Deskription2

Contents

1 Einführung

Outcome des Projekts *Surveyqualität* ist die Qualität von Fragebögen, welche sich aus der Reliabilität und Validität ($q^2 = r^2 * v^2$) zusammensetzt. Diese wurden durch ein Strukturgleichungsmodell *Multitrait-Multimethod* (MTMM) berechnet, vor allem mithilfe von großen Befragungen des *European Social Surveys* (ESS).

Die Qualität, sowie die Reliabilität und Validität sind hierbei stetige Merkmale, welche im Bereich $[0; 1]$ liegen.

Um die Qualität von neuen Surveyfragen zu prognostizieren wurden über die Zeit mehrere Modelle berechnet, *Survey Quality Predictor* (SQP). Die erste Version dieser Vorhersage wurde mittels linearer Regression berechnet. In der zweiten Version mittels *random forest*. Die dritte Version wiederum mittels *random forest* (Schweisthal).

Um bei der Erstellung neuer Fragebögen den Forschern unter die Arme zu greifen, soll nun ein Regressionsmodell berechnet werden, damit Aussagen wie "Falls Sie eine Einleitung zu Ihrer Frage hinzufügen erhöht sich die Qualität um xy". Also ein interpretierbares Modell.

2 Deskription

Im folgenden ein kurzer Überblick über alle Parameter, welche sich im Datensatz befinden und relevant zur Berechnung sind.

2.1 Gesamter Datensatz

Der Datensatz besteht aus **6074 Beobachtungen** (mit NAs: 6074), **60** zu verwendende **Variablen** (Outcome + Einfluss), wobei **42** Einflussvariablen **Nominal- und Ordinal** und **16** Einflussvariablen **Metrisch** skalliert sind.

Table 1: Variablen im gesamten Datensatz nach Skalenniveau

Datensatz	Anzahl
Gesamt	60
Nominal- / Ordinal	42
Metrisch	16

2.2 Nominal -/ Ordinalskallierte Einflussgrößen

Von den 42 Nominal -/ Ordinalskallierten Einflussgrößen sind die meisten **binär** kodiert (**29**), fast alle mit **weniger als 10 Kategorien** (**40**) und **2** mit **mehr als 10 Kategorien**.

Table 2: Häufigkeit von Nominal -/ Ordinalskallierten Variablen im Datensatz

Anzahl an Kategorien	Häufigkeit im Datensatz
2	29
3	5
4	3
5	1
6	2
11	1
29	1

Der Einfluss der Spalte mit **29** Kategorien (**Sprache**) ist von großer Wichtigkeit, da dieser verwendet werden soll, um ***random intercepts*** zu implementieren (hierarchische Struktur soll beachtet werden: Studien genestet in Experimenten in Ländern / Sprache).

Folgende Histogramme, sollen einen Überblick über die Datenlage liefern.

2.3 Filter

Im folgenden Abschnitt gehe ich etwas näher auf die Filter ein. Die Grafiken sind aus *Codebook Routing* entstanden, in dem beschrieben wird, in was für einer Reihenfolge einzelne Fragebögen bewertet werden sollten.

