Dokumentation einer reproduzierbaren Spickzettel-Prüfpipeline

Johannes Bleher

23. Juni 2026

Abstract. Dieses Dokument beschreibt eine reproduzierbare Pipeline zur Sichtung von eingereichten Spickzettel-Entwürfen. Die Pipeline zählt eingereichte Seiten, standardisiert heterogene Dateiformate, trennt Begründungsseiten von inhaltlichen Spickzettelseiten, misst die Ausnutzung der Seitenfläche, prüft auf identische oder sehr ähnliche Einreichungen und schätzt anschließend die inhaltliche Abdeckung der Vorlesungsthemen mit einem bildfähigen Sprachmodell. Didaktisch wird die Einreichung als früher formativer Lernanlass verstanden: Sechs Wochen vor der Klausur müssen Studierende Inhalte auswählen, strukturieren, verdichten und begründen. Ziel ist nicht die automatische Entscheidung über einzelne Studierende, sondern eine strukturierte, nachvollziehbare und überprüfbare Vorbereitung für eine manuelle Entscheidung. Die Auswertungen in diesem Dokument sind ausschließlich aggregiert und enthalten keine personenbezogenen Einzelfälle.

Schlagwörter: Prüfungsorganisation; reproduzierbare Datenpipeline; bild- und textgestützte Dokumentenanalyse; Texterkennung; Seitenkompression; Themenabdeckung; zeitlich verteiltes Lernen; gemischtes Üben; wünschenswerte Erschwernisse.

JEL-Codes: C80; C88; A22.

1 Motivation

In der Veranstaltung *Einführung in die Statistische Datenanalyse* wurden Studierende gebeten, Entwürfe ihrer handschriftlichen Spickzettel einzureichen. Die Einreichungen waren bewusst offen gehalten: Manche Studierende reichten ein PDF ein, andere einzelne Fotos, ZIP-Archive oder zusätzliche Begründungen für mehr als eine Vorder- und Rückseite. Für eine faire Prüfungsvorbereitung entsteht dadurch ein praktisches Problem: Die Einreichungen sind heterogen, die Flächenzahlen sind nicht unmittelbar vergleichbar, und ein großer Teil der Information liegt als Bild vor.

Die hier dokumentierte Pipeline übersetzt diese heterogenen Einreichungen in eine einheitliche, prüfbare Datenbasis. Sie soll drei Fragen beantworten: Erstens, wie viele inhaltliche Spickzettelseiten wurden eingereicht? Zweitens, wie dicht sind diese Seiten beschrieben? Drittens, wie breit decken die eingereichten Seiten die behandelten Vorlesungsinhalte ab?

Die resultierende Kennzahl ist bewusst als Entscheidungshilfe formuliert. Sie ersetzt nicht die fachliche Prüfung. Insbesondere Seiten mit schlechter Bildqualität, unsicherer Modellzuordnung oder fehlender Themenabdeckung werden für eine manuelle Nachsicht markiert.

2 Didaktisches Konzept

Die Spickzettelaufgabe ist nicht nur eine organisatorische Vorstufe zur Klausur, sondern ein formativer Lernanlass. Die Einreichung erfolgt sechs Wochen vor der Klausur. Dadurch wird die Auseinandersetzung mit dem Stoff aus der Endphase der Prüfungsvorbereitung herausgezogen und zeitlich vorverlagert. Studierende müssen bereits zu diesem Zeitpunkt entscheiden, welche Begriffe, Formeln, Darstellungen und Rechenschritte wichtig genug sind, um auf einem knappen handschriftlichen Artefakt Platz zu finden. Da der Stoff zu diesem Zeitpunkt noch nicht ganz in der Vorlesung behandelt wurde, muss dieser Prozess zwangsläufig gegen Ende der Vorlesung für die Klausur wiederholt werden.

Ausgangspunkt ist eine transparente in der ersten Vorlesung kommunizierte Lehrhypothese: Ein Blatt, vorne und hinten handschriftlich beschrieben, sollte für die Klausurvorbereitung ausreichen. Die Studierenden können dann in einen Verhandlung einsteigen, die datenbasiert geschieht. Von vorne herein ist dabei klar, dass maximal 4 Blatt DIN A4 zugelassen werden. Anders geframt: Die frühe Einreichung gibt Studierenden die Möglichkeit, die Lehrhypothese datengetrieben zu widerlegen. Wer mehr Platz benötigt, liefert mit dem eigenen Entwurf

beobachtbare Evidenz: Anzahl der beschriebenen Flächen, Verdichtung der Inhalte und thematische Breite. Die Auswertung ist deshalb keine versteckte Sanktion, sondern ein gemeinsamer empirischer Prüfrahmen für eine didaktische Ausgangsannahme.

