Claus Weihs *Hrsg.*

Statistische DatenanaIyse im Journalismus

Fallstudien und wissenschaftliche Anforderungen zum Einsatz fortgeschrittener statistischer Methoden

Springer VS

Statistische Datenanalyse im Journalismus

Claus Weihs
(Hrsg.)

Statistische Datenanalyse im Journalismus

Fallstudien und wissenschaftliche Anforderungen zum Einsatz fortgeschrittener statistischer Methoden

Springer VS

Hrsg.
Claus Weihs
Fakultät Statistik
TU Dortmund
Dortmund, Deutschland

ISBN 978-3-662-64692-2 ISBN 978-3-662-64693-9 (eBook)
https://doi.org/10.1007/978-3-662-64693-9

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Planung/Lektorat: Iris Ruhmann
Springer VS ist ein Imprint der eingetragenen Gesellschaft Springer-Verlag GmbH, DE und ist ein Teil von Springer Nature.
Die Anschrift der Gesellschaft ist: Heidelberger Platz 3, 14197 Berlin, Germany

Vorwort

> "While it is easy to lie with statistics, it is even easier to lie without them."
> (Frederick Mosteller)[1]

Meine langjährigen Erfahrungen mit Datenjournalisten, ihrer Liebe zu Visualisierungen und ihrer nicht so großen Neigung zur Verwendung weiterführender statistischer Methoden haben mich zu diesem Buch motiviert. Die Idee war, mit Hilfe von Fallstudien, denen journalistische Veröffentlichungen zugrunde liegen, zu zeigen, dass man mit weiterführenden statistischen Methoden tiefer liegende Ergebnisse erzielen kann, die objektiven Erkenntnissen näher kommen als vereinfachte Analysen.

Den Ausgangspunkt bildeten zwei Fallstudien, denen Veröffentlichungen von SPIEGEL.de zugrunde liegen (Kapitel 6 und 7). Danach habe ich mich entschieden, dass die Fallstudien die wichtigsten statistischen Methoden abdecken sollten, nämlich Verteilungen und Tests, Klassifikation, Regression, Zeitreihenanalyse, Clusteranalyse, Analyse von sequentiellen Daten ohne direkten Zeitbezug, Verwendung von Vorwissen und geplante Studien. Die statistischen Begriffe sollten in diesen Kapiteln selbst auf möglichst anschauliche Weise eingeführt werden und die Ideen und Ergebnisse der Fallstudien sollten von Datenjournalisten kommentiert werden. Das führte zu den Kapiteln 6 bis 13.

Später kamen dann noch Grundlagenkapitel hinzu, nämlich eine Diskussion der Anwendbarkeit des generellen datenanalytischen Vorgehens CRISP-DM im Datenjournalismus (Kapitel 2) sowie Kapitel über Datenkompetenz (data literacy) (Kapitel 3) und über ethische Richtlinien und Qualitätsstandards für Datenjournalisten (Kapitel 14 und 15).

Insbesondere bei den Fallstudien fiel auf, dass Visualisierungen wieder eine sehr große Rolle spielten, allerdings dieses Mal nicht der Rohdaten, sondern der Ergebnisse der weiterführenden statistischen Methoden. Zum Vergleich, und um den Kreis quasi zu schließen, dient schließlich ein Kapitel zur Datenvisualisierung (Kapitel 4). Daneben spielten in den Fallstudien auch Algorithmen eine wichtige Rolle, um die unbekannten Parameter der verwendeten Modelle zu bestimmen. Das motivierte schließlich ein allgemeines Kapitel zu Algorithmen (Kapitel 5).

Dortmund, im August 2021 *Claus Weihs*

[1] Mosteller, Frederick zitiert nach Murray, Charles: How to Accuse the Other Guy of Lying with Statistics. In: Institute of Mathematical Statistics (Hrsg.): Statistical Science 20, Nr. 3. Hayward, Kalifornien, USA 2005. S. 240

Danksagung

Danken möchte ich Wiebke Ahlers, Lilia Michailov, Manuela Vida-Mannl, Emily Weidle, Andreas Weilinghoff und Lisa Westermayer für ihre Unterstützung beim Korrekturlesen der Texte.

Inhaltsverzeichnis

Autoren

Anna Behrend [Kapitel 2]
Freiberufliche Daten- und Wissenschaftsjournalistin, Hamburg,
office@annabehrend.de

Christina Elmer [Kapitel 4, 5]
TU Dortmund, Institut für Journalistik, Dortmund,
christina.elmer@tu-dortmund.de

Ursula Garczarek [Kapitel 14]
Cytel, Strategic Consulting, Berlin,
Ursula.Garczarek@cytel.com

Tanja Hernández Rodríguez [Kapitel 12, Anhang A.12]
Technische Hochschule Ostwestfalen-Lippe, Lemgo,
tanja.hernandez@th-owl.de

Katja Ickstadt [Kapitel 3]
TU Dortmund, Lehrstuhl Mathematische Statistik und biometrische Anwendungen, Dortmund,
ickstadt@statistik.tu-dortmund.de

Hendrik Lehmann [Kapitel 11]
Head of Tagesspiegel Innovation Lab, Berlin,
Hendrik.Lehmann@tagesspiegel.de

Lilia Michailov [Kapitel 11, Anhang A.11]
TU Dortmund, Fakultät Statistik, Dortmund,
l.michailov@web.de

Ana Moya [Kapitel 8, Anhang A.8]
FUNKE Data & CX Management, Essen,
A.Moya@funkemedien.de

Henrik Müller [Kapitel 3]

TU Dortmund, Institut für Journalistik, Dortmund
henrik.mueller@tu-dortmund.de

Gerret von Nordheim [Kapitel 13]
Universität Hamburg, Journalistik und Kommunikationswissenschaft, Hamburg,
gerret.vonnordheim@uni-hamburg.de

Marcel Pauly [Kapitel 6]
Leiter Datenjournalismus, DER SPIEGEL, Hamburg,
marcel.pauly@spiegel.de

Patrick Stotz [Kapitel 7]
Redakteur Datenjournalismus, DER SPIEGEL, Hamburg,
patrick.stotz@spiegel.de

Detlef Steuer [Kapitel 14]
Helmut-Schmidt Universität, Computergestützte Statistik, Hamburg,
steuer@hsu-hh.de

Marie-Louise Timcke [Kapitel 8, 9]
Leitung Funke Interaktiv, FUNKE Zentralredaktion Berlin GmbH, Berlin,
Marie-Louise.Timcke@funkemedien.de

Claus Weihs [Herausgeber, Kapitel 1, 6, 7, 8, 9, 10, 11, 12, 13, Anhang A.6, A.7, A.8, A.9, A.10, A.11, A.12]
TU Dortmund, Fakultät Statistik, Dortmund,
claus.weihs@tu-dortmund.de