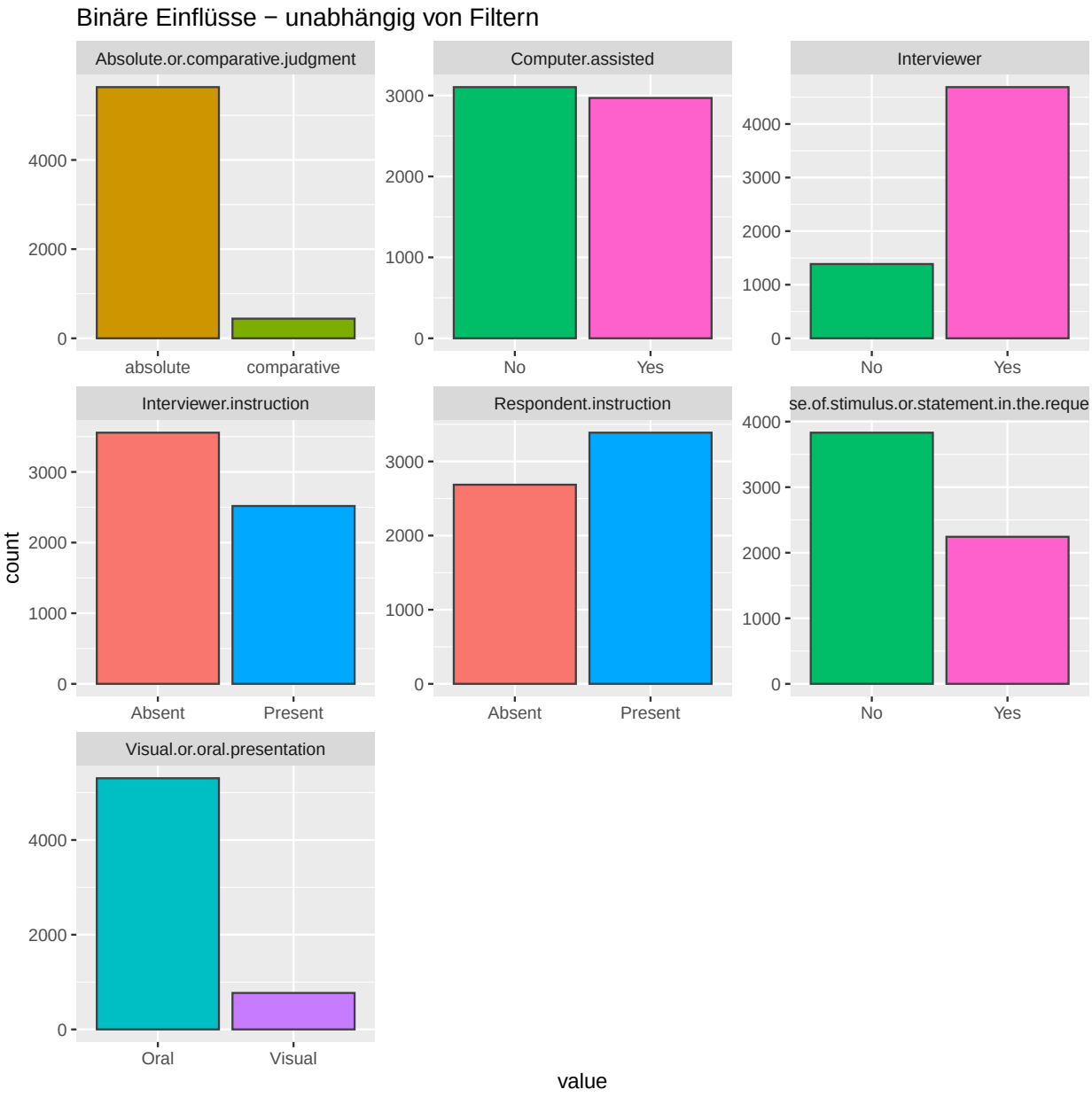


Figure 1: Binäre Einlussgrößen, welche nicht in Filtern vorkommen

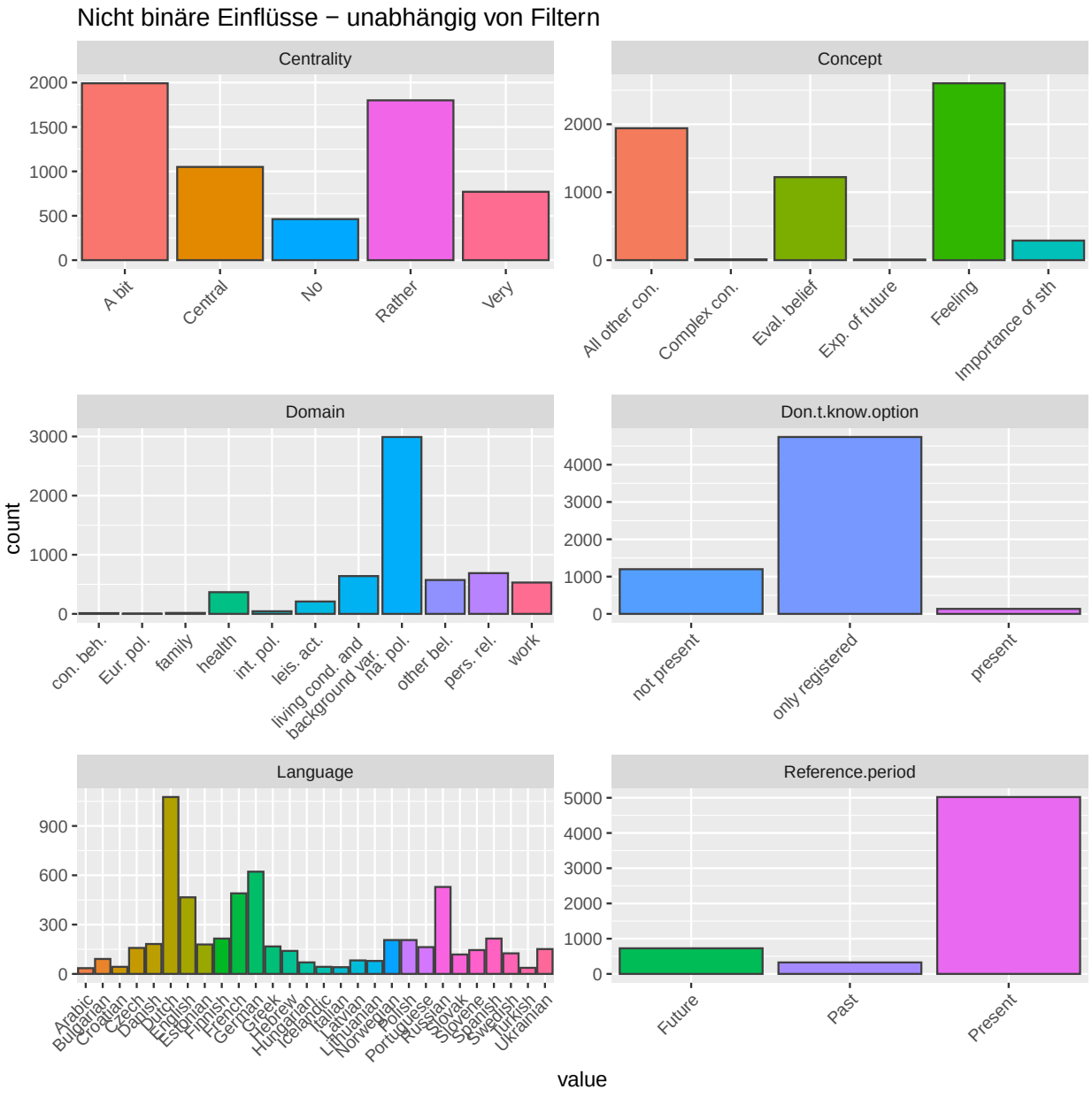
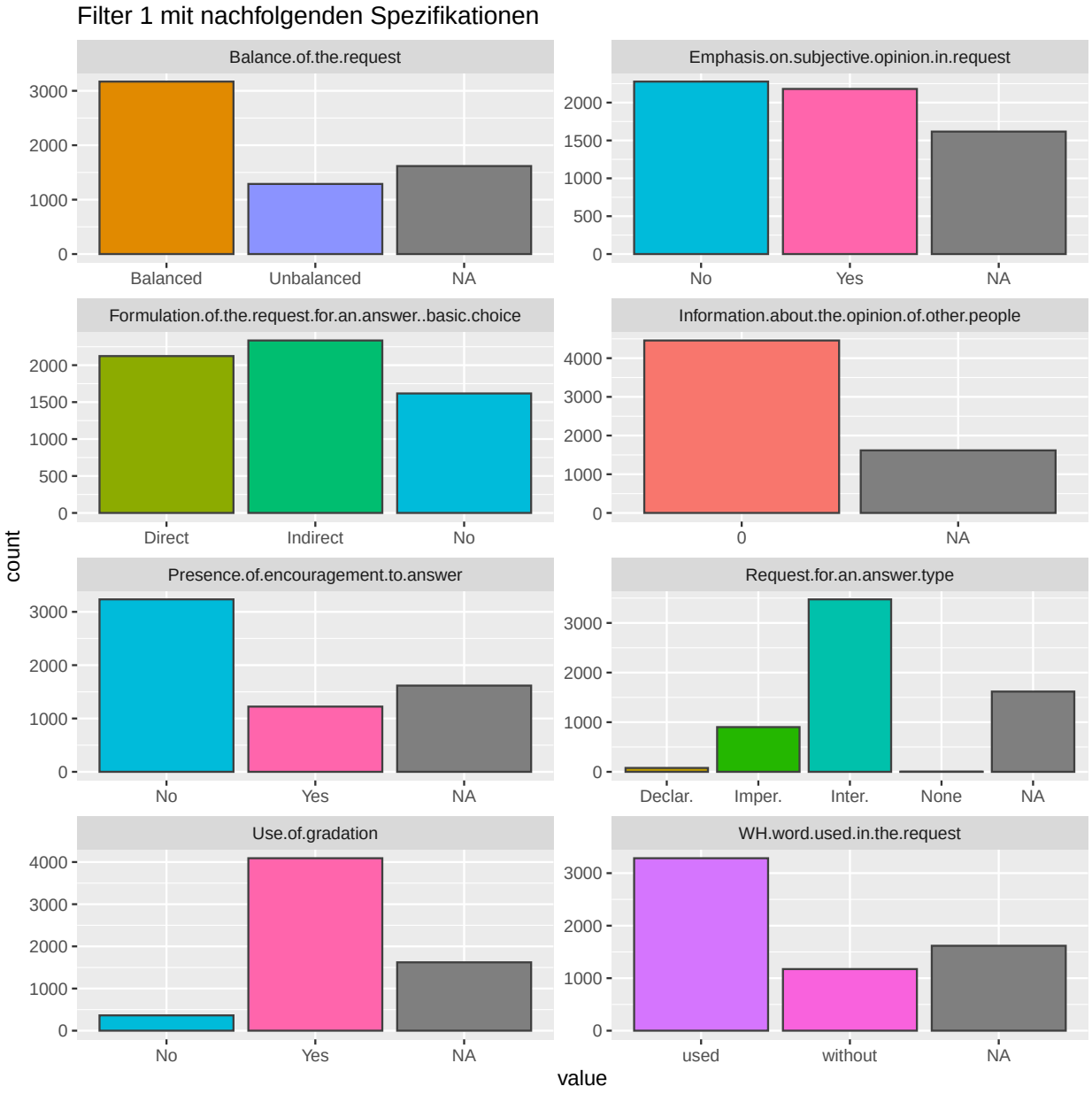


Figure 2: Nicht binäre Einlussgrößen, welche nicht in Filtern vorkommen

Figure 3: Übersicht über codebook routing mit eingezeichneten Filtern

2.3.1 Erster Filter

Der erste Filter ist **Formulation of the request for an answer basic choice**. Je nachdem ob ein Frageitem die Charakteristik **indirect request**, **direct request** oder **no request** aufweist, werden eventuell weitere Variablen ausgewertet. Für Items ohne Aufforderung werden einige Variablen übersprungen.



2.3.2 Zweiter Filter

Der zweite Filter ist **Response scale basic choice**. Je nachdem, ob das Item die Charakteristik **More than 2 categories scalec**, **More steps procedures**, **Magnitude estimation** oder **Line production** aufweist, werden weitere Variablen zu **response scales** abgefragt.

Bei diesem Filter handelt es sich jedoch um einen zwei-schichtigen Filter. Er teilt sich bei der Variable **Theoretical range of concept bipolar / unipolar** nochmals auf. Falls ein Item die Charakteristik **Theoretically unipolar** aufweist, werden die drei Variablen **Range of the used scale bipolar/unipolar**, **Symmetry of response scale** und **Neutral category** übersprungen.