Das didaktische Prinzip lässt sich als VHG beschreiben: *verteilt (zeitlich)*, bewusste *herausfordernd* mit *gemischten* Lehrinhalten. Zeitliche Verteilung bedeutet hier, dass die Spickzettelerstellung nicht erst unmittelbar vor der Klausur stattfindet. Verteilte Übung ist empirisch robust: Lernaktivitäten, die über Zeit gestreckt werden, führen im Durchschnitt zu besserer langfristiger Behaltenleistung als massiertes Lernen kurz vor der Prüfung (Cepeda et al., 2006; Dunlosky et al., 2013). Die frühe Abgabe erzeugt deshalb einen ersten verbindlichen Bearbeitungszeitpunkt. Wenn Studierende ihre Entwürfe später überarbeiten, entsteht zusätzlich eine Wiederaufnahme des Stoffes mit zeitlichem Abstand.

herausfordernd meint, dass die Aufgabe bewusst eine produktive Beschränkung enthält. Ein guter Spickzettel entsteht nicht durch bloßes Abschreiben von Folien. Er verlangt Auswahl, Verdichtung, Priorisierung, Notation und die Entscheidung, welche Rechenschemata ohne ausführliche Erklärung noch verständlich sind. Solche Anforderungen passen zur Idee "wünschenswerten Erschwernissen": Lernbedingungen können kurzfristig anstrengender sein, aber langfristiges Behalten und Transfer unterstützen, wenn die Schwierigkeit bewältigbar bleibt (Bjork, 1994; Soderstrom und Bjork, 2015). Wenn Studierende Inhalte für den Spickzettel ohne bloßes Kopieren rekonstruieren, schließt die Aufgabe zudem an Befunde zur Bedeutung von aktiver Abrufübung an (Karpicke und Roediger, 2008). Die Begrenzung der zugelassenen Blätter macht diese Schwierigkeit sichtbar. Sie hat ausdrücklich didaktische Gründe: Studierende sollen lernen, sich kurz zu fassen, Relevantes von weniger Relevantem zu trennen und ein knappes Hilfsmittel bewusst zu gestalten.

Gemischte Lehrinhalte wird inhaltlich als nicht kapitelweise Vorbereitung verstanden. In ESDA müssen Studierende nicht nur Formeln kennen, sondern erkennen, welche Methode zu welcher Daten- und Fragestellung passt. Gemischte Übung fördert genau diese Unterscheidungsleistung: In mathematischen Aufgaben zeigte sich, dass gemischte Problemformate gegenüber blockierter Übung spätere Testleistung verbessern können (Rohrer und Taylor, 2007); auch beim induktiven Lernen kann das räumlich oder zeitlich getrennte Betrachten verwandter Beispiele hilfreicher sein, als es sich beim Lernen subjektiv anfühlt (Kornell und Bjork, 2008). Das Themenraster der Pipeline setzt diesen Gedanken administrativ um: Nicht die Länge eines einzelnen Kapitels, sondern die Breite über verschiedene Vorlesungsblöcke

geht in die Abdeckungsschätzung ein.

Die Auswertung selbst ist ebenfalls Teil des didaktischen Designs. Die Spickzetteldaten werden mit Methoden analysiert, die in ESDA behandelt werden: Datenbereinigung, Standardisierung heterogener Beobachtungen, Operationalisierung latenter Konzepte, Verteilungen, Verhältniskennzahlen, Streudiagramme, Normalisierung und regelbasierte Entscheidungshilfen. Damit wird die Prüfungsvorbereitung zu einem Datensatz, an dem sichtbar wird, wie statistische Datenanalyse aus unscharfen, realen Beobachtungen eine strukturierte Entscheidungsgrundlage erzeugt.

Damit in der Kohorte ein gewisser Anreiz besteht selbstständig tätig zu werden und daran teilzunehmen einen Spickzettelentwurf abzugeben, wird ein Quorum von 50 Spickzetteln als Grundlage festgelegt ab der überhaupt darüber nachgedacht wird die zugelassene Seitenanzahl der Spickzettel zu überprüfen.

3 Daten und Grundprinzipien

Die Ausgangsdaten bestehen aus Studierendenordnern mit eingereichten Dateien. Die Ordnernamen enthalten pseudonymisierte Kürzel und interne Nummern. Für eine Veröffentlichung oder Vorlesungsdemonstration werden diese Identifikatoren nicht ausgewertet und nicht angezeigt. Alle Abbildungen in diesem Dokument zeigen Aggregatverteilungen.

