Newline-Charakter als Zeilentrennzeichen aus, so dass alle Semikolons am Ende der Zeilen gelöscht werden können.
2. Jede Zeile ist in Anführungszeichen eingeschlossen. Diese Anführungszeichen haben keine Bedeutung und können gelöscht werden.

[4] http://www.barebones.com/products/textwrangler/

3. Alle Umlaute sind falsch kodiert. Mit einem geeigneten Suchausdruck wird nach Sonderzeichen in der Textdatei gesucht. Aus dem Kontext der gefundenen Zeichen können die falsche Kodierung der Umlaute erkannt und die entsprechenden Ersetzungen durchgeführt werden.
4. Kommata werden in der csv-Datei als Feldtrennzeichen verwendet, können aber auch innerhalb von Textfeldern auftreten. Um Verwechslungen beim Umgang mit den Daten zu vermeiden, wird als neues Trennzeichen der Semikolon ausgewählt. Dafür müssen zuerst die Semikolons innerhalb von Textfeldern entfernt werden. Bei der Suche nach vorkommenden Semikolons innerhalb der gesamten Datei fällt auf, dass diese nur in einem Textfeld vorkommen und problemlos durch Kommata ersetzt werden können.
5. Da Textfelder von Anführungszeichen umschlossen werden und alle anderen Datentypen (Zahlen, Datumsangaben und boolsche Werte) keine Kommata enthalten, können alle Kommata außerhalb von Anführungszeichen durch Semikolons ersetzt werden. Diese Änderung kann allerdings nicht durch Ersetzen auf der Basis von regulären Ausdrücken umgesetzt werden, so dass dafür ein eigenständiges Programm erstellt wurde, das im Abschnitt 6.4.1 vorgestellt wird.
6. Im letzten Schritt werden die Anführungszeichen um die Textfelder in der gesamten Datei gelöscht.

Zusätzlich zu diesen syntaktischen Anpassungen wurde die Spalte der behandelnden Klinik bearbeitet. Die „klinr“-Spalte kodiert die behandelnde Klinik bis auf die Ebene der behandelnden Abteilungen. Da für die Auswertung nur die Kliniken miteinander verglichen werden sollen, wurde die Klinik-Spalte auf folgende Werte reduziert:

- 01 – Diakonieklinikum Stuttgart[5]
- 02 – Marienhospital[6]
- 03 – Robert-Bosch-Krankenhaus[7]
- 04 – Katharinenhospital Stuttgart[8]
- 05 – Bürgerhospital[9]
- 06 – Krankenhaus Bad Cannstatt[10]
- 10 – Frauenklinik[11]
- 12 – Klinik Schillerhöhe[12]

[5] http://www.diakonie-klinik.de
[6] http://www.marienhospital-stuttgart.de
[7] http://www.rbk.de
[8] http://www.klinikum-stuttgart.de/kh
[9] http://www.klinikum-stuttgart.de/bh
[10] http://www.klinikum-stuttgart.de/kbc
[11] http://www.klinikum-stuttgart.de/frauenklinik
[12] http://www.rbk.de/standorte/klinik-schillerhoehe.html

- 21 – Bethesda Krankenhaus[13]
- 23 – Karl-Olga-Krankenhaus[14]
- 26 – Krankenhaus vom Roten Kreuz[15]
- 27 – St. Anna Klinik[16]

	Suchen	Ersetzen	Vorkommen
1	;;;;\r	\r	16020
	;;;\r	\r	267
	;;\r	\r	44
	;\r	\r	9
2	^‘‘(.*)‘‘$	\1	16342
3	[^A-Z0-9‘‘,\.,!_/÷ ()$:\r;ÿ+]	–	–
	¸	ü	8131
	‰	ä	39540
	÷	Ö	84
	fl	ß	16686
	ˆ	ö	7735
	f	Ä	15
	<	Ü	480
4	(‘‘‘plu[^,;]*?);([^,;]*‘‘‘,)	\1,\2	284
	‘‘;	‘‘,	48
	;‘‘	,‘‘	48
	;	,	10
5	Eigene Anwendung csvConverter		
6	‘‘		7915750

Tabelle 3.1.: Vorabänderungen in der Brustkrebsdatei.

[13] http://www.bethesda-stuttgart.de
[14] http://www.karl-olga-krankenhaus.de
[15] http://www.rkk-stuttgart.de
[16] http://www.st-anna-klinik.de

4. Eigenes Konzept

In diesem Kapitel wird das Lösungskonzept erklärt, das auf den Informationen aus dem Grundlagenkapitel basiert und dessen detaillierte Implementierung in Kapitel 5 besprochen wird.

Aufbauend auf den Abschnitten 2.1 und 2.5 soll die erstellte Anwendung sowohl Visualisierungs- als auch Data Mining-Funktionen enthalten und die vier Schritte des Visual Analytics Mantras unterstützen.

4.1. Analyse first

Nach dem Starten der Anwendung und dem Importieren von Daten sollen Data Mining-Analysen gestartet werden. Der erstellte Prototyp soll exemplarisch eine Ausreißeranalyse durchführen. Der „Analyse first"-Schritt soll vom Benutzer deaktivierbar sein, so dass die Datenexploration vom Benutzer gesteuert werden kann.

4.2. Show the Important/Overview first

Auffälligkeiten, die durch die Data Mining-Algorithmen entdeckt werden, sollen dem Benutzer signalisiert und in einer Liste eingetragen werden. Diese Liste soll nach dem Maß der Auffälligkeit (beispielsweise nach dem Abstand bei der Ausreißererkennung) sortiert werden. Der Benutzer hat die Möglichkeit, einzelne dieser Auffälligkeiten auszuwählen. Er wird dadurch zu einer passenden Ansicht der Daten geführt.

Wurde der Schritt „Analyse first" übersprungen, so deckt sich der aktuelle Schritt mit dem Schritt „Overview first" des Visual Information Seeking Mantras. Der Benutzer soll dann die Möglichkeit haben, sich eine Übersicht über die Daten zu verschaffen. Dazu gehört ein Navigationsfenster, in dem der Benutzer seine Position innerhalb der Daten erkennen kann. Ebenfalls soll der Benutzer frei durch die kompletten Daten navigieren können.

4.3. Die w^3-Prämisse

Die zur Verfügung stehenden Daten enthalten zusätzlich zu den demografischen Informationen und zu den Informationen über den Krankheitsverlauf der Patienten auch Angaben über die Herkunft, die behandelnde Klinik und die Zeitpunkte verschiedener Ereignisse im Verlauf der Krankheit. Um eventuelle geografische Muster erkennen, die behandelnden Kliniken miteinander vergleichen und Qualitätsänderungen im Verlauf der Zeit erkennen zu können, sollen geografische und zeitliche Eigenschaften getrennt von den restlichen Dimensionen betrachtet und damit interagiert werden. In [LAMF05] wird die sogenannte w^3-Prämisse vorgestellt. Sie bezeichnet das Vorhandensein der Eigenschaften „Was?“ („What?“), „Wann?“ („When?“) und „Wo?“ („Where?“) in den betrachteten Datensätzen. [Gap] zeigt eine Visualisierung in der diese drei Aspekte berücksichtigt werden. Darin werden unterschiedliche Länder anhand zweier frei auswählbaren, quantitativen Eigenschaften im Verlauf der Zeit miteinander verglichen (Abb. 4.1). Da die zur Verfügung stehenden Daten der Brustkrebsdatei unterschiedliche Eigenschaften zu allen drei Komponenten der w^3-Prämisse beinhalten, werden die Visualisierungen und die Interaktionsmöglichkeiten ebenfalls in diese drei Bereiche getrennt.

Was?

Unter diesen Bereich fallen die meisten Dimensionen der Daten der Brustkrebsdatei. Bei der verwendeten Visualisierung sollen beliebige Eigenschaften der Datensätze angezeigt werden können.

Wann?

Die Änderungen der Qualität der medizinischen Versorgung im Laufe der 20-jährigen Laufzeit der Studie dürften sich auch in den Daten der Brustkrebsdatei widerspiegeln. Um Entwicklungen von Eigenschaften im Verlauf der Zeit erkennen zu können, soll es möglich sein, das betrachtete Zeitintervall zu beschränken und unterschiedliche Zeitintervalle miteinander zu vergleichen.

Wo?

Die Informationen zur Adresse und zur behandelnden Klinik der Patienten sollen verwendet werden, um die Datensätze auf einer Karte zu positionieren und den Benutzer bei der Erkennung geografischer Muster und bei dem Vergleich unterschiedlicher Kliniken zu unterstützen.

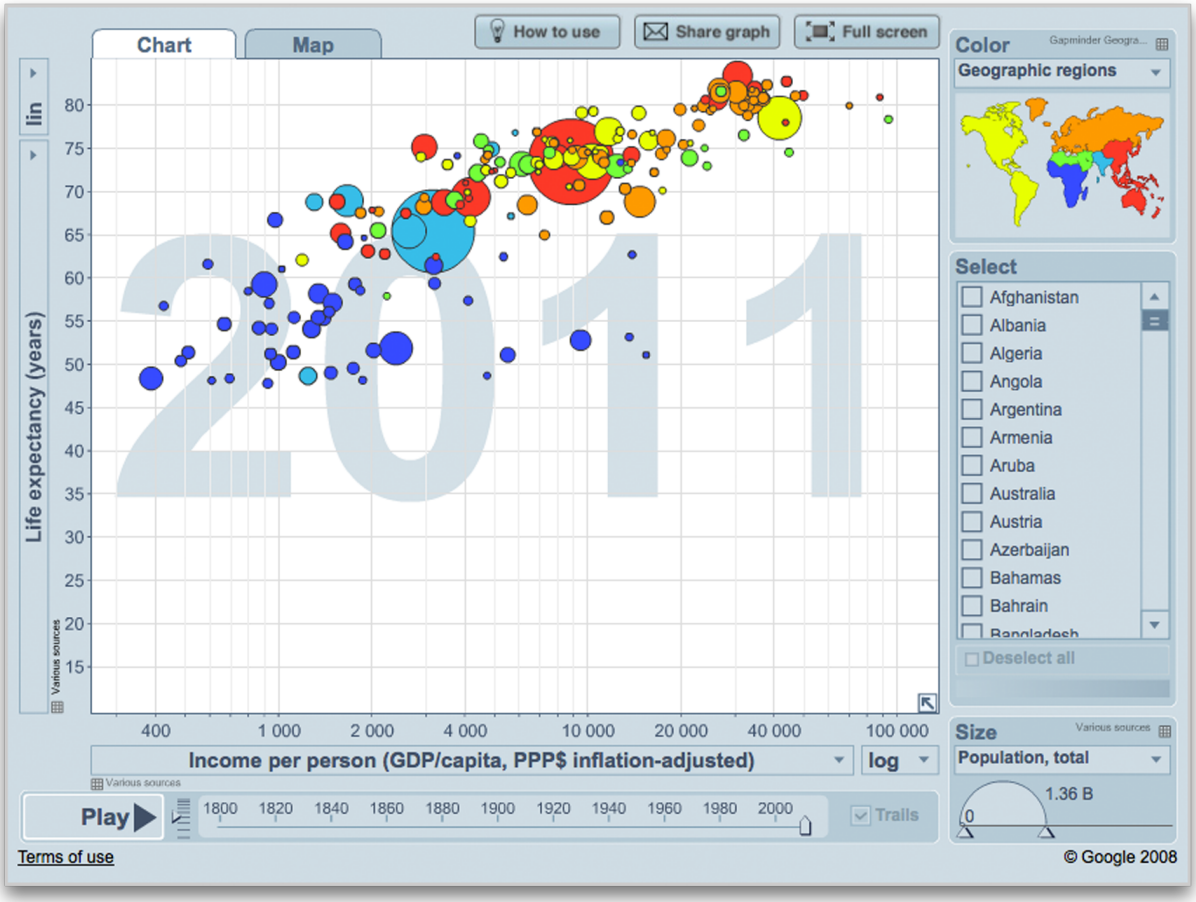


Abbildung 4.1.: „Wealth & Health of Nations“, Grafik erstellt mit GapMinder [Gap].

4.4. Filter, Zoom and Analyse Further

Für die Umsetzung dieses Schrittes soll die Möglichkeit bestehen, mehrere Datensätze interaktiv zu gruppieren bzw. wieder zu trennen, um sowohl detaillierte, als auch übersichtliche Ansichten zu ermöglichen. Dazu sollen mehrere geografische Detailstufen ([LAMF05]) durch Gruppierung nach Adresse und behandelnde Klinik der Patienten zur Verfügung gestellt werden. Die Kartenanzeige soll schwenk- und zoombar sein, um die Exploration dieser Daten zu ermöglichen. Außerdem soll auch die zeitliche Komponente interaktiv zoom- und verschiebbar sein.

Darüberhinaus sollen Funktionen für das manuelle Veranlassen weiterer Ausreißererkennungen durch den Benutzer zur Verfügung gestellt werden. Der Benutzer soll hier den gewünschten Umfang der Analyse anhand verschiedener Parameter anpassen können.

4.5. Details on Demand

Durch Interaktion mit der verwendeten Visualisierung soll der Benutzer weitergehende Informationen über die ausgewählte Gruppe erfahren. Kurzinformationen, die der Orientierung dienen (beispielsweise der Name der Gruppe), sollen als Tooltips beim Überfahren der Grafiken angezeigt werden. Weitergehende Informationen sollen durch Anklicken der jeweiligen Grafiken eingeblendet werden. Eine Such- und Filterfunktion soll das Finden der gewünschten Gruppen ermöglichen.

Der Benutzer soll außerdem die Möglichkeit haben, die einer Grafik zugrundeliegenden Daten zu visualisieren und diese gegebenenfalls für weitere Analysen in Statistikprogramme zu exportieren.

4.6. Datenimport und Persistenz

Bedingt durch die große Anzahl an – teils redundanten oder nicht verwendbaren – Spalten der Brustkrebsdatei und des damit verbundenen Leistungseinbruchs soll beim Importieren der Daten die Möglichkeit gegeben sein, einzelne Spalten von dem Importvorgang auszuschließen. Darüber hinaus soll das Speichern der geladenen Daten und der durchgeführten Analysen möglich sein, um die Wiederholung schon durchgeführter automatischer Analysen zu vermeiden.

5. Eigene Umsetzung

In diesem Kapitel wird die entwickelte Software beschrieben. Diese basiert auf dem Konzept, das in Kapitel 4 beschrieben wurde. Der Reihe nach werden alle Elemente der Benutzeroberfläche, deren Zusammenarbeit und Hintergründe erklärt. Angelehnt an dem Titel dieser Arbeit wird die erstellte Anwendung im Folgenden „MVA“ genannt.

Die Standardansicht der MVA-Anwendung ist in Abb. 5.1 dargestellt. Die Benutzeroberfläche besteht aus der Menüleiste am oberen Rand des Fensters, der Statusleiste am unteren Rand und aus den Arbeitsbereichen, die dazwischen positioniert sind.

Die Arbeitsbereiche ermöglichen die Anzeige und die Anpassung der geladenen Daten. Hier sind die drei Aspekte der w^3-Prämisse implementiert. Die Arbeitsbereiche sind in Tabs organisiert, so dass mehrere (maximal acht) Instanzen gleichzeitig geöffnet werden können. Mindestens ein Tab muss immer geöffnet sein, alle weiteren können auch wieder geschlossen werden. Die Arbeitsbereiche besitzen im linken Bereich der Oberfläche eine Reihe von Steuerelementen mit denen die meisten Anpassungen der Daten durchführbar sind. Diese Leiste implementiert den „Was?“-Teil der w^3-Prämisse zusammen mit der Datentabelle im unteren Teil des Arbeitsbereiches, die die aktuell geladenen Daten beinhaltet. Den größten Teil der Benutzeroberfläche nimmt die Karte ein (der „Wo?“-Teil), darunter befindet sich die Zeitleiste, der „Wann?“- und letzte Teil der w^3-Prämisse. Die Menüleiste ist ein „globales“ Element. Die darin durchgeführten Aktionen wirken sich auf alle Tabs aus.

5.1. Geoinformationen

Für die Implementierung des „Wo?“-Teils der w^3-Prämisse werden die Patientendaten auf einer Karte dargestellt. Die Position der jeweiligen Daten wird mit Hilfe der Spalte „plz5“ der Brustkrebsdatei berechnet. Diese Spalte enthält die Heimatadresse jedes Patienten im Format „Postleitzahl Stadt (Land)“. Dadurch wird eine relativ genaue Lokalisierung der Patienten unter Bewahrung der Anonymität der Daten ermöglicht. Da der überwiegende Teil der Patienten in Deutschland wohnhaft ist, wurden in der Analyse nur deutsche Postleitzahlen berücksichtigt.

5.1.1. Administrative Struktur der Bundesrepublik

Die Verteilung der Postleitzahlen in der Bundesrepublik entspricht nicht der administrativen Einteilung nach Bundesländern und Landkreisen. Zusammengehörende Postleitzahlen werden nach den Verkehrsflughäfen für die Auslieferung der Post zu Regionen und mehrere

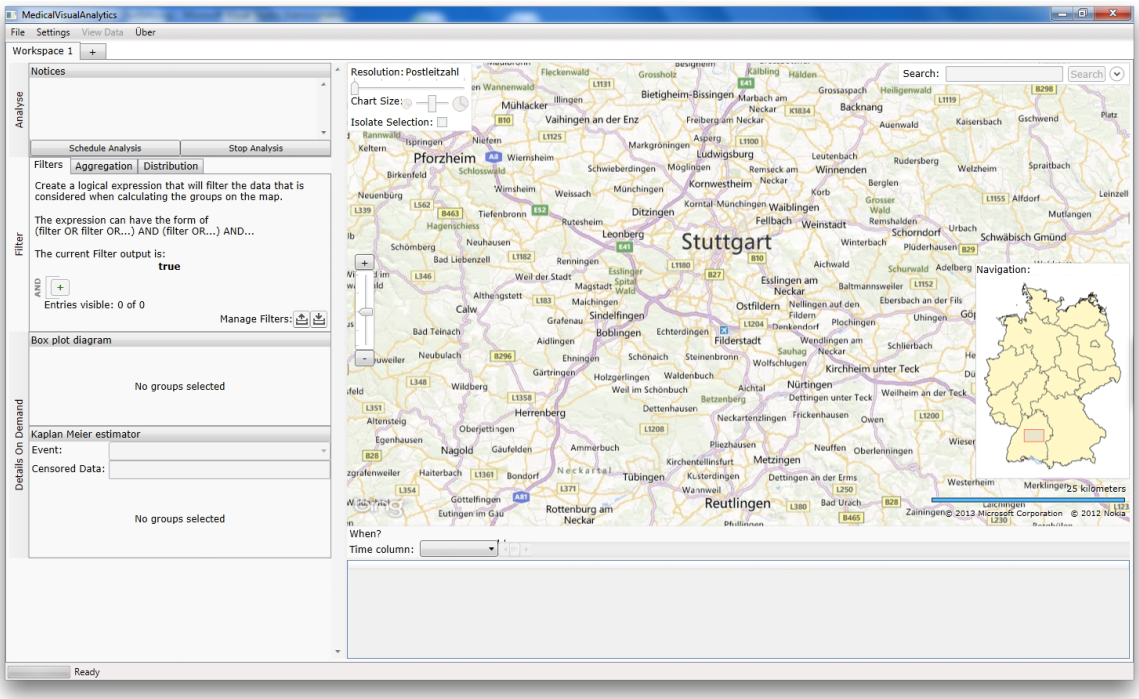


Abbildung 5.1.: Standardansicht der MVA-Anwendung.

Regionen zu Zonen zusammengefasst [Pos]. Dadurch gehören Ortschaften in unterschiedlichen Landkreisen (beispielsweise Tübingen und Reutlingen in Region 72) und sogar Bundesländern (beispielsweise Blaubeuren und Heidenheim an der Brenz in Region 89) zu gleichen Postleitregionen [Des]. Eine Zusammenfassung mehrerer Patientendaten anhand der Postleitzahlenregionen und -zonen würde somit nicht der von den Benutzern erwarteten Gruppierung entsprechen. Aus diesem Grund wurde nach einer hierarchischen Struktur der administrativen Gebiete Deutschlands gesucht. Folgende Alternativen wurden dabei verglichen:

- Das Statistische Bundesamt stellt im Bereich Länder und Regionen ein Gemeindeverzeichnis mit Informationen zur Fläche, Bevölkerungszahlen, geografische Position und anderen administrativen Details zur Verfügung [Des].
- OpenGeoDB ist ein Teil des ehrenamtlich geführten GISWiki[1] und bietet eine Datenbank mit Geokoordinaten zu allen Orten und Postleitzahlen der Bundesrepublik an [Ope].
- Unter [Man] wird eine auf die OpenGeoDB basierende csv-Datei der deutschen Postleitzahlen kostenfrei angeboten. Jeder Eintrag zu einer Postleitzahl enthält den dazugehörenden Ort, Landkreis und das Bundesland. In dieser Datei fehlen die Geokoordinaten der einzelnen Einträge.

[1] http://www.de.giswiki.org/index.php/Projekt:Über_GISWiki

Da Postleitzahlen nicht zur administrativen Einteilung der Bundesrepublik gehören, enthält die Datenbank des statistischen Bundesamtes keine vollständigen Informationen dazu. Für die Gruppierung der Datensätze auf Ebene der Postleitzahlen müssen die geografischen Koordinaten gesondert gesucht werden. Auch die openGeoDB-Datei enthält keine geografische Koordinaten der Postleitzahlen. Sie enthält aber zusätzlich zur Datenbank des statistischen Bundesamtes auch Angaben zu Stadtbezirken. Da Postleitzahlengebiete über Stadtbezirksgrenzen hinweg reichen, wäre keine eindeutige Zuordnung der Patientendaten zu Stadtbezirken möglich. Aus diesem Grund wurde die Ebene der Stadtbezirke nicht verwendet. Diese ersten beiden Datenbanken setzen die hierarchische Struktur ähnlich einer Datenbanktabelle um. Alle geografischen Bereiche werden in einer Spalte eingetragen und mittels Schlüssel auf den Eintrag der höheren Ebenen verlinkt. Das reduziert die Redundanz der Daten, benötigt im Gegensatz mehrere Abfrageschritte bei der Suche nach einem bestimmten Eintrag.

Aus diesen Gründen wurde für die Positionierung der Datensätze auf der Karte die dritte der oben genannten Optionen ausgewählt. Zwar enthält diese gar keine geografischen Koordinaten, dafür sind die übergeordneten Bereiche der Postleitzahlen in der gleichen Zeile angegeben und können in einer einzigen Abfrage eingelesen werden.

Jeder Datensatz der Brustkrebsdatei beinhaltet außerdem einen Code, der die behandelnde Klinik des Patienten identifiziert. Nach der Entschlüsselung dieses Codes (siehe Abschnitt 3.5) können die Patientendaten auch anhand der behandelnden Klinik zusammengefasst werden. Der Benutzer hat somit die Möglichkeit, die Daten nach einem der folgenden Kriterien zu gruppieren:

- Postleitzahl
- Ort
- Landkreis
- Bundesland
- Behandelnde Klinik

Für die Berechnung der Anzeigeposition auf der Karte wird die Bing Maps-API verwendet (für Details zur Syntax der Abfragen wird auf den Abschnitt 6.4.2 verwiesen). Um die optimale Herangehensweise bei der Abfrage der geografischen Koordinaten zu finden, wurden zwei Ansätze verglichen, die im Folgenden beschrieben werden.

5.1.2. Echtzeitabfrage der Koordinaten

Im ersten Ansatz wurde bei jeder eingelesenen Zeile aus der Brustkrebsdatei eine Anfrage an die Bing Maps-API gesendet. Die Antworten wurden in einer Liste zwischengespeichert, so dass für weitere Datensätze aus dem gleichen geografischen Bereich keine Anfrage mehr benötigt wurde. Für jeden Datensatz werden jeweils vier Abfragen für die Postleitzahl, den Ort, den Landkreis und das Bundesland gemacht. Durch die Zwischenspeicherung der schon

abgefragten Daten werden nur etwa 2.000 Anfragen benötigt. Bei einer Durchschnittsdauer von 0,06 Sekunden pro Abfrage, dauert die Abfrage aller Gruppen etwa zwei Minuten.