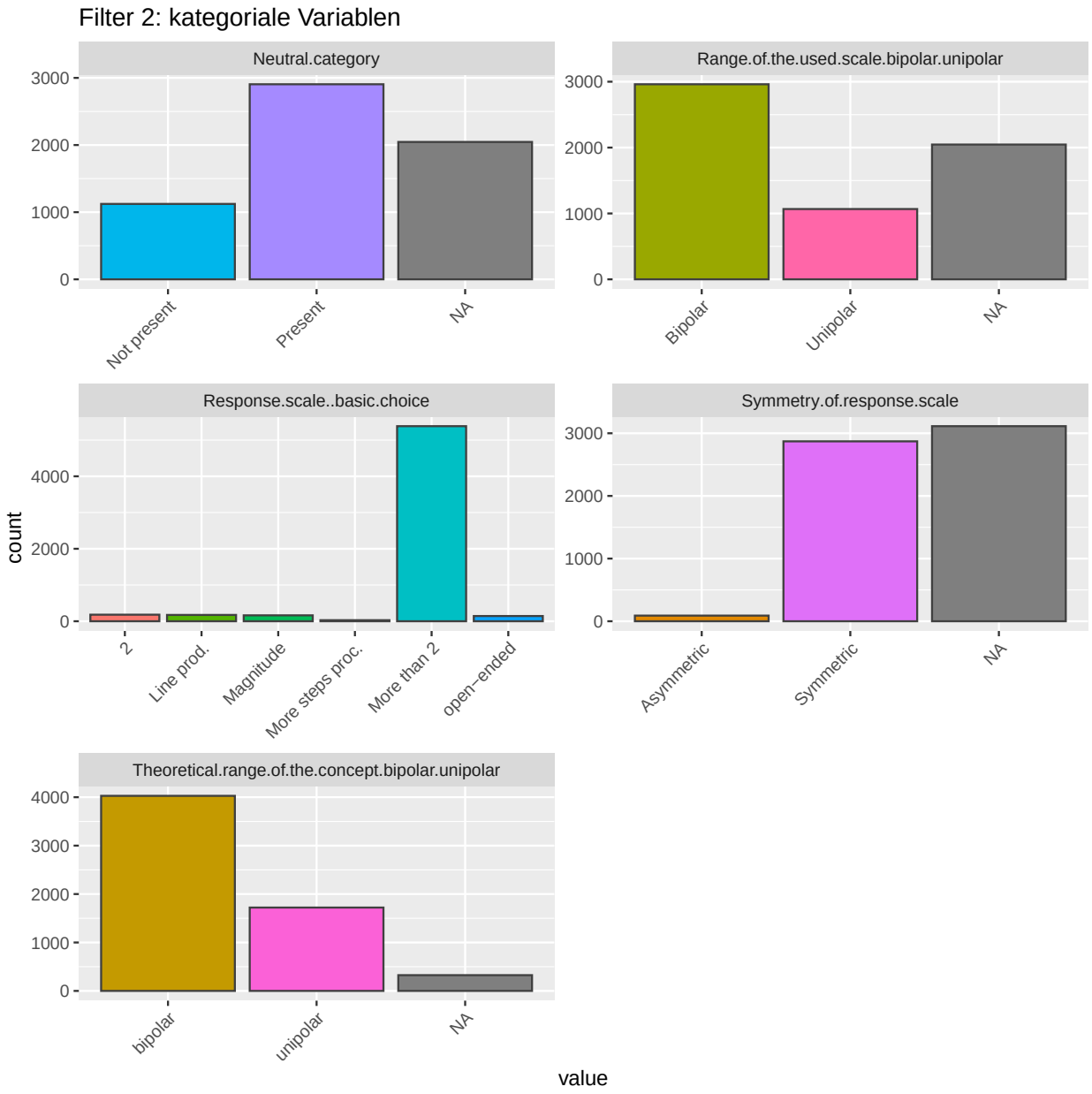


Figure 4: Variablen innerhalb des zweiten Filters: kategoriale Variablen

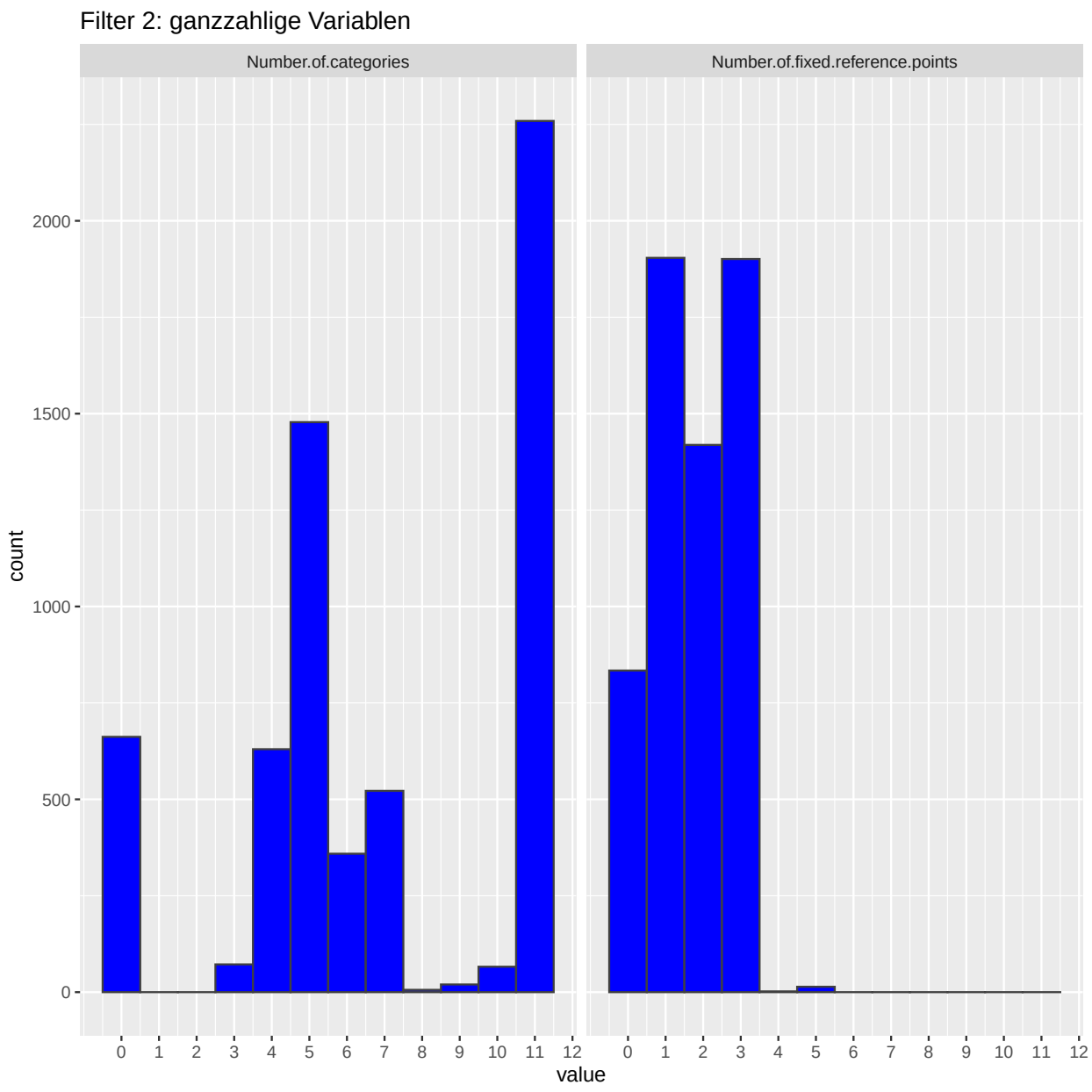
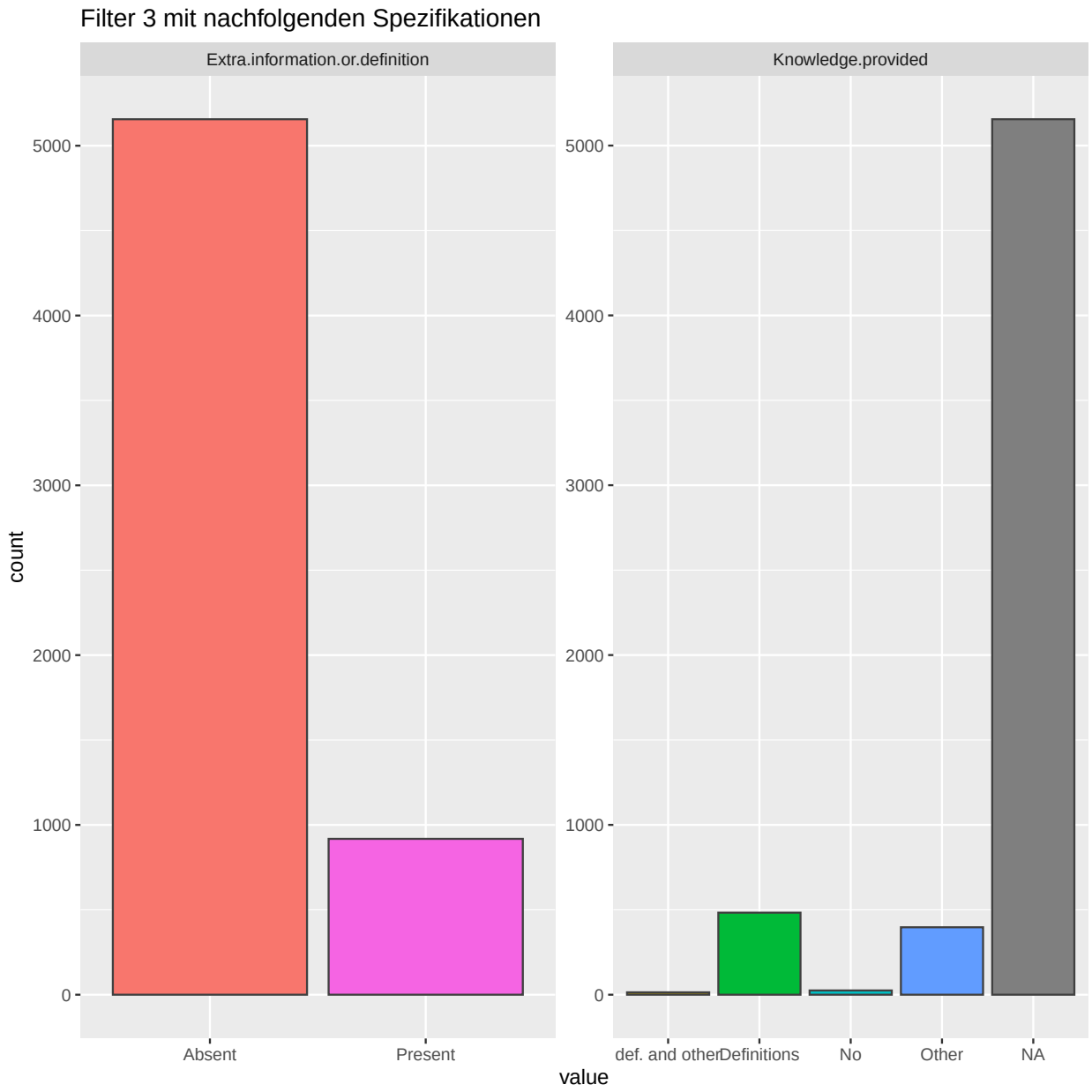


