KI für Alle 2: Verstehen, Bewerten, Reflektieren

Themenblock Clustering: vom Sortieren bis zum Explorieren: 06_02Daten_Fehlerarten

Daten- und Fehlerarten

Erarbeitet von
Dr. Katarina Boland

Lernziele

- Du kannst die Unterschiede zwischen „designed data" und „found data" erklären
- Du kannst Beispiele für „designed data" nennen
- Du kannst Beispiele für „found data" nennen
- Du kannst erklären, welche Verzerrungen in Umfragedaten entstehen können
- Du kannst erklären, welche Verzerrungen in digitalen Verhaltensdaten entstehen können
- Du kannst einordnen, welche Verzerrungen in den beiden Datentypen einander entsprechen

Projekt ai4all „KI für Alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

Inhalt

Einstieg: designed data vs. found data

In den Sozial- und Geisteswissenschaften, aber auch anderen Feldern wie Wirtschaft und Marketing, hat das Erheben von Daten mit eigens entwickelten Experimenten und Instrumenten wie Fragebögen eine lange Tradition.
Einstellungen und Verhaltensweisen können gezielt unter kontrollierten Bedingungen gemessen werden.

Das experimentelle Design ist dabei entscheidend für die Qualität der resultierenden Daten, denn während durch ein geschicktes Design der Einfluss von Störfaktoren minimiert werden kann, kann das künstliche Setting auch zu irreführenden und verzerrten Daten führen. Bei Daten, die in kontrollierten Erhebungen gewonnen werden, spricht man auch von „designed data".

Es gibt aber auch eine andere Art von Daten, die zunehmend für Forschungszwecke an Bedeutung gewinnt: die sogenannten „found data", also Daten, die nicht aus kontrollierten Erhebungen stammen, aber deren Auswertung trotzdem Rückschlüsse auf Einstellungen und Verhaltensweisen liefern kann.

Quelle [1]

Beispiele sind Loggingdaten von Webseitenaufrufen oder Klicks, Likes und Shares von Nachrichten in sozialen Medien oder Verkaufszahlen. Allerdings gibt es auch hier viele mögliche Fehlerarten und Verzerrungen, die sich bei der Auswahl und dem Sammeln der Daten einschleichen können.

Quelle [2]

Umfragedaten vs. digitale Verhaltensdaten

„Digitale Verhaltensdaten" bezeichnet alle Arten von Spuren, die Nutzer*innen bei der Interaktion mit Inhalten im Internet hinterlassen, z. B. Kommentare, die sie in sozialen Medien hinterlassen oder Posts, die sie teilen oder liken.
Bei digitalen Verhaltensdaten muss man sich nicht auf korrektes Erinnern und Wiedergabe durch die untersuchten Personen verlassen und kann Verhalten, Dynamiken und Wechselwirkungen mit realen Ereignissen in Echtzeit untersuchen. Zudem spart man sich die Kosten von großen Befragungen.

Quelle [2]

Umfragedaten und digitale Verhaltensdaten können einen anderen Blickwinkel auf Ereignisse bieten, z. B. bilden digitale Verhaltensdaten besser kurzzeitige Fluktuationen in Verhalten oder Einstellungen von Personen ab als Umfragen, die in längeren Zeitabständen durchgeführt werden. In Umfragen können die Gründe für Verhalten direkt erfragt werden,

dafür ist Verhalten nicht direkt, sondern nur über Selbstauskünfte messbar. Bei digitalen Verhaltensdaten verhält es sich genau andersherum. Hier kann Verhalten beobachtet, aber auf zugrundeliegende Beweggründe nur geschlossen werden.

Quelle [1]