5.1.3. Vorababfrage der Koordinaten

Um die Dauer des Datenimports zu verkürzen, werden im zweiten Ansatz die Koordinaten aller administrativen Bereiche der Bundesrepublik im Voraus abgefragt und in einer Datei gespeichert. Beim Importieren neuer Daten werden die Koordinaten der Bereiche dieser Datei entnommen. Dazu wurde die im Abschnitt 5.1.1 erwähnte Datei der administrativen Hierarchie der Bundesrepublik um folgende Spalten erweitert:

- GeoPosPlzLat
- GeoPosPlzLong
- GeoPosOrtLat
- GeoPostOrtLong
- GeoPosKreisLat
- GeoPosKreisLong
- GeoPosLandLat
- GeoPosLandLong

Das Ausfüllen dieser Spalten erfolgt ebenfalls mit Hilfe der Bing Maps API. Dabei werden die Koordinaten aller administrativen Bereiche abgefragt und anschließend in den jeweiligen Spalten gespeichert. Dazu wurde ein separates Programm geschrieben, dessen Funktionsweise im Abschnitt 6.4.2 beschrieben wird. Die 27.736 Abfragen, die nötig sind, um diese Datei zu vervollständigen, werden in etwa einer halben Stunde durchgeführt. Die entstandene Datei muss im Programmordner der MVA-Anwendung abgelegt werden. Durch die Verwendung dieser Datei werden die Daten der Brustkrebsdatei bei einem Importieren in etwa sieben Sekunden eingelesen.

5.2. Programmstart

Während des Startvorgangs wird der in Abb. 5.2 dargestellte Splashscreen angezeigt.

Nach dem ersten Start der Anwendung wird die Standardansicht aus Abb. 5.1 gezeigt. Nur die Navigation auf der Karte, sowie das Hinzufügen weiterer Tabs, ist möglich. Alle weiteren Schaltflächen sind deaktiviert und werden erst nach dem Laden von Daten aktiviert.

Wurden in einer früheren Sitzung des Programms schon Daten importiert und ist die Einstellung „Save State on Quit“ (siehe 5.3.15) aktiviert, so sind nach dem Start die Daten bereits geladen und alle Schaltflächen aktiviert.

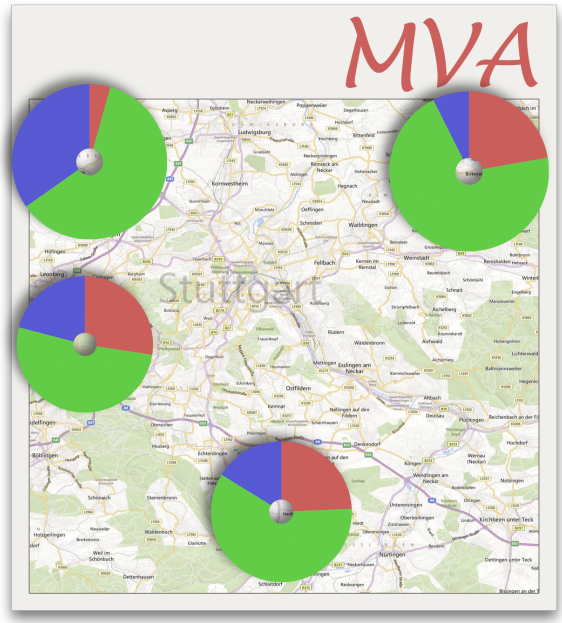


Abbildung 5.2.: Splashscreen der MVA-Anwendung.

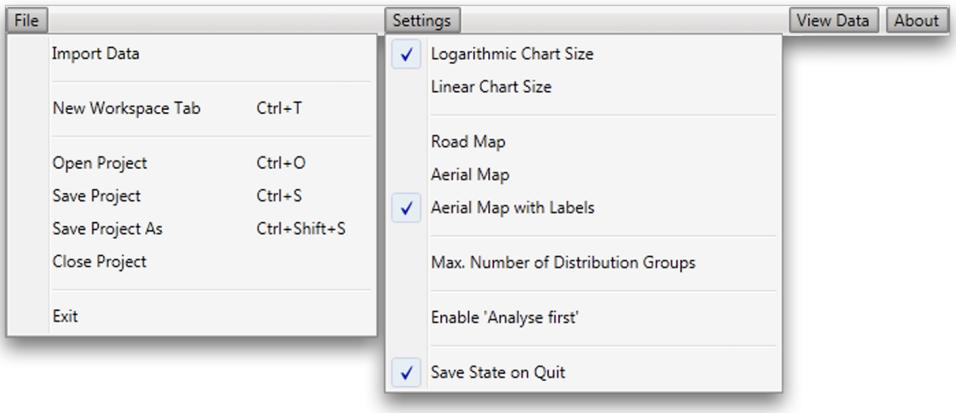


Abbildung 5.3.: Menüoptionen.

5.3. Die Menüleiste

Im Folgenden werden die Einträge in den Menüs am oberen Rand des Anwendungsfensters erläutert. Abb. 5.3 zeigt eine Übersicht aller Menüoptionen.

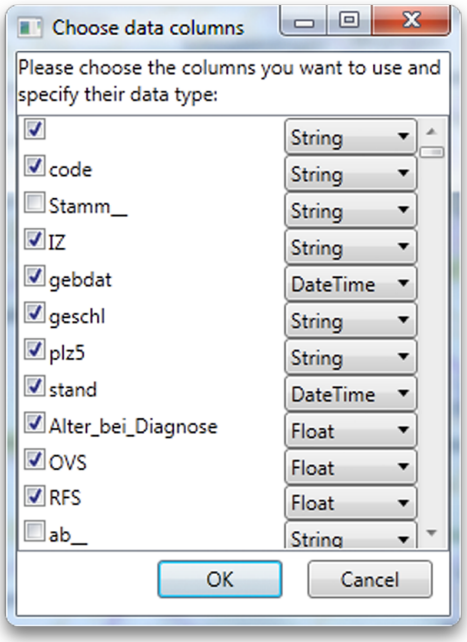


Abbildung 5.4.: Auswahlfenster für die zu importierenden Spalten.

5.3.1. File – Import Data

Das Import-Menü ist der Einstiegspunkt in der Arbeit mit der MVA-Anwendung. Darüber können Daten aus einer csv-Datei importiert werden. Nach dem Aufrufen des Import-Befehls wird ein standardisierter FileOpen-Dialog angezeigt, über welchem der Benutzer eine csv-Datei auswählen kann.

Die eingelesene csv-Datei muss folgende Kriterien erfüllen:

- die erste Zeile muss die Spaltennamen der Tabelle beinhalten. Diese werden in der weiteren Arbeit mit den Daten benötigt.
- als Trennzeichen wird der Semikolon erwartet.

Die Brustkrebsdatei erfüllt nach den Bearbeitungen aus Abschnitt 3.5 die genannten Kriterien.

Nach der Auswahl der csv-Datei wird das Spaltenauswahlfenster aus Abb. 5.4 geöffnet. Das Fenster zeigt alle Spalten der gewählten csv-Datei an und erlaubt dem Benutzer die Auswahl deren, die für die Datenanalyse verwendet werden sollen. Darüber hinaus kann der Benutzer den Datentyp jeder Spalte angeben. Dadurch werden Funktionen, die nur mit numerischen oder Datumswerten arbeiten können, ermöglicht. Da die csv-Datei über 400 Spalten beinhaltet, darunter viele Hilfsspalten, sollten nicht alle in einer Sitzung importiert werden, da das Ansprechverhalten der Anwendung dadurch stark abnehmen würde.

Nach dem Bestätigen des Spaltenauswahlfensters wird zunächst ein `DataTable`-Objekt erstellt (siehe Abschnitt 6.6.1). Vor dem Hinzufügen der Zeilen müssen die Spalten der Tabelle definiert werden. Dafür werden die Angaben des Datentyps und der zu importierenden Spalten aus dem davor ausgefüllten Spaltenauswahlfenster verwendet. Zum Schluss wird die ausgewählte csv-Datei zeilenweise eingelesen und in der DataTable gespeichert.

Die importierte Tabelle wird schließlich im unteren Teil der Benutzeroberfläche angezeigt (siehe Abschnitt 5.12). Eine größere Ansicht des Datenfensters kann über den Menüpunkt „View Data“ (Abschnitt 5.3.16) geöffnet werden.

5.3.2. File – New Workspace Tab

Mit Hilfe dieser Funktion kann der Anwendung ein weiterer Arbeitsbereich hinzugefügt werden. Wenn Daten in der Anwendung bereits geladen sind, dann werden diese im neuen Arbeitsbereich übernommen.

Diese Funktion ist auch über die Tastenkombination „CTRL+T“ aufrufbar.

5.3.3. File – Open Project

Über diesen Menüeintrag kann ein zuvor gespeichertes Projekt wieder geladen werden. Dabei werden alle aktuellen Arbeitsbereiche mit den Arbeitsbereichen und den dazugehörenden Einstellungen des gespeicherten Projektes ersetzt. Die technischen Details der Speicherung werden im Abschnitt 6.6.4 erläutert.

Diese Funktion ist auch über die Tastenkombination „CTRL+O“ aufrufbar.

5.3.4. File – Save Project As

Die MVA-Anwendung erlaubt die Speicherung einer Instanz der kompletten Anwendung. Beim Aufruf des „Save Project As“-Befehls wird ein Dialogfenster angezeigt, in dem der Benutzer den gewünschten Namen des Projektes eingeben und einen Pfad für die Speicherung angeben kann. Projektdateien werden mit der Dateiendung „.mvaproj“ gespeichert. Nach dem Speichern des Projektes zeigt die Titelzeile des Anwendungsfensters den Namen des Projektes an.

Diese Funktion ist auch über die Tastenkombination „CTRL+SHIFT+S“ aufrufbar.

5.3.5. File – Save Project

Wurde die aktuelle Arbeitsumgebung bereits in einem Projekt gespeichert, wird beim Ausführen dieses Befehls das Projekt erneut gespeichert, in dem die alte Projektdatei durch eine neue ersetzt wird. Handelt es sich dagegen um ein noch nicht gespeichertes Projekt, so wird der Befehl „Save Project As“aufgerufen.

Diese Funktion ist auch über die Tastenkombination „CTRL+S“ aufrufbar.

5.3.6. File – Close Project

Mit dem „Close“-Befehl kann die Arbeitsfläche in den Ursprungszustand zurückgesetzt werden. Die geladenen Daten werden gelöscht, alle Benutzerschaltflächen werden deaktiviert und alle Arbeitsbereiche geschlossen. Zusätzlich ändert sich der Titel des Anwendungsfensters zurück zu dem Wert „MedicalVisualAnalytics“. Die Werte des Menüs „Settings“ werden von diesem Befehl nicht beeinflusst und bleiben unverändert.

5.3.7. File – Quit

Der Quit-Befehl schließt das Programmfenster inklusive aller geöffneten Tabs.

Abhängig von dem Wert der Einstellung „Save state on quit“ werden die geladenen Daten und die geöffneten Tabs gemäß den Angaben aus dem Abschnitt 5.3.15 gespeichert.

Diese Funktion ist auch über die Tastenkombination „CTRL+X“ aufrufbar.

5.3.8. Settings – Logarithmic Chart Size

Um eine größere Flexibilität bei dem Vergleich unterschiedlicher geographischer Bereiche miteinander zu ermöglichen, kann die Größe der Diagramme auf der Karte nach einer logarithmischen Skala berechnet werden. Insbesondere, wenn die Größe eines Diagramms die Anzahl der Patienten in einem geografischen Bereich darstellt, können bei einer linearen Skala manche Diagramme zu klein sein, um erkannt zu werden.

Abb. 5.5(a) zeigt zwei geografische Bereiche mit 6.579 Patienten (links) und mit einem Patienten (rechts) auf einer logarithmischen Skala. Abb. 5.5(b) zeigt die gleichen Bereiche mit einer linearen Skala. Das rechte Diagramm ist hier nicht erkennbar.

Die logarithmische Skala ist die Standardeinstellung der Anwendung, da hiermit keine Bereiche übersehen werden. Will man gezielt feinste Unterschiede zwischen verschiedenen Bereichen direkt erkennen, kann die lineare Skala eingestellt werden.

Die Größe der Diagramme liegt standardmäßig zwischen 10 und 40 Pixeln.

5.3.9. Settings – Linear Chart Size

Die lineare Skala für die Größe der Diagramme eignet sich für Fälle, in denen die Werte der betrachteten Variablen auf einen kleinen Bereich verteilt sind. Der Durchschnittswert der Überlebenszeit der Patienten ist ein solches Beispiel, da er zwischen 0 und der maximalen Beobachtungsdauer von 20 Jahren liegt. Bei der linearen Skala sind auch kleine Unterschiede in der Größe der Diagramme mit bloßem Auge erkennbar. In Abb. 5.5(c) scheinen die zwei Bereiche auf einer logarithmischen Skala die gleiche Größe zu haben. Auf der linearen Skala in Abb. 5.5(d) ist der Größenunterschied leichter erkennbar.

Die Größe der Diagramme liegt hier standardmäßig zwischen 0 und 40 Pixeln. Die Diagramme werden auf die Größe von maximal 40 Pixeln normalisiert, indem nach der Berechnung des eigentlichen Wertes das Maximum ermittelt und daraus der Umrechnungsfaktor berechnet wird:

(5.1) $PieScale = 40/MaxValue$

Dieser Umrechnungsfaktor wird mit der tatsächlichen Größe jedes Diagramms multipliziert, um so den normalisierten Wert zu erhalten.

Zusätzlich zu dieser automatischen Berechnung der Größe kann der Benutzer die Größe aller Diagramme auch manuell steuern. Informationen dazu finden sich in Abschnitt 5.4.3.

5.3.10. Settings – Road Map

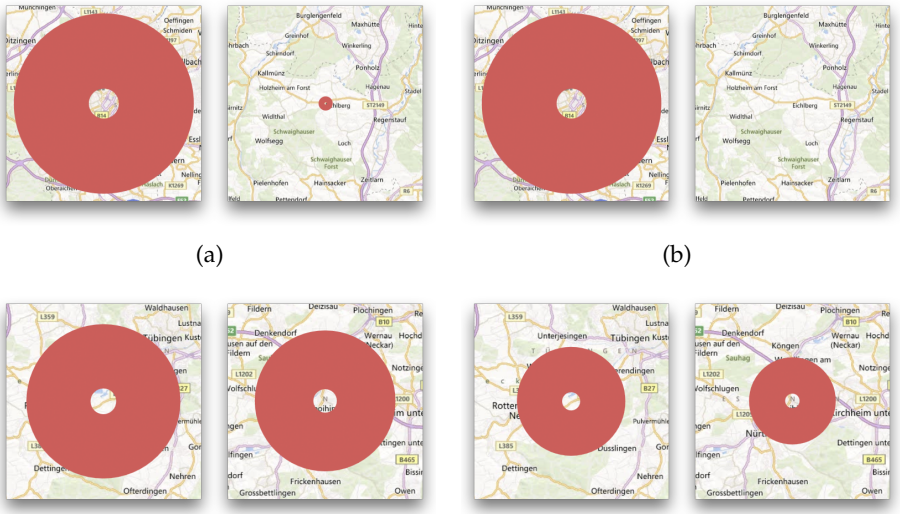
Die Einstellung „Road Map“ wechselt die Karte (siehe Abschnitt 5.4) in die Straßendarstellung (Abb. 5.6(a)).

5.3.11. Settings – Aerial Map

Die Einstellung „Aerial Map“ wechselt die Karte (siehe Abschnitt 5.4) in die Satellitendarstellung (Abb. 5.6(b)).

5.3.12. Settings – Aerial Map with Labels

Die Einstellung „Aerial Map with Labels“ wechselt die Karte (siehe Abschnitt 5.4) in die Hybriddarstellung, eine Kombination aus der Straßen- und der Satellitendarstellung (Abb. 5.6(c)).



(a) (b)

(c) (d)

Abbildung 5.5.: Unterschiedliche Skalendarstellungen für die Größe der Diagramme: (a) zeigt eine logarithmische Skala für zwei geografische Bereiche mit den Werten 6.579 (links) und 1 (rechts); (b) zeigt die gleichen Bereiche aus (a) unter Verwendung der linearen Skala; (c) zeigt zwei geografische Bereiche mit den Werten 8,75 (links) und 7,00 (rechts) unter Verwendung der logarithmischen Skala; (d) zeigt dieselben Bereiche aus (c) auf einer linearen Skala.

(a) (b) (c)

Abbildung 5.6.: Auswählbare Kartendarstellungen: (a) zeigt die Straßendarstellung („RoadMode“); (b) zeigt die Satellitendarstellung („AerialMode“); (c) zeigt die Hybriddarstellung („AerialWithLabels“).

5.3.13. Settings – Max. Number of Distribution Groups

Bei der Anzeige von Distributionsgruppen wird die Anzeige bei Überschreitung einer bestimmten Anzahl an Gruppen abgebrochen (siehe Abschnitt 5.9). Der Benutzer hat über diese Einstellung die Möglichkeit, die Grenze ab der die Gruppen nicht mehr angezeigt werden, zu verändern. Beim Aufruf dieses Menüeintrages wird ein Fenster geöffnet, in dem der neue Grenzwert eingegeben werden kann.

5.3.14. Settings – Enable 'Analyse first'

Mit Hilfe dieser Einstellung wird festgelegt, ob nach dem Import von Daten die Data Mining-Algorithmen automatisch gestartet werden sollen. Standardmäßig ist die Einstellung deaktiviert, so dass Analysen nur auf Wunsch des Benutzers gestartet werden (siehe Abschnitt 4.1).

5.3.15. Settings – Save State on Quit

Zusätzlich zur Möglichkeit der Speicherung der geladenen Daten und der geöffneten Tabs in einem benannten Projekt, besitzt die Anwendung ein Standard-Projekt, in dem die Daten automatisch gespeichert werden können. Das Speichern in dem Standardprojekt wird über die Einstellung „Save State on Quit“ gesteuert.

Ist diese Einstellung aktiviert, so werden die Projektinformationen beim Schließen der Anwendung in die Datei „default.mvaproj“ gespeichert und beim nächsten Start wieder geladen. Diese Speicherung geschieht unabhängig davon, ob das Projekt schon unter einem anderen Namen gespeichert wurde.

Ist diese Einstellung deaktiviert, wird beim Schließen der Anwendung die genannte Datei gelöscht. Beim nächsten Start werden keine Daten geladen und die Anwendung befindet sich im gleichen Zustand, wie beim ersten Öffnen (siehe Abschnitt 5.2).

Die Werte aller Einstellungen im „Settings“-Menü der Menüleiste werden beim Schließen der Anwendung unabhängig von dieser Einstellung gespeichert und beim nächsten Programmstart wieder geladen. Die technische Umsetzung dieser Speicherung wird im Abschnitt 6.6.4 beschrieben.

5.3.16. View Data

Dieser Menüpunkt öffnet ein separates Fenster, in dem die geladenen Daten eingesehen werden können. Dabei handelt es sich um eine globale Ansicht der Daten, die nicht gefiltert ist. Diese Ansicht dient lediglich der genaueren Analyse der Daten. Die einzelnen Spalten können sortiert und in der gesamten Tabelle kann nach bestimmten Werten gesucht werden.

5.3.17. About

Dieser Menüeintrag öffnet ein Fenster in welchem eine Kurzinformation über den erstellten Prototyp präsentiert wird.

5.4. Die Karte

Die Kartendarstellung verwendet Kartenmaterial der Microsoft Bing Maps. Die Einbindung erfolgt über die von Microsoft bereitgestellte Bibliothek `Microsoft.Maps.MapControl.WPF`. Die `MapControl.WPF`-Bibliothek stellt ein Kartenelement zur Verfügung, das die Anzeige, das Schwenken und die Vergrößerung der Karte sowie die Darstellung weiterer Elemente und Formen an geographischen Positionen erlaubt. Außerdem bietet die Bibliothek auch einen Webdienst, mit dessen Hilfe die Koordinaten geografischer Punkte abgefragt werden können. Weitere Informationen zu diesem Webdienst finden sich im Abschnitt 6.4.2.

Standardmäßig wird die Karte im „RoadMode“ angezeigt, um eine bessere Orientierung des Benutzers auf der Karte zu ermöglichen. In diesem Modus werden Straßen, Ortschaften und Reliefformen auf der Karte dargestellt. Die alternativen Darstellungen „AerialMode“ und „AerialWithLabels“, die Satellitenbilder der Erde zeigen, können im Einstellungsmenü (siehe Abschnitte 5.3.11 und 5.3.12) ausgewählt werden. Da man davon ausgehen kann, dass die Auswahl des Kartenmodus nur selten benötigt wird, wurde sie nicht direkt in die Karte eingebaut.

Der sichtbare Ausschnitt der Karte kann durch Anklicken und Bewegen der Maus verschoben werden. Die Karte kann mit Hilfe des Scrollrades einer Maus oder über das Zoom-Control vergrößert oder verkleinert werden. Das Zoom-Control erfüllt die in [HS05] beschriebenen Anforderungen und erlaubt ein stufenweises Zoomen über die Schaltflächen „+“ und „-“ oder ein direktes Springen auf eine bestimmte Zoomstufe durch Anklicken der gewünschten Position des Schiebereglers.

Auf das Kartenmaterial können eigene Objekte und Controls gezeichnet werden, indem sie in einer zusätzlichen Ebene oberhalb der Kartenebene hinzugefügt werden. Unter Angabe der Längen- und Breitengrade sowie der Höhe über Normal-Null können diese Elemente geographisch genau positioniert werden. Da die Karte nur in einer zweidimensionalen Sicht angezeigt wird, wird die Höhenangabe nicht benötigt. Bei ineinander verschachtelten Elementen wird nur das äußerste Element mit Hilfe der geografischen Koordinaten positioniert. Die inneren Elemente befinden sich in dem Koordinatenraum des äußersten Elementes und benötigen keine Kenntnis über die eigene Positionierung auf der Karte. Die `Maps`-Klasse behandelt beim Schwenken der Karte die Verschiebung der darauf gezeichneten Objekte selbst, so dass eine neue Positionierung der angezeigten Elemente nicht nötig ist.

An den Rändern der Karte befinden sich sämtliche Bedienelemente mit denen die Navigation und die Anzeige der Ergebnisse auf der Karte angepasst werden kann. Diese werden in den folgenden Abschnitten vorgestellt. Abb. 5.7 zeigt das Kartenelement und die erwähnten Bedienelemente.

Abbildung 5.7.: Kartendarstellung mit Bedienelemente.

5.4.1. Miniaturkarte

Um die Orientierung auf der Karte weiter zu verbessern wurde der Karte ein Navigationsfenster mit einer Miniaturkarte hinzugefügt. Dieses erlaubt dem Benutzer, neben den Detailinformationen innerhalb des ausgewählten Kartenausschnittes, auch den Überblick über die Position des betrachteten Ausschnitts zu behalten. Ein rotes Rechteck verdeutlicht die Position und die Größe des aktuellen Kartenausschnittes. Durch Klicken und Bewegen der Maus im Navigationsfenster wird die Position des Kartenausschnittes festgelegt. Dadurch werden schnelle Sprünge zwischen unterschiedlichen Bereichen der Karte ermöglicht.

Da man dadurch die Details und den Überblick in einer einzigen Anzeige vereint, spricht man vom „Focus+Context"-Prinzip [Fco02] oder auch „Local-global orientation" [HS05].

Daher, dass der verwendete Datensatz nur Patienten aus Deutschland beinhaltet, enthält das Navigationsfenster nur die Karte Deutschlands. Positionen des Kartenausschnitts außerhalb des Navigationsfensters werden durch dessen Ausgrauung verdeutlicht. Ein Klick auf das Miniaturfenster verschiebt den Kartenausschnitt zurück auf deutsches Gebiet.

Die Position des Rechteckes auf der Miniaturkarte wird aus der Position der Eckpunkte des aktuellen Kartenausschnittes und den Extrempunkten Deutschlands, die bei folgenden Koordinaten liegen [Sta], berechnet:

- N: $55°03'3''(55,0591°)$
- S: $47°16'15''(47,2708°)$
- W: $5°52'01''(5,8669°)$
- O: $15°02'37''(15,0436°)$

Aufgeführt werden die benötigten Transformationen in den Formeln 5.2 und 5.3. Die Variablen P_{tl} und P_{br} stellen die oberen linken bzw. unteren rechten Eckpunkte des Bildschirmausschnittes dar. Die Variable P_{nav} stellt die linke obere Ecke des Positionsrechteckes im Navigationsfenster dar, die Variablen B_{nav} und H_{nav} dessen Breite und Höhe. Die Variablen N, S, W und O haben die oben genannten Werte der Extrempunkte Deutschlands. Für die Berechnungen werden ebenfalls die Maße des Navigationsfensters benötigt. Diese liegen bei 175 x 240 Pixeln (B x H).