Henrike Weinert [Kapitel 3]
TU Dortmund, DaCoNet (Data Competence Network), Dortmund,
henrike.weinert@tu-dortmund.de

Holger Wormer [Kapitel 12, 15]
TU Dortmund, Institut für Journalistik, Dortmund,
holger.wormer@tu-dortmund.de

Notation

Der Herausgeber dieses Buches hat den Autoren und Autorinnen freigestellt, wie sie mit dem Geschlechteraspekt in Bezug auf eine Gruppe von Personen umgehen wollen. Deshalb werden in diesem Buch verschiedene Formulierungen z. B. für Autor und Autorinnen verwendet wie etwa Autoren, Autor/innen und Autor*innen. Eine gemeinsame Darstellung war nicht zu erreichen. Ich denke, diese Vielfalt spiegelt die augenblickliche Heterogenität der Einstellungen in der Gesellschaft zu diesem Thema angemessen wider.

Abweichend von Standardrechtschreibregeln haben wir im gesamten Buch

- einen Dezimalpunkt verwendet und kein Dezimalkomma (also z. B. 1.23 statt 1,23) und
- die deutschen Anführungsstriche auch bei englischen Ausdrücken (also z. B. „english“ und nicht ‘english’).

Teil I
Einführung, Konzept und Grundlagen

Kapitel 1
Einführung

Claus Weihs

Zusammenfassung Dieses Kapitel motiviert das Buch und gibt eine kurze Inhaltsangabe. Außerdem werden kurz verschiedene datenanalytische Studientypen diskutiert sowie statistische Modelle und ihre Beurteilung.

1.1 Motivation

Datenjournalistik ist eine immer bedeutsamer werdende journalistische Disziplin. Neben der Erfassung und Aufbereitung von Daten hängt die Qualität und Aussagekraft datenjournalistischer Veröffentlichungen wesentlich von einer adäquaten Datenanalyse ab. Sehr oft werden Visualisierungen von Rohdaten für solche Analysen genutzt. In diesem Buch wollen wir aber insbesondere die Bedeutung von statistischen Analysemethoden für die Qualität der Ergebnisse untersuchen.

Es gibt leider nur wenige wissenschaftliche Studien zu der Qualität von statistischen Auswertungen oder Datenvisualisierungen im Journalismus (z.B. Young (2018) und die dort angegebenen Zitate). Aus praktischer Sicht stellen viele relevante analytische Methoden oft Neuland für Datenjournalisten dar und obwohl die Qualität datenjournalistischer Arbeiten in den letzten Jahren deutlich zugenommen hat, gibt es immer noch methodische Schwächen. Dazu passt, dass man in datenjournalistischen Lehrbüchern und Kursen immer noch Hinweise zum Gebrauch so grundlegender statistischer Methoden wie Median und Mittelwert usw. findet und dass man Korrelation nicht mit Kausalität verwechseln sollte (s. z.B. Cairo (2016)).

Wenn man berücksichtigt, dass Zeitungs- und Zeitschriftenleser meist insofern die Allgemeinheit repräsentieren, dass sie sehr unterschiedliche mathematisch/statistische Vorkenntnisse haben, sollten datenjournalistische Analysen relativ leicht verständlich sein. Deshalb nehmen Datenjournalisten z. B. manchmal an, dass sogar die Interpretation von Streudiagrammen zu Verständnisproblemen füh-

C. Weihs (Hrsg.), *Statistische Datenanalyse im Journalismus*,
https://doi.org/10.1007/978-3-662-64693-9_1

ren kann. Der wohlmeinende Datenjournalist möchte also Fehlinterpretationen vermeiden, aber natürlich auch das Wichtigste zeigen, das aus den verfügbaren Daten gelernt werden kann. Datenjournalisten sollten daher zumindest die wichtigsten fortgeschrittenen statistischen Methoden kennen und ihre Ergebnisse einfach erklären können. Das wiederum schafft für einen Statistiker die Notwendigkeit, solche Methoden möglichst einfach einzuführen und Wege aufzuzeigen, wie man die Ergebnisse einfach erklären kann. Das sind die Hauptmotivationen für dieses Buch.

Die für dieses Buch grundlegenden Fragen sind: Wie statistisch arbeiten Datenjournalisten heute und wie statistisch können oder sollten sie arbeiten? Diese Fragen werden auch innerhalb der aktiv im Feld des Datenjournalismus arbeitenden Journalisten viel diskutiert. Zum Beispiel hat die Datenjournalistin Anna Behrend dazu die Standardmethode der statistischen Datenanalyse CRISP-DM (CRoss Industry Standard Process for Data Mining) mit Datenjournalisten diskutiert. Die Ergebnisse stellt sie in Kapitel 2 dieses Buches vor. Auch dabei wurden die Methodenkenntnisse der Datenjournalisten als Schwachstelle identifiziert. In diesem Buch soll nun das besondere Augenmerk auf den statistischen Analyse- und Darstellungs-Methoden in datenjournalistischen Anwendungen liegen. Dabei soll es darum gehen, den Vorteil weiterführender statistischer Methoden gegenüber statistischen Basismethoden an Hand von Beispielen herauszuarbeiten. Dazu werden relevante statistische Begriffe anhand von Fallstudien erklärt. Daneben werden die Begriffe aber auch formal definiert, um eine exakte Argumentation zu ermöglichen.

Das Buch diskutiert also ausführlich Anforderungen und Beispiele von statistischer Datenanalyse im Datenjournalismus in den Kapiteln 2 und 6-13. Außerdem setzt das Buch aber auch einen Anforderungsrahmen für die datenjournalistische Arbeit auf Gebieten, in denen Daten eine entscheidende Rolle spielen: der Datenkompetenz und -visualisierung, dem Einsatz von Algorithmen sowie bei daten-ethischen Anforderungen und der Überprüfung externer Studien (Kapitel 3-5 und 14-15).

1.2 Daten und Studientypen

Statistik und Datenjournalismus haben gemeinsam, dass sie nur mit Daten sinnvoll möglich sind. Diese Daten resultieren aus Beobachtungen der Realität. Wenn der Beobachter nicht in die Realität eingreift, spricht man von *Beobachtungsstudien*. Wenn die Realität bewusst durch Aktionen des Beobachters beeinflusst wird, spricht man von *geplanten Studien* mit Hilfe von Versuchsanordnungen (Design of Experiments). Letzteres ist z. B. bei Laborversuchen der Fall, wo die Realität unter idealisierten, festgelegten Laborbedingungen untersucht wird. Darauf werden wir in Kapitel 13 des Buches noch eingehen. Die Standardsituation eines Datenjournalisten ist die des passiven Beobachters. Dabei soll z. B. ein Datensatz analysiert werden, der einer Datenbank entnommen wird.