Die Pipeline folgt vier Grundprinzipien:

1. **Trennung von Inhalt und Begründung.** Begründungsseiten erklären, warum mehr Platz benötigt wird. Sie sollen nicht in die Kompressionsmessung und nicht in den Ähnlichkeitsvergleich eingehen.

2. **Standardisierung vor Bewertung.** Fotos, Scans und PDFs werden zunächst in ein einheitliches Seitenformat überführt. Dadurch beziehen sich spätere Kennzahlen auf vergleichbare Seitenbilder.

3. **Konservative Automatisierung.** Unsichere Fälle werden nicht automatisch entschieden, sondern als manuell zu prüfen markiert.

4. **Nachvollziehbarkeit.** Jede Stufe erzeugt Zwischenergebnisse, so dass spätere Kennzahlen auf vorherige Stufen zurückgeführt werden können.

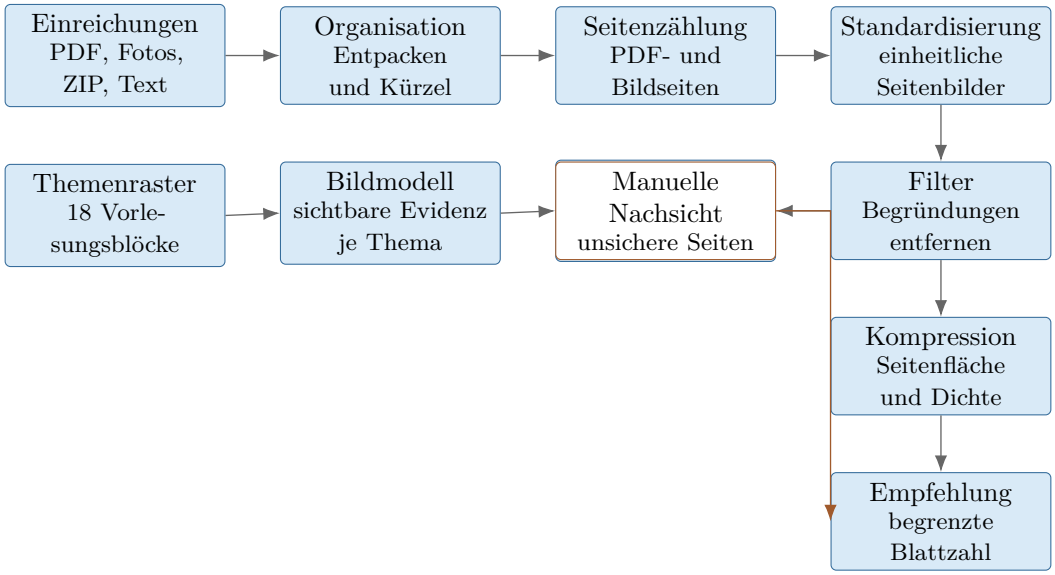
4 Software-Setup

Die Pipeline ist als lokale, nummerierte Auswertung aufgebaut. Die Vorverarbeitung nutzt Python für Dateizählung, Bildstandardisierung, Ähnlichkeitsprüfungen, Texterkennung, Schnittstellenaufrufe und die finale Excel-Ausgabe. Externe Werkzeuge werden dort eingesetzt, wo sie robuste Standardfunktionalität bereitstellen: PDF-Seiten werden gerendert, Bilder werden normalisiert, und die Ergebnisse werden als CSV- und Excel-Dateien gespeichert. Für die inhaltliche Abdeckung wird ein vision-fähiges Sprachmodell über eine OpenAI-kompatible HTTP-Schnittstelle auf dem AIDAHO Server der Uniersität Hohenheim angesprochen.

Die statistische Reportauswertung ist in R implementiert. Sie liest die Zwischenergebnisse der Pipeline, erzeugt die aggregierten Tabellen für dieses Dokument und rendert die Auswertungsgrafiken. Diskrete Häufigkeitsverteilungen werden als Stabdiagramme dargestellt. Kumulierte relative Häufigkeiten werden als empirische Verteilungsfunktionen gezeichnet. Für stetig interpretierte Kennzahlen wie die geschätzte Themenabdeckung werden Kerndichteschätzer verwendet; der Zusammenhang von Abdeckungsrate und Kompressionsrate wird zusätzlich als bivariate Kerndichte gezeigt: einmal als Draufsicht mit Konturen, Flächen und überlagerter Punktwolke, und einmal als dreidimensionale Dichteoberfläche.

Abbildung 1: Ablauf der Spickzettel-Pipeline.