(5.2a)

$$P_{nav}(X) = (P_{tl}(X) - W)\frac{175}{O - W} \qquad P_{nav}(Y) = (N - P_{tl}(Y))\frac{240}{N - S}$$

$$B_{nav} = (P_{br}(X) - P_{tl}(X))\frac{175}{O - W} \qquad H_{nav} = (P_{br}(Y) - P_{tl}(Y))\frac{240}{N - S}$$

(5.3a)

$$P_{tl}(X) = W + \frac{P_{nav}(X)}{\frac{175}{O-W}} \qquad P_{tl}(Y) = N - \frac{P_{nav}(Y)}{\frac{240}{N-S}}$$

Um die Orientierung und den Überblick zu verbessern, werden die erstellten Kartendiagramme als rote Punkte auch auf der Miniaturkarte angezeigt. Herausgefilterte Gruppen (siehe 5.7.4) werden dabei grau angezeigt. Abb. 5.8 zeigt die vorgestellte Miniaturkarte, auf der alle Gruppen mit einer mittleren Überlebenszeit unter sieben Jahren herausgefiltert wurden und somit auf der Karte ausgegraut sind.

5.4.2. Auflösung der Datensätze

Unter Auflösung der Datensätze ist die adminsitrative Hierarchieebene der Bundesrepublik zu verstehen, auf der die geladenen Datensätze zusammengefasst werden.

Der Benutzer hat im linken oberen Bereich der Kartendarstellung die Möglichkeit, diese Auflösung über einen Schieberegler zu verändern. Das Bedienelement wird im oberen Bereich der Abb. 5.9 dargestellt. Zur Verfügung stehen die in Abschnitt 5.1.1 genannten Ebenen: „Postleitzahl“, „Ort“, „Landkreis“, „Bundesland“ und „Klinik“.

Bei Änderung des Auflösungswertes werden die geografischen Gruppen unter Berücksichtigung der aktiven Filter neu erstellt. Die Aggregationsfunktion und gegebenenfalls die Distributionsfunktion werden neu ausgewertet. Um die Reaktionsgeschwindigkeit der Anwendung zu erhöhen, wurden diese Funktionen als parallele Prozesse implementiert. Die technische Umsetzung der parallelen Ausführung wird in Abschnitt 6.6.3 beschrieben.

Abbildung 5.8.: Miniaturkarte.

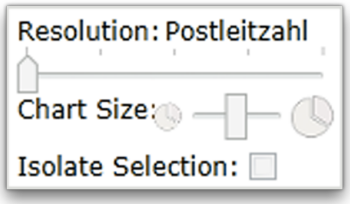


Abbildung 5.9.: Bendienelement für die Anpassung der Datendarstellung.

5.4.3. Größe der Diagramme

Unterhalb des Bedienelementes für die Auflösung befindet sich ein weiterer Schieberegler (siehe Abb. 5.9), der die Anpassung der Größe der Diagramme erlaubt. Durch Verschieben dieses Reglers wird ein Faktor geändert, der mit der normalisierten Größe der Diagramme multipliziert wird. Der Wert dieses Faktors kann zwischen 20 und $\frac{1}{20}$ gewählt werden. Der Standardwert liegt bei 1.

5.4.4. Isolierung der Auswahl

Die Auswahl dieses Kontrollkästchens bewirkt das Ausblenden aller nicht ausgewählten Diagramme. Dadurch kann bei überlappten Diagrammen kurzfristig ein Überblick verschafft werden.

5.4.5. Zoom der Karte

Am linken Rand der Karte befindet sich das Bedienelement zur Anpassung der Zoomstufe der Karte. Über die zwei Schaltflächen „+“ und „-“ kann die Karte schrittweise vergrößert und verkleinert werden. Der Schieberegler zwischen den beiden Schaltflächen erlaubt eine direkte Steuerung des Zoomstufe.

5.4.6. Suchfeld

Im rechten oberen Teil des Kartenelementes kann über einen Suchfilter nach dem Namen einer geografischen Gruppe gesucht werden. Die Suche funktioniert mit einem einfachen Textabgleich. Werden eine oder mehrere geografische Gruppen gefunden, deren Name den gesuchten Text enthalten, wird die Position und die Zoomstufe der Karte so ausgerichtet, dass sich alle gefundenen Gruppen im sichtbaren Bereich befinden. Zusätzlich werden die gefundenen Gruppen ausgewählt, um dem Benutzer deren Position zu verdeutlichen.

5.4.7. Gruppenfilter

Auf der rechten Seite des Suchelementes befindet sich eine Schaltfläche, mit der das Gruppenfilterelement geöffnet werden kann. Die Bedienung und die Funktionsweise dieses Elementes wird in Abschnitt 5.7.4 beschrieben.

5.5. Die Kartendiagramme

Die Hauptinformationsquelle bei der Visualisierung der Daten sind die auf der Karte angezeigten Diagramme. Es handelt sich dabei um Kreisdiagramme, die im Zentrum des jeweiligen geografischen Bereiches positioniert werden und die eine bivariate Analyse der geladenen Daten ermöglichen. Dazu wird der Radius und die Fläche des Kreisdiagramms als Informationsträger verwendet.

Die Auswahl der Diagramme wurde von der Übersichtlichkeit und der Einfachheit der Kreisdiagramme beeinflusst. Während andere Visualisierungsmethoden (beispielsweise parallele Koordinaten) deutlich komplexere Darstellungen mit mehreren betrachteten Eigenschaften ermöglichen, lassen der Radius und die Färbung der Kreisflächen einen schnellen Vergleich zwischen vielen nebeneinander positionierten Diagrammen auf der Karte zu. Abb. 5.10

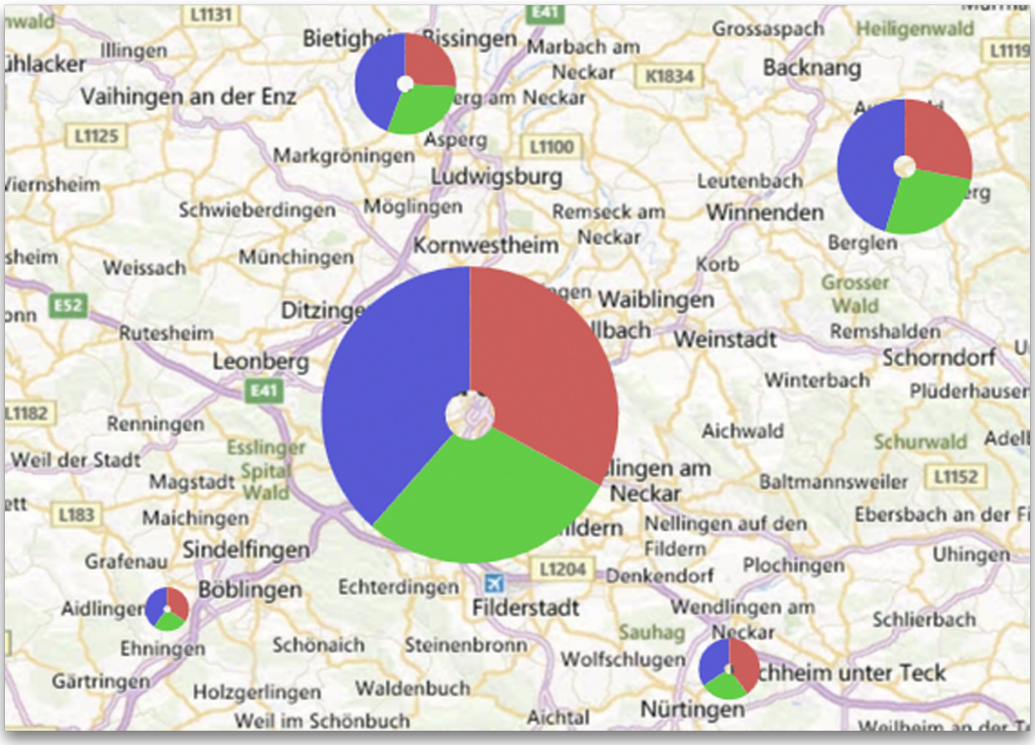


Abbildung 5.10.: Beispiel für Kartendiagramme. Jedes Diagramm beinhaltet die Datensätze eines Landkreises. Der Radius der Diagramme ist direkt proportional mit der Anzahl der darin enthaltenen Datensätzen. Die einzelnen Teile des Diagrammes stellen den Anteil der unterschiedlichen Klassen der Brusterhaltung, „f“, „n“ und „j“, dar.

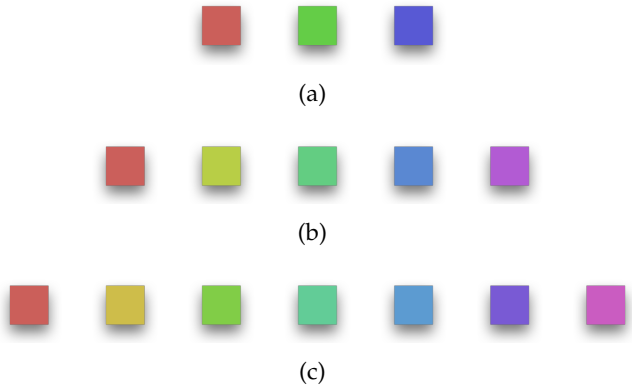
zeigt unterschiedliche Diagramme, die sich in dem Radius und in ihren Teilelementen unterscheiden.

5.5.1. Farben

Standardmäßig werden die Diagramme in der Farbe „indian red“ (RGB-Wert: 205, 92, 92) angezeigt. Bei Verwendung der Distributionsfunktion (siehe Abschnitt 5.9) wird die Fläche der Kreisdiagramme in Sektoren eingeteilt und unterschiedlich gefärbt. Die Auswahl der Farben verfolgt die Richtlinien aus [HB11] und verwendet unterschiedliche Methoden je nach Art der angezeigten Kategorien.

Die Farben numerischer Kategorien unterscheiden sich in der Farbhelligkeit. Niedrige Werte werden durch eine helle, hohe Werte durch eine dunkle Farbe gekennzeichnet („Dark equals more“ [HB11], „Dunkel ist mehr“). Die Farbhelligkeiten der Kategorien mit dem kleinsten und dem größten Wert sind immer dieselben. Bei unterschiedlicher Anzahl von Gruppen

Abbildung 5.11.: Farbskala für zehn numerische Kategorien.



(a)

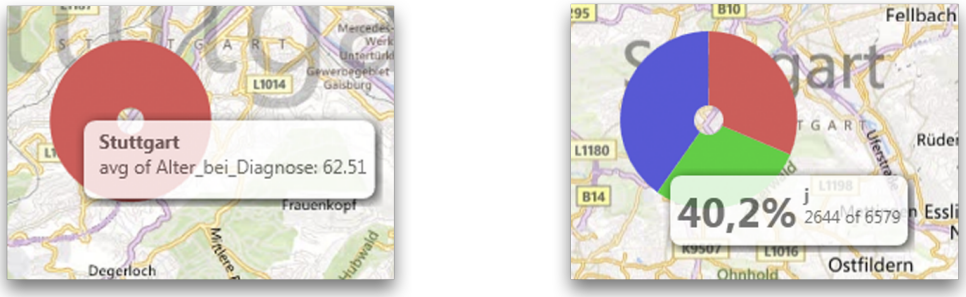
(b)

(c)

Abbildung 5.12.: Farbskalen für unterschiedliche Anzahlen von qualitativen Kategorien: (a) zeigt die Farben für drei Kategorien; (b) zeigt die Farben für fünf unterschiedliche Kategorien; (c) zeigt die Farben für sieben unterschiedliche Kategorien.

variieren die Farbskalen in der Abstufung zwischen diesen zwei Extremen. Abb. 5.11 zeigt die verwendete Farbskala für zehn numerische Kategorien.

Die verwendeten Farben dürfen bei kategorischen Daten keine Abstufung suggerieren. Deshalb werden in solchen Fällen die Farben durch Veränderung des Farbwertes anstatt der Farbhelligkeit generiert. Der erste Entwurf der Farbgenerierung basierte auf einer festen Anzahl an Farben, die auf die vorhandenen Kategorien verteilt wurden. [Col] bietet Farbschemata mit bis zu zwölf Farben zum Herunterladen an. Bei der Anzeige von mehr als zwölf Kategorien müssten die Farben zyklisch wiederholt werden. Das kann allerdings zu Verwechslungen führen, insbesondere dann, wenn manche Kategorien ausgeblendet werden und gleichfarbige Kategorien dadurch nebeneinander angezeigt werden. Um dieses Problem zu umgehen, wurden ähnlich wie bei den numerischen Kategorien, die Farbwerte gleichmäßig auf dem zur Verfügung stehenden Bereich verteilt. Da der Farbwert im HLV-Farbraum als Winkelmaß angegeben wird, beträgt der Farbwertunterschied zweier benachbarter Kategorien $360° / n$, wobei n die Anzahl der darzustellenden Kategorien ist. Auch wenn in diesem Fall keine Kategorie die gleiche Farbe verwenden erhöht sich die Ähnlichkeit der Farben bei steigender Anzahl der Kategorien. Bei Unsicherheit über die genaue Zuordnung einer Kategorie sollten die Tooltips verwendet werden. Abb. 5.12 zeigt die entstandenen Farben bei unterschiedlicher Anzahl an Kategorien.



(a) (b)

Abbildung 5.13.: Tooltips der Kartendiagramme: (a) zeigt den Tooltip eines kompletten geografischen Bereiches; (b) zeigt den Tooltip einer Gruppe innerhalb eines geografischen Bereiches.

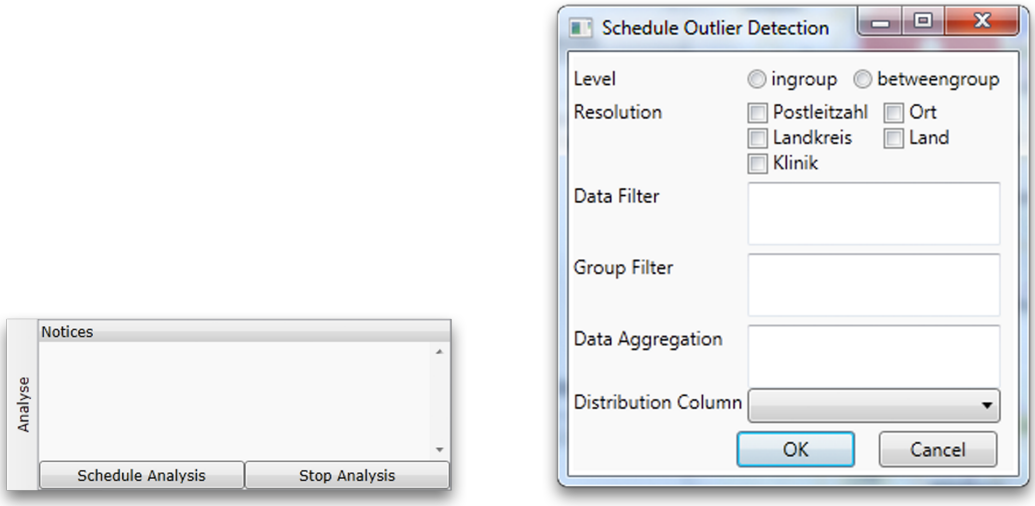
5.5.2. Tooltips

Beim Positionieren der Maus über den Kartendiagrammen werden Tooltips mit Informationen zu den darunterliegenden geografischen Gruppen eingeblendet. Es wird dabei der Name der Gruppe und das Ergebnis der Aggregationsfunktion (Abschnitt 5.8), das den Radius des Kartendiagramms bestimmt, angezeigt. Bei Verwendung der Distributionsfunktion (Abschnitt 5.9) zeigt der Tooltip Informationen zu der Distributionsgruppe unterhalb des Mauszeigers. Es werden dabei erneut der Name der geografischen Gruppe zusammen mit dem prozentualen und dem zahlenmäßigen Anteil der Distributionsgruppe an der Gesamtgröße der Gruppe angezeigt. Abb. 5.14 zeigt beide erwähnten Tooltips.

5.6. Die Ausreißererkennung

Die MVA-Anwendung implementiert einen einfachen Algorithmus für die Ausreißererkennung. Basierend auf [KN98] wird eine abstandsbasierte Methode verwendet, um Werte, die mehr als um das dreifache der Standardabweichung vom Mittelwert abweichen, als mögliche Ausreißer zu identifizieren. Die Ausreißererkennung wird entweder beim Start der Anwendung (siehe Abschnitt 5.3.14) oder auf Wunsch des Benutzers über die „Schedule Analysis“-Schaltfläche (Abb. 5.14(a)) gestartet. Ist die Einstellung „Enable ’Analyse first’“ aktiviert, wird die Analyse stets im Visualisierungskontext des Benutzers durchgeführt.

Der Algorithmus sucht nach Ausreißern in unterschiedlichen Dimensionen, abhängig von den eingegebenen Parametern. Dabei können Auflösungen der Daten, Daten- und Gruppenfilter, Aggregationsfunktionen und Distributionsspalten angegeben werden, die bei der Ausreißererkennung berücksichtigt werden sollen. Abb. 5.14(b) zeigt die Eingabemaske, in welcher der Umfang der Analyse definiert werden kann. Die Ausreißererkennung wird parallel zu dem Anwendungsthread durchgeführt (siehe Abschnitt 6.6.3), um die Reaktionsgeschwindigkeit der Anwendung nicht zu blockieren. Durch den hohen Rechenaufwand und



(a) (b)

Abbildung 5.14.: Steuerelement der Ausreißererkennung: (a) zeigt die Ergebnisliste der Ausreißererkennung mit den Schaltflächen für den Start und den Abbruch der Analyse; (b) zeigt die Eingabemaske, in welcher der Umfang der Ausreißererkennung bestimmt wird.

die dadurch verursachte Belastung des Prozessors sollten umfangreiche Analysen vermieden werden. Laufende Berechnungen können bei Bedarf über die „Stop Analysis“-Schaltfläche abgebrochen werden.

Gefundene Auffälligkeiten werden in der „Notices“-Liste im linken oberen Bereich der Arbeitsfläche aufgelistet. Falls es die aktuellen Anzeigeeinstellungen der Karte erlauben, werden Auffälligkeiten auch auf der Karte mit Hilfe eines gelben Warnzeichens neben der betroffenen geografischen Gruppe gekennzeichnet. Durch Anklicken eines Eintrages in der „Notices“-Liste wechselt die Ansicht zu der entsprechenden Gruppe. Der Benutzer hat die Möglichkeit, mit Hilfe der „Details on Demand“-Elemente die gefundenen Auffälligkeiten zu untersuchen und gegebenenfalls verfeinerte Analysen zu starten, um so den Zyklus des Visual Analytics-Mantras weiterzuführen.

5.7. Die Filterfunktion

Ein wichtiger Bestandteil des Visual Analytics-Mantras ist die Möglichkeit der Datenfilterung. Um einen Überblick in der großen Anzahl an Datensätzen zu gewinnen und um die gesuchten Informationen hervorheben zu können, implementiert die Anwendung eine Filterfunktion.

Die Entwicklung der Filterfunktion fand in zwei Schritten statt. Nach der ersten Implementierung wurde klar, dass eine mächtigere und flexiblere Filterung nötig ist, so dass

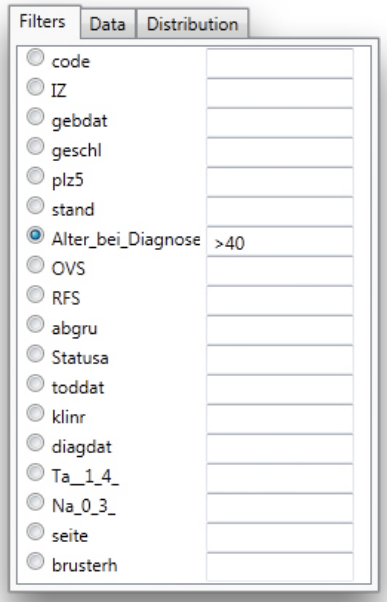


Abbildung 5.15.: Erster Ansatz der Filterfunktion.

in der Implementierung des zweiten Ansatzes beliebige Filter definierbar sind. Im folgenden werden die zwei Entwicklungsetappen und die zwei Einsatzgebiete der Filterfunktion erläutert.

5.7.1. Erster Ansatz

Der erste Ansatz der Filterfunktion ist in Abb. 5.15 dargestellt. Die Benutzeroberfläche besteht aus einer Liste aller geladenen Datenspalten, aus der eine Spalte für die Filterung ausgewählt werden kann. In dem Textfeld neben dem Spaltennamen kann eine Bedingung eingegeben werden. Aus dieser Kombination wird ein Filterausdruck der Form „Spaltenname Ausdruck“ erstellt, wobei der Ausdruck selbst aus einem Operator und einem Wert bestehen muss.

Da immer nur eine Spalte auswählbar ist, kann die Liste zu einer Drop-Down-Liste umprogrammiert werden, um Platz zu sparen. Alternativ dazu können mehrere Spalten auswählbar gemacht und somit zusammengesetzte Filter ermöglicht werden. Diese würden die Flexibilität der Filterungsmöglichkeiten erhöhen. Dabei muss entschieden werden, ob die einzelnen Filter mittels einer Konjunktion oder einer Disjunktion zusammengesetzt werden sollen. Diese Entscheidung kann auch dem Benutzer überlassen werden, so dass je nach Bedarf der Filter als eine Konjunktion oder eine Disjunktion mehrerer Terme definiert werden kann. Abb. 5.16 zeigt eine solche Implementierung in der Emailverwaltungssoftware Mail auf dem Betriebssystem MacOS X. Hier hat der Benutzer die Möglichkeit, das gleiche Kriterium (z.B. „Betreff“) mehrmals zu verwenden. Dadurch werden auch Ausdrücke wie „Spaltenname > 5

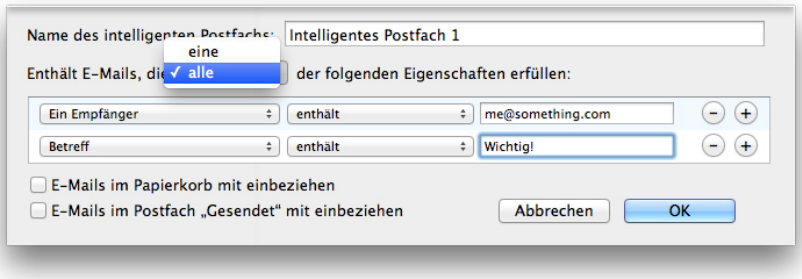


Abbildung 5.16.: Filterfunktion in Apple Mail.

UND Spaltenname < 10“ ermöglicht, die mit dem hier beschriebenen Ansatz nicht möglich sind.

Allerdings sind auch in diesem Fall Problemstellungen wie „Patienten älter als 60 Jahre im T-Stadium 0 oder N-Stadium 1“ nicht möglich. Um auch solche Filter zu ermöglichen wurde ein neuer Ansatz verfolgt.

5.7.2. Zweiter Ansatz: die konjunktive Normalform

Die größte Flexibilität bei der Erstellung von Filtern wird erreicht, wenn der Benutzer diese selbst in Textform eingeben kann. Da diese Lösung in der Benutzung sehr fehleranfällig ist, wurde ein Ansatz bevorzugt, in dem der Benutzer den Filterausdruck mit Hilfe einer grafischen Benutzeroberfläche (im folgenden „Filter-Control“ genannt) definieren kann. Durch die visuelle Darstellung des Filterausdrucks wird eine schrittweise Erweiterung des Filters erleichtert. Die Klammerung des Filters wird visuell durch Einrücken der entsprechenden Elemente verdeutlicht.

Basierend auf den Spalten der geladenen Datentabelle kann der Benutzer Filter in konjunktiver Normalform (KNF) [Knf] erstellen. Da jede logische Aussage in eine konjunktive Normalform umgewandelt werden kann ist damit die Erstellung beliebiger Filter möglich. Die Umwandlung des gewünschten Filters in KNF wird dabei dem Benutzer überlassen.

Abb. 5.17 zeigt ein ähnliches Element, das die Erstellung eines Filters in dem Standarddateiverwaltungsprogramm „Finder“ auf dem Betriebssystem MacOS X ermöglicht. Die Maske erlaubt die Erstellung von Filterausdrücken als Konjunktion von beliebigen anderen Termen. In dieser Maske wurden folgende Nachteile identifiziert:

- Dem Benutzer wird keine Zusammenfassung des erstellten Filters gezeigt. Dadurch muss er auf die Korrektheit des Filters besonders achten.
- Alle Schaltflächen zum Hinzufügen weiterer Terme befinden sich an der gleichen Stelle. Das erschwert die Orientierung in der Maske, da nicht sofort ersichtlich ist, auf welche Ebene ein neuer Term hinzugefügt wird.

Abbildung 5.17.: Beispiel eines Datei-Suchfilters im Apple Finder.