Definition 1.1. *[Beobachtungsstudie] Für die Analyse einer* Beobachtungsstudie *nimmt der Statistiker typischerweise an, dass nicht alle Daten einer Gesamtpopulation (Grundgesamtheit) vorliegen, sondern nur ein Ausschnitt, eine sogenannte*

Stichprobe. *Dabei werden ein oder mehrere Merkmale (Variablen) gemeinsam an mehreren Merkmalseinheiten (Objekten / Subjekten) erhoben. Die Repräsentativität der Stichprobe wird angenommen, d. h. die Stichprobe sollte die relevanten Teile der Gesamtpopulation widerspiegeln. Aus der Stichprobe versucht der Statistiker, auf Eigenschaften der Gesamtpopulation zu schließen.*

Beispiel 1.1 Stichprobe aus der Gesamtbevölkerung *Aus einer Stichprobe aus der Gesamtbevölkerung eines Staates mit Beobachtungen zu den Merkmalen Alter, Geschlecht, Bruttoeinkommen und Familienstand soll z. B. auf die Verteilung des Bruttoeinkommens für eine bestimmte Altersklasse, Geschlecht und Familienstand geschlossen werden. Dazu muss die Stichprobe natürlich diese Eigenschaften möglichst in ihrer gesamten Breite repräsentieren. Es müssen also z. B. genügend Daten für alle betrachteten Altersklassen, Geschlechter und Familienstände zur Verfügung stehen. Außerdem sollte die Stichprobe die verschiedenen Bruttoeinkommen in der richtigen Häufigkeit repräsentieren. Nur dann können Aussagen über die Gesamtbevölkerung akzeptabel sein.*

Definition 1.2. *[Geplante Studie] Eine* geplante Studie *besteht aus mehreren sogenannten* Versuchen*. Für eine geplante Studie werden möglichst alle Merkmale (sogenannte Faktoren), die einen Einfluss auf eine Zielgröße haben können, für jeden einzelnen Versuch festgelegt. Jeder* Versuch *wird durch eine bestimmte Kombination von Merkmalswerten repräsentiert. Die Idee bei geplanten Studien ist, mit so wenigen Versuchen wie möglich die maximale Information über den untersuchten Zusammenhang zu erhalten, da die einzelnen Versuche häufig aufwendig durchzuführen sind. Dazu wurden sogenannte* statistische Versuchspläne *entwickelt, die nach unterschiedlichen Kriterien optimal sind.*

Beispiel 1.2 Wahrnehmungskluft zwischen ökonomischen Experten und Laien *(vgl. Kapitel 13) Vergleichende Befragungsstudien belegen eine fundamentale Wahrnehmungskluft zwischen ökonomischen Experten und Laien. Das Phänomen ist von hoher Relevanz, weil die Einstellungen der Laien typischerweise im diametralen Gegensatz zu den wissenschaftlich fundierten und übereinstimmenden Empfehlungen der ökonomischen Experten stehen. Mittels kontrollierter Laborexperimente können nun journalistische Strategien und Faktoren identifiziert werden, die diese Wahrnehmungskluft verringern. Dazu könnten z. B. Zeitungs-, Online- und Fernsehbeiträge gezielt variiert werden. Dabei interessieren u. a. die Effekte*
a) unterschiedlicher Argumentations- und Einstiegsarten,
b) der Vermittlung von Nebenwirkungen wirtschaftspolitischer Handlungen,
c) der Nutzung einer „Make-it-fun“-Strategie,
d) der Auswahl unterschiedlicher ökonomischer Experten als Quellen.
Insbesondere können dabei mehrere Einflussfaktoren gleichzeitig variiert und in ihren Wechselwirkungen untersucht werden. Das Verständnis der journalistischen Beiträge kann z. B. mit Hilfe von Fragebögen überprüft werden.

1.3 Modelle und ihre Beurteilung

Die statistische Datenanalyse beschäftigt sich im Allgemeinen mit **Modellen** für Zusammenhänge zwischen Merkmalen. Solche Modelle erheben nicht den Anspruch, den Zusammenhang exakt wiederzugeben, noch nicht einmal für die beobachteten Daten. Es sind ja nur Modelle der Realität, nicht die Realität selber.

Warum nimmt man aber einen solchen „Realitätsverlust“ in Kauf? Tatsächlich werden die Modelle z. B. für die Vorhersage von Werten verwendet, die nicht beobachtet wurden. Genauer geht es oft um die Vorhersage unbekannter Werte einer so genannten Zielgröße aus den bekannten Werten so genannter Einflussgrößen. Zum Beispiel könnte man versuchen, das Wahlergebnis einer Partei (Zielgröße) aus der Einkommensstruktur (Einflussgröße) eines Wahlkreises vorherzusagen. Dazu benötigt man einen möglichst allgemein gültigen Zusammenhang zwischen der Zielgröße und den Einflussgrößen und nicht einen Zusammenhang, der zwar auf den beobachteten Daten exakt gilt, aber viel schlechter bei nicht beobachteten Werten der Zielgröße.

Um zu beurteilen, wie gut eine solche Vorhersage gelingt, verwendet man häufig sogenannte **Resampling Verfahren** (vgl. Kapitel 7). Tatsächlich liegt im Allg. nur ein einziger Datensatz vor, bei dem die Werte aller beteiligten Merkmale bekannt sind. Dieser Datensatz muss deshalb aufgeteilt werden, wenn man eine echte Vorhersagesituation schaffen will. Dann verwendet man nur einen Teil der Daten zur Modellbestimmung und den anderen Teil für die Vorhersage. Diese Methode hat den großen Vorteil, dass die vorherzusagenden Werte der Zielgröße bekannt sind und die Vorhersagegüte deshalb bestimmbar ist.