Die Pipeline trennt zunächst Eingabeorganisation, Standardisierung und Messung. Erst danach folgen Ähnlichkeitsprüfung, Themenraster, bildgestützte Abdeckungsschätzung und die finale Seitenempfehlung. Manuelle Nachsicht ist als eigener Prüfpfad eingezeichnet, weil einzelne Modell- oder Bildqualitätsprobleme bewusst nicht automatisch entschieden werden.



5 Messlogik

Die Seitenzählung unterscheidet zwischen allen eingereichten Seiten und den Seiten, die für die eigentliche Spickzettelbewertung verwendet werden. Wenn eine Seite als Begründung erkannt wird, wird sie aus der Kompressionsmessung und aus der Ähnlichkeitsprüfung entfernt. Damit wird verhindert, dass ein zusätzlicher Erklärungstext die Dichte des Spickzettels künstlich senkt oder Ähnlichkeitsprüfungen verzerrt.

Die Kompressionsrate ist auf den am stärksten verdichteten beobachteten Spickzettel normiert. Sie liegt zwischen 0 und 1. Ein Wert nahe 1 bedeutet nicht, dass eine Seite schwarz oder vollständig gefüllt ist. Er bedeutet, dass die Seite im beobachteten Datensatz zu den dichtesten inhaltlichen Spickzettelseiten gehört.

Die inhaltliche Abdeckung basiert auf einem reduzierten Raster aus 18 Vorlesungsblöcken. Das Raster fasst eng verwandte Detailthemen zusammen, etwa Häufigkeiten und empirische Verteilungen, Lage- und Streuungsmaße, Kovarianz/Korrelation, Regression oder Preisindizes. Für jede standardisierte Seite prüft das Bildmodell, ob sichtbare Evidenz für einen Themenblock vorliegt. Als Evidenz zählen etwa Formeln, eindeutig benannte Konzepte, Tabellen, Diagramme oder beschriftete Rechenschemata. Reine Treffer der Texterkennung reichen nicht aus, wenn die Bildseite keine erkennbare fachliche Evidenz zeigt.

Die Pipeline misst zunächst beschriebene DIN-A4-Flächen. Die didaktische Regel wird aber in beidseitig beschreibbaren Blättern formuliert: Ein Blatt entspricht zwei Flächen. Die finale Empfehlung verwendet daher zunächst den Flächenwert

$$\text{erwartete Flächen} = \frac{\text{inhaltliche Flächen} \times \text{Kompressionsrate}}{\text{Abdeckungsrate}}.$$

Die berechnete Blattzahl für eine einzelne Einreichung ist anschließend

$$\text{berechnete Blätter} = \max\left\{1, \min\left\{4, \left\lceil \frac{\text{erwartete Flächen}}{2} \right\rceil\right\}\right\}.$$

Damit nimmt sie exakte diskrete Werte im Bereich von 1 bis 4 Blättern an, also bis zu 8 beschreibbaren DIN-A4-Flächen. Die obere Begrenzung ist nicht rein technisch motiviert, sondern didaktisch: Sie hält den Auftrag knapp und zwingt zur Auswahl. Wenn die Abdeckungsrate null ist, aber inhaltliche Flächen vorhanden sind, wird der Wert auf 4 begrenzt und als Sonderfall markiert. Wenn keine Abdeckungsschätzung vorliegt, bleibt

die finale Empfehlung leer und wird ebenfalls markiert. Einreichungen ohne verwertbaren Inhalt werden nicht aus der Grundgesamtheit entfernt, sondern als Mindestfall mit einem Blatt gezählt. Dadurch bleibt die Auswertung vollständig, ohne leere Abgaben inhaltlich aufzuwerten.

Diese Einzelempfehlungen sind noch nicht mit der endgültigen Zulassungsregel gleichzusetzen. Die zulässige Seitenzahl sollte als aggregierter Wert festgelegt werden, beispielsweise über ein Quantil der berechneten Blattzahlen. In der aktuellen Auswertung ergibt das 70-Prozent-Quantil drei Blätter; das 80-Prozent-Quantil liegt bei vier Blättern.

6 Aggregierte Ergebnisse

Die aktuelle Auswertung enthält 159 Einträge in der finalen Liste. Für 159 Einträge liegt eine modellgestützte Abdeckungsschätzung vor. Insgesamt wurden 751 standardisierte, nicht als Begründung markierte Seiten in der Abdeckungsstufe bewertet.

Abbildung 2: Die meisten ausgewerteten Einreichungen bestehen aus wenigen inhaltlichen Flächen.

Die Abbildung zeigt die Anzahl der inhaltlichen DIN-A4-Flächen je Einreichung nach dem Entfernen von Begründungsseiten. Ein Wert von null bedeutet, dass die Pipeline für diesen Eintrag keine inhaltliche Spickzettelfläche in der standardisierten Auswertung verwendet hat. Da die Flächenzahl diskret ist, wird die Verteilung als Stabdiagramm und nicht als Histogramm dargestellt.