- Verschachtelte Terme werden durch einen Klick auf der „+“-Schaltfläche bei gedrückter ALT-Taste hinzugefügt. Diese Funktion ist in der Maske nicht dokumentiert.
- Der erstellte Filter verwendet keine explizite Klammerung. Die Klammerung kann aus der Einrückung der einzelnen Elemente impliziert werden. Allerdings können auf einer Einrückungsebene in unterschiedlichen Termen unterschiedliche Funktionen („UND“ bzw. „ODER“) vermischt werden, was ebenfalls eine erhöhte Aufmerksamkeit des Benutzers erfordert.

In der Entwicklung des Filter-Controls wurde versucht, diese Nachteile zu beseitigen. Das Filter-Control wurde als wiederverwendbares Benutzersteuerelement (siehe Abschnitt 6.5.2) programmiert. Abb. 5.18 zeigt dessen Standardansicht.

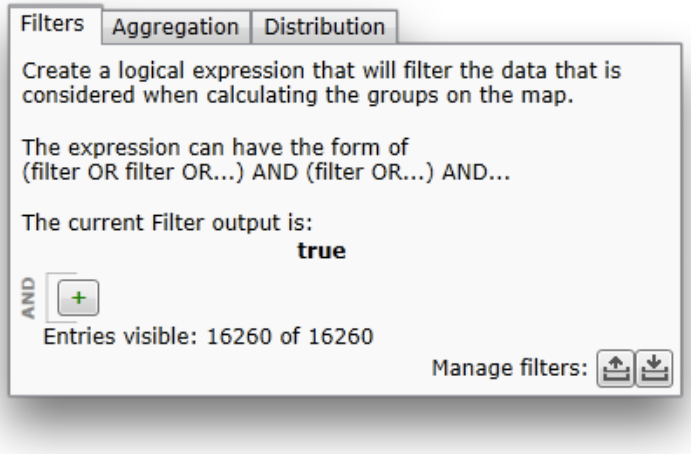


Abbildung 5.18.: Filter-Control.

Name	Typ	Anmerkung
Negation	Drop-Down-Liste	optional
Aggregationsfunktion	Drop-Down-Liste	optional
Spaltenname	Drop-Down-Liste	
Bedingung	Textfeld	

Tabelle 5.1.: Änderbare Benutzerelemente des Filter-Controls.

Der obere Bereich enthält einen Informationstext zur Funktionsweise des Filters. Je nach Einsatzgebiet des Filter-Elements kann dieser Text variiert werden.

Darunter wird der erstellte Filter in Textform angezeigt. Dadurch soll dem Benutzer die Überprüfung des eingegebenen Filters ermöglicht werden. Wenn das Filter-Element nicht ausgewertet werden kann, wird dieser Text in roter Farbe angezeigt. Dies kann der Fall sein, wenn der Filter Fehler enthält oder wenn neue Terme hinzugefügt wurden, die dazugehörenden Ausdrücke aber noch nicht ausgefüllt sind.

Den Hauptteil des Filter-Elements bestimmen die Bedienelemente mit denen der Filterausdruck definiert werden kann. Eine genaue Erklärung der Vorgehensweise bei der Erstellung eines Filters wird in Abschnitt 5.7.2 präsentiert.

Im unteren Teil des Filter-Controls wird die Anzahl der nach dem Filtern noch aktiven Elemente aus der Gesamtanzahl der Elemente angezeigt.

Das letzte Element des Filter-Controls erlaubt die Speicherung und das Laden von Filtern.

Funktionsweise

Das Filter-Control wird erst nach dem Laden eines Datensatzes aktiviert. Davor sind keine Schaltflächen anklickbar.

Durch Drücken der Schaltfläche „+“ kann eine erste Bedingung hinzugefügt werden. Da die Filter in KNF erstellt werden, wird der Term als Teil einer Disjunktion – die ihrerseits Teil der umschließenden Konjunktion ist – erstellt.

Die änderbaren Benutzerelemente der Terme werden in Tabelle 5.1 aufgelistet. Abb. 5.19 zeigt zwei Beispiele von Filtertermen, jeweils unter Verwendung der optionalen Negations- und Aggregationsspalte. In diesen Abbildungen sind auch die Textversionen der zwei Filter dargestellt.

Durch wiederholtes Drücken der gleichen Schaltfläche „+“ können der äußeren Konjunktion weitere Disjunktionen hinzugefügt werden. Der Filter in Abb. 5.20(a) kann somit erstellt werden. Durch Drücken der Schaltfläche „+“ innerhalb eines „OR“-Blocks können dagegen in den jeweiligen Disjunktionen weitere Terme hinzugefügt werden. Die Positionierung der unterschiedlichen „+“-Schaltflächen erlaubt das intuitive Hinzufügen von Termen oder Disjunktionen. Abb. 5.20(b) zeigt den entstandenen Filter.

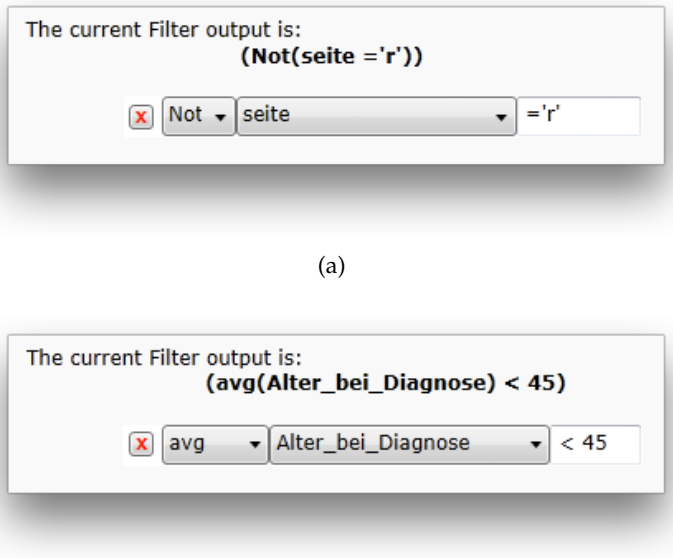


Abbildung 5.19.: Beispiel zweier Filtertermen mit unterschiedlichen, auswählbaren Feldern: (a) enthält die Negationsspalte, wie sie bei der Filterung der Datensätze verwendet wird (siehe 5.7.3); (b) enthält die Spalte für die Aggregationsfunktion, wie sie bei der Filterung der geografischen Gruppen verwendet wird (siehe 5.7.4).

Die Terme einer Disjunktion sowie die Disjunktionen einer Konjunktion werden durch einen farbig geänderten Hintergrund und einen Teilrahmen zusammengefasst. Der jeweilige Verknüpfungsoperator wird an der linken Seite des Rahmens um 90° nach links gedreht angezeigt. Somit wird die Klammerung der einzelnen Elemente dem Benutzer verdeutlicht.

Alle erstellten Terme und Disjunktionen können auch gelöscht werden. Auf der linken Seite der Elemente befindet sich dazu eine Schaltfläche, die mit einem roten Buchstaben „x“ beschriftet ist. Beim Löschen des letzten Terms einer Disjunktion wird auch die Disjunktion selbst gelöscht, da eine leere Disjunktion nicht ausgewertet werden kann.

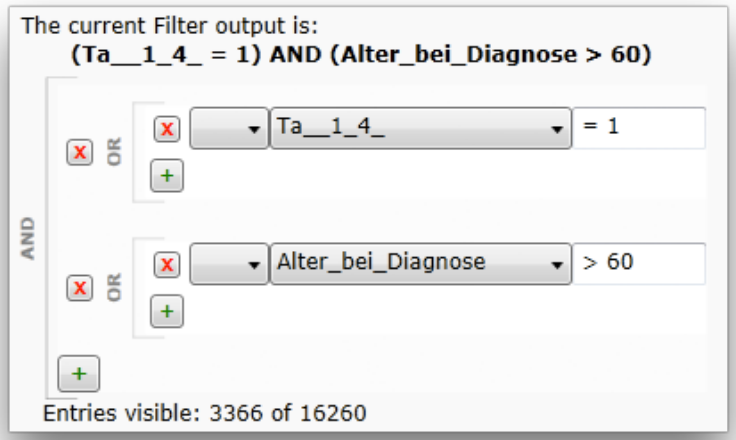
Die Zusammenfassung des erstellten Filters hilft dem Benutzer zu überprüfen, ob das Ergebnis dem gewünschten Filter entspricht.

Um die Arbeit mit den Filtern zu erleichtern gibt es die Möglichkeit, erstellte Filter zu speichern und später wieder zu laden. Dazu werden die zwei Schaltflächen in der rechten unteren Ecke des Filter-Controls verwendet. Ein Filter kann nur dann gespeichert werden, wenn er keine Fehler enthält.

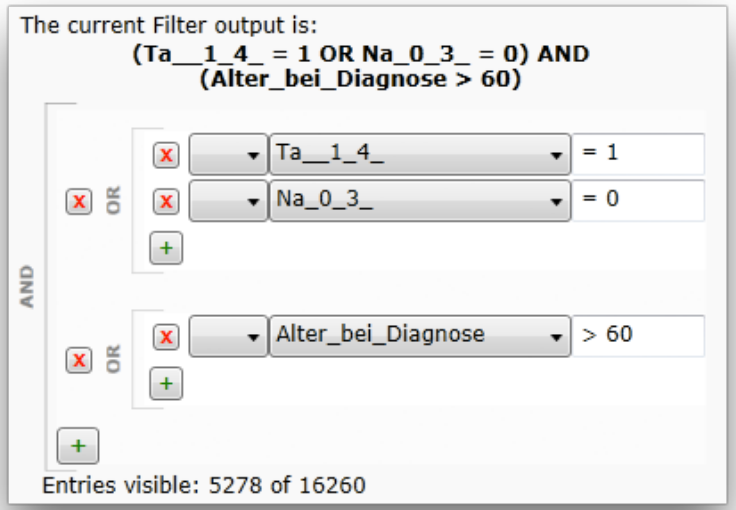
Beim Speichervorgang eines Filters wird ein Fenster geöffnet, in welchem dem Filter ein Name gegeben werden kann. In diesem Fenster kann auch der Text des Filters im rechten Textfeld geändert werden. Dadurch können Filter aus anderen Quellen manuell eingegeben

werden. Durch Drücken der „Speichern“-Schaltfläche wird der geänderte Filter erneut überprüft und im aktuellen Projekt gespeichert. Bei Fehlern im Filter wird der Speichervorgang verweigert.

Beim Betätigen der „Laden“-Schaltfläche wird ein Fenster geöffnet, in dem alle bisher gespeicherten Filter angezeigt werden. Hier können bestehende Filter gelöscht werden.



(a)



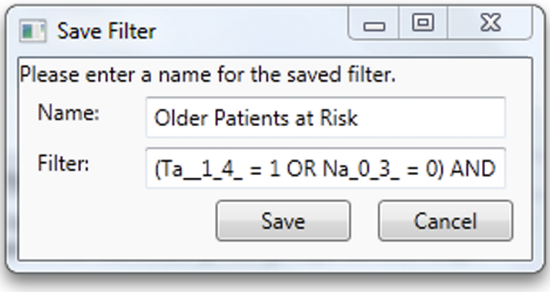
(b)

Abbildung 5.20.: Beispiel zweier Filter mit unterschiedlichem Umfang: (a) enthält zwei Disjunktionen bestehend aus jeweils einem Term; (b) enthält zwei Disjunktionen von denen eine aus zwei Termen besteht.

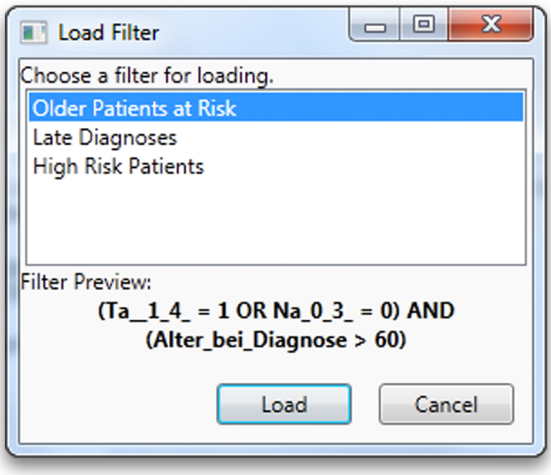
Durch die Auswahl eines Filters und das anschließende Drücken der Schaltfläche „Laden“ wird dieser Filter geladen und der aktuelle überschrieben.

Gespeicherte Filter stehen in allen offenen Arbeitsbereichen zur Verfügung und werden beim Speichern des Projektes gesichert.

Abb. 5.21 zeigt die zwei Fenster zum Speichern und Laden von Filtern.



(a)



(b)

Abbildung 5.21.: Sichern und Laden von Filtern: (a) zeigt das Fenster zum Speichern der Filter; (b) zeigt das Fenster zum Laden der Filter.

5.7.3. Filterung der Datensätze

Der erste Einsatzort des Filter-Controls ist die Filterung der geladenen Datensätze. Über die oben beschriebenen Methoden hat der Benutzer die Möglichkeit, die Daten, die bei der Berechnung der Diagramme berücksichtigt werden, zu filtern. Im Datenansichtsfenster werden die gefilterten Daten entsprechend markiert (siehe Abschnitt 5.12).

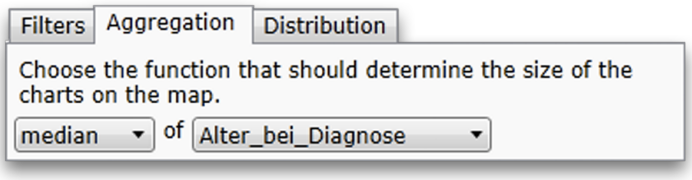


Abbildung 5.22.: Bedienelemente der Aggregationsfunktion.

Diese Instanz des Filter-Controls befindet sich im Bereich „Filter and Zoom“ im linken Teil des Arbeitsbereiches.

5.7.4. Filterung der Gruppen

Zusätzlich zu der Filterung der Datensätze ist auch die Filterung der entstandenen Gruppen möglich, die das Ausblenden bestimmter Gruppen erlaubt. Damit können Gruppen mit einer zu kleinen Anzahl an Patienten ausgeblendet werden, wenn diese nicht repräsentativ sind oder Zielgruppen mit einem bestimmten Durchschnittsalter analysiert werden.

5.8. Die Aggregationsfunktion

Diese Funktion erlaubt die Anwendung einer Aggregationsfunktion auf die Werte einer bestimmten Spalte des Datensatzes. Für die einzelnen geografischen Gruppen erfolgt die Auswertung getrennt. Die daraus resultierende Kennzahl beeinflusst den Radius der Diagramme auf der Karte.

Das Benutzerelement zur Definition der Aggregationsfunktion wird in Abb. 5.22 gezeigt. Es enthält zwei Drop-Downs über die die Aggregationsfunktion und die Spalte, auf der diese angewandt werden soll, ausgewählt werden können.

Zur Verfügung stehen folgende Aggregationsfunktionen: „exists“, „count“, „avg“, „median“, „var“, „stdev“ und „sum“. Die Funktion „exists“ gibt als Ergebnis die Werte 0 oder 1 zurück. Je nachdem, ob in dem entsprechenden administrativen Bereich Datensätze vorhanden sind oder nicht, wird der Bereich auf der Karte angezeigt oder ausgeblendet. Diese Funktion eignet sich für die Darstellung einer ersten Übersicht der Verteilung der geladenen Datensätze. Insbesondere bei der Verwendung eines Gruppenfilters (siehe Abschnitt 5.7.4) kann schnell überprüft werden, welche geografische Bereiche herausgefiltert wurden. Die Funktion „count“ gibt die Anzahl der Datensätze in einem bestimmten Bereich zurück. Zusammen mit der „exists“-Funktion ist es die einzige Funktion, die auf beliebige Spalten angewandt werden kann. Alle restlichen Funktionen können nur auf numerische Spalten angewandt werden. Bei Auswahl einer dieser Funktionen werden deshalb alle nicht-numerischen Spalten aus der Spaltenauswahlliste ausgeblendet. Den Mittelwert berechnet die Funktion „avg“, die Funktion „median“ berechnet das Median der ausgewählten Spalte für die jeweiligen

Bereiche. Die Funktion „var“ berechnet die Varianz, die Funktion „stdev“ die Standardabweichung und die Funktion „sum“ die Summe der Werte der ausgewählten Spalte. In der zweiten Drop-Down-Liste werden alle importierten Spalten des Datensatzes angezeigt.

Standardmäßig werden die aus der Aggregationsfunktion resultierenden Größen in einer logarithmischen Skala auf die Größe der Diagramme abgebildet. Über das Menü „Settings“ kann zwischen einer „logarithmischen“ und einer „linearen“ Skala ausgewählt werden (siehe Abschnitte 5.3.8 und 5.3.9).

5.9. Die Distributionsfunktion

Die Distributionsfunktion ermöglicht es dem Benutzer, die Verteilung der Datensätze in einem geografischen Bereich nach ausgewählten Kriterien anzuzeigen. Dafür kann aus einer Drop-Down-Liste eine der geladenen Spalten ausgewählt werden, nach der die Verteilung der Untergruppen berechnet wird. Je nach Datentyp der ausgewählten Spalte werden weitere Bedienelemente eingeblendet, die die Anpassung der Distributionsgruppen ermöglichen. Die entstandenen Gruppen werden in einer Liste unterhalb dieser Bedienelemente angezeigt. Auf der Karte werden sie in Form eines Kreisdiagramms dargestellt, in dem die Länge der Kreisbögen der einzelnen Kreissektoren der prozentualen Größe der einzelnen Distributionsgruppen entsprechen.

Bei der Auswahl einer Spalte mit dem Datentyp „Float“ wird der Wertebereich der Spalte in mehrere Gruppen eingeteilt, deren Größe von dem Benutzer angegeben werden kann. Dazu wird ein Textfeld eingeblendet, in dem die gewünschte Gruppengröße eingegeben werden kann. Standardmäßig wird der Wert „2“ verwendet. Ermittelt werden die Gruppen, indem der kleinste Wert, der in der Spalte vorkommt, abgerundet und als untere Grenze der ersten Gruppe verwendet wird. Die obere Grenze liegt um den Wert der Gruppengröße höher als die Untergrenze. Alle weiteren Gruppen werden in gleichmäßigen Schritten fortgeführt.

Bei der Auswahl von Spalten des Datentyps „DateTime“ werden die vorhandenen Werte nach Zeiteinheiten gruppiert. Dazu kann in einer eingeblendeten Drop-Down-Liste zwischen Jahren, Monaten, Tagen und Wochentagen ausgewählt werden, um die Datensätze nach der entsprechenden Zeiteinheit zu gruppieren.

Bei der Auswahl einer Spalte mit dem Datentyp „String“ werden alle unterschiedlichen Werte dieser Spalte als mögliche Distributionsgruppen herausgesucht. Weitere Gruppierungen sind hier nicht möglich.

Unabhängig von dem Datentyp der ausgewählten Spalte werden die erstellten Distributionsgruppen in einer Liste unterhalb der Spaltenauswahlliste angezeigt (siehe Abb. 5.23). Diese Liste ermöglicht dem Benutzer mittels Kontrollkästen einzelne dieser Gruppen ein- oder auszublenden. Dabei wird die Verteilung der eingeblendeten Gruppen in Echtzeit auf der Karte aktualisiert. Um die Identifikation der einzelnen Spalten zu vereinfachen, wird neben jeder der erstellten Gruppen, die im Diagramm verwendete Farbe angezeigt. In Fällen, in denen die entstandene Verteilung eine zu hohe Anzahl an Distributionsgruppen enthält, wird eine Warnung ausgegeben und die Anzeige der Gruppen abgebrochen. Somit

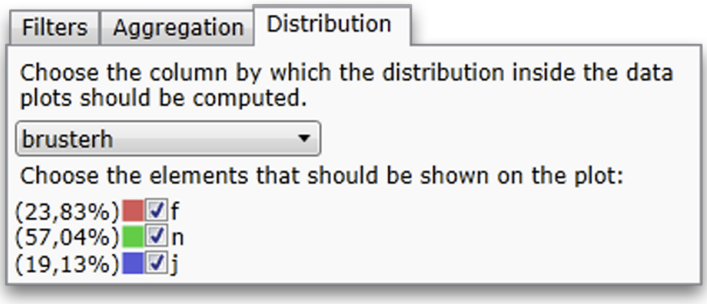


Abbildung 5.23.: Bedienelemente der Distributionsfunktion.

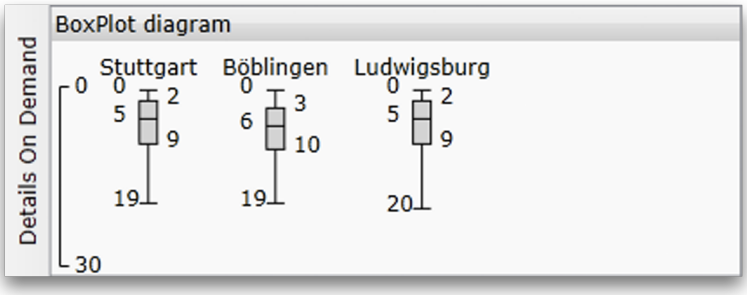


Abbildung 5.24.: Verteilung der Überlebenszeiten von Patienten dreier Landkreise.

soll vermieden werden, dass Spalten, die einmalige Werte beinhalten, angezeigt werden. In solchen Fällen können keine einzelnen Gruppen mehr identifiziert werden und die Analyse ist somit nicht mehr möglich. Über das Menü „Settings“ (siehe Abschnitt 5.3.13) kann die Anzahl der maximal angezeigten Distributionsgruppen angepasst werden. Standardmäßig liegt diese bei 25 Gruppen.

5.10. Die Boxplot-Diagramme

Im linken Teil des Arbeitsbereiches befindet sich unterhalb der „Filter and Zoom“-Elemente der Bereich „Details on Demand“. Dieser Bereich ermöglicht die Anzeige von detaillierten Informationen zu ausgewählten geografischen Gruppen.

Das erste Steuerelement dieses Bereiches zeigt Boxplot-Diagramme der ausgewählten Gruppen nebeneinander an. Die in den Boxplots angezeigten Werte werden aus der Variablen berechnet, die in der Aggregationsfunktion ausgewählt wurde. Mit dem Namen der jeweiligen Gruppe sind die einzelnen Diagramme beschriftet und durch Bewegen der Maus über eines dieser Diagramme wird die dazugehörende Gruppe auf der Karte hervorgehoben. In Abb. 5.24 werden drei Boxplots angezeigt, die die Verteilung der Überlebenszeiten für Patienten aus den Landkreisen Stuttgart, Böblingen und Ludwigsburg darstellen.

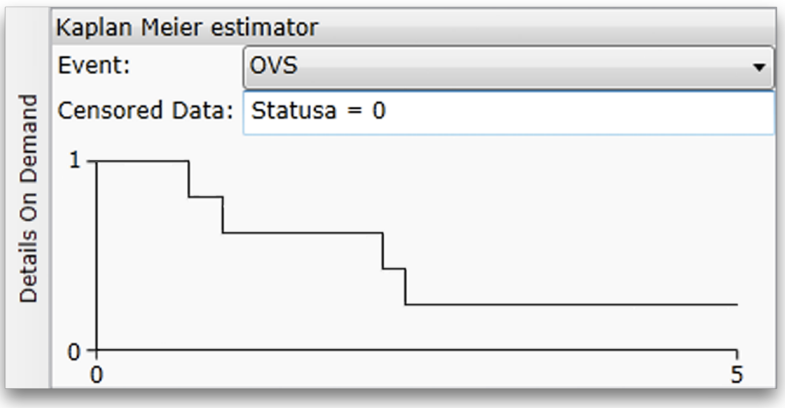


Abbildung 5.25.: Standardansicht des Kaplan-Meier-Steuerelementes.

5.11. Der Kaplan-Meier-Schätzer

Der Kaplan-Meier-Schätzer ist ebenfalls Teil des „Details on Demand"-Bereiches. Wie bei den Boxplot-Diagrammen lassen sich damit unterschiedliche geografische Bereiche miteinander vergleichen. Abb. 5.25 zeigt das benutzerdefinierte Steuerelement, in dem die Kaplan-Meier Kurve angezeigt wird.

Im oberen Teil des Steuerelementes hat der Benutzer die Möglichkeit, die Daten auszuwählen, die bei der Berechnung der Grafik verwendet werden sollen. Aus der „Event"-Liste wird die Spalte ausgewählt, die die Dauer von Beginn der Beobachtung bis zum Eintreten des Ereignisses enthält. In dem „Censored Data"-Textfeld wird die Bedingung eingegeben, die die zensierten Daten beschreibt (siehe 2.3.3). Diese wird folgendermaßen interpretiert: ein Ereignis ist genau dann eingetreten, wenn die Bedingung erfüllt ist, ansonsten gilt der Datensatz als zensiert.