Definition 1.3. *[Kreuzvalidierung] Ein Beispiel für so ein Vorgehen ist die so genannte Kreuzvalidierung, z. B. die Leave-One-Out (LOO)-Kreuzvalidierung. Echte Vorhersagen werden dabei mit Hilfe des so genannten „Resampling-Tricks“ erreicht, um nicht einige der Beobachtungen ganz aus der Bestimmung des Modells herauszulassen. Beim Resampling werden aus dem einen beobachteten Sample (Stichprobe) neue (künstliche) Stichproben gezogen, bei LOO n neue Stichproben mit jeweils (n-1) Elementen, so dass jede Beobachtung je einmal weggelassen wird. Die (n-1) Beobachtungen werden zur Modellbildung verwendet, die eine ausgelassene Beobachtung zur Vorhersage. Damit werden alle Beobachtungen (n-1) mal zur Modellbestimmung verwendet und jede Beobachtung einmal vorhergesagt. Diese Vorhersagen können dann mit den echten Beobachtungen verglichen werden. Daraus kann man ein Maß für die Vorhersagegüte eines Modells bestimmen.*

Verschiedene Arten von statistischen Modellen in diesem Buch: Schon die statistischen Verteilungen sind oft nur ein Modell der Realität. Verteilungen kommen in vielen Kapiteln vor, werden aber insbesondere in den Kapiteln 6 und 12 diskutiert. Das Arbeitspferd unter den statistischen Modellen ist das *lineare Modell*, das wir z. B. bei der Regression, der Glättung und der Versuchsplanung vorstellen (Kapitel 8, 9, 10, 13). Bei der Regression verwenden wir auch verallgemeinerte Varianten, wie das generalisierte lineare Modell in Kapitel 8 und das nichtlineare

Regressionsmodell in Kapitel 9. Ein weiteres sehr beliebtes Modell ist das *Baummodell*, das wir sowohl für die Klassifikation als auch für die Regression diskutieren (Kapitel 7, 8). Dieses Modell hat den großen Vorteil der einfachen Interpretierbarkeit. Die in diesem Buch vorgestellten *Zeitreihenmodelle* lassen sich oft auch als lineare Modelle interpretieren, allerdings wird dabei die zeitliche Dynamik mit modelliert. Wichtigste Vertreter sind die autoregressiven und die Moving-Average Modelle (Kapitel 9).

1.4 Inhalt

Anforderungen an Datenjournalisten: Neben dem grundlegenden Kapitel 2 zum Aufbau einer datenjournalistischen Analyse beinhaltet das Buch noch weitere grundlegende Kapitel zu Anforderungen an datenjournalistische Arbeit. Kapitel 3 diskutiert die für einen Datenjournalisten geradezu existentiell wichtige Datenkompetenz, Kapitel 4 führt Standards der Datenvisualisierung ein, Kapitel 5 diskutiert den Einsatz von Algorithmen im Datenjournalismus, Kapitel 14 ethische Probleme im Datenjournalismus und Kapitel 15 schlägt Checklisten für die praktische Überprüfung externer Studien vor.

In diesem Buch werden außerdem die aus Sicht eines Statistikers wichtigsten statistischen Begriffe und Methoden an Hand von Fallstudien diskutiert und die Erzeugung der Ergebnisse anhand von Computerprogrammen demonstriert.

Statistische Datenanalyse: Aufbauend auf dem grundlegenden Begriff der *Verteilung* (Kapitel 6) gehen wir zu den wichtigsten Analyseverfahren über, nämlich *statistisches Testen* (Kapitel 6), *Klassifikation* (Kapitel 7), *Regression* (Kapitel 8), *Zeitreihenanalyse* (Kapitel 9) und *Clusteranalyse* (Kapitel 10). Als weitergehende Verfahren wird eine Kombination von *Glätten* bei Zeitreihen, Clusteranalyse und Klassifikation vorgestellt (Kapitel 10), eine *Analyse von sequentiellen Daten in Verkehrsnetzen* (Kapitel 11) und die Berücksichtigung von *Vorinformation* bei statistischen Analysen (Kapitel 12). Als abschließende Methode stellen wir eine mögliche Anwendung von geplanten Studien im Datenjournalismus vor (Kapitel 13). Häufig haben nicht sämtliche Variablen, für die Beobachtungen vorliegen, einen relevanten Einfluss auf die zu vorhersagende Zielgröße. Die dann nötige *Variablenselektion* wird anhand der Entscheidungsbäume in den Kapiteln 7, 8 demonstriert.

Visualisierungen: In Kapitel 4 werden grundsätzliche Überlegungen zur Datenvisualisierung im Datenjournalismus diskutiert. In diesem Buch werden *grafische Darstellungen* insbesondere zur Darstellung der Ergebnisse der statistischen Analyseverfahren verwendet. Verschiedene *Dichte- und Verteilungsfunktionsdarstellungen* finden sich in den Kapiteln 6, 7, 10, 12. *Darstellungen von Baummodellen* werden in den Kapiteln 7, 8, 10 gezeigt. Darstellungen zur Überprüfung der *Modellanpassung bei Zeitreihen* finden sich in Kapitel 9, 10 und *Glättungsdarstellungen* in Kapitel 10. *Clusterdarstellungen* mit Hilfe von *Dendrogrammen* werden in

Kapitel 10 vorgestellt und Darstellungen von *Unsicherheitsintervallen* in den Kapiteln 10, 12. Kapitel 11 zeigt Darstellungen der Ergebnisse von Analysen von sequentiellen Daten in Verkehrsnetzen.

Fallstudien: Wir verwenden grundsätzlich zwei verschiedene Arten von Fallstudien. Entweder ist die Basis eine journalistische Veröffentlichung und das Kapitel schlägt eine weitergehende statistische Analyse vor, die danach mit Journalisten diskutiert wird (Kapitel 6, 7, 8, 13). Oder das Kapitel verwendet Daten aus der journalistischen Praxis für eine eigene Analyse, die dann mit Journalisten diskutiert wird (Kapitel 9, 10, 11). Kapitel 12 beruht schließlich auf Daten, die mit der Unterstützung von Datenjournalisten selbst erhoben wurden.

Die Themen der Fallstudien umfassen sehr unterschiedliche journalistische Teilgebiete wie die Altersstruktur von Parlamentariern (Kapitel 6), Wahlanalysen (Kapitel 7, 8), Infizierte bei der Corona-Pandemie (Kapitel 9), Buchbestselleranalysen (Kapitel 10), Radwegnutzungsanalysen (Kapitel 11), Arbeitszeitanalysen (Kapitel 12) und Wirkungsanalysen von journalistischen Texten in verschiedenen Darstellungen (Kapitel 13).

Computerprogramme: Das Vorgehen bei den datenanalytischen Analysen wird nicht nur in den Kapiteln zu den Fallstudien dokumentiert, sondern auch anhand von Programmen in der Software R (R (2019)) im Anhang. Damit sollte es für einen Datenjournalisten möglich sein, die Analysen an ähnlichen Datensätzen teilweise selbst zu wiederholen.

Insgesamt deckt das Buch also sowohl grundlegende Überlegungen zur Arbeit von Datenjournalisten ab als auch Beispiele für relevante Typen von Datenanalysen.

Literaturverzeichnis

Cairo A. (2016) *The Truthful Art: Data, Charts, and Maps for Communication (Voices That Matter)*. New Riders

R Core Team (2019). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. https://www.R-project.org/.