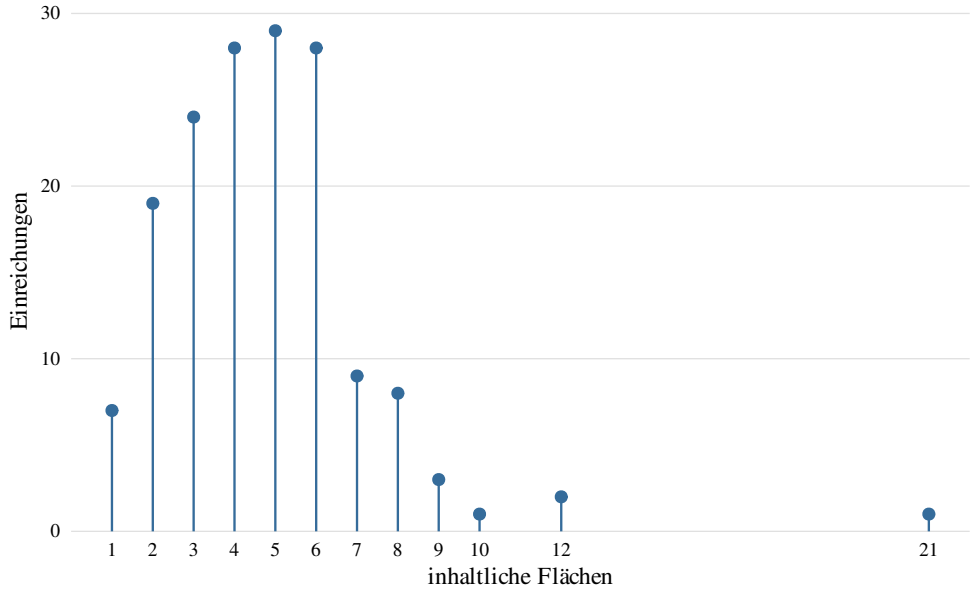


Abbildung 3: Die kumulierte Verteilung der inhaltlichen Flächen ist eine Treppenfunktion.

Die empirische Verteilungsfunktion zeigt die kumulierte relative Häufigkeit der inhaltlichen Flächenzahlen. Die Treppenform ist angemessen, weil die Flächenzahl nur diskrete Werte annehmen kann. Die Sprungstellen sind im Sinne von `stats::ecdf()` markiert.

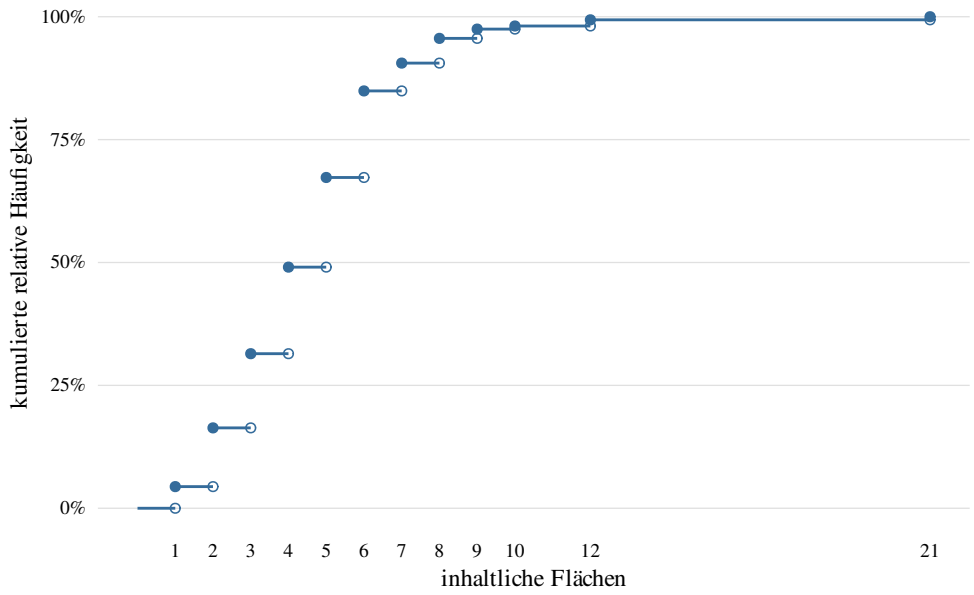


Abbildung 4: Die geschätzte Themenabdeckung konzentriert sich im mittleren Bereich.