Sofern sie beim Importieren der Daten ausgewählt wurden, werden standardmäßig die Spalten der Brustkrebsdatei eingestellt, die für die Analyse der Überlebenszeiten der Patienten benötigt werden. Die „Event"-Auswahlliste beinhaltet die „OVS"-Spalte, die die Dauer zwischen der Diagnose und dem Eintreten des Todes bzw. des Abbruches der Beobachtung des Patienten angibt. Das „Censored Data"-Textfeld enthält die Formel „$Statusa = 0$". Die Spalte „Statusa" enthält die Werte 1 für verstorbene Patienten und 0 für alle anderen. Alternativ dazu könnte auch die Bedingung „$abgru = t$" verwendet werden. Die Spalte „abgru" beinhaltet den Grund für den Abbruch der Beobachtung der Patienten.

Der Kaplan-Meier-Schätzer kann nicht nur für die Überlebenszeitanalyse, sondern für beliebige Analysen der Zeit bis zum Eintreten eines Ereignisses verwendet werden. So kann durch die Auswahl der Spalte „RFS" („rezidivfreies Überleben") in der „Event"-Liste beispielsweise die Zeit von der Diagnose bis zum Wiederauftreten der Krankheit analysiert werden. Die Bedingung, die die zensierten Daten beschreibt, ist dabei „$RFS < OFS$". Der

RFS-Wert wird bei Nichtwiederauftreten der Krankheit vor dem Ende der Beobachtungszeit auf den Wert der beobachteten Überlebenszeit gesetzt.

Im unteren Teil des Steuerelementes werden die Kaplan-Meier-Kurven angezeigt. Bei Auswahl einer geografischen Gruppe auf der Karte wird die Kurve unter Berücksichtigung der ausgewählten Ereignisspalte und der eingegebenen Definition der zensierten Daten aus den Datensätzen, die zu der ausgewählten Gruppe gehören, berechnet und in das Diagramm eingezeichnet. Dabei wird auf der Ordinate das Verhältnis der zu einem Zeitpunkt t der Abszisse noch beobachteten lebenden Patienten zu allen Patienten, deren Beobachtung nicht unterbrochen wurde, angezeigt. Bei Auswahl mehrerer Gruppen werden die Kaplan-Meier-Kurven übereinander angezeigt um den direkten Vergleich der Gruppen zu ermöglichen. Wird die Maus über eine der ausgewählten Gruppen bewegt, wird die dazugehörende Kaplan-Meier-Kurve hervorgehoben.

5.12. Die Datenansicht

Ein weiteres Element, das zum Bereich „Details on Demand“ gehört, ist die Datenansicht. Diese wird aufgrund ihrer Maße separat von den zwei schon erwähnten Funktionen, im unteren Teil des Fensters positioniert. Darin werden die importierten Daten in einer Tabelle angezeigt. Durch Drücken der Spaltenkopfzeilen können die Daten darin nach den entsprechenden Spalten sortiert werden. Die durch die Filterfunktion (siehe Abschnitt 5.7) herausgefilterten Datensätze werden durch Ausgrauung und Kursivdruck gekennzeichnet. Datensätze, die zu den auf der Karte ausgewählten geografischen Gruppen gehören, werden durch Fettschrift hervorgehoben. Sowohl die Filterung, als auch die Auswahl der Datensätze, wird durch eine entsprechende Markierung in den Spalten „fil“ und „sel“ gekennzeichnet. Deshalb kann die Tabelle durch Drücken der Spaltenkopfzeilen auch nach diesen Kriterien sortiert werden. Abb. 5.26 zeigt einen Ausschnitt dieser Tabelle, in dem die Datensätze nach dem Geburtstag der Patienten sortiert sind und in dem sowohl herausgefilterte, als auch ausgewählte Datensätze, sichtbar sind.

Durch Anklicken einzelner Zeilen können Datensätze zusätzlich hervorgehoben werden. Diese Hervorhebung wird durch einen blauen Hintergrund gekennzeichnet. Bei gedrückt gehaltener „STRG“-Taste und mehrmaligem Anklicken werden mehrere Zeilen ausgewählt. Die so hervorgehobenen Datensätze können mit dem Tastaturkürzel „STRG+C“ kopiert und in andere Anwendungen eingefügt werden.

5.13. Die Zeitleiste

Wie im Kapitel 4 bereits erwähnt, soll dem Benutzer auch die Möglichkeit gegeben werden, mit den zu analysierenden Daten aus zeitlicher Perspektive zu interagieren. Dazu wurde die Zeitleiste entwickelt, die unterhalb der Kartendarstellung positioniert ist und dem Benutzer erlaubt, die angezeigten Daten zeitlich zu filtern. Das dazu gehörende Benutzersteuerelement wird in Abb. 5.27 dargestellt.

sel	fil	Column1	code	IZ	gebdat	geschl	plz5	stand	Alterbei_Diagnose	OVS	RFS	abgru	St
☑	☑	*351*	*C01210260*	*2571*	*5/30/1903 12:00:00 AM*	*w*	*70199*	*12/5/1995 12:00:00 AM*	*90.03*	*2.49*	*2.49*	*t*	*1*
☐	☑	*1678*	*C02212839*	*3739*	*5/31/1903 12:00:00 AM*	*w*	*71522*	*2/12/1996 12:00:00 AM*	*91.69*	*1.02*	*1.02*	*t*	*1*
☑	☐	**3188**	**C02410523**	**666**	**6/7/1903 12:00:00 AM**	**w**	**70193**	**7/1/1995 12:00:00 AM**	**86.97**	**5.1**	**4.38**	**t**	**1**
☑	☐	**3002**	**C02410121**	**239**	**6/20/1903 12:00:00 AM**	**w**	**70186**	**9/1/1995 12:00:00 AM**	**85.81**	**6.39**	**6.39**	**t**	**1**
☑	☑	*11690*	*C04218923*	*8268*	*8/7/1903 12:00:00 AM*	*w*	*70192*	*3/11/2003 12:00:00 AM*	*96.07*	*3.52*	*0.95*	*t*	*1*
☐	☐	11247	C04115580	3290	9/9/1903 12:00:00 AM	w	71254	1/12/2005 12:00:00 AM	88.48	12.86	6.87	t	1
☐	☐	11505	C04216075	1866	10/7/1903 12:00:00 AM	w	74385	11/7/1997 12:00:00 AM	88.86	5.23	5.23	t	1
☑	☑	*6962*	*C02416382*	*12586*	*11/4/1903 12:00:00 AM*	*w*	*70599*	*1/11/2005 12:00:00 AM*	*100.41*	*0.78*	*0.78*	*t*	*1*
☑	☑	*13942*	*C04316699*	*762*	*1/20/1904 12:00:00 AM*	*w*	*70191*	*6/29/1990 12:00:00 AM*	*86.21*	*0.23*	*0.23*	*m*	*0*
☑	☑	*3190*	*C02410529*	*777*	*2/20/1904 12:00:00 AM*	*w*	*70619*	*4/22/1992 12:00:00 AM*	*86.3*	*1.87*	*1.87*	*t*	*1*
☑	☑	*1938*	*C02214691*	*8504*	*2/28/1904 12:00:00 AM*	*w*	*70599*	*12/8/2002 12:00:00 AM*	*95.75*	*3.02*	*3.02*	*t*	*1*
☑	☑	*11480*	*C04215877*	*1490*	*3/2/1904 12:00:00 AM*	*w*	*70186*	*12/2/1991 12:00:00 AM*	*87.75*	*0*	*0*	*m*	*0*
☐	☐	1462	C02211739	1646	3/15/1904 12:00:00 AM	w	70734	8/3/2005 12:00:00 AM	88.24	13.15	13.15	t	1

Abbildung 5.26.: Datenansicht.

Alle zeitliche Filter, die mit Hilfe der Zeitleiste erstellbar sind, können ebenfalls über die Filterfunktion aus Abschnitt 5.7 erstellt werden. Die Filterfunktion erlaubt sogar die Definition komplexerer zeitlicher Einschränkungen, als es mit Hilfe der Zeitleiste möglich ist. Allerdings bietet diese eine übersichtlichere Bedienung bei der Anzeige von Zeitintervallen und ermöglicht eine einfachere und schnellere Interaktion.

Abbildung 5.27.: Zeitachse mit Anzeige der Verteilung der Geburtsjahre der Patienten.

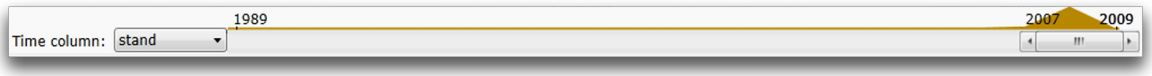


Abbildung 5.28.: Zeitachse mit Anzeige des Datums der letzten Beobachtung der Patienten.

Bei der Verwendung der Zeitachse muss im ersten Schritt aus der „Time column“-Liste das zu beobachtende Merkmal der Datensätze ausgewählt werden. Dieses Merkmal stellt den Zeitpunkt dar, zu dem ein bestimmtes Ereignis eingetreten ist. Dazu sind in der Auswahlliste nur die Spalten der Brustkrebsdatei auswählbar, die beim Importieren der Daten als „DateTime“-Format definiert wurden. Nach Auswahl einer Spalte wird die Zeitleiste aktiviert und umfasst die Zeitspanne von dem Jahr des frühesten bis zu dem Jahr des spätesten Zeitpunktes, der in der ausgewählten Spalte enthalten ist. Die zwei Schieberegler am linken und rechten Rand der Zeitachse bestimmen das Zeitintervall, das die geladenen Daten filtert. Für die Berechnung der Aggregations- und der Distributionsfunktion werden nur Datensätze berücksichtigt, deren ausgewähltes Ereignis innerhalb dieses Zeitintervalls liegt. Durch Verschieben der zwei Schieberegler am linken und rechten Rand der Zeitachse kann die Länge des definierten Zeitintervalls angepasst werden. Die Schieberegler können jahresgenau positioniert werden. Die Leiste zwischen den Schiebereglern kann direkt durch Anklicken und Bewegen oder durch Anklicken der Bereiche links oder rechts davon verschoben werden, was eine Verschiebung der Endpunkte des Intervalls bewirkt, ohne dessen Länge zu verändern.

Oberhalb der Zeitachse wird die Verteilung der Ereignisse in einer Grafik verdeutlicht. Dadurch soll dem Benutzer eine Übersicht der Ereignishäufungen gegeben sowie die Erkennung dieser unterstützt werden. Die Grafik in Abb. 5.27 zeigt die Verteilung der Geburtsjahre aller Patienten. In Abb. 5.28 ist die Spalte „stand“ (das Datum der letzten Beobachtung eines Patienten) ausgewählt. Bedingt durch die Laufzeit der OSP-Datenbank erstreckt sich die Zeitachse von dem Jahr 1988 bis zum Jahr 2009. Dadurch, dass die Daten aller nicht verstorbenen und weiterhin unter Beobachtung stehenden Patienten regelmäßig aktualisiert werden, befinden sich die meisten Angaben der letzten Beobachtung gegen Ende der Zeitspanne, was die Häufung der Ereignisse im Jahr 2008 erklärt.

6. Prototyp

Im folgenden Kapitel werden zunächst die technischen Grundlagen der MVA-Anwendung erklärt. Danach werden die Architektur und die Bestandteile der Anwendung vorgestellt. Im letzten Abschnitt wird die Implementierung verschiedener Funktionen der Anwendung erklärt und die bei der Entwicklung aufgetretenen Herausforderungen und Probleme sowie deren Lösung beschrieben.

6.1. Voraussetzungen für die Ausführung

Die MVA-Anwendung kann nur unter dem Betriebssystem Microsoft Windows gestartet werden. Außerdem benötigt es das .NET 4.0-Framework, das unter [Dot11] kostenlos heruntergeladen werden kann.

6.2. Programmierumgebung

Die MVA-Anwendung wurde auf einer virtuellen Maschine unter dem Betriebssystem Microsoft Windows 7 Professional programmiert. Der virtuellen Maschine wurden 4 Prozessorkerne und 2 Gigabyte Arbeitsspeicher zugewiesen.

Die verwendete Programmierumgebung war Visual Studio 2010, die zu Beginn der Arbeit aktuellste Version von Visual Studio. Bei der Entwicklung wurde die .NET 4.0-Bibliothek verwendet. Als Programmiersprache kam C# zur Anwendung.

6.3. Architektur der Anwendung

Die MVA-Anwendung wurde nach dem MVVM-Muster implementiert. Detaillierte Informationen über das MVVM-Muster werden in Abschnitt 6.5.1 gegeben.

6.4. Aufbau der Projektmappe

Die Projektmappe der MVA-Anwendung besteht aus folgenden Projekten:

- csvConverter
- GetLocations
- MedicalVisualAnalytics
- MVAGraphControlLib
- WPFHelper

Diese Projekte werden im Folgenden beschrieben.

6.4.1. csvConverter

Das csvConverter-Projekt ist eine Hilfsanwendung, die als Consolenanwendung ohne graphische Benutzeroberfläche programmiert wurde. Sie dient der Umwandlung der Kommata aus der ursprünglichen csv-Datei in Semikolons. Diese Umwandlung ist Teil der Vorverarbeitung der Brustkrebsdatei, die in Abschnitt 3.5 beschrieben wird.

Bei der Umwandlung dürfen nur Kommata ersetzt werden, die als Feldbegrenzer dienen. Kommata, die innerhalb eines Strings vorkommen, dürfen nicht ersetzt werden. Da diese Anforderung nicht durch einen regulären Ausdruck umgesetzt werden konnte, wurde diese Anwendung geschrieben.

Die Funktionsweise der csvConverter-Anwendung wird in Listing 6.1 dargestellt.

Die Anwendung erwartet die zu bearbeitende csv-Datei als Kommandozeilenparameter beim Aufruf. Die angegebene Datei wird zeilenweise eingelesen und jede Zeile wird in einen Array geteilt. Dabei dienen die Anführungszeichen als Trennzeichen. In dem so entstandenen Array sind alle geraden Elemente Teil eines Textfeldes, alle ungeraden befinden sich außerhalb von Textfeldern. In den ungeraden Elementen können alle Kommata durch einen Semikolon ersetzt werden. Zum Schluss wird die neue Datei unter einem neuen Namen gespeichert.

Algorithmus 6.1 Pseudocode der csvConverter-Anwendung

```
inFile ← readFile(argument1)
for all line ∈ inFile.Lines do
    lineArray ← line.split('"')
    for i = 1 to lineArray.Length step 2 do
        lineArray[i].Replace(',';')
    end for
end for
writeFile(argument1.Replace(".csv","_replaced.csv"))
```

Gesuchter Bereich	Benötigte Parameter
Postleitzahl	&postalCode=*postleitzahl*
Ortschaft	&admindistrict=*Landkreis*&locality=*Ortschaft*
Landkreis	&admindistrict=*Landkreis*
Bundesland	&admindistrict=*Bundesland*

Tabelle 6.1.: Benötigte Parameter bei der Koordinatensuche mit der Bing Maps API.

6.4.2. GetLocations

Das GetLocations-Projekt ist ebenfalls eine eigenständige Konsolenanwendung ohne Benutzeroberfläche. Es wird für die Vervollständigung der geografischen Koordinaten in der Liste der Postleitzahlen, Ortschaften, Landkreise und Bundesländer verwendet.

Bei der Anreicherung der Daten mit Geokoordinaten (siehe Abschnitt 5.1) wird die Bing Maps-API verwendet. Diese stellt eine Webadresse zur Verfügung, über die die geografischen Koordinaten gesuchter Begriffe abgefragt werden können. Die Details der Anfragen werden als GET-Parameter hinter der Adresse angefügt. Listing 6.1 zeigt eine mögliche Anfrage an die Programmierschnittstelle.

(6.1)
```
http://dev.virtualearth.net/REST/v1/Locations?o=xml&maxResults=1
                &countryRegion=DE&postalCode=70173&key=XXXX
```

Der Webdienst antwortet bei Angabe des Parameters „o=xml“ mit einer XML-Datei. Ohne Verwendung dieses Parameters wird eine JSON-Datei zurückgegeben. Da eine manuelle Revision der Ergebnisse durch den Benutzer nicht gewünscht ist, wird die Anzahl der zurückgegebenen Ergebnisse durch den Parameter „maxResults=1“ auf ein Ergebnis beschränkt. Jeder Anfrage muss der Lizenzschlüssel in dem Parameter „key“ angefügt werden. Dadurch, dass nicht die Koordinaten einer genauen Adresse, sondern die höherer administrativer Bereiche gesucht werden, wird bei der Suche eine sogenannte „unstrukturierte URL“ verwendet [Msdd]. Tabelle 6.1 listet die Parameter auf, die bei der Suche nach diesen Bereichen angegeben werden.

6.4.3. MedicalVisualAnalytics

Dieses WPF-Projekt beinhaltet den Hauptteil des Prototyps. Dazu gehören alle erstellten Views, Models und ViewModels (siehe Abschnitt 6.5.1).

6.4.4. MVAGraphControlLib

Das MVAGraphControlLib-Projekt erzeugt als Ausgabe eine Klassenbibliothek, die in anderen Projekten importiert werden kann. Es beinhaltet alle benutzerdefinierten Steuerelemente

Klasse	Beschreibung
BoxPlot	Implementierung eines Boxplot-Diagrammes.
BoxPlotList	Ein Container für die Darstellung mehrerer Boxplot-Diagramme.
KaplanMeier	Implementierung der Kaplan-Meier-Kurve.
LineGraph	Implementierung eines Kurvendiagramms.
PieChart	Implementierung eines Kuchendiagramms bestehend aus mehreren Teilen.
PiePiece	Implementierung eines Teils eines Kuchendiagrammes.
RangeSlider	Implementierung eines Schiebereglers für Wertebereiche.

Tabelle 6.2.: Suchparameter der Anfragen an die Bing Maps-API.

(siehe 6.5.3), die in der MVA-Anwendung benötigt werden. Eine Liste dieser Steuerelemente und deren Beschreibungen findet sich in Tabelle 6.2.

6.4.5. WPFHelper

Das WPFHelper-Projekt erzeugt ebenfalls eine Klassenbibliothek und enthält eine Reihe von Hilfsklassen, die sowohl im Projekt MedicalVisualAnalytics, als auch in der MVAGraphControlLib-Bibliothek benötigt werden. Beide Projekte verweisen auf die WPFHelper.dll-Datei und können somit auf diese Hilfsklassen zugreifen.

Die Klassen und die Namensräume dieses Projektes, sowie deren Beschreibung, werden in Tabelle 6.3 vorgestellt.

6.5. Grundlagen

Dieser Abschnitt erklärt die wichtigsten Konzepte und Techniken, die bei der Entwicklung der MVA-Anwendung verwendet wurden. Genaue Details der Implementierung werden in den folgenden Abschnitten beschrieben.

6.5.1. MVVM und Datenbindung

Das MVVM-Muster schreibt ähnlich des MVC-Musters [MVC07] eine Trennung zwischen der grafischen Benutzeroberfläche (die View) und der Speicherung der Daten (das Model) vor. Das MVVM-Muster ist eine Spezialisierung des allgemeinen Presentation Model-Musters [MVV09, Fow04] von Martin Fowler, in dem eine Abstraktion der View beschrieben wird, die Zustand und Funktionalität der View enthält. MVVM wurde im Jahr 2005 als Muster für eine flexible Entwicklung von Benutzeroberflächen mit WPF vorgestellt [MVV05]. Anders als beim MVC-Muster steuert die dritte Komponente des MVVM-Musters – das

Klasse oder Namensraum	Beschreibung
PrismEventAggregator (Namensraum)	Enthält alle Ereignisse und benutzerdefinierte Ereignisparameter für die Kommunikation zwischen verschiedenen Bereichen der Anwendung.
ColorListProvider	Erstellt Farbenlisten, die für die Darstellung der Diagramme verwendet werden.
ExtendedObservableCollection	Erweitert die ObservableCollection-Klasse, so dass Änderungen in den Elementen der Sammlung über „PropertyChanged“-Ereignisse bekannt gegeben werden. Benutzerelemente mit Datenbindung an die Sammlung werden dadurch automatisch aktualisiert.
GeoDataGroup	Modell für die geografischen Gruppen, die auf der Karte angezeigt werden.
RelayCommand [MVV09]	Ermöglicht die Behandlung von Ereignissen der Benutzeroberfläche im ViewModel anstatt in der Codebehind-Datei. Das MVVM-Muster wird dadurch aufrecht erhalten.

Tabelle 6.3.: Elemente des WPFHelper-Projektes.

ViewModel – die Aktualisierung der Elemente in der View nicht selbst. Aktualisierungen werden über Datenbindungen zwischen den Elementen der View und den Eigenschaften des ViewModels durchgeführt. Dazu muss das ViewModel bei Änderungen von Eigenschaften ein „PropertyChanged“-Ereignis auslösen und den Namen der geänderten Eigenschaft als Parameter des Ereignisaufrufes übermitteln. Die View-Klasse überwacht diese Ereignisse des ViewModels und aktualisiert die Elemente, die an die geänderte Eigenschaft gebunden sind. Datenbindungen werden im XAML-Code definiert. Listing 6.1 zeigt ein Beispiel für eine einfache Datenbindung, in der der Inhalt einer Textbox an der Eigenschaft „GeoResolution“ des ViewModels gebunden wird.

Listing 6.1 Beispiel einer einfachen Datenbindung

```
<TextBlock Text="{Binding GeoResolution}" />
```

Neben der Bindung an eine Eigenschaft des ViewModels ist es in XAML möglich, Eigenschaften von Elementen der View direkt untereinander zu binden. So lässt sich beispielsweise die Anzeige eines Benutzerelementes an den Status eines Kontrollkästchens binden. Für Datenbindungen zwischen inkompatiblen Datentypen (beispielsweise Integer und Boolean) können Konvertoren geschrieben werden, die bei der Definition der Datenbindungen angegeben und bei der Aktualisierung der Datenbindungen aufgerufen werden.

6.5.2. Benutzersteuerelemente

Benutzersteuerelemente („User Controls“) sind neue Elemente, die durch die Gruppierung bestehender Steuerelemente entstehen. Dadurch wird die Wiederverwendbarkeit und die Übersichtlichkeit komplexer Oberflächen verbessert. User Controls bieten keine zusätzliche Funktionalität und keine neuen Eigenschaften an, die in Datenbindungen verwendet werden können. Datenbindungen müssen über geeignete ViewModels, an denen die Eigenschaften der Teilelemente gebunden sind, implementiert werden.

Das Filter-Control ist das komplexeste Benutzersteuerelement der MVA-Anwendung.

6.5.3. Benutzerdefinierte Steuerelemente

Benutzerdefinierte Steuerelemente („Custom Controls“) sind neue Elemente, die durch Vererbung von einer Control-Klasse definiert werden. Sie können um neue Abhängigkeitseigenschaften („Dependency Properties“) erweitert werden, die auch in Datenbindungen verwendet werden können. Außerdem erlauben sie die Implementierung neuer Funktionen.

Das TimeLine-Control der MVA-Anwendung wurde als benutzerdefiniertes Steuerelement entwickelt.

6.6. Implementierungsbeispiele

6.6.1. Datenstruktur

Die zentrale Datenstruktur der MVA-Anwendung ist die `DataTable`-Klasse [Msdc]. Beim Importieren einer neuen Datei werden alle Datensätze in einem `DataTable`-Objekt gespeichert. Eine `DataTable` eignet sich für die Speicherung tabellarischer Daten und erlaubt die bequeme Berechnung von Aggregationsfunktionen einzelner Spalten mittels der Mitgliedsfunktion `Compute`. Die `Compute`-Funktion erwartet zwei Textparameter:

- die zu berechnende Aggregationsfunktion im Format „fct(Spaltenname)“
- eine Filterfunktion die angibt, auf welche Datensätze die Aggregationsfunktion angewandt werden soll

Die `Compute`-Funktion unterstützt außerdem die wichtigsten statistischen Werte [Dato1]:

- Minimum („min“)
- Maximum („max“)
- Anzahl („count“)
- Mittelwert („avg“)

- Summe („sum“)
- Varianz („var“)
- Standardabweichung („stdev“)

Weitere Berechnungen können der `Compute`-Funktion allerdings nicht hinzugefügt werden, so dass nicht unterstützte Berechnungen manuell durchgeführt werden müssen. Vor dem Aufruf der Compute-Funktion wird der Wert der `DataAggregation.SelectedDataFunction`-Variablen überprüft. Handelt es sich dabei um eine unterstützte Berechnung, wird die Compute-Funktion aufgerufen. Ansonsten wird eine selbst entwickelte Funktion aufgerufen. Im Listing 6.2 wird diese Lösung am Beispiel der Median-Funktion erläutert. Bei Auswahl der „Median“-Funktion werden die Datensätze der aktuellen geografischen Gruppen nach der ausgewählten Spalte sortiert und als Median entweder das mittlere Element (Zeile 8) oder das arithmetische Mittel der zwei mittleren Elemente (Zeile 10) zurückgegeben. Bei Auswahl einer unterstützten Funktion wird die Funktion `Compute` ausgeführt (Zeile 13).