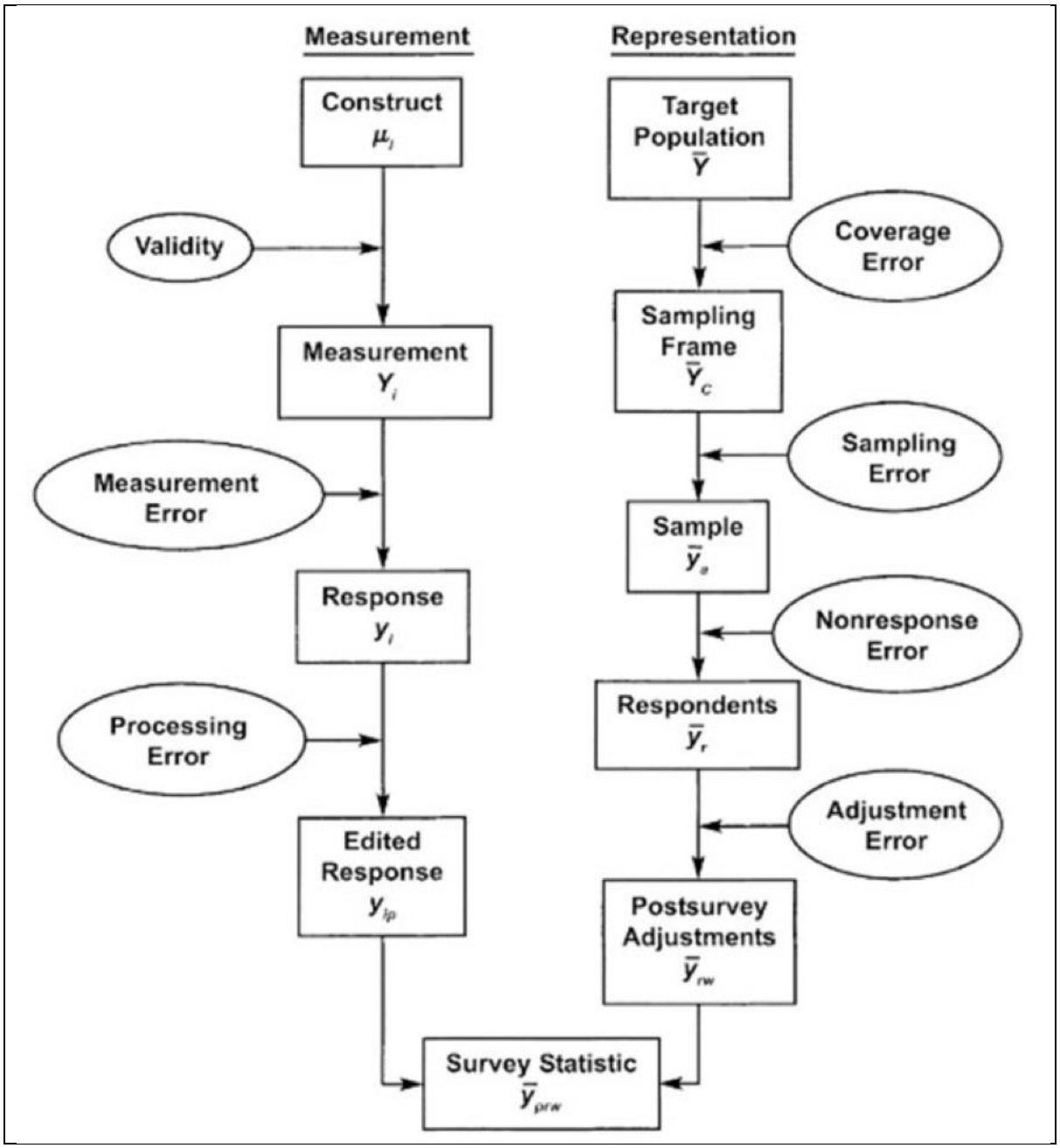
Beide Datenarten sollten also als komplementär und nicht als besser oder schlechtere Alternative angesehen werden.

Im Folgenden schauen wir uns statistische Fehler an, die in Umfragedaten und in digitalen Verhaltensdaten enthalten sein können, und wie diese zu verzerrten Ergebnissen führen können.

Mögliche Verzerrungen bei der Erhebung von Umfragedaten

Eine bekannte Systematik zur Beschreibung von statistischen Fehlern in Umfragen ist das „Total Error Framework''.

Quelle [3]



Total Error Framework. Quelle: [2]

Das Total Error Framework unterscheidet zwei Arten von Fehlern: Messfehler (Measurement Errors) und Repräsentationsfehler (Representation Errors). Wir gehen hier nur auf die Fehler in Bezug auf die Datenerhebung ein, also die Fehler, die nach der Erhebung in den Daten zu finden sind, nicht jene, die bei der Verarbeitung der Daten entstehen können.

Zu Messfehlern kommt es, wenn die Messinstrumente, also hier die Fragen im Fragebogen, nicht das messen, was sie sollen oder ihre Ergebnisse nicht korrekt aufgezeichnet oder übersetzt werden. Der beobachtete oder aufgezeichnete Wert weicht also vom wahren Wert ab. Dies kann auch passieren, wenn Befragte die Unwahrheit sagen, sei es absichtlich oder unabsichtlich.