Figure 5: Variablen innerhalb des zweiten Filters: ganzzahlige Variablen

2.3.3 Dritter Filter

Der dritte Filtervariable lautet **Extra information or definition**. Nur für Items mit dem Label **Present Extra information**, wird mittels der Variable **Knowledge provided**, die Art der extra-Information erfasst.



2.3.4 Vierter Filter

Der vierte Filter lautet **Introduction available**. Nur für Items mit dem Label **Available**, gibt es mehrere Variablen, welche die Merkmale der Einleitung erfassen.

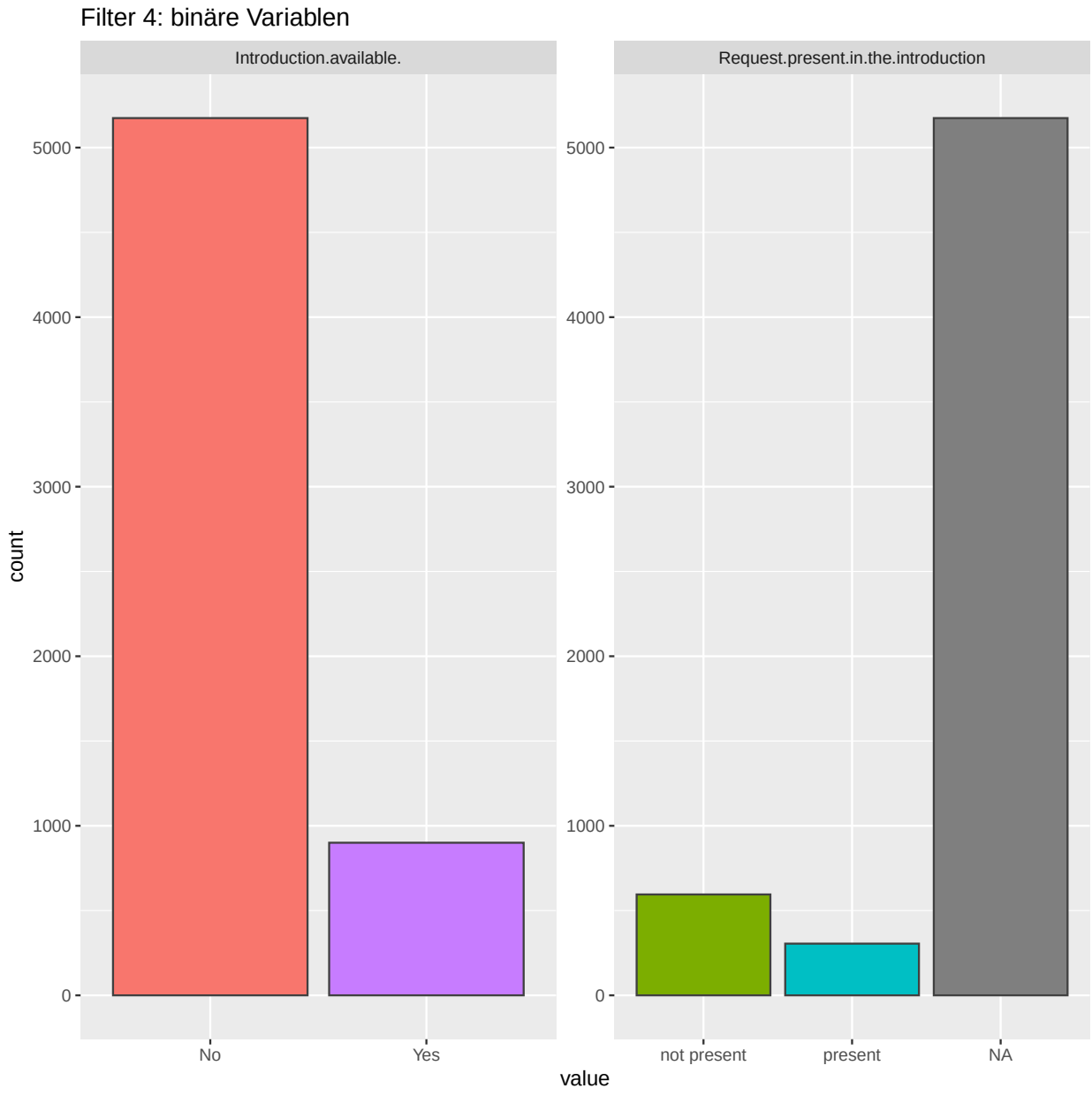


Figure 6: Variablen innerhalb des vierten Filters: binäre Variablen

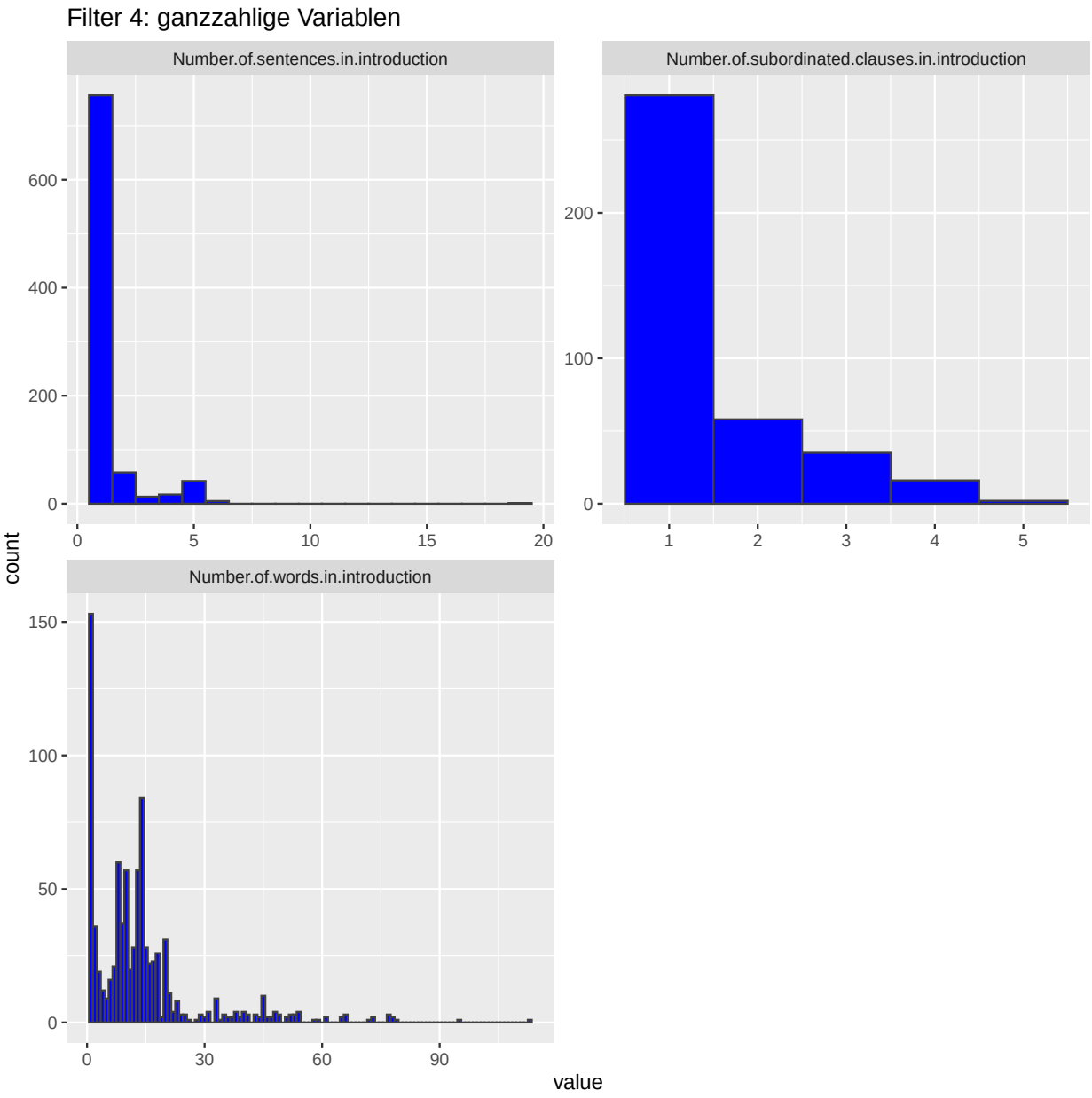
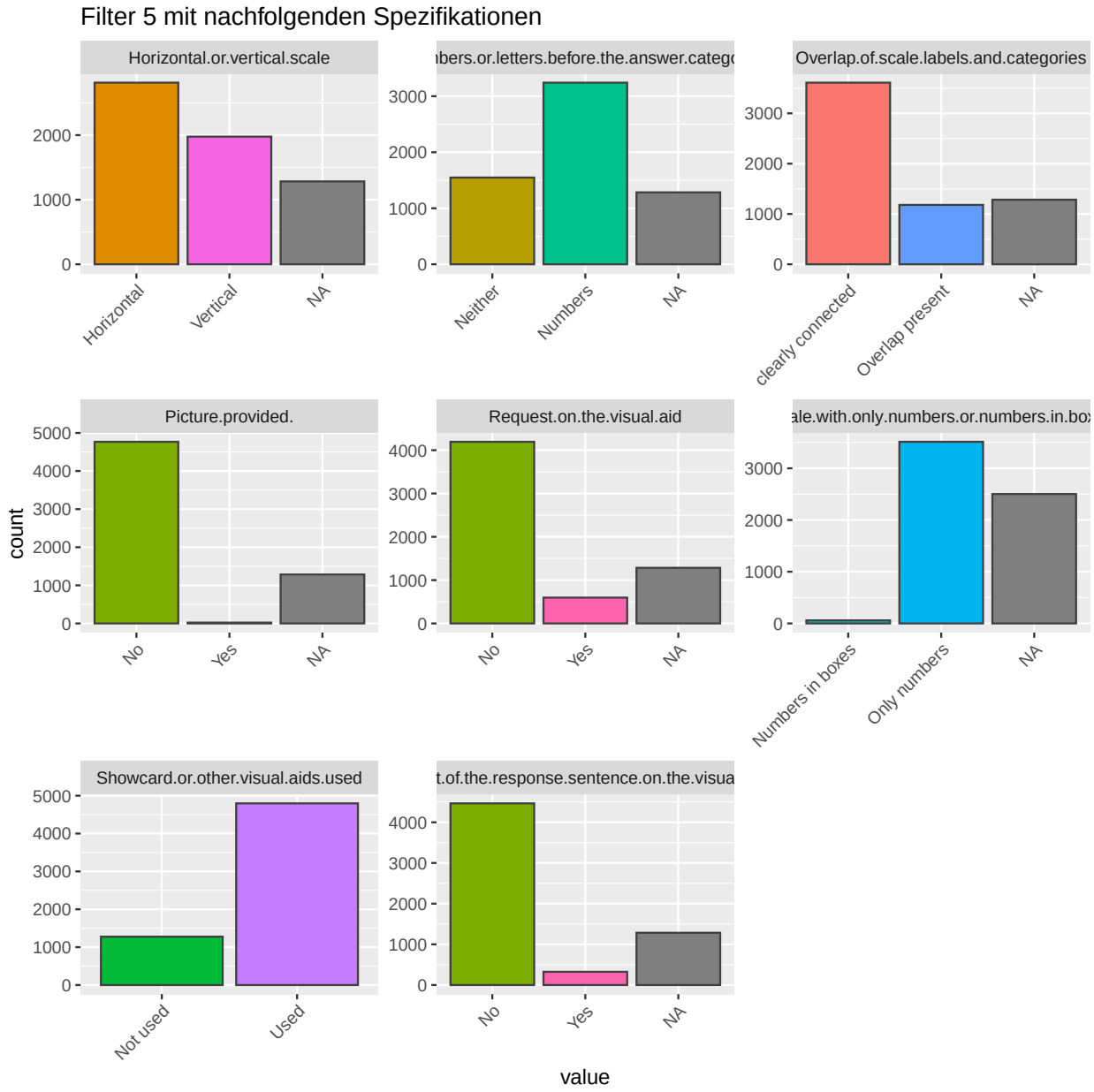


Figure 7: Variablen innerhalb des vierten Filters: ganzzahlige Variablen

2.3.5 Fünfter Filter

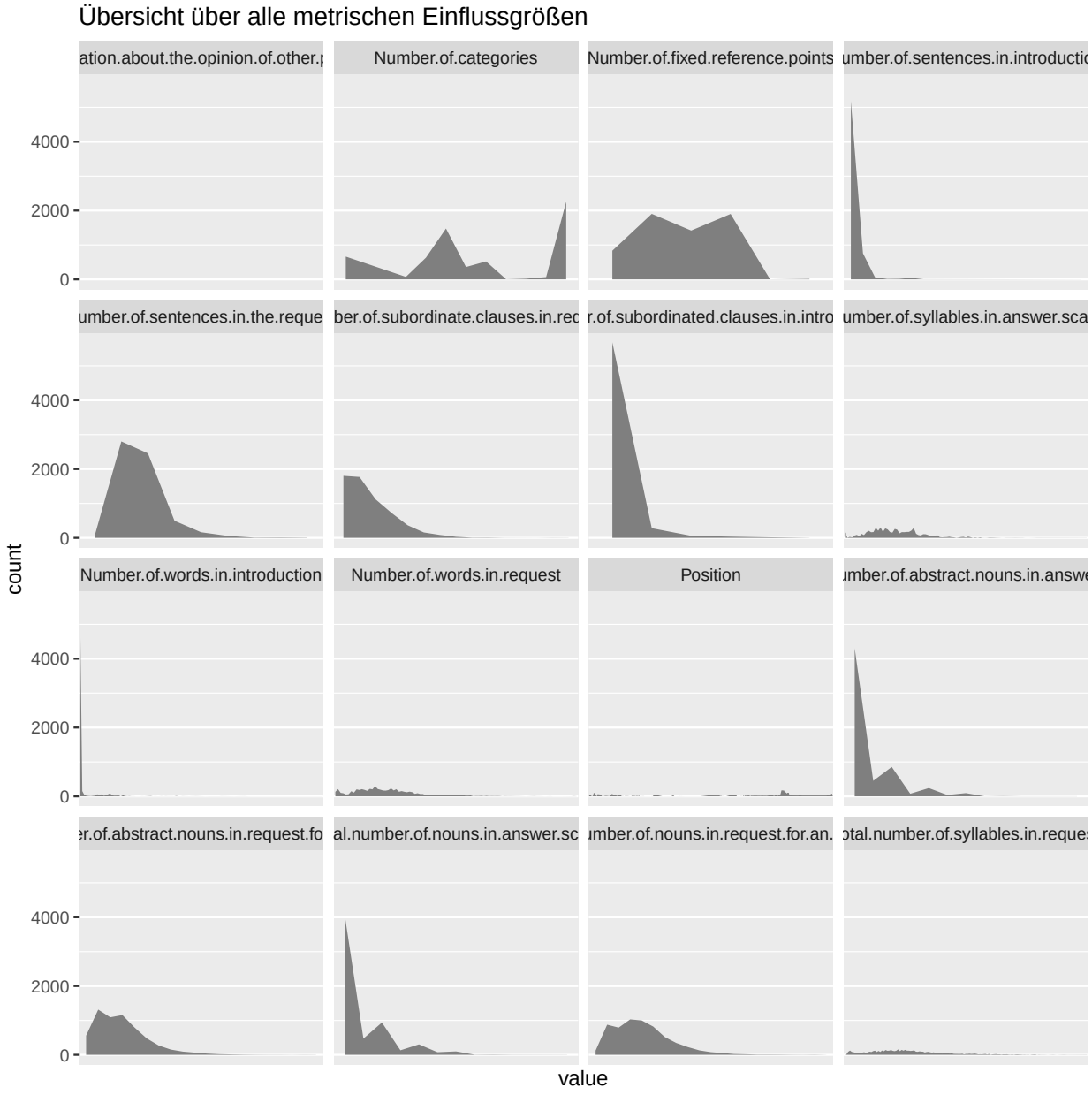
Der fünfte Filter lautet **Showcard or other visual aids**. Nur für Items mit dem Label **Used showcard** werden mehrere Variablen zur Art und Weise der benutzten visuellen Hilfsmittel erfasst.