Listing 6.2 Erweiterung der Aggregationsfunktionen des `DataTable`.

```
1  if(DataAggregation.SelectedAggregationFunction.Key == "median")
2  {
3    String selectedColumn = DataAggregation.SelectedColumn.Key;
4    DataRow[] rows = group.DataTable.Select(
5      getAllFilters(),
6      selectedColumn + " ASC");
7      if (rows.Count() % 2 == 1)
8        tempSize = (double)rows[rows.Count() / 2][selectedColumn];
9      else
10       tempSize = ((double)rows[rows.Count() / 2][selectedColumn] +
             (double)rows[rows.Count() / 2 - 1][selectedColumn]) / 2;
11 }
12 else
13   tempSize =
         Double.Parse(group.DataTable.Compute(DataAggregation.StringDataAggregation,
         getAllFilters()).ToString());
```

6.6.2. FilterView

Die „FilterView“-Komponente ist als wiederverwendbares UserControl implementiert. Die Implementierung des Filter-Controls ist auf folgende Klassen verteilt:

- `BasicFilterModel`
- `DisjunctiveFilterModel`
- `ConjunctiveFilterModel`

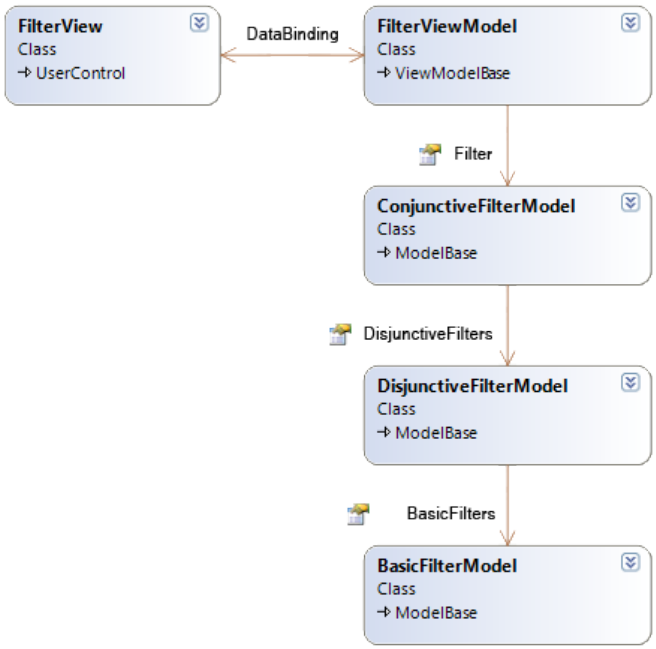


Abbildung 6.1.: Abhängigkeitsverhältnisse der Klassen des Filter-Controls.

- `FilterViewModel`

Abb. 6.1 verdeutlicht die Zusammenarbeit dieser Klassen. In den nächsten Abschnitten wird die Implementierung der Klassen beschrieben. Die relevanten Eigenschaften der BasicFilterModel-, der DisjunctiveFilterModel- und der ConjunctiveFilterModel-Klasse sind in den Listings A.1, A.2 und A.3 dargestellt.

BasicFilterModel

Die Klasse `BasicFilterModel` implementiert die einzelnen Filterterme. Sie erlaubt die Erstellung von Termen, bestehend aus Negation (optional), Aggregationsfunktion (optional), Spaltenname und Bedingung. Je nach Einsatz des Filter-Controls können die optionalen Elemente angezeigt oder versteckt werden. Die Aggregationsfunktionen werden zum Beispiel bei der Filterung der Daten nicht angezeigt, da die Filterung nicht auf mehrere Datensätze, deren Daten aggregiert werden können, sondern auf einzelne angewandt wird.

Die Negation, die Aggregationsfunktion und der Spaltenname sind als Drop-Down-Listen implementiert. Sie werden mit vorgegebenen Werten (bei der Negation und Aggregationsfunktion) bzw. mit der Liste der geladenen Spalten des Datensatzes (beim Spaltennamen) belegt. Die Bedingung des Filters ist als Textfeld implementiert und kann vom Benutzer frei eingegeben werden.

Die Werte, die in der Liste der Spalten angezeigt werden, können bei bestimmten ausgewählten Aggregationsfunktionen weiter eingeschränkt werden. Die Aggregationsfunktionen „avg“ (Mittelwert) oder „var“ (Varianz) können nur auf numerische Spalten angewandt werden. Bei der Auswahl dieser Funktionen werden alle nichtnumerischen Spalten ausgeblendet.

Bei jeder Änderung der Elemente der `BasicFilterModel`-Klasse wird der neu entstandene Filter auf Korrektheit geprüft. Das Ergebnis dieser Überprüfung wird in der Variablen „Error“ gespeichert. Wurde ein fehlerhafter Filter entdeckt, so wird das Bedingungstextfeld mit einem roten Rahmen hervorgehoben, da Fehler nur vom Benutzer eingegeben werden können. Ist der erstellte Filter korrekt, so wird die Änderung über das Auslösen des „FilterChanged“-Ereignisses bekanntgegeben. Klassen, die dieses Ereignis überwachen, können mit Hilfe der `FilterString`-Eigenschaft eine Textversion des Filters abrufen, die folgendermaßen aussieht:

(6.2) `Not(Aggregationsfunktion(Spaltenname)) Bedingung`

Die farbig hervorgehobenen Teile „`Not(...)`“ und „`Aggregationsfunktion(...)`“ können je nach Konfiguration des Filters oder der Auswahl des Benutzers fehlen.

Die Überprüfung eines Filters auf Korrektheit ist in C# nicht vor dem eigentlichen Filtervorgang möglich. Aus diesem Grund werden Filter nach Änderungen auf ein leeres DataTable-Objekt angewandt und aufgetretene Ausnahmen über einen „try/catch“-Block abgefangen.

DisjunctiveFilterModel

Die Klasse `DisjunctiveFilterModel` enthält in der `_basicFilters`-Eigenschaft eine Liste von `BasicFilterModel`-Objekten, die die einzelnen Terme einer Disjunktion darstellen. Die Eigenschaften `_showNegationColumn` und `_showAggregationColumn` geben an, ob diese optionalen Spalten in den Termen dieser Disjunktion angezeigt werden sollen.

Objekte der Disjunktionsklasse überwachen das Ereignis `FilterChanged` der untergeordneten `BasicFilterModel`-Objekten. Bei einem Aufruf dieses Ereignises wird die `Error`-Eigenschaft aller Terme überprüft. Enthält eines davon einen Fehler, dann wird auch die `Error`-Eigenschaft des Disjunktionsobjektes auf dem boolschen Wert „wahr“ gesetzt. Da diese Klasse nur eine Zusammensetzung der Terme zu einer Disjunktion durchführt und es somit keine andere mögliche Fehlerquellen gibt, muss keine Fehlerüberprüfung der entstandenen Disjunktion erfolgen.

Unabhängig von dem Ergebnis der Fehlerüberprüfung lösen auch `DisjunctiveFilterModel`-Objekte ein `FilterChanged`-Ereignis aus. Die `FilterString`-Eigenschaft stellt die Disjunktion der Terme als Textversion zur Verfügung. Das Ergebnis sieht folgendermaßen aus:

(6.3) `(`*term*1.*FilterString*`) OR (`*term*2.*FilterString*`) OR ...`

ConjunctiveFilterModel

Die `ConjunctiveFilterModel`-Klasse ähnelt der `DisjunctiveFilterModel`-Klasse. Auch sie enthält eine Liste untergeordneter Elemente und die zwei Eigenschaften zur Anzeige der optionalen Spalten in der Termen.

Die Fehlerüberprüfung findet nach demselben Prinzip wie bei der `DisjunctiveFilterModel`-Klasse statt. Die `Error`-Eigenschaften der enthaltenen Disjunktionen werden überprüft und davon abhängig die eigene `Error`-Eigenschaft gesetzt. Diese wird in der Datenbindung der View-Komponente des Filter-Controls für die farbige Hervorhebung des Filters im Falle eines Fehlers verwendet.

Die `FilterString`-Eigenschaft dieser Klasse wird ebenfalls in der Datenbindung der View für die Anzeige des entstandenen Filters verwendet. Diese Eigenschaft verknüpft die untergeordneten DisjunctiveFilterModel-Objekten mit Hilfe des „UND“-Operators zu einer Konjunktion:

(6.4) `(`*disjunktion*1.*FilterString*`) AND (`*disjunktion*2.*FilterString*`) AND ...`

FilterViewModel

Die `FilterViewModel`-Klasse stellt die höchste Ebene der Datenbindungshierarchie dar. Diese Klasse wird direkt an der View des Filter-Controls gebunden und enthält neben einem `ConjunctiveFilterModel`-Objekt noch Eigenschaften für den Informationstext, die Anzahl der gesamten und der gefilterten Datensätze.

Das `FilterChanged`-Ereignis, das sich von der `BasicFilterModel`-Klasse nach oben propagiert hat, wird auch hier abgefangen. Er wird von der `FilterViewModel`-Klasse nur im Falle eines korrekten Filters ausgelöst. Somit wird vermieden, dass die Datensätze der Anwendung durch einen fehlerhaften Filter gefiltert werden.

6.6.3. Parallelität

Standardmäßig wird eine WPF-Anwendung in einem einzigen Thread ausgeführt. Das hat zur Folge, dass während langlaufenden Abläufen die Benutzerschnittstelle auf Befehle des Benutzers nicht reagieren kann, da jeder Funktionsaufruf beendet werden muss, bevor neue Funktionen aufgerufen werden können. Um das allgemeine Ansprechverhalten des entwickelten Prototyps zu erhöhen, wurden an verschiedenen Stellen asynchrone Abläufe implementiert. Diese werden in separaten Threads parallel zu der Benutzerschnittstelle ausgeführt und blockieren diese dadurch nicht.

Die einfachste Version der Parallelisierung wird mit Hilfe der Klasse `BackgroundWorker` aus dem Namensraum `System.ComponentModel` [Msdb] ermöglicht. Diese führt eine angegebene Aufgabe parallel zu dem Thread der Benutzerschnittstelle aus. Einzelne Schritte der auszuführenden Aufgabe selbst werden nicht parallel zueinander ausgeführt. Somit verkürzt

sich die Zeit der Bearbeitung nicht, lediglich das Reaktionsverhalten der Anwendung wird verbessert. Die `BackgroundWorker`-Klasse wird beim Importieren der Datensätze verwendet. Da die Daten dabei aus einer Datei stammen, deren Zeilen der Reihe nach eingelesen und bearbeitet werden müssen, ist eine parallele Ausführung einzelner Schritte dieser Aufgabe nicht möglich.

Listing 6.3 beinhaltet die Definition einer Instanz der `BackgroundWorker`-Klasse, wie sie beim Importieren der Daten verwendet wird. Nach der Instantiierung der Klasse wird in Zeile 2 die Delegat-Funktion für die auszuführende Arbeit bestimmt. Diese Funktion wird in einem separaten Thread ausgeführt. Da die Übermittlung des Fortschrittes dieser Funktion erwünscht ist, muss die Eigenschaft `BackgroundWorker.WorkerRepoersProgress` auf den Wert `true` gesetzt werden. Die Funktion `readData` meldet mittels des Aufrufs `BackgroundWorker.ReportProgress` den prozentualen Fortschritt der Ausführung. Beim Melden eines neuen Fortschrittswertes wird die Funktion `progressChanged` ausgeführt, die die Anzeige des Fortschrittbalkens aktualisiert. Schließlich wird nach Ende der `readData`-Funktion die Funktion `runWorkerCompleted` ausgeführt, die die Fortschrittsanzeige ausblendet und dem Benutzer eventuelle Fehler beim Import der Daten meldet.

Listing 6.3 Beispiel einer Instanz der `BackgroundWorker`-Klasse.

```
1  _bgWorker = new BackgroundWorker();
2  _bgWorker.DoWork += readData;
3  _bgWorker.WorkerReportsProgress = true;
4  _bgWorker.ProgressChanged += progressChanged;
5  _bgWorker.RunWorkerCompleted += runWorkerCompleted;
6  _bgWorker.RunWorkerAsync();
```

Für Anwendungsfälle, in denen durch eine parallele Ausführung von Aufgaben eine Erhöhung der Berechnungsgeschwindigkeit möglich ist, bietet das .NET-Framework ab der Version 4.0 die Klasse `System.Threading.Tasks.Task` [Msde] an. Diese ermöglicht die Definition von Aufgaben, die parallel in mehreren Threads auf allen zur Verfügung stehenden Prozessorkernen ausgeführt werden. Die Verwendung dieser Aufgaben bei der Implementierung der Parallelität wird anstatt der Verwendung von „ThreadPools“ [Msdf] empfohlen, da sie einen effizienteren Umgang mit Systemressourcen und eine bessere Kontrolle über die erstellten Aufgaben bietet. Aufgaben, die mit Hilfe der `Task`-Klasse erstellt werden, können bei Bedarf abgebrochen werden und die Ausführung weiterer Aufgaben kann nach dem Beenden der parallelen Ausführung aller Aufgaben eingeplant werden [Msda].

Diese zwei Vorteile werden bei der Aktualisierung der Kartendiagramme in Anspruch genommen. Bei schnellem Wechsel der Datenauflösung werden laufende Berechnungen abgebrochen und der Berechnung der Daten auf die neu gewählten Ebene gestartet. Für die Normalisierung der Größe der Kartendiagramme wird der maximale Wert aller geografischen Gruppen benötigt. Dieser ist erst nach vollständiger Berechnung aller Gruppen ermittelbar, so dass die Normalisierung erst nach dem Beenden aller Aufgaben durchgeführt werden kann.

Listing 6.4 veranschaulicht eine vereinfachte Version der Funktion `rebuildGroups` des entwickelten Prototyps. Zunächst werden noch aktive Berechnungen über die Funktion `Cancel`

des Objektes `CancellationTokenSource` abgebrochen. Danach wird eine Liste definiert, die die erstellten Aufgaben beinhalten wird. Die Schleife in den Zeilen 4 bis 12 iteriert durch die Tabelle `tempGeoDataGroups`, die die unterschiedlichen geografischen Gruppen beinhaltet und erstellt für jede Zeile eine asynchrone Aufgabe. Davor wird in Zeile 6 eine Kopie der Laufvariablen row erstellt, die für die Berechnung der Gruppe verwendet wird. Wird anstelle der lokalen Kopie die Laufvariable selbst verwendet, so werden alle Aufgaben den letzten Wert der Variablen row verwenden, da die Schleife bis zur Erstellung der Aufgaben beendet sein wird. In Zeile 10 wird der neu erstellten Aufgabe mitgeteilt, dass die Variable `_groupCts` für das Abbrechen der Berechnungen verwendet wird. In Zeile 11 wird die erstellte Aufgabe in die davor definierte Liste eingefügt. Zum Schluss wird nach dem Beenden aller Aufgaben die Funktion `normalizeGroups` ausgeführt.

Listing 6.4 Beispiel für die Verwendung der `Task`-Klasse.

```
1  _groupCts.Cancel();
2  _groupCts = new CancellationTokenSource();
3  List<Task> tasks = new List<Task>();
4  foreach (DataRow row in tempGeoDataGroups.Rows)
5  {
6    DataRow tempRow = row;
7    var rebuildGroupsTask = Task.Factory.StartNew(() =>
8    {
9     rebuildGroup(tempRow);
10   }, _groupCts.Token);
11   tasks.Add(rebuildGroupsTask);
12 }
13 Task.Factory.ContinueWhenAll(tasks.ToArray(),
14 result =>
15 {
16   normalizeGroups();
17 }, CancellationToken.None, TaskContinuationOptions.None, _ui);
```

Durch die Implementierung der Nebenläufigkeit konnte die Zeit für die Aktualisierung der Kartendiagramme beim Wechsel der Datenauflösung auf die Ebene der Postleitzahlen von 15 Sekunden auf 3,5 Sekunden reduziert werden.

6.6.4. Persistenz

Die MVA-Anwendung erlaubt das Speichern der geladenen Daten und der Arbeitsbereiche inklusive dem Zustand aller Bedienelemente in einer Projektdatei. Das Speichern und das Laden der Projekte kann vom Benutzer gesteuert (siehe Abschnitte 5.3.3 bis 5.3.5) oder von der Anwendung automatisch durchgeführt werden (siehe Abschnitt 5.3.15).

Durch die Implementierung des MVVM-Musters (siehe Abschnitt 6.5.1) muss für die Umsetzung dieser Funktion nur der Zustand der ViewModel-Klassen gespeichert werden. Die Datenbindungen zwischen den Views und den ViewModels erlauben beim erneuten Laden der ViewModel-Klassen eine automatische Aktualisierung der Benutzeroberfläche. C# bietet die Möglichkeit komplette Klassen automatisch im Binärformat oder in XML-Dateien zu speichern und daraus wieder zu laden. Die Speicherung wird „Serialisierung“ und das Laden „Deserialisierung“ genannt. Um die Deserialisierung zu ermöglichen muss jede

Klasse einen sogenannten „Default Constructor“ (parameterlosen Konstruktor) beinhalten. Außerdem müssen schreibgeschützte Eigenschaften (Eigenschaften, die über keinen Setter verfügen) von der Serialisierung ausgeschlossen werden. Das kann durch das Setzen des „[XmlIgnore]“-Tags vor der Deklaration der Eigenschaft festgelegt werden.

Wegen der großen Anzahl an Klassen, die um parameterlose Konstruktoren erweitert werden müssen und den benötigten Erweiterungen von Klassen, die keine Serialisierung unterstützen (wie beispielsweise die Klasse „KeyValuePair“ [Msd05]), wird nicht die komplette „WorkspaceViewModel“-Klasse serialisiert. Lediglich die Einstellungen, die den Zustand der Anwendung definieren, werden serialisiert. Beim Filter-Control wird somit statt der Klasse „FilterViewModel“ und den untergeordneten Klassen „ConjunctiveFilterModel“, „DisjunctiveFilterModel“ und „BasicFilterModel“ lediglich die Textdarstellung des Filters gespeichert, aus der beim Laden eines gespeicherten Projektes der Zustand des Filter-Controls wiederhergestellt werden kann. Diese Funktionalität wird sowohl beim Laden eines Projektes, als auch beim Laden gespeicherter Filter im Filter-Control verwendet (siehe Abschnitt 5.7.2).

Im Folgenden werden die Daten, die beim Speichern des Zustandes der Anwendung gesichert werden, aufgeführt.

Das DataTable-Objekt, das die geladenen Patientendaten enthält, wird mit Hilfe der Funktionen „WriteXmlSchema()“ und „WriteXml()“ bzw. „ReadXmlSchema()“ und „ReadXml()“ im XML-Format gespeichert und daraus wieder eingelesen. Dabei werden zwei XML-Dateien erstellt, die das XML-Schema bzw. die eigentlichen Daten der gespeicherten DataTable beinhalten. Die genannten Funktionen werden von der .NET-Bibliothek zur Verfügung gestellt. Ein Nachteil dieses Speicherungsformates ist der erhöhte Platzbedarf von XML-Daten. Dieser steigt im Vergleich zur Speicherung im csv-Format durch die Verwendung der XML-Tags um ein 4,5-faches von 4,6 MB auf 20,6 MB.

Vor dem Speichern eines Projektes wird eine Liste von „WorkspaceDataStub“-Objekten erstellt. Diese Objekte enthalten in ihren Eigenschaftsfeldern die zu speichernden Informationen der offenen Arbeitsbereiche. Nach dem Laden eines Projektes werden die Arbeitsbereiche wieder erstellt und die Informationen zum Zustand dieser aus den jeweiligen WorkspaceDataStub-Objekten eingelesen. Durch die geringe Menge an gesicherten Informationen ist der Speicherbedarf hierbei minimal. Die Daten eines Arbeitsbereiches belegen weniger als ein Kilobyte.

Folgende Daten werden für jeden geöffneten Arbeitsbereich gespeichert:

- Textdarstellung des aktuellen Filters
- Textdarstellung der aktuellen Aggregationsfunktion
- Die für die Distributionsfunktion ausgewählte Spalte
- Auflösungsstufe der geografischen Gruppen
- Skalierungsfaktor der Diagramme
- Einstellung „Isolation der Auswahl“
- Suchfilter

- Textdarstellung des aktuellen Gruppenfilters
- Sichtbarer Ausschnitt der Karte

Die Liste der gespeicherten Filter ist nicht Teil der einzelnen Arbeitsbereiche und wird in einer separaten Datei gespeichert.

Die vier genannten Dateien werden in einem temporären Ordner erstellt und anschließend zu einem Archiv komprimiert. Das Archiv trägt den bei der Speicherung angegebenen Namen mit der Dateiendung „.mvaproj“. Für die Komprimierungs- sowie die Dekomprimierungsfunktion wurde die .NET-native Klasse `System.IO.Packaging.Package` verwendet. Durch die Archivierung wird die Größe einer Projektdatei mit drei Arbeitsbereichen auf 1,3 MB verkleinert.

7. Expertenmeinung

Um die Tauglichkeit des entwickelten Prototyps und die Qualität der entdeckten Ergebnisse zu beurteilen, wurden bei einem Treffen mit einem ärztlichen Experten, Herrn Dr. Peter Fritz, die Daten der Brustkrebsdatei analysiert.

Herr Dr. Fritz ist Facharzt für Pathologie und molekulare Pathologie und war bis zu seiner Berentung im Jahr 2007 Chef des Pathologischen Instituts am Robert Bosch Krankenhaus in Stuttgart. Herr Dr. Fritz war nach seiner Approbation als Arzt im Jahr 1970 zunächst sechs Jahre als Militärarzt und anschließend als praktischer Arzt in der Entwicklungshilfe tätig. Während seiner Ausbildung zum Arzt für Pathologie (1976-1983) begann er seine wissenschaftliche Arbeit auf dem Gebiet des Mammakarzinoms und arbeitete am Onkologischen Schwerpunkt Stuttgart seit dessen Gründung. Seit seiner Berentung konzentriert sich Dr. Fritz auf die Nutzung der anonymisierten Daten des OSP in wissenschaftlichen Arbeiten.

7.1. Anwendungsszenarien

Im Vorfeld des Experteninterviews wurde eine Reihe von Fragen erstellt, die mit Hilfe der Funktionen der MVA-Anwendung beantwortet werden sollten. Da die meisten dieser Fragen aber sehr spezifisch waren und auch durch Statistikprogramme beantwortet werden könnten, wurde dieses Vorgehen verworfen und mit einer explorativen Analyse der Daten begonnen.

Nach dem Start der Anwendung und dem Einlesen der Daten wurde die Verteilung der Datensätze innerhalb der Bundesrepublik Deutschland analysiert. Dazu wurde die Auflösung (Abschnitt 5.4.2) auf die Position „Land“ gesetzt, um die Datensätze nach den Bundesländern zu gruppieren. Auf der Miniaturkarte (Abschnitt 5.4.1) wurde sichtbar, dass die Datensätze in ganz Deutschland, mit Ausnahme der Bundesländer Sachsen und Sachsen-Anhalt, verteilt sind. Um die Anzahl der Patienten aus den jeweiligen Bundesländern anzuzeigen, wurde die Aggregationsfunktion (Abschnitt 5.8) „count(code)“ ausgewählt. Die Unterschiede in der Anzahl der Datensätze aus den einzelnen Bundesländer war dabei so groß, dass bei linearer Größenskala der Diagramme nur die Fälle in Baden-Württemberg angezeigt wurden. Nach der Umstellung der Diagrammgröße auf die logarithmische Skala (Abschnitt 5.3.8) wurden auch die anderen Bundesländer angezeigt. Abb. 7.1 zeigt die Einstellungen, die für die Ermittlung der Verteilung der Datensätze verwendet wurden.