Young M.L., Hermida A., Fulda J (2018) *What makes for great data journalism? A content analysis of data journalism awards finalists 2012 – 2015*. Journalism Practice 12. 115–135, doi: 10.1080/17512786.2016.1270171

Kapitel 2
CRISP-DM - Ein Konzept für die journalistische Datenanalyse?

Anna Behrend *(mit Kommentaren des Herausgebers)*

Zusammenfassung Das folgende Kapitel behandelt die Frage, inwiefern sich Methoden des industriellen Data Mining sinnvoll auf Arbeitsabläufe im Datenjournalismus übertragen lassen. Der Data-Mining-Standardprozess „CRISP-DM“ weist viele Gemeinsamkeiten mit für den Datenjournalismus typischen Arbeitsschritten auf. Ein wesentlicher Unterschied besteht jedoch in der Wahl der Analysemethoden: Im Datenjournalismus sind diese in der Regel wesentlich simpler als im Data Mining. Das zeigen sowohl die Literatur als auch Experteninterviews zum Thema. Eine starre Datenjournalismus-Leitlinie im Stil von „CRISP-DM“ sehen die meisten befragten Experten kritisch. Dafür sind die Projekte und Arbeitsabläufe zu divers. Damit ein inhaltlich richtiges und relevantes Endprodukt entsteht, wird zusätzlich vom Datenjournalisten viel eigenes Einschätzungsvermögen und Fachwissen verlangt, welches er in seiner Ausbildung – vielleicht mit Hilfe einer Leitlinie – erlernt und verinnerlicht hat.

2.1 Motivation und Methodik

Seit einigen Jahrzehnten werden immer mehr maschinenlesbare Daten verfügbar. Dies hat dazu geführt, dass unterschiedliche gesellschaftliche Akteure aus diesen Daten Erkenntnisse gewinnen wollen: Unternehmen nutzen Daten zum Beispiel verstärkt, um das Verhalten ihrer Kunden zu analysieren und aus den gewonnen Informationen wirtschaftlichen Profit zu ziehen. Dabei kann es um die Identifikation von Käuferprofilen gehen, um Warenkorbanalysen, die Beurteilung von Mitarbeitern oder die Prognose von Vertragslaufzeiten (Gabler Wirtschaftlexikon, Stand 2018). Journalisten suchen in Datensätzen nach Antworten auf gesellschaftlich relevante Fragen. Sie untersuchen etwa das Bewertungssystem der Schufa, den Lärmpegel in den Straßen Berlins oder wie sich die Debatte im Bundestag mit dem Einzug der AfD verändert hat.

C. Weihs (Hrsg.), *Statistische Datenanalyse im Journalismus*,
https://doi.org/10.1007/978-3-662-64693-9_2

Im Rahmen meiner Bachelorarbeit aus dem Jahr 2013 (Behrend (2013)) habe ich untersucht, inwiefern sich Methoden des industriellen Data Mining sinnvoll auf Arbeitsabläufe im Datenjournalismus übertragen lassen. Die zentrale **Forschungsfrage** der Arbeit lautet: „Ist der Data-Mining-Standardprozess CRISP-DM als Ausgangspunkt für die Entwicklung eines datenjournalistischen Leitfadens geeignet?“

Damit einhergehend wurden folgende **Leitfragen** behandelt:

Leitfrage 1: Kann für Datenjournalismus überhaupt eine Leitlinie entwickelt werden?
Leitfrage 2: Wie hoch ist der Grad der Übereinstimmung zwischen CRISP-DM und dem datenjournalistischen Arbeitsprozess?
Leitfrage 3: Würde eine höhere Übereinstimmung des datenjournalistischen Arbeitsprozesses mit CRISP-DM gemäß den Qualitätskriterien des Datenjournalismus eine Verbesserung bedeuten?
Leitfrage 4: Welche Gründe sprechen für oder gegen CRISP-DM als Ausgangspunkt für die Entwicklung einer Leitlinie?

Dieses Kapitel fasst die wichtigsten Ergebnisse der Arbeit zusammen, ergänzt um zwei zusätzliche Experteninterviews aus dem Jahr 2015 sowie Literaturquellen neueren Datums.

Zunächst werden zentrale Begrifflichkeiten erläutert, anschließend der Standardprozess „CRISP-DM“ sowie typische datenjournalistische Arbeitsschritte vorgestellt. Der folgende Vergleich von Data-Mining-Standardprozess und datenjournalistischer Arbeitsweise stützt sich auf die Literatur zum Thema sowie auf Leitfadeninterviews mit namhaften Datenjournalisten. Abschließend werden die Ergebnisse im Hinblick auf die Forschungsfrage sowie die Leitfragen diskutiert.

2.2 Begriffsbildung

2.2.1 Datenjournalismus

Viele Praktiker der Branche dürften es halten wie der Datenjournalimus-Pionier Adrian Holovaty. In einem Blogpost beantwortete er die Frage nach der Definition von Datenjournalismus kurzerhand mit der Gegenfrage: „Wen interessiert’s?“ (Rogers (2013)). Da die Begriffsklärung für die Journalismusforschung jedoch durchaus relevant ist, wurde in den vergangenen Jahren eine Vielzahl von unterschiedlichen Definitionen für datengetriebene Formen des Journalismus aufgestellt. Coddington spricht gar von einer „Kakophonie der sich überlappenden und unscharfen Definitionen“ (Coddington (2015)).

Laut Ausserhofer wird Datenjournalismus in manchen Fällen eher als Prozess gesehen, in anderen eher als Produkt (Ausserhofer et al. (2017)). Dem Prozess-Ansatz nach geht es darum, Geschichten mittels quantitativer Methoden in Daten zu finden. Liegt der Fokus auf Datenjournalismus als Produkt, so wird er häufig als spezielle Form der Präsentation durch interaktive Visualisierungen beschrieben.

In diesem Kapitel wird folgende Definition verwendet:

Datenjournalismus ist eine Recherchemethode, bei der Daten Quelle und Berichterstattungsgegenstand sind. Die oft großen Datenmengen werden einer computergestützten Analyse – inklusive Bereinigung, Strukturierung und gegebenenfalls Verknüpfung – unterzogen. Durch diese Analyse in Kombination mit klassischen journalistischen Recherchemethoden, zum Beispiel der Befragung, entsteht ein Mehrwert. Dieser kann im Endprodukt in Form einer ausformulierten journalistischen Geschichte enthalten sein. Alternativ oder ergänzend dazu können die Daten dem Rezipienten zum individuellen Erkunden in interaktiver Form angeboten werden. Datenjournalistische Produkte sind hinsichtlich ihrer Veröffentlichungsform nicht festgelegt, auch wenn Online-Artikel mit teils interaktiven Visualisierungen typisch sind. Die Datenquellen werden im fertigen Produkt so weit wie möglich transparent gemacht und die Rohdaten im Idealfall veröffentlicht.