Ähnlich zum zweiten Filter, teilt dieser sich mit Hilfe der Variable **Numbers or letters before the answer categories** mit den Labels **numbers** oder **letters** auf.



2.4 Metrisch Skallierte Einflussgrößen

Im Gegensatz zu den Nominal -/ Ordinalskallierten Einflussgrößen gibt es wenige **metrischskallierte** Einflüsse (**Anteil** von **0.27**). Diese sind ausschließlich diskrete Zählvariablen.



2.5 Outcome: Qualität

Die Qualität setzt sich aus dem Produkt von Reliabilität und Validität zusammen. In folgendem Scatterplot lässt sich gut erkennen, dass für Validität und Reliabilität nur diskrete Werte angenommen werden. Dies lässt sich dadurch erklären, dass Multitrait-Multimethod-Methoden zur Schätzung dessen verwendet wurden.

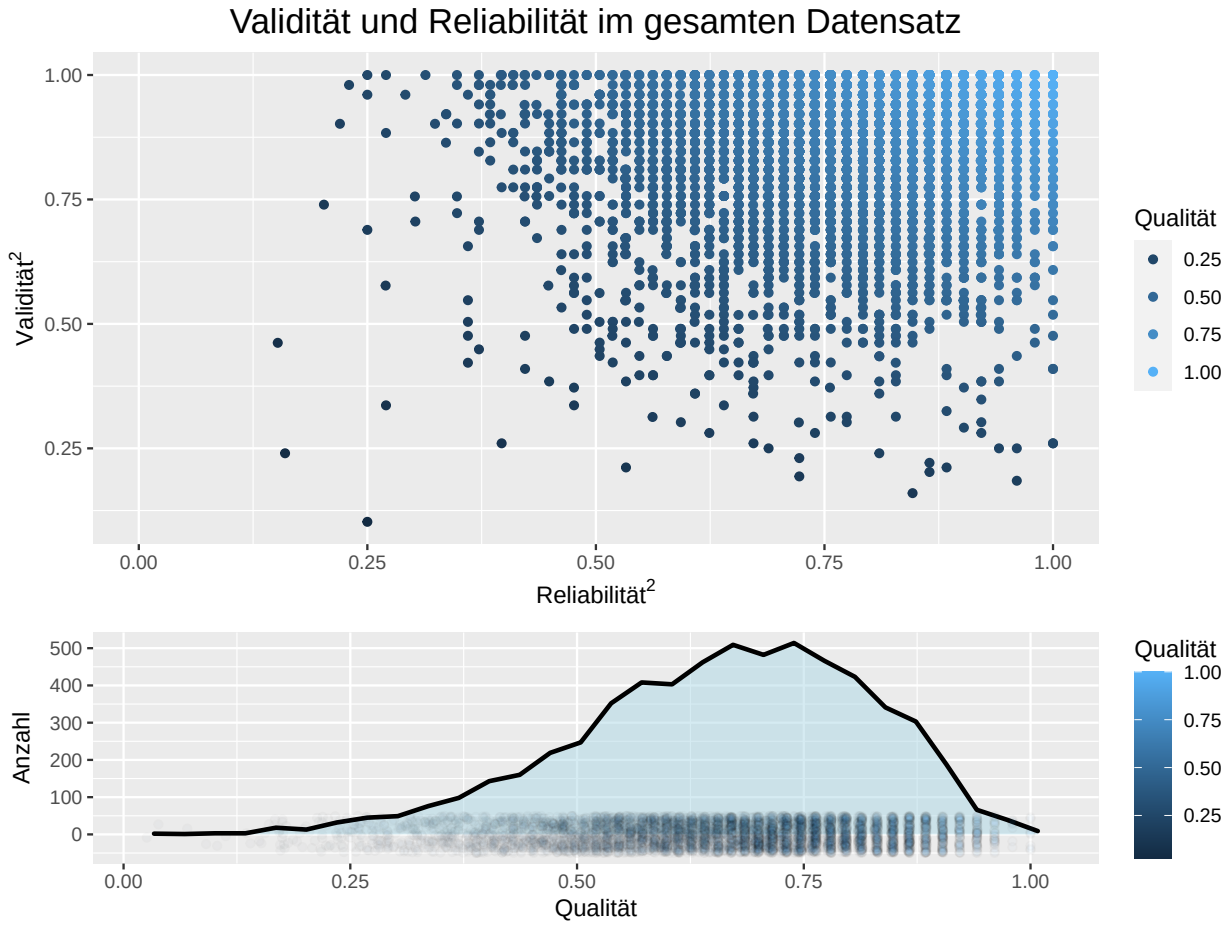


Figure 8: Verteilung der Outcome Variable, sowie den zugehörigen Einflüssen Reliabilität und Validität

3 Modelle

Das primäre Ziel ist es, Handlungsempfehlungen für Anwender zu entwickeln. Dementsprechend ist es notwendig, dass das verwendete Model zur Schätzung der Variableneinflüsse gut interpretierbar ist. Regressionsmodelle sind demnach eine sinnvolle Modellklasse für das vorliegende Problem.

Da der Stichprobenumfange groß ist, kann ein Modell - im Optimalfall - nach der Daumenregel *eine Einflussvariable pro 10 Beobachtungen* insgesamt ~ 600 Variablen enthalten.

Die hierarchische Struktur der Daten sollte für die Einflüsse **Studien**, **Experimente** und **Sprache** beibehalten werden. Die Häufigkeitsverteilungen sollen im Folgenden dargestellt werden.