Die Abdeckungsrate misst den Anteil der 18 groben Vorlesungsblöcke, für die sichtbare Evidenz auf den eingereichten Spickzettelseiten gefunden wurde. Die Verteilung ist bewusst konservativ: unsichere oder nicht eindeutig lesbare Seiten werden nicht automatisch als abgedeckt gezählt. Da die Abdeckungsrate als stetige Verhältniskennzahl interpretiert wird, wird ihre Verteilung mit einem Kerndichteschätzer dargestellt.

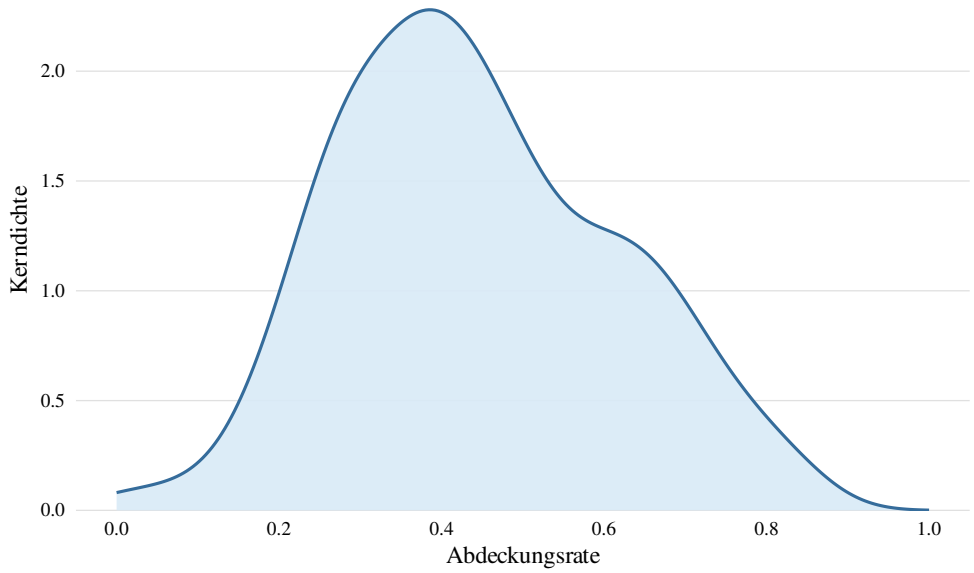


Abbildung 5: Kompression und Themenabdeckung erfassen verschiedene Dimensionen.

Jeder Punkt steht für eine Einreichung mit vorhandener Abdeckungsschätzung. Die Konturen und Flächen zeigen eine bivariate Kerndichteschätzung für Kompressionsrate und Abdeckungsrate in der Draufsicht; die überlagerte Punktwolke lässt die einzelnen Beobachtungen sichtbar.

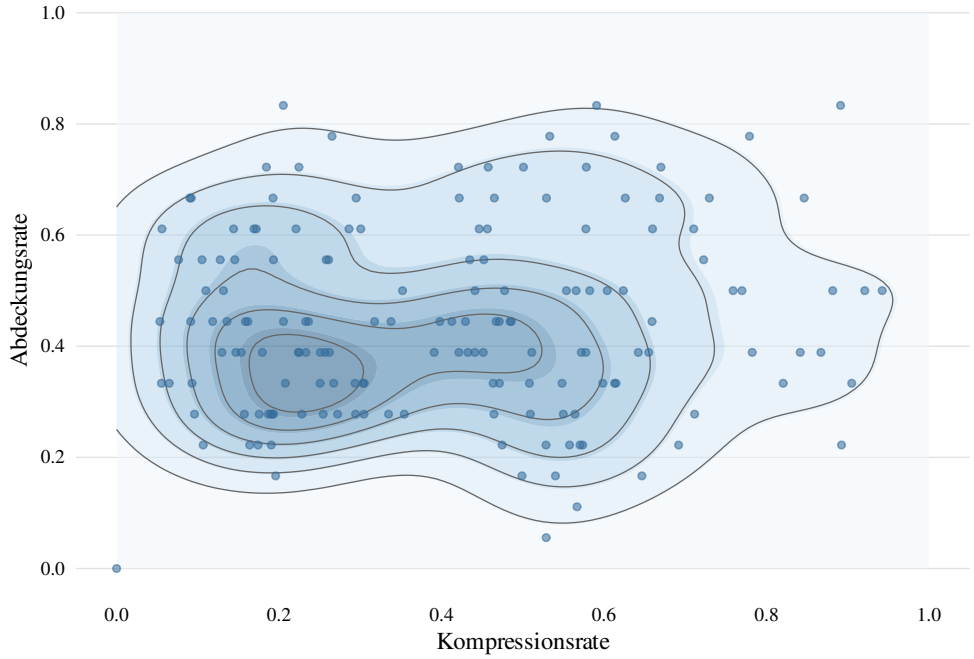


Abbildung 6: Die bivariate Kerndichte kann auch als 3D-Oberfläche gelesen werden.