Die Beschaffung von Daten wird in der Definition implizit vorausgesetzt und daher nicht extra erwähnt. In der Realität kann sie aber einen erheblichen Aufwand bedeuten und bestimmte, für den Datenjournalismus typische Techniken erfordern. Dazu zählen etwa das Sammeln von Daten aus dem Internet mittels selbst erstellter Programme (sogenannten „Scrapern“) oder das automatisierte Auslesen von Daten aus PDF-Dokumenten.

2.2.2 Data Mining und KDD

Ähnlich wie beim Datenjournalismus finden sich auch zu den Begriffen Data Mining und Knowledge Discovery in Databases (kurz: KDD, zu Deutsch: Wissensentdeckung in Datenbanken) unterschiedliche Definitionen. Häufig ist jedoch von Data Mining als „Kern“ (Gabler, Wirtschaftslexikon, 2018) oder „eigentlichem Analyseschritt“ (Ester und Sander (2000)) im KDD-Prozess die Rede.

Motivation und Aufgabe der Wissensentdeckung in Datenbanken und somit auch des Data Mining ist es, durch „(semi)automatische Extraktion“ gültiges, bisher unbekanntes und potenziell nützliches Wissen aus Daten zu gewinnen (Ester und Sander (2000)).

In einem bis heute viel zitierten Aufsatz aus dem Jahr 1996 werden neun Schritte des KDD-Prozesses beschrieben, durch die aus Daten Wissen generiert wird. Der

Prozess führt von der Identifikation der Ziele über Auswahl, Bereinigung und Analyse der Daten bis hin zur Auswertung und Anwendung der gewonnenen Ergebnisse (**?**). Die beschriebenen Schritte ähneln stark den in Abschnitt 2.3.1 beschriebenen Elementen des "CRoss-Industry Standard Process for Data Mining" (CRISP-DM). Nach der Begriffsdefinition von nach Fayyad et al. beschreibt CRISP-DM also eher einen KDD- als einen Data-Mining-Standardprozess.

2.2.3 Qualität im Datenjournalismus

Um zu beurteilen, ob sich eine Leitlinie in Anlehnung an den Standardprozess CRISP-DM positiv auf die Qualität des Datenjournalismus auswirken könnte, muss definiert werden, an welchen Qualitätskriterien dieser zu messen ist.

Datenjournalismus ist in erster Linie Journalismus. Daher gelten für ihn auch die typischen journalistischen Qualitätsdimensionen Aktualität, Relevanz, Richtigkeit und Vermittlung (Rager (1994)). Dabei spielen Aspekte wie der Neuigkeitswert einer Information, Überprüfbarkeit, Sorgfalt, Transparenz oder die Bewertung von Relevanz gemäß der Nachrichtenwerttheorie (Rager (1994), S.196) eine Rolle.

Da Datenjournalismus sich Methoden bedient, die teilweise denen der Wissenschaft ähneln, lässt sich die Frage stellen, ob Datenjournalismus auch wissenschaftlichen Qualitätskriterien genügen muss.

Was das Streben nach intersubjektiver Information angeht, gibt es zwischen Wissenschaft und Journalismus durchaus eine große Übereinstimmung. Der entscheidende Unterschied ist jedoch, welche Maßstäbe angelegt werden, um eine Information als einigermaßen gesichert einzustufen: Bedarf es in der Wissenschaft dafür meist eines Peer-Review-Verfahrens, genügen im Journalismus in der Regel schon zwei unabhängige Quellen. Ein stark reglementierter Prozess zur Qualitätssicherung wie in der Wissenschaft würde aufgrund des Aktualitätsdrucks jede Art von Journalismus – auch den Datenjournalismus – unmöglich machen. Dürfen datenjournalistische Ergebnisse also nicht veröffentlicht werden, weil sie keinen derartigen Prozess durchlaufen haben? Nach Auffassung der Autorin ist dies klar zu verneinen: Sie dürfen lediglich nicht so präsentiert werden, als wären sie wissenschaftlich geprüft.

So wie die Pflicht zur Sorgfalt im Journalismus die Pflicht zur Wahrheit ersetzt, sollte die Pflicht zur Transparenz die Pflicht zur wissenschaftlichen Vorgehensweise ersetzen. Dies gilt nicht nur, aber insbesondere für den Datenjournalismus.

2.3 Beschreibung der Arbeitsabläufe

2.3.1 Data-Mining-Prozess CRISP-DM

CRISP-DM ist ein Akronym für "CRoss-Industry Standard Process for Data Mining", was in etwa übersetzt werden kann mit „Industrieübergreifender Standard-Prozess für Data-Mining". CRISP-DM wurde Ende der 90er Jahre von drei Pionieren des Data-Mining-Marktes (NCR, SPSS und Daimler AG) entwickelt. Durch die Einführung eines Standards sollte verhindert werden, dass jeder Neuling im Bereich Data Mining immer wieder alles durch Versuch und Irrtum von neuem herausfinden muss. CRISP-DM zählt bis heute zu den am meisten genutzten Prozessmodellen Göpfert und Breiter (2015)). Die Erfinder des Konzepts schreiben dessen Erfolg dem starken Anwendungsbezug zu.

Das CRISP-DM-Konzept ist in vier Ebenen gegliedert, die den Data-Mining-Prozess vom allgemeinen bis hin zum speziellen Vorgehen beschreiben. Auf der ersten, allgemeinsten Ebene stehen die Phasen des Data-Mining-Prozesses. Sie bestehen jeweils aus mehreren generischen Aufgaben, die möglichst alle denkbaren Data-Mining-Prozesse und -Anwendungen abdecken sollen und somit die zweite Ebene bilden. Die dritte Ebene beschreibt, wie diese Aufgaben sich in unterschiedlichen Situationen unterscheiden. Die vierte und detaillierteste Ebene wird als Prozessinstanz bezeichnet. Sie ist „entsprechend den Aufgaben strukturiert, die auf den höheren Ebenen definiert sind, stellt jedoch dar, was in einem bestimmten Projekt tatsächlich geschehen ist, statt allgemein zu bleiben." (IBM: CRISP-DM 1.0. Data Mining Schritt für Schritt - Ein Leitfaden., S. 3)

Zu der Phase „Datenaufbereitung" gehört zum Beispiel die generische Aufgabe „Daten bereinigen". Bei den speziellen Aufgaben muss dann unterschieden werden, ob zum Beispiel „numerische oder kategoriale Werte bereinigt werden sollen"(IBM: CRISP-DM 1.0. Data Mining Schritt für Schritt - Ein Leitfaden., S. 3.). Auf der vierten Ebene könnte dann zum Beispiel festgehalten werden, dass in dem vorliegenden Projekt Beobachtungen mit fehlenden oder unsinnigen Werten ausgeschlossen wurden.