Repräsentationsfehler beziehen sich auf die Generalisierbarkeit oder externe Validität einer Erhebung. Wenn diese die Gruppe der befragten Personen adäquat repräsentiert, nicht aber die zu untersuchende Grundgesamtheit, sind keine Rückschlüsse auf die Grundgesamtheit zulässig und die Studienergebnisse können nicht über die Gruppe der Befragten hinaus verallgemeinert werden.

Die Grundgesamtheit ist jene Gruppe, über die mithilfe der Untersuchung eine Aussage getroffen werden soll, z. B. wahlberechtigte Personen in einem Land.
Zunächst muss eine Möglichkeit gefunden werden, diese Gruppe bzw. ihre Mitglieder zu identifizieren bzw. eine bestmögliche Annäherung zu erreichen. Hier kommen beispielsweise Wählerverzeichnisse, Melderegister und Telefonverzeichnisse in Frage. Dies nennt man den Sampling Frame.
Da es meist nicht möglich ist, 100 % der Personen im Sampling Frame zu befragen, muss eine Stichprobe gezogen werden. Beispielsweise eine Zufallsauswahl von 10 % des Sampling Frames.

In beiden Schritten kann es zu Fehlern kommen: der Sampling Frame kann bereits einige Mitglieder unter- oder überrepräsentieren. Dies nennt man Coverage Error. Waren beispielsweise früher Telefone vor allem in einkommensstarken Haushalten zu finden, konnten einkommensschwache Haushalte in Telefonverzeichnissen unterrepräsentiert sein.

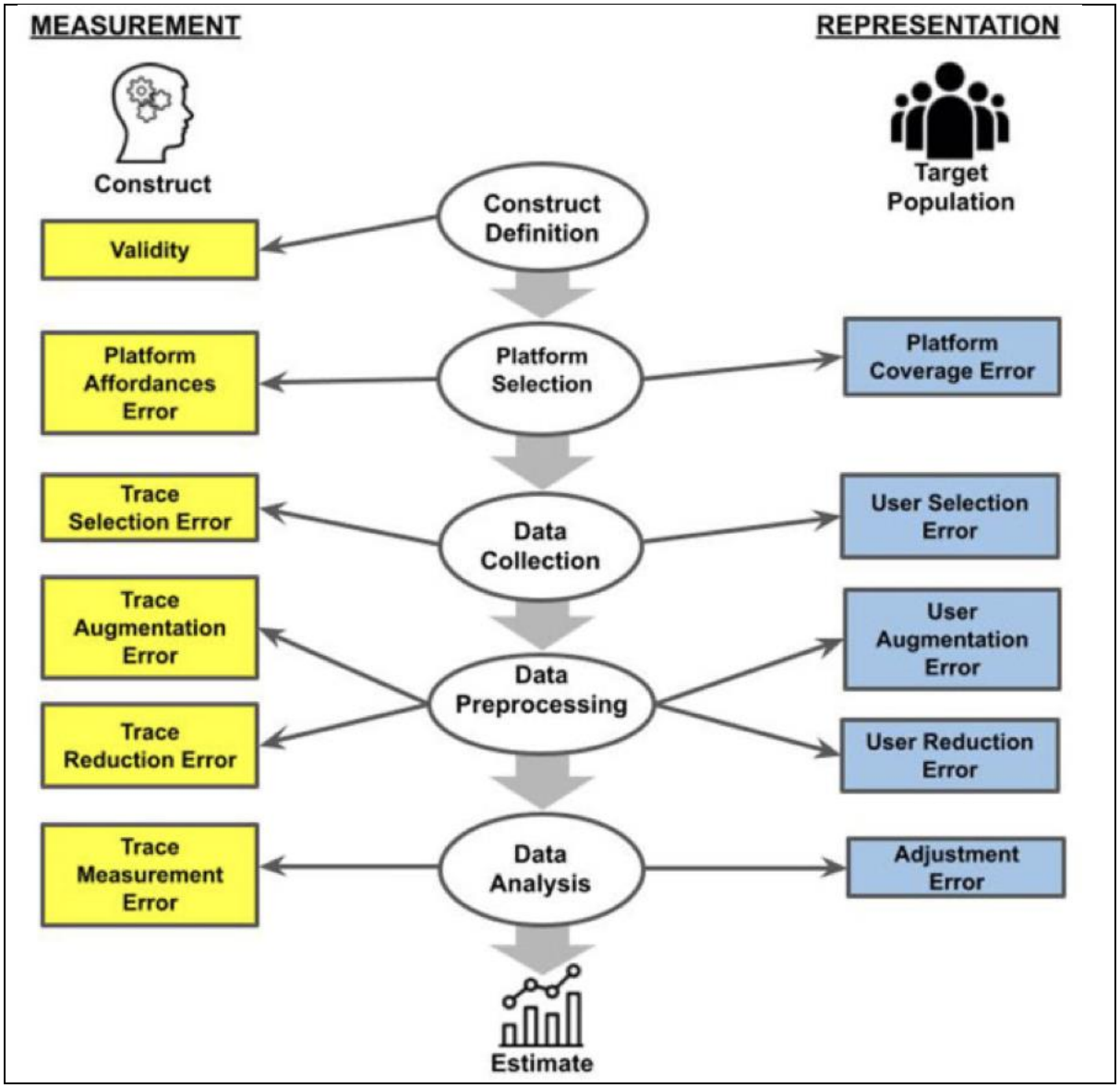
Auch die Stichprobenziehung kann weitere Fehler mit sich bringen, wenn der Auswahlprozess dazu führt, dass bestimmte Gruppen mit höherer Wahrscheinlichkeit inkludiert werden als andere, die Ziehung also zum Beispiel aus praktischen Gründen nicht randomisiert ist.