Als nächstes wurde der Grund für den Abbruch der Beobachtung der Patienten analysiert. Dazu wurde in der Distributionsfunktion die Spalte „abgru“ ausgewählt. Nach der Anzeige der Verteilung der unterschiedlichen Abbruchgründe in den Landkreisen ist das hohe

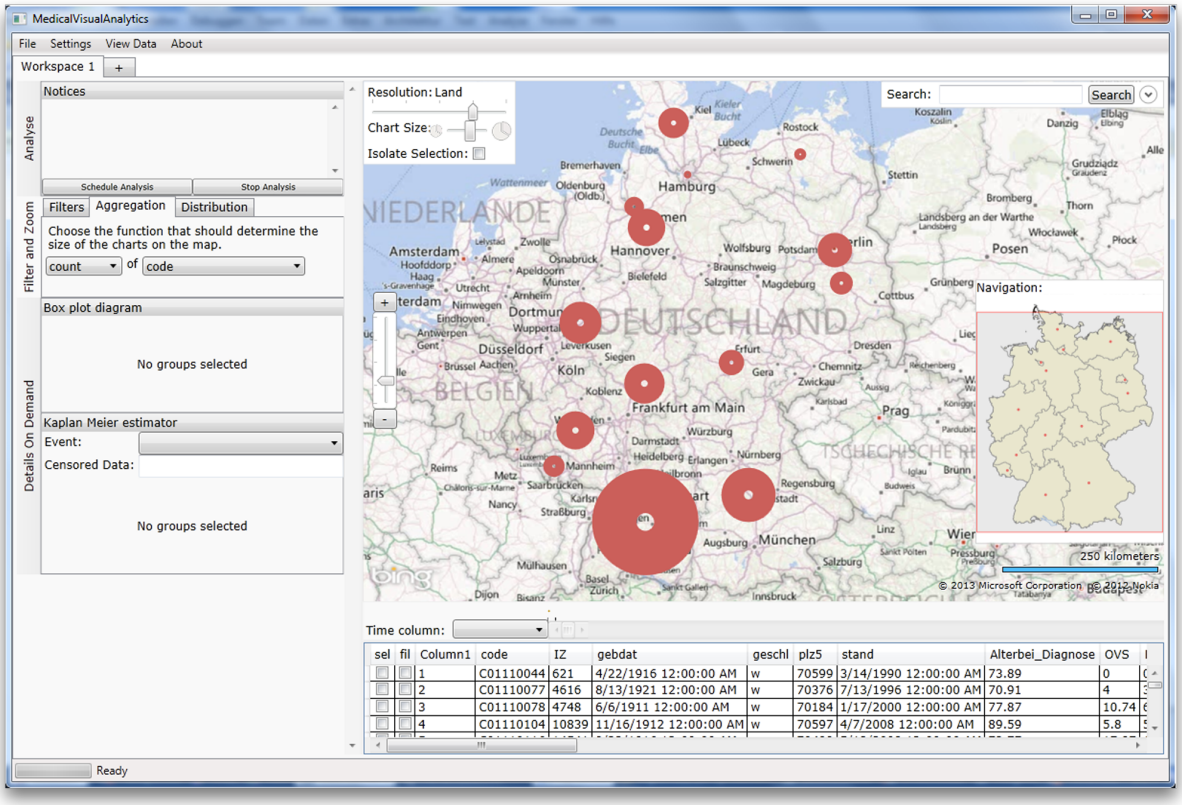


Abbildung 7.1.: Einstellungen für die Anzeige der Verteilung der Brustkrebsdatensätze innerhalb Deutschlands. Der Resolution-Schieberegler gruppiert die Datensätze nach Bundesländer. Durch die Aggregationsfunktion „count(code)“ entsprechen die Diagrammradien der Anzahl der Datensätze aus dem entsprechenden Bundesland. In der Miniaturkarte ist zu erkennen, dass es in den Bundesländern Sachsen und Sachsen-Anhalt keine Datensätze gibt.

prozentuale Vorkommen des Wertes „u“ (unbekannt verzogen) in den Landkreisen außerhalb Baden-Württembergs aufgefallen. Gründe dafür können sowohl der seltene Kontakt mit Patienten, als auch die fehlende Möglichkeit, die Daten von Patienten außerhalb von Baden-Württemberg mit den Melderegistern abzugleichen. Abb. 7.2 zeigt die beschriebene Ansicht und die dazu benötigten Einstellungen.

Um eine mögliche Verbesserung der Betreuung entfernter Patienten im Verlauf der Zeit zu erkennen, wurde die Zeitleiste (siehe Abschnitt 5.13) auf einen Intervall von fünf Jahren beschränkt. Dieses Intervall wurde über die komplette Zeitachse von 1989/1994 bis 2004/2009 verschoben. Tatsächlich wurde mit fortschreitendem Diagnosedatum eine Senkung der Fälle mit unbekanntem Status beobachtet, wie in Abb. 7.3 ersichtlich.

Wegen der großen Anzahl an Patienten in Baden-Württemberg wurden in den nächsten Untersuchungen die Landkreise in Baden-Württemberg miteinanander verglichen. Dazu wurde ein Filter mit der Bedingung „GeoPosLand = ’Baden-Württemberg’“ erstellt (siehe

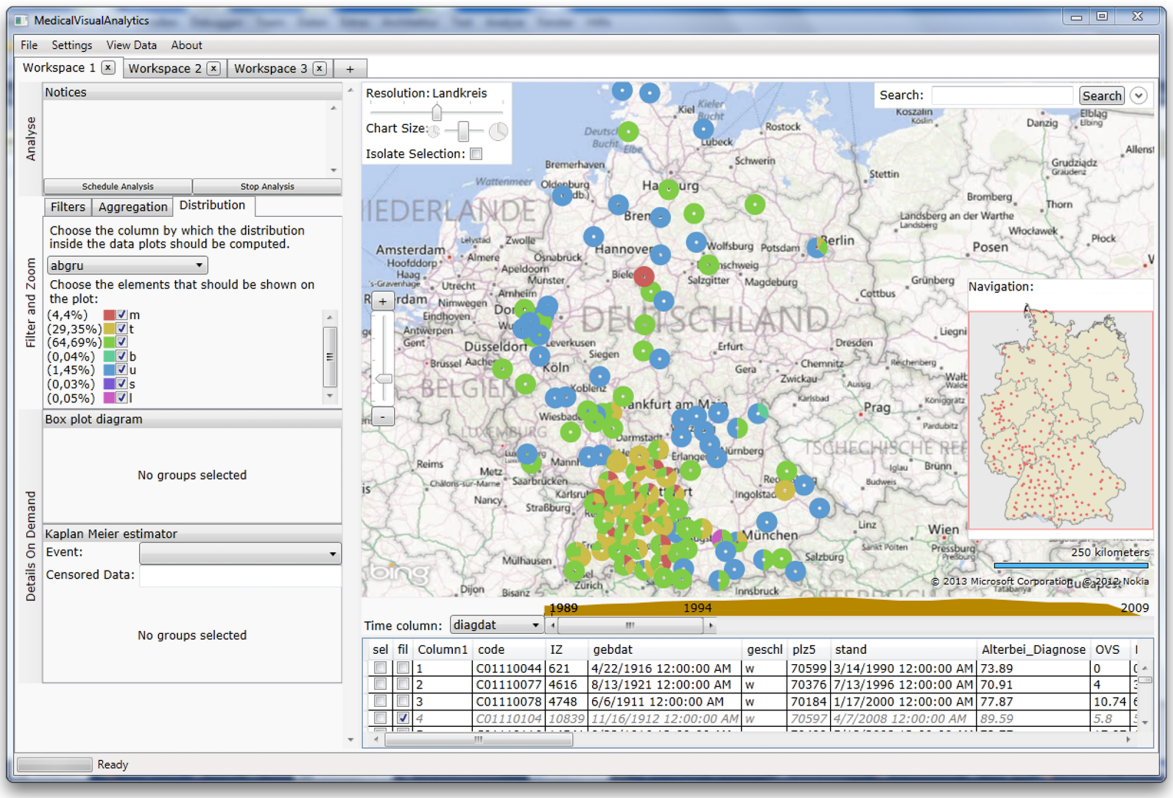


Abbildung 7.2.: Einstellungen für die Anzeige des Abbruchgrundes für die Beobachtung in ganz Deutschland. Die Daten sind nach Landkreisen gruppiert. Die Distributionsfunktion basiert auf den Werten der Eigenschaft „abgru“ (Abbruchgrund der Beobachtung). Der deutliche Anteil an Datensätzen mit unbekanntem Stand außerhalb von Baden-Württemberg ist klar ersichtlich.

Abschnitt 5.7), um alle Datensätze außerhalb von Baden-Württemberg herauszufiltern. Auch hier zeichnete sich das gleiche Bild ab: die meisten Datensätze stammen aus Landkreisen in der Nähe von Stuttgart. Diese Erkenntnis stellte keine Überraschung dar, da sich alle behandelnden Kliniken in der Umgebung der Stadt Stuttgart befinden. Im nächsten Schritt wurde nach Auffälligkeiten in der mittleren Beobachtungszeit unterschiedlicher Regionen gesucht. Dazu wurde die Aggregationsfunktion „median(OVS)“ eingestellt und ein Gruppenfilter (siehe Abschnitt 5.7.4) mit der Bedingung „count(code) > 10“ erstellt, um Gruppen mit weniger als zehn Datensätzen als Ausreißer auszuschliessen. Die erzeugte Visualisierung ergab unerwartete Ergebnisse, da der Median der Überlebenszeiten mit steigender Entfernung zur Landeshauptstadt Stuttgart ebenfalls steigt (Abb. 7.4). Zwar ist ein möglicher Grund die höhere Fluktuationsrate der Bevölkerung in großen Städten, die zu einer Unterbrechung der Beobachtung führt, eine bewiesene Erklärung hierfür wurde aber noch nicht gefunden. Um die Ergebnisse dieser und ähnlicher Visualisierungen besser interpretieren zu können, wurde von dem Experten die Einbindung von Metainformationen über die jeweiligen geografischen Gebiete empfohlen. Angaben über Bevölkerungszahlen, Bildungsstand oder andere sozio-

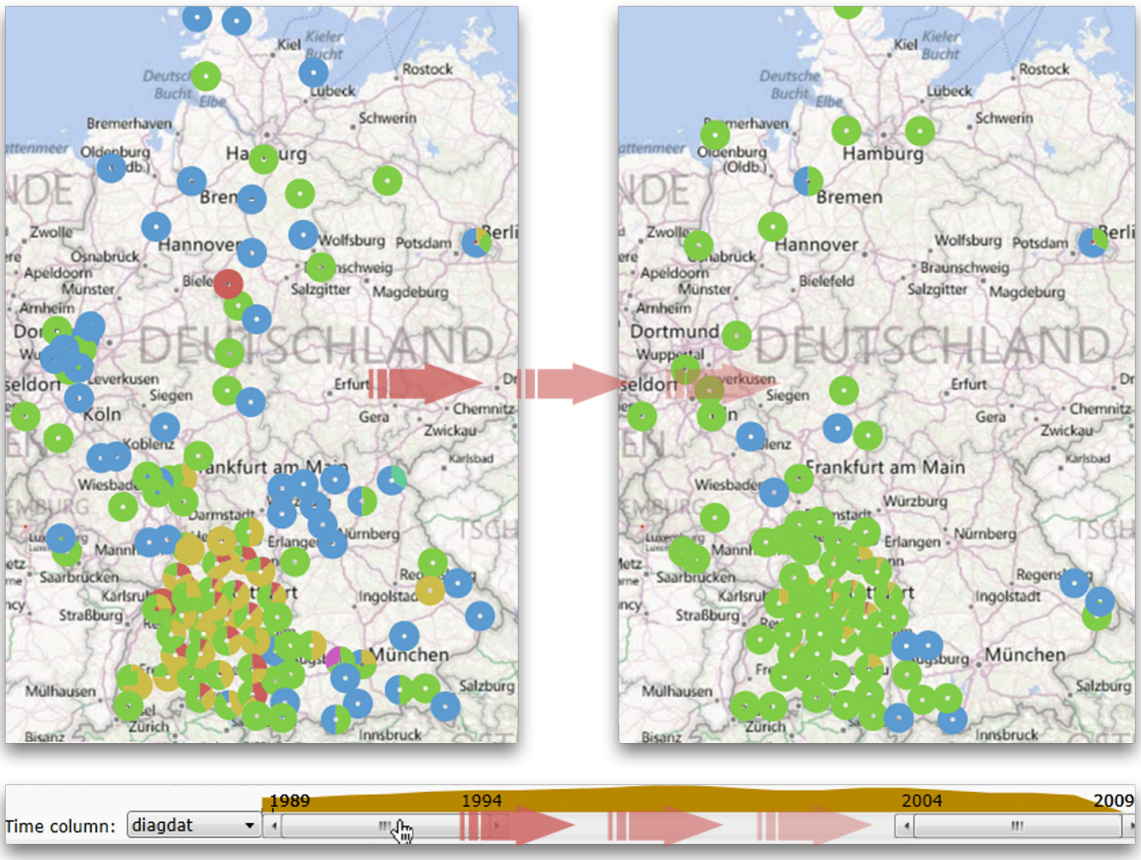


Abbildung 7.3.: Entwicklung des Anteils der Patienten mit unbekanntem Status außerhalb von Baden-Württemberg.

ökonomische Merkmale sollten bei der Interpretation der Ergebnissen einbezogen werden können.

Weitere entdeckte Informationen im Verlauf des Experteninterviews waren eine verbesserte Bilanz von Brustamputationen seit Beginn der Datensammlung 1989 und die unterschiedlichen Zeitpunkte, ab denen der CerbB2-Wert von den einzelnen Kliniken aufgenommen wurde. Während die ersten Kliniken schon im Jahr 2000 die ersten Werte aufgenommen haben und im Jahr 2001 den Anteil der Patienten, bei denen der Wert gemessen wurde, erhöht haben, wurde der Wert erst 2005 in allen Kliniken aufgenommen. Ab 2006 erfolgte die Aufnahme bei fast allen Patienten.

7.2. Ergebnisse

Das durchgeführte Experteninterview und die vorgestellten Fragestellungen haben gezeigt, dass der entwickelte Prototyp dem Motto „Erwartetes herausfinden und Unerwartetes ent-

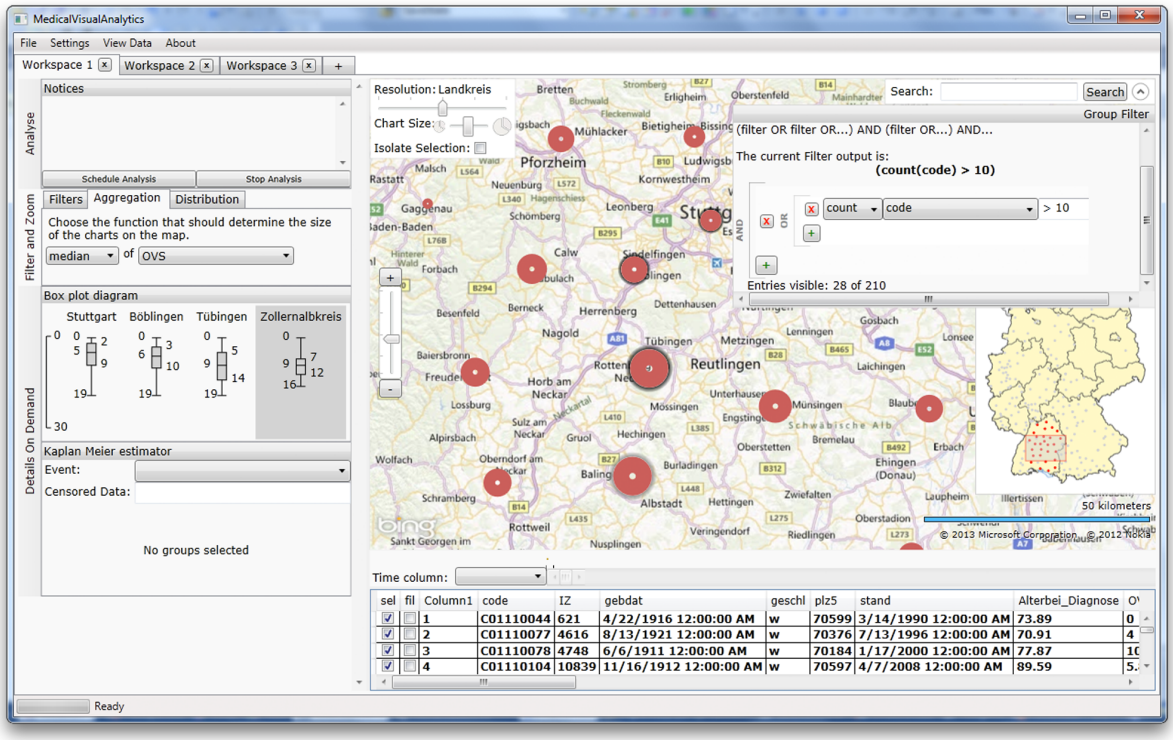


Abbildung 7.4.: Vergleich der mittleren Beobachtungszeit in den Landkreisen in Baden-Württemberg. Über den Gruppenfilter im oberen rechten Teil des Anwendungsfensters wurden Gruppen mit weniger als zehn Datensätzen herausgefiltert. Vier Landkreise sind ausgewählt und die Verteilung der Beobachtungszeit mit Hilfe der Box Plot-Diagramme verdeutlicht.

decken“ (siehe Abschnitt 2.1) durch den flüssigen Übergang zwischen vordefinierten und neuen Fragestellungen, die erst durch die Visualisierung der Daten angestoßen werden, gerecht wird. Gleichzeitig hat die Vorbereitung des Experteninterviews auch die Abgrenzung zwischen Visual Analytics und Statistik hervorgehoben. Während die Statistik eine „bestätigende Datenanalyse“ darstellt, verfolgt Visual Analytics eine „explorative Datenanalyse“ [KMT10]. So wurden die im Voraus vorbereiteten Fragestellungen für das Experteninterview teilweise wegen ihrer großen Spezifität, teilweise wegen Trivialität verworfen. Für komplexere Fragestellungen reicht die Flexibilität der einstellbaren Parameter der MVA-Anwendung noch nicht aus. Der Median der Überlebenszeiten – der als Zeitpunkt, zu dem 50% der Patienten verstorben sind, definiert ist – kann mit den aktuell implementierten Funktionen nicht angezeigt werden. Da dieser Wert in der Regel für keine geografische Gruppe erreicht wird, da die Sterberate aller Patienten bei unter 30% liegt, wurde von dem Experten die Möglichkeit der Anzeige eines beliebigen p-Quantils der Überlebenszeit gewünscht.

Die bei der Visualisierung entdeckten Ergebnisse sollten stets kritisch bewertet werden, da sie nicht immer relevante Fakten wiedergeben. Zusätzlich zu der erwähnten medianen Beobachtungszeit wurde eine weitere unerwartete Entdeckung bei der Verteilung der Dia-

gnosezeitpunkte im Verlauf des Jahres gemacht. So zeigte diese ein erhöhtes Aufkommen von Diagnosen in der Mitte des Jahres. Bei der Suche nach einer plausiblen Erklärung wurde herausgefunden, dass bei fehlendem Diagnosedatum in den Patientendaten, dieses standardmäßig auf den 15. Juni gesetzt wird. Bei Überprüfung der Verteilung der Diagnosezeitpunkte innerhalb eines Monats wurde der 15. ebenfalls als häufigstes Datum entdeckt, was diese Erklärung unterstützt.

8. Zusammenfassung und Ausblick

In diesem Kapitel wird eine Zusammenfassung der Ergebnisse dieser Arbeit präsentiert und Vorschläge für die Weiterentwicklung diskutiert.

8.1. Zusammenfassung

Im Rahmen dieser Arbeit wurde ein interaktiver Ansatz für die Exploration und Analyse von historischen Patientendaten entwickelt.

Der Arbeitsablauf des erstellten Prototypen verfolgt die Schritte des Visual Analytics Mantra: „Analyse first“, „Show the Important“, „Filter, Zoom and Analyse further“ und „Details on Demand“. Aufbauend auf der w^3-Prämisse wurde die Darstellung der Daten in den drei Dimensionen „Was?“, „Wann?“ und „Wo?“ aufgeteilt. Nach dem Laden der Daten wird eine Ausreißererkennung durchgeführt und die Ergebnisse dem Benutzer präsentiert. Parallel dazu hat der Benutzer die Möglichkeit, die Daten explorativ zu analysieren. Die Datensätze werden dazu auf einer Karte nach Herkunft oder behandelnder Klinik gruppiert dargestellt und das Zeitintervall der Beobachtungen kann dynamisch angepasst werden. Dem Benutzer stehen die Filterung der Daten nach beliebigen Kriterien, Aggregationsfunktionen für alle Gruppen von Datensätzen sowie die Anzeige der Verteilung verschiedener Merkmale innerhalb dieser Gruppen zur Verfügung. Für ausgewählte Gruppen können darüber hinaus statistische Auswertungen angezeigt werden, die den Vergleich unterschiedlicher Gruppen ermöglichen.

8.2. Ausblick

Der erstellte Prototyp ermöglicht einen ersten Einblick in die Verwendung von Visual Analytics bei der Analyse von medizinischen Daten und hat keinesfalls den Anspruch, die Möglichkeiten der angebotenen Funktionen ausgeschöpft zu haben. Sowohl der Bereich der Visualisierung als auch der Bereich des Data Minings können durch das Hinzufügen weiterer und die Erweiterung bestehender Funktionen weiterentwickelt werden.

Im Bereich des Data Minings können schnellere Algorithmen für die Ausreißererkennung ([CSM02]) oder Clusteringalgorithmen implementiert werden. Im Bereich Details on Demand können Boxplots um die Anzeige weiterer Informationen ([Poto6, Kamo8]) oder der Kaplan-Meier-Schätzer um verwandte Funktionen, wie die Hazardrate erweitert werden. Im Bereich

der Geovisualisierung bieten [KPSN04b], [KPSN04a] und [SBM08] eine Übersicht über weitere Ansätze für Visualisierungen.

Auch bei den bestehenden Funktionen besteht noch Verbesserungspotential. Die Reaktionsgeschwindigkeit bei der Verarbeitung der Daten sollte unter 100 Millisekunden liegen [SLZ03]. Zwar wurde durch Parallelisierung die benötigte Zeit um den Faktor vier verbessert, durch den Einsatz von schlankeren Strukturen, wie beispielsweise Listen anstelle von DataTables, sollte die Geschwindigkeit weiterhin verbessert werden.

Weitere Entwicklungsmöglichkeiten wurden im Rahmen des Experteninterviews in Kapitel 7 vorgestellt.