In Abbildung 2.1 sind die sechs Phasen von CRISP-DM in einem Flussdiagramm dargestellt. Die Pfeile bezeichnen dabei die häufigsten Beziehungen zwischen ihnen. Die Entwickler des Prozesses betonen jedoch, dass Beziehungen prinzipiell zwischen allen Phasen bestehen können. Im Folgenden werden die sechs Phasen mit ihren zugehörigen generischen Aufgaben erläutert.

Phase I) Untersuchung der Geschäftsziele
Generische Aufgaben: Geschäftsziele bestimmen, Situation beurteilen, Data-Mining-Ziele bestimmen, Projektplan erstellen

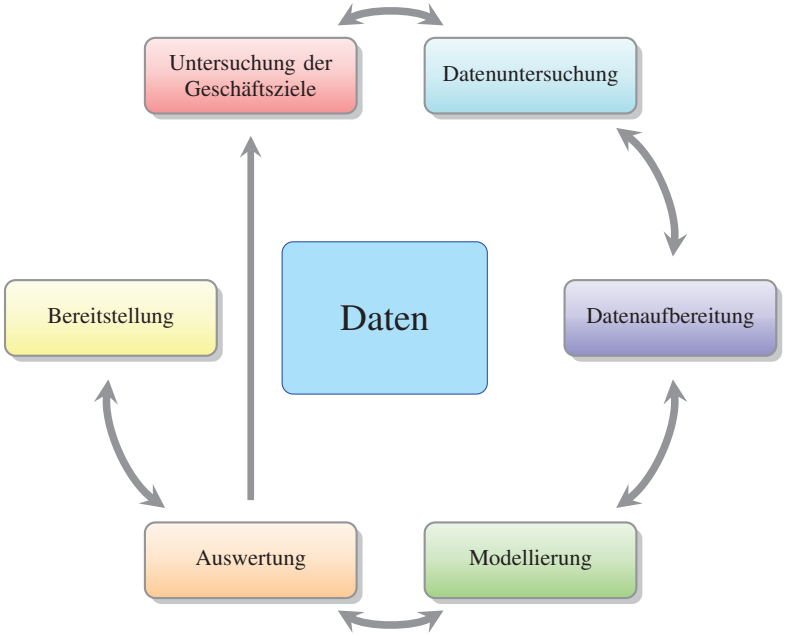

Abb. 2.1 Die sechs iterativen Phasen von CRISP-DM

In dieser Phase müssen die Ziele und Anforderungen des Projekts aus geschäftlicher Perspektive ("business perspective") verstanden werden. Aus diesem Verständnis heraus wird das konkrete Datenanalyse-Problem definiert und ein vorläufiger Plan zum Erreichen der Ziele formuliert.

Phase II) Datenuntersuchung
Generische Aufgaben: Anfangsdaten erfassen, Daten beschreiben, Daten untersuchen, Datenqualität prüfen

Diese Phase beginnt mit der Datenerfassung. Die weiteren Arbeitsschritte dienen dazu, mit den Daten vertraut zu werden, die Qualität der Daten zu beurteilen, erste Einblicke in die Daten zu gewinnen und interessante Teilmengen in den Daten zu entdecken. Unter der Aufgabe „Daten beschreiben" verstehen Chapman et al., dass die oberflächlichen Eigenschaften der Daten wie Format, Umfang, Art der Einträge untersucht und dokumentiert werden. Beim Erkunden der Daten werden erste einfache Analysen durchgeführt, die sich entweder direkt auf die Datenanalyse-Aufgabe des Projekts beziehen, oder der Aufbereitung der Daten dienen. Erste Visualisierungen können hierbei hilfreich sein. Bei der Überprüfung der Datenqualität sollte geklärt werden, ob die Daten vollständig sind sowie ob und - wenn ja - wie häufig Fehler auftreten.

Phase III) Datenaufbereitung
Generische Aufgaben: Daten auswählen, Daten bereinigen, Daten erstellen, Daten integrieren, Daten formatieren

Diese Phase umfasst alle Schritte, die nötig sind, um aus den Rohdaten den Datensatz zu erstellen, der letztendlich zur Modellierung dienen soll. Zu den Aufgaben gehören die Auswahl von Tabellen, Datensätzen und Attributen sowie die Transformation und Bereinigung von Daten für Modellierungstools.

Phase IV) Modellierung
Generische Aufgaben: Modellierungsverfahren auswählen, Testdesign generieren, Modell erstellen, Modell beurteilen

In dieser Phase werden Modellierungstechniken wie zum Beispiel Clustering, künstliche neuronale Netzwerke oder Nächste-Nachbarn-Algorithmen entsprechend des Problemtyps ausgewählt und angewendet. Mit ihrer Hilfe wird ein Modell erstellt und dessen Parameter so gut wie möglich angepasst. Da manche Techniken bestimmte Anforderungen an die Daten stellen, ist häufig eine erneute Datenaufbereitung erforderlich.

Phase V) Auswertung
Generische Aufgaben: Ergebnisse auswerten, Prozess prüfen, Weitere Schritte festlegen

Um sicherzugehen, dass das in IV) entwickelte Modell die geschäftlichen Anforderungen erfüllt, wird es in dieser Phase sorgfältig ausgewertet und die einzelnen Schritte, die zu seiner Entwicklung geführt haben, werden überprüft. Dabei wird insbesondere kontrolliert, ob ein wichtiges Geschäftsziel nicht genug berücksichtigt wurde. Am Ende dieser Phase wird entschieden, wie die Ergebnisse der Datenanalyse genutzt werden sollen.

Phase VI) Bereitstellung
Generische Aufgaben: Bereitstellung planen, Monitoring und Wartung planen, Schlussbericht erstellen, Projekt überprüfen

Das vom Modell erzeugte Wissen wird in dieser Phase so organisiert und präsentiert, dass der Kunde es verwerten kann. In manchen Fällen reicht es, einen Bericht über das Projekt zu schreiben. In anderen Fällen muss vielleicht ein Computerprogramm geschrieben werden, mit dem der Datenanalyse-Prozess jederzeit im Unternehmen wiederholbar ist. Oft nimmt nicht der Datenanalyst, sondern der Kunde die Bereitstellung vor.

Im CRISP-DM-Prozess spielt insgesamt die Dokumentation der einzelnen Arbeitsschritte eine wichtige Rolle. Es ist vorgesehen, dass Entscheidungen, Ergebnisse und Erfahrungen sorgfältig protokolliert werden.