Auch die Befragten selbst können Verzerrungen auslösen: Nicht jede Person, die zu einer Befragung eingeladen wird, nimmt auch an dieser teil. Dies führt dann zu Problemen, wenn die Entscheidung zur Teilnahme von Faktoren abhängig ist, die mit dem Untersuchungsgegenstand in Verbindung stehen und es somit systematische Unterschiede zwischen den Personen gibt, für die Antworten vorliegen und denen, für die keine Antworten vorliegen. Sind Personen beispielsweise politisch interessiert und engagiert, so werden sie vielleicht mit höherer Wahrscheinlichkeit an einer freiwilligen Befragung zu politischen Fragestellungen teilnehmen als andere Personen. Genauso kann es auch vorkommen, dass Personen an einer Befragung teilnehmen, allerdings einzelne Fragen nicht beantworten. Auch dieser Nonresponse Error führt zu einer eingeschränkten Generalisierbarkeit der Ergebnisse, weil die Antworten nicht repräsentativ für die Grundgesamtheit sind.

Mögliche Verzerrungen bei der Auswahl von digitalen Verhaltensdaten

Für digitale Verhaltensdaten müssen wir ähnliche Fehlerquellen beachten, allerdings gibt es hier einige Unterschiede. Diese erklären wir anhand eines Ausschnitts des „TED-On" (Total Error Framework for Digital Traces of Human Behavior on Online Platforms) Frameworks.

Quelle [2]



TED-On Framework. Quelle: [2]

Anders als bei designed data müssen keine Messinstrumente entworfen werden, wenn digitale Verhaltensdaten als found data, also ohne eigens erstelltes Experiment, genutzt werden. Trotzdem muss sichergestellt werden, dass das beobachtete Verhalten tatsächlich ein Indikator ist für das Konstrukt, das untersucht werden soll. Nur dann lassen sich aus den Befunden gültige, auch genannt valide, Schlussfolgerungen ziehen.

Quelle [5]

Wenn Personen beispielsweise in sozialen Medien positiv über eine Politikerin reden, bedeutet dies nicht zwangsläufig, dass sie diese auch wählen würden. Das Messen von Sentimenten in Bezug auf Politiker*innen wäre also allein kein gutes Instrument zur

Vorhersage von Wahlverhalten. Eine Umfrage, die Sentimente als alleinige Indikatoren für individuelles Wahlverhalten verwendet, besitzt also eine geringe Validität.

Welche Indikatoren verwendet werden, wird bei found data oft erst bei Exploration der verfügbaren Daten entschieden.

Im nächsten Schritt müssen geeignete Plattformen ausgesucht werden, deren Nutzer*innen man analysieren möchte. Diese sollten, wie bei Umfragedaten, eine möglichst gute Annäherung an die zu untersuchende Grundgesamtheit darstellen. Ein Platform Coverage Error steht analog zum Coverage Error, wenn die Plattformnutzer*innen nicht repräsentativ für die Grundgesamtheit sind. Wählt man beispielsweise die ehemalige Plattform Twitter, sollte man die Demografie der Twitter-Nutzenden beachten, die auch noch regional unterschiedlich war. Beispielsweise waren Twitter-Nutzer*innen in Großbritannien im Schnitt jünger, besser gebildet und liberaler eingestellt als die Gesamtbevölkerung.