Die Abbildung zeigt dieselbe gemeinsame Dichte von Kompressionsrate und Abdeckungsrate als dreidimensionale Kerndichteschätzung. Die Höhe der Oberfläche entspricht der geschätzten lokalen Konzentration der Einreichungen.

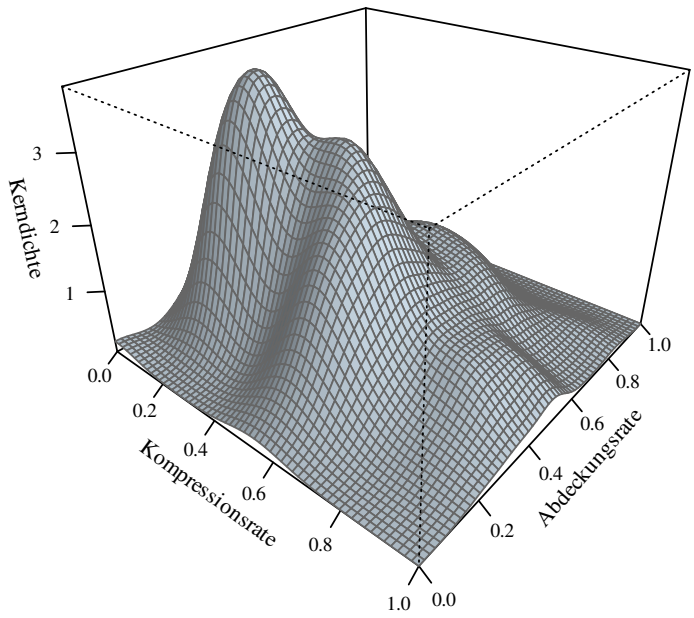


Abbildung 7: Die berechnete Blattzahl nimmt exakte diskrete Werte an.

Die Einzelempfehlung kombiniert inhaltliche Seitenflächen, Kompressionsrate und Abdeckungsrate, rechnet anschließend von Flächen auf beidseitig beschreibbare Blätter um und begrenzt das Ergebnis auf den Bereich von 1 bis 4. Die Verteilung wird daher als Stabdiagramm über den exakten Werten 1, 2, 3 und 4 dargestellt.

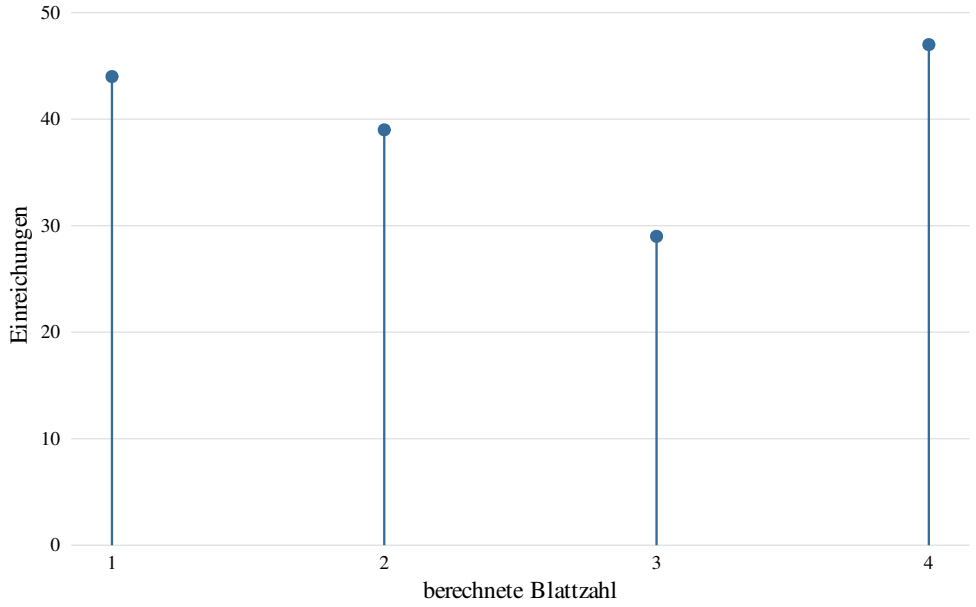
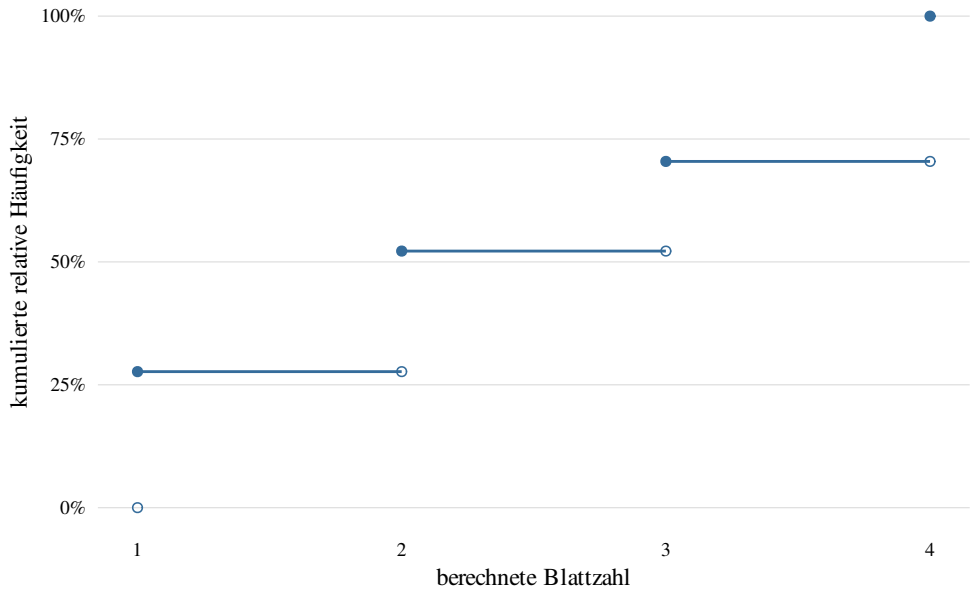