2.3.2 Typische datenjournalistische Arbeitsschritte

Interviews mit Datenjournalisten, Blogbeiträge, Schulungsunterlagen, Forschungs- und Fachliteratur geben einen Überblick über Arbeitsschritte, die typisch für datenjournalistische Projekte sind. Die im Folgenden aufgelisteten Arbeitsschritte wurden aus den zuvor genannten Quellen abgeleitet und stellen keinen starren Prozess dar: Je nach Projekt können manche dieser Arbeitsschritte wegfallen, mehrfach oder in unterschiedlicher Reihenfolge vorkommen.

I) Ausgangsfrage oder -hypothese festlegen: Häufig werden in der Literatur zwei verschiedene Ausgangspunkte für eine datenjournalistische Recherche genannt. Entweder es wird von Beginn an eine bestimmte Fragestellung verfolgt, die anhand von Daten überprüft werden soll, oder die Recherche wird „durch den Datensatz selbst angestoßen". Laut Thibodeaux (2011) beginnen großartige datenjournalistische Projekte meist nicht mit großartigen Datensätzen, sondern mit bedeutsamen Fragen. Meyer (2002) geht so weit zu sagen, dass man ohne Theorie in den ungeordneten Rohdaten zu ersticken droht.

II) Daten beschaffen: Die Beschaffung von Daten kann auf die unterschiedlichsten Weisen geschehen, sei es durch öffentlich zugängliche Quellen im Internet, durch Inanspruchnahme des Informationsfreiheitsgesetzes oder durch Informanten. Manche Datenjournalisten lesen auch mit Hilfe von „Scraper"-Programmen gezielt Daten aus dem Netz aus. Manchmal kann es sogar nötig sein, die Daten für eine bestimmte Fragestellung selbst zu erheben. Dabei kann in manchen Fällen das Crowdsourcing, also die Mitarbeit vieler Freiwilliger an einem Projekt, von Nutzen sein. Liegt der Recherche eine Hypothese zugrunde, sollte bei der Datenbeschaffung stets geprüft werden, ob die Daten geeignet sind, um die Hypothese zu überprüfen.

III) Daten vorbereiten und bereinigen: Liegen die Daten vor, so gilt es, sie in eine „einheitliche, maschinenlesbare Form"zu bringen und zu bereinigen. Beispielsweise müssen uneinheitliche Benennungen innerhalb eines Datensatzes geändert werden. Leßmöllmann (2012) spricht davon, die Daten „journalistisch urbar" zu machen und zitiert damit den britischen Datenjournalisten Simon Rogers. Zu diesem Arbeitsschritt gehört auch, dass die Daten in das richtige Format gebracht werden. Manchmal müssen Daten beispielsweise erst aus einem PDF extrahiert und in ein CSV-Format überführt werden, bevor mit dem eigentlichen Bereinigen begonnen werden kann.

IV) Daten verstehen: Vor der weiteren Verarbeitung ist es wichtig, die Daten und ihren Kontext genau zu verstehen. Dazu gehört beispielsweise, sich zu vergegenwärtigen, welche Definitionen bei der Erhebung verwendet und welche Fälle dadurch ein- bzw. ausgeschlossen wurden. Im Idealfall gibt es entsprechende Metadaten zu dem Datensatz. In den Bereich „Daten verstehen" fällt auch der von Elmer (2013) angegebene Arbeitsschritt „Aussagekraft und Vollständigkeit prüfen". Dabei kann auch bereits ersichtlich werden, dass für ein besseres Verständnis des Themas

weitere Daten beschafft werden müssen.

V) Daten analysieren und verknüpfen: Die Analyse und Auswertung von Daten ist vielleicht das Kernstück des datenjournalistischen Arbeitsprozesses. Sie scheint für die Erzeugung eines Mehrwerts unentbehrlich. „Analysieren heißt, die Daten zu sortieren, Veränderungen, Durchschnitte, Verhältnisse, Kennzahlen oder Mittelwerte auszurechnen und nach Auffälligkeiten zu suchen“, schreibt Bons (2012). Um solche Auffälligkeiten zu finden, werden Daten häufig visualisiert (siehe auch Schritt X). Auch das Berechnen vergleichbarer Parameter und das Filtern von Daten sind in dieser Phase von Bedeutung. In manchen Fällen kann es nötig sein, mehrere Datensätze miteinander zu verknüpfen. Typisches Beispiel für solch eine Verknüpfung ist das Zusammenführen von ortsabhängigen Informationen mit den entsprechenden Geodaten.

VI) Daten in einen Kontext setzen und interpretieren: Laut Leßmöllmann (2012) ist die Einordnung und Interpretation der Daten unerlässlich für die Erarbeitung einer journalistischen Geschichte: „Dazu gehört auch, sie [die Daten] mit anderen Daten, aber auch mit Kontextinformationen und Hintergrundwissen in Beziehung zu setzen, sie also einzuordnen [...], um dann schließlich Schlüsse daraus zu ziehen, die gesellschaftlich relevant sind.“

VII) Die journalistische Geschichte in den Daten identifizieren: Die Identifikation der journalistischen Geschichte in den Daten ist nicht klar von Aspekten wie Daten verstehen, analysieren und in einen Kontext setzen abzugrenzen. Es handelt sich hier vermutlich eher um einen Findungsprozess innerhalb des Arbeitsprozesses als um einen einzelnen Schritt. Obwohl er schwer fassbar ist, scheint dieser Aspekt doch zentral zu sein. In Experteninterviews wurden immer wieder Aspekte genannt wie „journalistische Themen finden“, „Recherche-Thesen identifizieren“ oder „gesellschaftlich relevante Schlüsse ziehen“. Bradshaw und Rohumaa (2011) schreiben, wenn man die Daten erst einmal habe, gelte es herauszufinden, ob sich darin eine Geschichte verberge.

VIII) Recherche jenseits der Daten: Dieser Arbeitsschritt ist von dem Punkt „Daten in einen Kontext setzen und interpretieren“ nicht unbedingt zu trennen. Er wird jedoch einzeln aufgeführt, um die journalistische Komponente des datenjournalistischen Arbeitens hervorzuheben. Oft stößt die Analyse von Daten eine Recherche mit klassischen journalistischen Methoden wie Interviews und Befragungen an.

IX) Ergebnisse überprüfen: Wichtiger Teil des datenjournalistischen Arbeitsprozesses ist die Überprüfung der Ergebnisse und Berechnungen. Wann immer möglich solle man als Journalist die eigenen Ergebnisse mit Hilfe eines Statistikers aus dem jeweiligen Fachgebiet überprüfen, schreiben Bradshaw und Rohumaa (2011). Zur Überprüfung kann auch das Hinzuziehen einer zweiten Quelle gehören oder die Überprüfung eines Ergebnisses vor Ort.