Studien

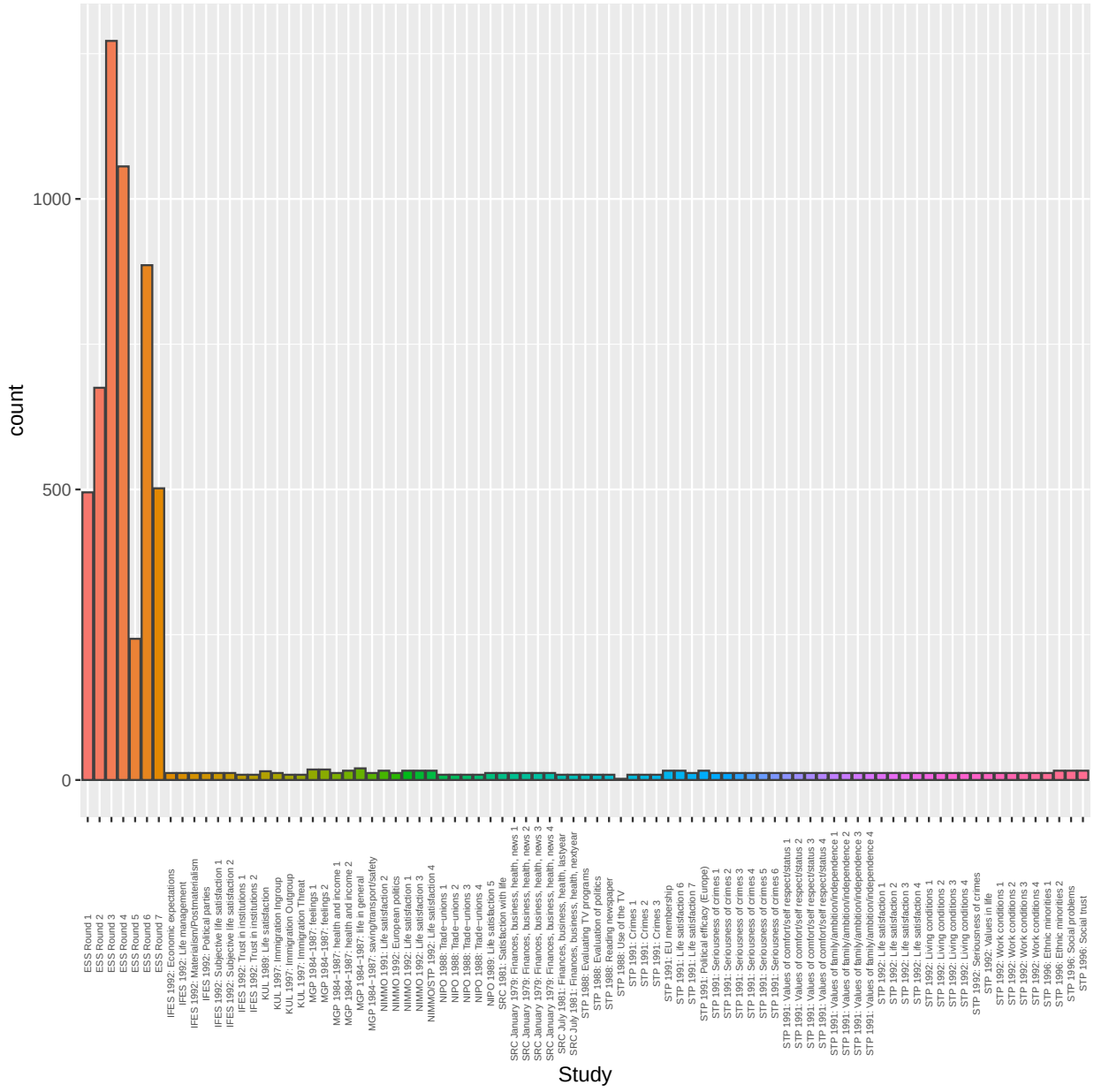


Figure 9: Häufigkeitsverteilung der Studien

Es ist klar ersichtlich, dass die Studien des **European Social Survey** den Großteil der Beobachtungen ausmachen. Andere Studien sind allesamt wesentlich kleiner.

Sprachen

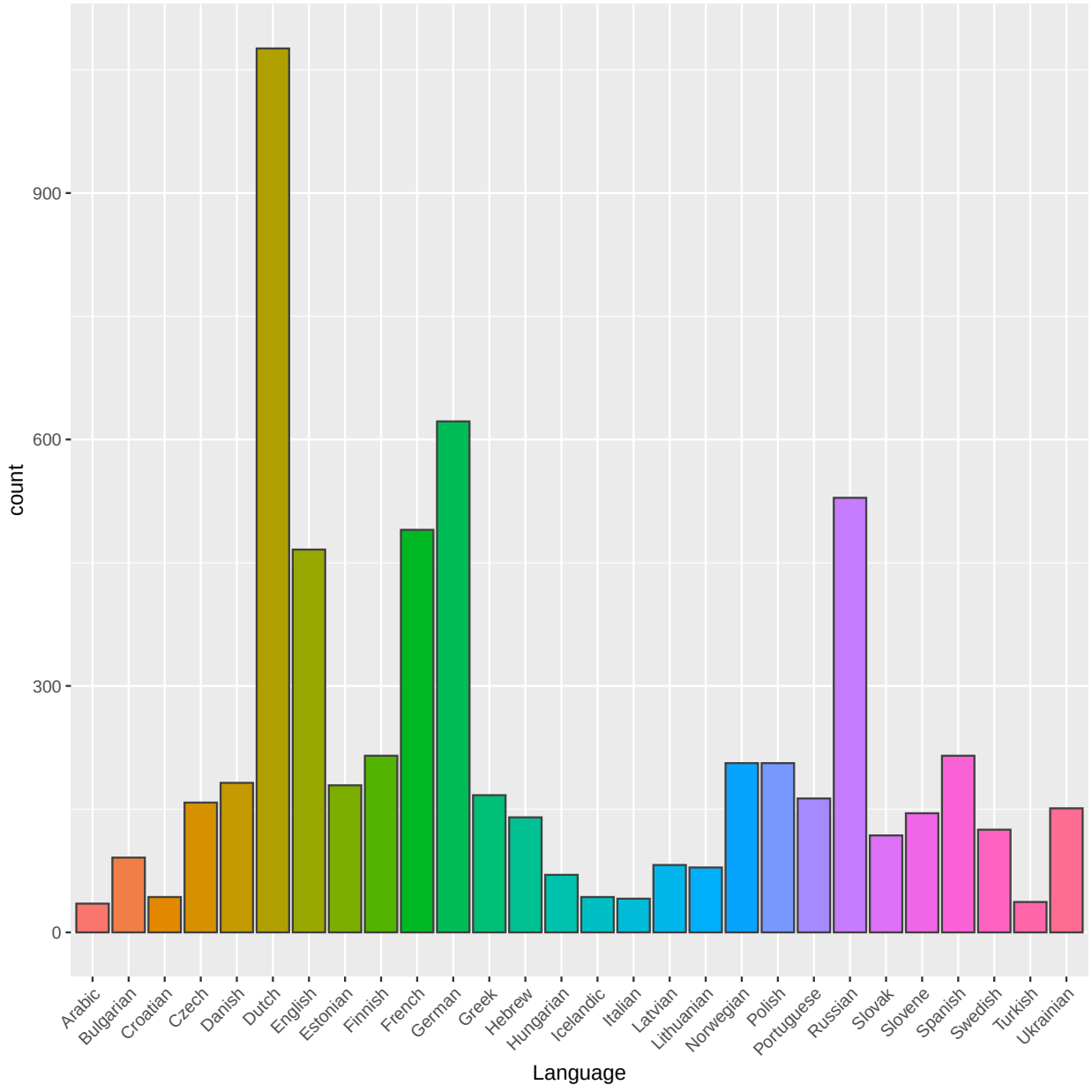


Figure 10: Häufigkeitsverteilung der Sprachen

Die Sprachen erscheinen relativ gleichverteilt. Jedoch gibt es 5 Sprachen, welche über 300 mal auftreten. Der Anteil dieser Sprachen macht 52% der Beobachtungen aus.

3.1 Genestete Variablen

Im Modell soll später berücksichtigt werden, dass eine hierarchische Struktur der Daten existiert. Die Frageitems sind hierbei in Experimente genestet. Die Experimente sind wiederum in Studien genestet und

diese in Sprachen. Diese sollten heterogen verteilt sein, d.h. es sollte beispielsweise für **ESS 1** nicht nur ein Land repräsentiert sein.