Abbildung 8: Die kumulierte Verteilung der berechneten Blattzahl steigt an vier Stellen.

Die empirische Verteilungsfunktion zeigt die kumulierte relative Häufigkeit der berechneten Blattzahl. Die Sprungstellen sind offen vor und gefüllt nach dem Sprung markiert. Ein aggregierter Beschluss über die zugelassene Blattzahl kann anschließend als Quantil dieser Verteilung formuliert werden.



7 Interpretation und Grenzen

Die Pipeline erzeugt eine strukturierte Datengrundlage, aber keine vollautomatische Prüfungsentscheidung. Dafür gibt es drei Gründe. Erstens sind handschriftliche Fotos nicht immer zuverlässig lesbar. Zweitens kann eine Abdeckungsschätzung bei stark verdichteten Formelsammlungen einzelne Themen übersehen oder zu streng bewerten. Drittens ist die normative Frage, wie viele Seiten zugelassen werden sollen, keine rein technische Frage.

Die Stärke der Pipeline liegt deshalb in der Konsistenz. Jede Einreichung wird nach denselben Regeln gezählt, standardisiert, gefiltert und bewertet. Die manuellen Prüfhinweise machen sichtbar, wo die automatische Auswertung nicht ausreicht. Für die Lehre kann die Pipeline außerdem genutzt werden, um mit Studierenden über Messung, Standardisierung, Bilddaten, Modellunsicherheit und die Übersetzung qualitativer Regeln in quantitative Kennzahlen zu sprechen.

8 Reproduzierbarkeit

Die Auswertung ist in nummerierte Schritte gegliedert. Jeder Schritt liest die Ergebnisse der vorherigen Stufe und schreibt eigene Zwischenergebnisse. Dadurch kann die Pipeline nach einer Änderung an einer Stufe gezielt ab dieser Stufe erneut ausgeführt werden. Die finale Excel-Liste enthält die Berechnungsformeln direkt in den Zellen, so dass die Seitenempfehlung beim Öffnen der Datei nachvollzogen und bei Bedarf manuell angepasst werden kann.

Für eine Veröffentlichung sollten nur aggregierte Tabellen und Abbildungen bereitgestellt werden. Rohdateien, Ordnernamen, interne Kürzel und Einzelfallnotizen können personenbezogene oder prüfungsbezogene Informationen enthalten und gehören nicht in eine öffentliche Dokumentation.

Literatur

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice

in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14(1), 4–58. https://doi.org/10.1177/1529100612453266

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, 319(5865), 966–968. https://doi.org/10.1126/science.1152408

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science*, 19(6), 585–592. https://doi.org/10.1111/j.1467-9280.2008.02127.x

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, 35, 481–498. https://doi.org/10.1007/s11251-007-9015-8

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, 10(2), 176–199. https://doi.org/10.1177/1745691615569000