Quelle [4]

Im digitalen Raum kann das Verhalten von Personen auch von der ausgewählten Plattform beeinflusst werden. Zum Beispiel gibt Twitter eine Höchstanzahl von Zeichen vor, und diese änderte sich im Laufe der Zeit, und Facebook beeinflusst die Bildung von sozialen Netzwerken durch das Vorschlagen von potentiellen Freund*innen. Diese Art von Fehler wird „Platform Affordances" Fehler genannt und entspricht einem Measurement Error in Umfragen, wo beispielsweise die Antwortskala das Antwortverhalten verzerren kann.

Nach der Wahl der Plattform müssen, analog zur Ziehung einer Stichprobe, individuelle Datenpunkte ausgewählt werden. Hier kann es zu User Selection und Trace Selection Fehlern kommen. Diese Fehler bedeuten, dass relevante Datenpunkte fehlen oder irrelevante Datenpunkte in der Stichprobe enthalten sind.

Angenommen, du möchtest Debatten über eine Politikerin analysieren und hast dich für Twitter als Plattform entschieden. Wenn du nun alle Tweets heraussuchst, die ihren natürlichen Namen enthalten, verpasst du jene, die sie mit ihrem Amt oder nur mit ihrem Nachnamen referenzieren. Dies wäre ein Trace Selection Fehler.

Es könnte auch passieren, dass verschiedene Bevölkerungsgruppen verschiedene Begriffe verwenden, zum Beispiel Jugendliche einen bestimmten Spitznamen. Wenn du diesen nicht mit einbeziehst, verlierst du die Daten dieser Gruppe.

Das Filtern mit bestimmten Begriffen kann so auch zu einem User Selection Fehler führen, durch den Bevölkerungsgruppen unter- oder überrepräsentiert werden. Dieser Fehler kann auch entstehen, wenn der Sampling Frame durch explizites Hinzufügen oder Entfernen von Daten bestimmter Nutzer*innen gebildet wird, beispielsweise unter Verwendung von Profilinformationen, die unvollständig oder inkorrekt sein können.
Auch Plattformbeschränkungen können zu Trace oder User Selection Fehlern führen, wenn die Plattform beispielsweise nur den Zugriff auf bestimmte Daten erlaubt und diese keine

randomisierte Stichprobe darstellen. Dies entspräche einem Sampling Error im Total Error Framework.

Abschluss

Du siehst also, dass die Erhebung von Daten komplex und in der Tat ein eigenes Forschungsfeld an sich ist. Selbst wenn du nicht eigene Daten erhebst, sondern Daten nachnutzt, die andere erhoben haben, solltest du genau prüfen, wie diese entstanden sind, ob und wie weit sie zur Untersuchung deiner Fragestellung geeignet sind und welche Einschränkungen du ggf. bei der Interpretation deiner Ergebnisse beachten und auch dokumentieren musst.

Umfragedaten und digitale Verhaltensdaten haben ihre eigenen Charakteristika und Vor- und Nachteile. Keine Datenart ist dabei in Gänze der anderen überlegen und teilweise bilden sie schlicht andere Informationen ab. Du solltest also immer überlegen, welche Datenart für deine Fragestellung die beste ist oder ggf. wie du beide miteinander kombinieren kannst.

Quellen

Quelle [1] Beuthner, C., Breuer, J., & Jünger, S. (2021). Data Linking—Linking survey data with geospatial, social media, and sensor data (GESIS Survey Guidelines) (Version 1.0). GESIS - Leibniz Institute for the Social Sciences. https://doi.org/10.15465/GESIS-SG_EN_039

Quelle [2] Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). A Total Error Framework for Digital Traces of Human Behavior on Online Platforms. Public Opinion Quarterly, 85(S1), 399–422. https://doi.org/10.1093/poq/nfab018

Quelle [3] Groves, R. M., & Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. Public Opinion Quarterly, 74(5), 849–879. https://doi.org/10.1093/poq/nfq065

Quelle [4] Mellon, J., & Prosser, C. (2017). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. Research & Politics, 4(3), 2053168017720008. https://doi.org/10.1177/2053168017720008

Quelle [5] Validität im Dorsch Lexikon der Psychologie. (2021). https://dorsch.hogrefe.com/stichwort/validitaet

Disclaimer

Transkript zu dem Video „06 Clustering: vom Sortieren bis zum Explorieren: Daten- und Fehlerarten“, Katarina Boland.