A. Anhang

A.1. FilterView

Listing A.1 Eigenschaften der BasicFilterModel-Klasse

```
public class BasicFilterModel : ModelBase
{
#region Properties
    /// Store whether to show the negation column
    bool _showNegationColumn;
    public bool ShowNegationColumn {...}
    /// Predefined list of the values in the negation column
    public string[] NegationValues {...}
    /// Negation value selected by the user
    bool _selectedNegation;
    public string SelectedNegation {...}
    /// An empty DataTable containing the column definitions. Used to test the filter.
    DataTable _dummyDataTable;
    public DataTable DummyDataTable {...}
    /// Store whether to show the aggregation column
    bool _showAggregationColumn;
    public bool ShowAggregationColumn {...}
    /// Predefined list of data aggregation functions that the users can choose from
    ObservableCollection<SKeyValuePair<String, String>> _aggregationFunctions;
    public ObservableCollection<SKeyValuePair<String, String>> AggregationFunctions {...}
    /// The data aggregation function selected by the user
    SKeyValuePair<String, String> _selectedAggregationFunction;
    public SKeyValuePair<String, String> SelectedAggregationFunction {...}
    /// The columns of the loaded data
    Collection<String> _columns;
    public Collection<String> Columns {...}
    /// The column selected by the user
    String _selectedColumn;
    public String SelectedColumn {...}
    /// The filtering expression entered by the user
    String _expression;
    public String Expression {...}
    /// Stores whether there is an error in the current filter
    bool _error;
    public bool Error  {...}
    /// The string equivalent of the current filter
    public String FilterString {...}
#endregion Properties

    ...
}
```

Listing A.2 Eigenschaften der DisjunctiveFilterModel-Klasse

```
public class DisjunctiveFilterModel : ModelBase
{
#region Properties
    /// A collection of BasicFilterModel objects that are combined to a disjunctive term
    ExtendedObservableCollection<BasicFilterModel> _basicFilters;
    public ExtendedObservableCollection<BasicFilterModel> BasicFilters {...}
    /// Store whether to show the negation column
    /// (this will be passed down to all BasicFilterModels of this object)
    bool _showNegationColumn;
    public bool ShowNegationColumn {...}
    /// Store whether to show the aggreagtion column
    /// (this will be passed down to all BasicFilterModels of this object)
    bool _showAggregationColumn;
    public bool ShowAggregationColumn {...}
    /// An empty DataTable containing the column definitions. Used to test the filter.
    /// (this will be passed down to all BasicFilterModels of this object)
    DataTable _dummyDataTable;
    public DataTable DummyDataTable {...}
    /// Stores whether there is an error in the current filter
    bool _error;
    public bool Error {...}
    /// The string equivalent of the current filter
    public String FilterString {...}
#endregion Properties

    ...
}
```

Listing A.3 Eigenschaften der ConjunctiveFilterModel-Klasse

```
public class ConjunctiveFilterModel : ModelBase
{
#region Properties
    /// A collection of DisjunctiveFilterModel objects that are combined to a conjunctive statement
    ExtendedObservableCollection<DisjunctiveFilterModel> _disjunctiveFilters;
    public ExtendedObservableCollection<DisjunctiveFilterModel> DisjunctiveFilters {...}
    /// Store whether to show the negation column
    /// (this will be passed down to all DisjunctiveFilterModels of this object)
    bool _showNegationColumn;
    public bool ShowNegationColumn {...}
    /// Store whether to show the aggreagtion column
    /// (this will be passed down to all DisjunctiveFilterModels of this object)
    bool _showAggregationColumn;
    public bool ShowAggregationColumn {...}
    /// An empty DataTable containing the column definitions. Used to test the filter.
    /// (this will be passed down to all DisjunctiveFilterModels of this object)
    DataTable _dummyDataTable;
    public DataTable DummyDataTable {...}
    /// Stores whether there is an error in the current filter
    bool _error;
    public bool Error {...}
    /// The string equivalent of the current filter
    public String FilterString {...}
#endregion Properties

    ...
}
```

Literaturverzeichnis

[AAD+10] G. Andrienko, N. Andrienko, U. Demsar, D. Dransch, J. Dykes, S. I. Fabrikant, M. Jern, M. J. Kraak, H. Schumann, C. Tominski. Space, time and visual analytics. *Int. J. Geogr. Inf. Sci.*, 24(10):1577–1600, 2010. doi:10.1080/13658816.2010.508043. URL http://dx.doi.org/10.1080/13658816.2010.508043. (Zitiert auf Seite 15)

[Agg] Aggregation - Gabler Wirtschaftslexikon. URL http://wirtschaftslexikon.gabler.de/Archiv/55812/aggregation-v8.html. (Zitiert auf Seite 16)

[AY01] C. C. Aggarwal, P. S. Yu. Outlier detection for high dimensional data. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, SIGMOD '01, S. 37–46. ACM, New York, NY, USA, 2001. doi:10.1145/375663.375668. URL http://doi.acm.org/10.1145/375663.375668. (Zitiert auf Seite 21)

[Bar] File:Incarceration Rates Worldwide.svg.png - Wikimedia Commons. URL http://upload.wikimedia.org/wikipedia/commons/3/35/Incarceration_Rates_Worldwide_ZP.svg. (Zitiert auf Seite 19)

[BG05] I. Ben-Gal. OUTLIER DETECTION. In M. O., R. L., Herausgeber, *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, Kapitel 1, S. 117–132. Kluwer Academic Publishers, 2005. (Zitiert auf Seite 21)

[BL94] V. Barnett, T. Lewis. *Outliers in Statistical Data*. Wiley Series in Probability & Statistics. Wiley, 1994. (Zitiert auf Seite 21)

[Box] File:Elements of a boxplot.svg - Wikimedia Commons. URL http://commons.wikimedia.org/w/index.php?title=File:Elements_of_a_boxplot.svg&oldid=67368580. (Zitiert auf Seite 19)

[Bur04] H. B. Burke. Outcome Prediction and the Future of the TNM Staging System. *Journal of the National Cancer Institute*, 96(19):1408–1409, 2004. doi:10.1093/jnci/djh293. URL http://jnci.oxfordjournals.org/content/96/19/1408.short. (Zitiert auf Seite 27)

[CM02] K. J. Cios, G. W. Moore. Uniqueness of medical data mining. *Artificial Intelligence in Medicine*, 26(1-2):1–24, 2002. doi:10.1016/S0933-3657(02)00049-0. URL http://www.sciencedirect.com/science/article/pii/S0933365702000490. (Zitiert auf den Seiten 11 und 27)

[Col] Colorbrewer: Color Advice for Maps. URL http://colorbrewer2.org. (Zitiert auf Seite 52)

[CSM02] A. Chaudhary, A. S. Szalay, A. W. Moore. Very Fast Outlier Detection in Large Multidimensional Data Sets. In *DMKD*. 2002. (Zitiert auf den Seiten 21 und 89)

[Dat] Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization. URL http://www.datavis.ca/milestones/index.php?group=pre-1600. (Zitiert auf Seite 14)

[Dato1] Compute Feature of DataTable, 2001. URL http://www.c-sharpcorner.com/uploadfile/lcamlibel/datatablecomputelc11302005060506am/datatablecomputelc.aspx. (Zitiert auf Seite 74)

[Des] Länder Regionen - Gemeindeverzeichnis - Gemeindeverzeichnis-Informationssystem (GV-ISys) - Statistisches Bundesamt (Destatis). URL https://www.destatis.de/DE/ZahlenFakten/LaenderRegionen/Regionales/Gemeindeverzeichnis/Administrativ/Archiv/GV100ADQ/GV100AD3QAktuell.zip?__blob=publicationFile. (Zitiert auf Seite 36)

[DMKK12] J. Davey, F. Mansmann, J. Kohlhammer, D. Keim. Visual Analytics: Towards Intelligent Interactive Internet and Security Solutions. In F. Álvarez, F. Cleary, P. Daras, J. Domingue, A. Galis, A. Garcia, A. Gavras, S. Karnourskos, S. Krco, M.-S. Li, V. Lotz, H. Müller, E. Salvadori, A.-M. Sassen, H. Schaffers, B. Stiller, G. Tselentis, P. Turkama, T. Zahariadis, Herausgeber, *The Future Internet*, Band 7281 von *Lecture Notes in Computer Science*, S. 93–104. Springer Berlin Heidelberg, 2012. doi:10.1007/978-3-642-30241-1_9. URL http://dx.doi.org/10.1007/978-3-642-30241-1_9. (Zitiert auf den Seiten 11 und 13)

[Dot11] Download: Microsoft .NET Framework 4 (eigenständiger Installer) - Microsoft Download Center - Download Details, 2011. URL http://www.microsoft.com/de-de/download/details.aspx?id=17718. (Zitiert auf Seite 69)

[Few13] S. Few. Data Visualization for Human Perception. In M. Soegaard, R. F. Dam, Herausgeber, *The Encyclopedia of Human-Computer Interaction, 2nd Ed.* The Interaction Design Foundation, Aarhus, Denmark, 2013. URL http://www.interaction-design.org/encyclopedia/data_visualization_for_human_perception.html. (Zitiert auf Seite 14)

[FKG+10] P. Fritz, S. Klenk, S. Goletz, A. Gerteis, W. Simon, F. Brinkmann, E. Heidemann, E. Lüttgen, G. Ott, M. Alscher, M. Schwab, J. Dippon. Clinical Impacts of Histological Subtyping Primary Breast Cancer. *Anticancer Research*, 30(12):5137–5144, 2010. URL http://ar.iiarjournals.org/content/30/12/5137.abstract. (Zitiert auf Seite 27)

[Foco2] Focus+Context, 2002. URL http://www.infovis.net/printMag.php?num=85&lang=2. (Zitiert auf den Seiten 22 und 47)

[Fow04] Presentation Model, 2004. URL http://martinfowler.com/eaaDev/PresentationModel.html. (Zitiert auf Seite 72)

[GAK+11] T. Gschwandtner, W. Aigner, K. Kaiser, S. Miksch, A. Seyfang. CareCruiser: Exploring and visualizing plans, events, and effects interactively. In *Pacific Visualization Symposium (PacificVis), 2011 IEEE*, S. 43–50. 2011. doi:10.1109/PACIFICVIS.2011.5742371. (Zitiert auf Seite 23)

[Gap] Gapminder World. URL http://www.gapminder.org/world/. (Zitiert auf den Seiten 32 und 33)

[HB11] M. Harrower, C. A. Brewer. ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. In *The Map Reader*, S. 261–268. John Wiley Sons, Ltd, 2011. doi:10.1002/9780470979587.ch34. URL http://dx.doi.org/10.1002/9780470979587.ch34. (Zitiert auf Seite 51)

[HS05] M. Harrower, B. Sheesley. Designing Better Map Interfaces: A Framework for Panning and Zooming. *Transactions in GIS*, 9(2):77–89, 2005. doi:10.1111/j.1467-9671.2005.00207.x. URL http://dx.doi.org/10.1111/j.1467-9671.2005.00207.x. (Zitiert auf den Seiten 46 und 47)

[ID90] A. Inselberg, B. Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *Proceedings of the 1st conference on Visualization '90*, VIS '90, S. 361–378. IEEE Computer Society Press, Los Alamitos, CA, USA, 1990. URL http://dl.acm.org/citation.cfm?id=949531.949588. (Zitiert auf Seite 15)

[KAF+08] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, G. Melancon. Visual Analytics: Definition, Process, and Challenges. In A. Kerren, J. T. Stasko, J.-D. Fekete, C. North, Herausgeber, *Information Visualization*, Kapitel Visual Analytics: Definition, Process, and Challenges, S. 154–175. Springer-Verlag, Berlin, Heidelberg, 2008. doi:10.1007/978-3-540-70956-5_7. URL http://dx.doi.org/10.1007/978-3-540-70956-5_7. (Zitiert auf den Seiten 13 und 14)

[Kam08] P. Kampstra. Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets*, 28(1):1–9, 2008. URL http://www.jstatsoft.org/v28/c01. (Zitiert auf den Seiten 20 und 89)

[Kap] Kaplan-Meier-sample-plot.svg.png. URL http://upload.wikimedia.org/wikipedia/commons/thumb/f/f9/Kaplan-Meier-sample-plot.svg/2000px-Kaplan-Meier-sample-plot.svg.png. (Zitiert auf Seite 19)

[Kei02] D. A. Keim. Information Visualization and Visual Data Mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002. doi:10.1109/2945.981847. URL http://dx.doi.org/10.1109/2945.981847. (Zitiert auf den Seiten 11 und 15)

[KHDH02] D. A. Keim, M. C. Hao, U. Dayal, M. Hsu. Pixel Bar Charts: A Visualization Technique for Very Large Multi-Attribute Data Sets. *Information Visualization*, 1(1):20–34, 2002. doi:10.1057/palgrave.ivs.9500003. URL http://ivi.sagepub.com/content/1/1/20.abstract. (Zitiert auf Seite 18)

[KKEM10] D. A. Keim, J. Kohlhammer, G. Ellis, F. Mansmann, Herausgeber. *Mastering The Information Age - Solving Problems with Visual Analytics*. Eurographics, 2010. URL http://www.vismaster.eu/book/. (Zitiert auf den Seiten 13 und 15)

[KMS+08] D. Keim, F. Mansmann, J. Schneidewind, J. Thomas, H. Ziegler. Visual Analytics: Scope and Challenges. In S. Simoff, M. Böhlen, A. Mazeika, Herausgeber, *Visual Data Mining*, Band 4404 von *Lecture Notes in Computer Science*, S. 76–90. Springer Berlin Heidelberg, 2008. doi:10.1007/978-3-540-71080-6_6. URL http://dx.doi.org/10.1007/978-3-540-71080-6_6. (Zitiert auf Seite 22)

[KMSZ06] D. Keim, F. Mansmann, J. Schneidewind, H. Ziegler. Challenges in Visual Data Analysis. In *Proceedings of the conference on Information Visualization*, IV '06, S. 9–16. IEEE Computer Society, Washington, DC, USA, 2006. doi:10.1109/IV.2006.31. (Zitiert auf Seite 11)

[KMT10] D. A. Keim, F. Mansmann, J. Thomas. Visual analytics: how much visualization and how much analytics? *SIGKDD Explor. Newsl.*, 11(2):5–8, 2010. (Zitiert auf Seite 87)

[KN98] E. M. Knorr, R. T. Ng. Algorithms for Mining Distance-Based Outliers in Large Datasets. In *Proceedings of the 24rd International Conference on Very Large Data Bases*, VLDB '98, S. 392–403. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998. URL http://dl.acm.org/citation.cfm?id=645924.671334. (Zitiert auf Seite 53)

[Knf] Konjunktive Normalform – Wikipedia. URL http://de.wikipedia.org/wiki/Konjunktive_Normalform. (Zitiert auf Seite 56)

[KPSN04a] D. A. Keim, C. Panse, M. Sips, S. C. North. Pixel based visual data mining of geo-spatial data. *Computers Graphics*, 28(3):327–344, 2004. doi:10.1016/j.cag.2004.03.022. URL http://www.sciencedirect.com/science/article/pii/S0097849304000263. (Zitiert auf Seite 90)

[KPSN04b] D. A. Keim, C. Panse, M. Sips, S. C. North. Visual Data Mining in Large Geospatial Point Sets. *IEEE Computer Graphics and Applications*, 24(5):36–44, 2004. doi:10.1109/MCG.2004.41. (Zitiert auf Seite 90)

[Kre12] Krebs in Deutschland 2007/2008, 2012. URL http://www.rki.de/Krebs/DE/Content/Publikationen/Krebs_in_Deutschland/kid_2012/krebs_in_deutschland_2012.pdf?__blob=publicationFile. (Zitiert auf Seite 11)

[Kum05] R. Kumar. *Research Methodology*. Sage Publications, Inc., 2005. (Zitiert auf Seite 18)

[LAMF05] Y. Livnat, J. Agutter, S. Moon, S. Foresti. Visual correlation for situational awareness. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, S. 95 – 102. 2005. doi:10.1109/INFVIS.2005.1532134. (Zitiert auf den Seiten 32 und 33)

[LHFS11] L. Lins, M. Heilbrun, J. Freire, C. Silva. VISCARETRAILS: Visualizing Trails in the Electronic Health Record with Timed Word Trees, a Pancreas Cancer Use Case. In J. J. Caban, D. Gotz, Herausgeber, *Proceedings of the IEEE VisWeek Workshop on Visual Analytics in Healthcare: Understanding the Physicians Perspective*, S. 13–16. 2011. (Zitiert auf Seite 23)

[Man] manfrin-it: Postleitzahlen für Deutschland zum Download. URL http://www.manfrin-it.com/postleitzahlen/plz.html. (Zitiert auf Seite 36)

[MDR+10] R. Maciejewski, T. Drake, S. Rudolph, A. Malik, D. Ebert. Data Aggregation and Analysis for Cancer Statistics - A Visual Analytics Approach. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, S. 1–5. 2010. doi: 10.1109/HICSS.2010.128. (Zitiert auf Seite 23)

[MHK+10] R. May, P. Hanrahan, D. A. Keim, B. Shneiderman, S. K. Card. The state of visual analytics: Views on what visual analytics is and where it is going. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology, IEEE VAST 2010, Salt Lake City, Utah, USA, 24-29 October 2010, part of VisWeek 2010*, S. 257–259. IEEE, 2010. doi:http://dx.doi.org/10.1109/VAST.2010.5649078. (Zitiert auf Seite 13)

[Mil08] H. J. Miller. Geographic Data Mining and Knowledge Discovery. In J. P. Wilson, Fotheringham, Herausgeber, *The handbook of geographic information science*, Band 7 von *Blackwell companions to geography*, Kapitel 19, S. 352–366. Blackwell Pub., 2008. (Zitiert auf den Seiten 21 und 22)

[Min] File:Minard.png - Wikimedia Commons. URL http://commons.wikimedia.org/w/index.php?title=File:Minard.png&oldid=85122093. (Zitiert auf Seite 16)

[Msda] Aufgabenparallelität (Task Parallel Library). URL http://msdn.microsoft.com/de-de/library/dd537609.aspx. (Zitiert auf Seite 79)

[Msdb] BackgroundWorker-Klasse (System.ComponentModel). URL http://msdn.microsoft.com/de-de/library/system.componentmodel.backgroundworker.aspx. (Zitiert auf Seite 78)

[Msdc] DataTable-Klasse (System.Data). URL http://msdn.microsoft.com/de-de/library/system.data.datatable.aspx. (Zitiert auf Seite 74)

[Msdd] Find a Location by Address. URL http://msdn.microsoft.com/en-us/library/ff701714.aspx. (Zitiert auf Seite 71)

[Msde] Task-Klasse (System.Threading.Tasks). URL http://msdn.microsoft.com/de-de/library/system.threading.tasks.task.aspx. (Zitiert auf Seite 79)

[Msdf] ThreadPool-Klasse (System.Threading). URL http://msdn.microsoft.com/de-de/library/system.threading.threadpool.aspx. (Zitiert auf Seite 79)

[Msd05] Serializing an object of the KeyValuePair Generic class, 2005. URL http://blogs.msdn.com/b/seshadripv/archive/2005/11/02/488273.aspx. (Zitiert auf Seite 81)

[MVC07] The original MVC reports, 2007. URL http://heim.ifi.uio.no/~trygver/2007/MVC_Originals.pdf. (Zitiert auf Seite 72)

[MVV05] Introduction to Model/View/ViewModel pattern for building WPF apps - Tales from the Smart Client - Site Home - MSDN Blogs, 2005. URL http://blogs.msdn.com/b/johngossman/archive/2005/10/08/478683.aspx. (Zitiert auf Seite 72)

[MVV09] Das Model-View-Viewmovdel (MVVM)-Entwurfsmuster für WPF, 2009. URL http://msdn.microsoft.com/de-de/magazine/dd419663.aspx. (Zitiert auf den Seiten 72 und 73)

[OHS05] N. O'Rourke, L. Hatcher, E. J. Stepanski. *A Step-by-Step Approach to Using SAS® for Univariate and Multivariate Statistics, Second Edition*. SAS Institute Inc., Cary, NC, USA, 2005. (Zitiert auf den Seiten 14 und 15)

[OL03] M. Ferreira de Oliveira, H. Levkowitz. From visual data exploration to visual data mining: a survey. *Visualization and Computer Graphics, IEEE Transactions on*, 9(3):378 – 394, 2003. doi:10.1109/TVCG.2003.1207445. (Zitiert auf Seite 14)

[Ope] OpenGeoDb. URL http://opengeodb.org/wiki/OpenGeoDB. (Zitiert auf Seite 36)

[OSPa] Jahresbericht OSP 2011. URL http://www.osp-stuttgart.de/osp/Dokumente/OSP_Jahresbericht_2011.pdf. (Zitiert auf Seite 25)

[OSPb] Stuttgarter Krebsregister - Qualitätsbericht 2011. URL http://www.osp-stuttgart.de/tudok/Dokumente/Q_bericht_lang.pdf. (Zitiert auf den Seiten 11, 25 und 27)

[Pie] File:English_dialects1997.svg.png. URL http://upload.wikimedia.org/wikipedia/commons/thumb/d/db/English_dialects1997.svg/2000px-English_dialects1997.svg.png. (Zitiert auf Seite 19)

[PMS+98] C. Plaisant, R. Mushlin, A. Snyder, J. Li, D. Heller, B. Shneiderman, K. P. Colorado. LifeLines: Using Visualization to Enhance Navigation and Analysis of Patient Records. In *Proceedings of the 1998 American Medical Informatic Association Annual Fall Symposium*, S. 76–80. 1998. (Zitiert auf Seite 23)

[PMZ+08] S. Petushi, J. Marker, J. Zhang, W. Zhu, D. Breen, C. Chen, X. Lin, F. U. Garcia. A visual analytics system for breast tumor evaluation. *Anal. Quant. Cytol. Histol.*, 30(5):279–290, 2008. (Zitiert auf Seite 23)

[Pos] Liste der Postleitregionen in Deutschland – Wikipedia. URL http://de.wikipedia.org/wiki/Liste_der_Postleitregionen_in_Deutschland. (Zitiert auf Seite 36)

[Pot06] K. Potter. Methods for Presenting Statistical Information: The Box Plot. In H. Hagen, A. Kerren, P. Dannenmann, Herausgeber, *Visualization of Large and Unstructured Data Sets*, Band S-4 von *GI-Edition Lecture Notes in Informatics (LNI)*, S. 97–106. 2006. (Zitiert auf den Seiten 20 und 89)

[SBM08] S. J. Simoff, M. H. Böhlen, A. Mazeika, Herausgeber. *Visual Data Mining - Theory, Techniques and Tools for Visual Analytics*, Band 4404 von *Lecture Notes in Computer Science*. Springer, 2008. (Zitiert auf Seite 90)

[Shn96] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, S. 336 –343. 1996. doi:10.1109/VL.1996.545307. (Zitiert auf Seite 22)

[SLJ93] A. J. Sasco, A. B. Lowenfels, P. Pasker-de Jong. Review article: epidemiology of male breast cancer. A meta-analysis of published case-control studies and discussion of selected aetiological factors. *Int J Cancer*, 53(4):538–549, 1993. URL http://www.biomedsearch.com/nih/Review-article-epidemiology-male-breast/8436428.html. (Zitiert auf Seite 12)

[SLZ03] S. Shekhar, C.-T. Lu, P. Zhang. A Unified Approach to Detecting Spatial Outliers. *GeoInformatica*, 7:139–166, 2003. doi:10.1023/A:1023455925009. URL http://dx.doi.org/10.1023/A%3A1023455925009. (Zitiert auf den Seiten 21 und 90)

[SMG02] E. B. Steiner, A. M. Maceachren, D. Guo. Developing and assessing light-weight data-driven exploratory geovisualization tools for the web. *Advances in Spatial Data Handling Proceedings of the 10th International Symposium on Spatial Data Handling*, S. 487–500, 2002. URL http://www.geovista.psu.edu/publications/Beijing01/SteinerICA01/flash_db2.htm. (Zitiert auf Seite 15)

[Sno] Datei:Snow-cholera-map.jpg – Wikipedia. URL http://de.wikipedia.org/w/index.php?title=Datei:Snow-cholera-map.jpg&filetimestamp=20051106111039#file. (Zitiert auf Seite 17)

[Soz] Soziale Ungleichheit – Wikipedia. URL http://de.wikipedia.org/wiki/Soziale_Ungleichheit. (Zitiert auf Seite 18)

[Sta] Staatsgebiet Deutschlands. URL http://de.wikipedia.org/wiki/Geographie_Deutschlands#Staatsgebiet. (Zitiert auf Seite 47)

[TC05] J. J. Thomas, K. A. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr, 2005. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0769523234. (Zitiert auf Seite 13)

[TC06] J. Thomas, K. Cook. A visual analytics agenda. *Computer Graphics and Applications, IEEE*, 26(1):10 – 13, 2006. doi:10.1109/MCG.2006.5. (Zitiert auf Seite 13)

[Tod12a] Sterbefälle insgesamt 2011 nach den 10 häufigsten Todesursachen der ICD-10, 2012. URL https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/Gesundheit/Todesursachen/Tabellen/SterbefaelleInsgesamt.html. (Zitiert auf Seite 11)

[Tod12b] Sterbefälle weiblich 2011 nach den 10 häufigsten Todesursachen der ICD-10, 2012. URL https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/Gesundheit/Todesursachen/Tabellen/SterbefaelleWeiblich.html. (Zitiert auf Seite 12)

[TSWS05] C. Tominski, P. Schulze-Wollgast, H. Schumann. 3D information visualization for time dependent data on maps. In *Information Visualisation, 2005. Proceedings. Ninth International Conference on*, S. 175–181. 2005. doi:10.1109/IV.2005.3. (Zitiert auf Seite 15)

[Tuf83] E. R. Tufte. *Visual Display of Quantitative Information*. Graphics Press, 1983. (Zitiert auf den Seiten 15 und 16)

[WPK89] D. F. Williamson, R. A. Parker, J. S. Kendrick. The Box Plot: A Simple Visual Method to Interpret Data. *Annals of Internal Medicine*, 110(11):916–921, 1989. doi:10.7326/0003-4819-110-11-916. URL +http://dx.doi.org/10.7326/0003-4819-110-11-916. (Zitiert auf Seite 20)

[WPQ+08] T. D. Wang, C. Plaisant, A. J. Quinn, R. Stanchak, S. Murphy, B. Shneiderman. Aligning temporal data by sentinel events: discovering patterns in electronic health records. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, S. 457–466. ACM, New York, NY, USA, 2008. doi:10.1145/1357054.1357129. URL http://doi.acm.org/10.1145/1357054.1357129. (Zitiert auf Seite 23)

[Zha05] D. Zhang. Analysis of Survival Data. http://www4.stat.ncsu.edu/ dzhang2/st745/chap1.pdf, 2005. (Zitiert auf Seite 20)

Alle URLs wurden zuletzt am 28.01.2013 geprüft.

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

(Patricius-Samuel Albu)