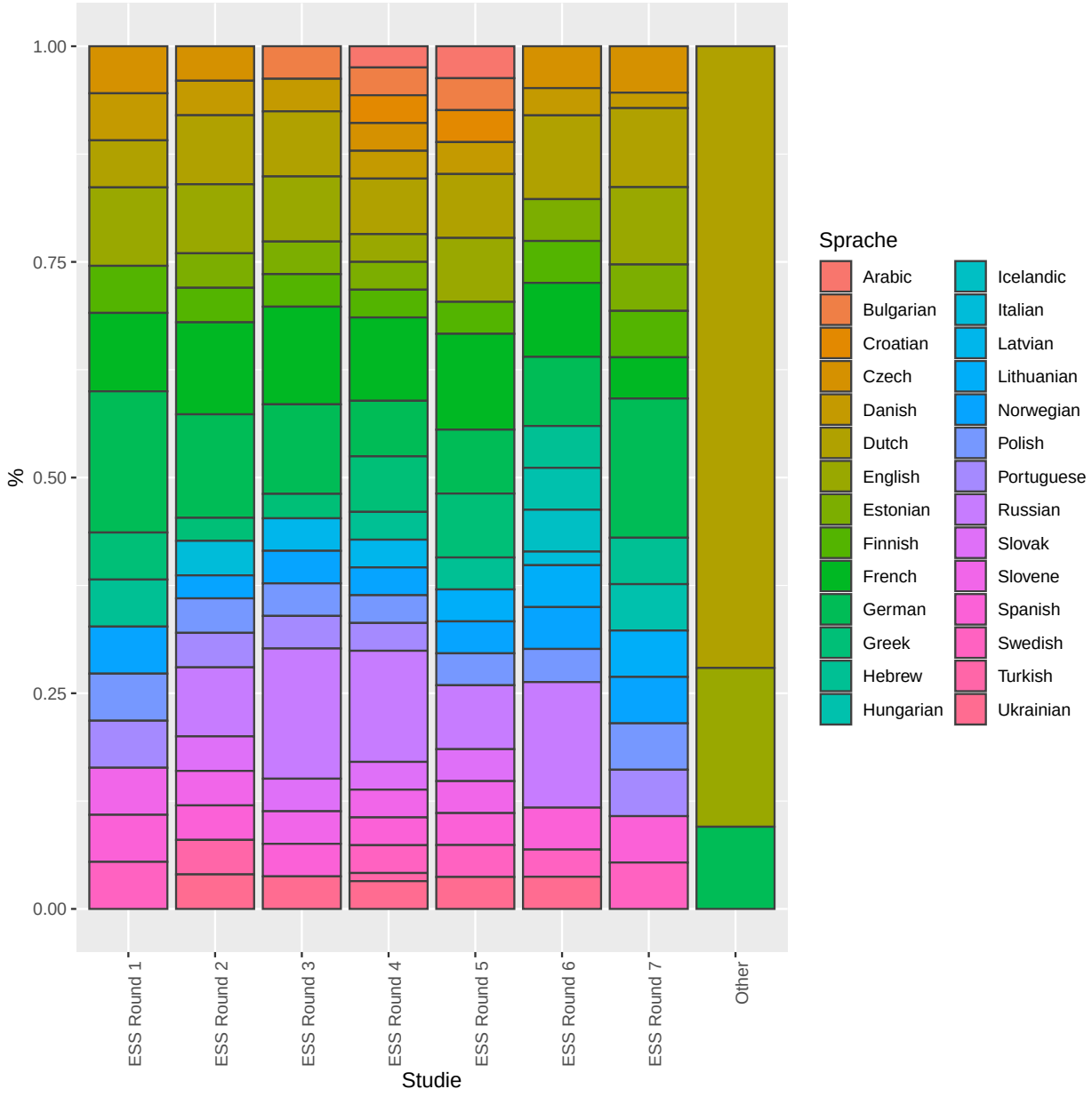


Figure 11: Verteilung der Länder in Studien

In der oberen Grafik erscheinen die Länder in den **ESS Studien** gleichverteilt zu sein. Studien der Kategorie "**Other**" sind wesentlich eintöniger. Sie umfassen lediglich drei Sprachen, wobei Niederländisch mit Abstand am meisten vertreten ist.

Um die Gleichverteilung ein wenig näher zu betrachten werden im folgenden die Lorenzkurven der einzelnen Studien visualisiert.

Wie in der vorherigen Grafik, scheinen die *ESS Studien* gleichverteilt zu sein. Studien welche nicht vom **European Social Survey** durchgeführt wurden erscheinen jedoch sehr heterogen.

Dies spiegelt sich auch in den einzelnen Gini-Koeffizienten wieder:

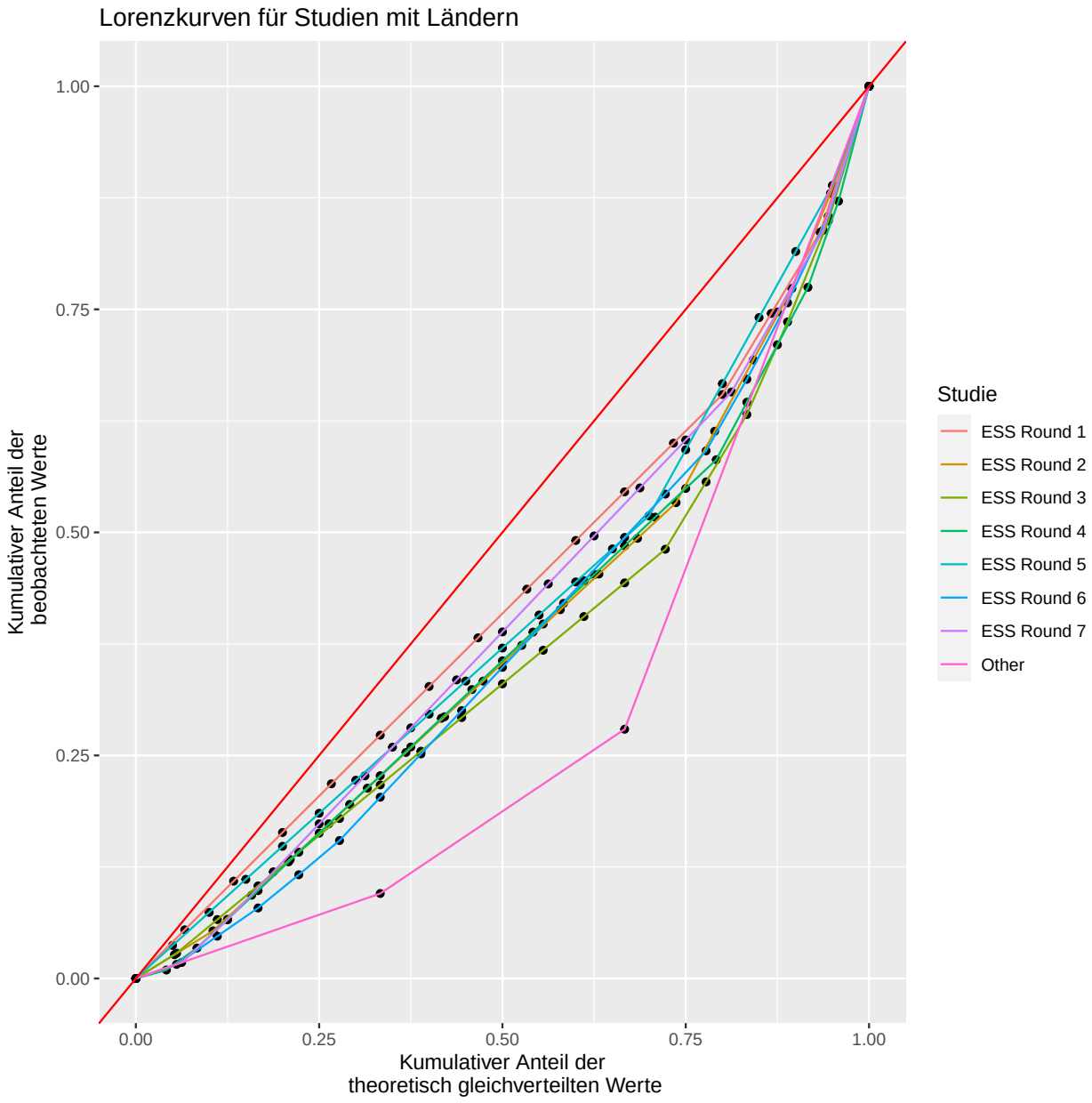


Figure 12: Lorenzkurven bezüglich der Häufigkeit von Ländern, seperat berechnet für jede Studie

Table 3: Gini-Koeffizient bezüglich der Häufigkeit von Ländern berechnet für jede Studie

Gini	Study
0.155	ESS Round 1
0.232	ESS Round 2
0.268	ESS Round 3
0.246	ESS Round 4
0.191	ESS Round 5
0.248	ESS Round 6
0.195	ESS Round 7
0.417	Other

3.2 Modelle

Nach bisherigem Stand haben wir die Möglichkeit zwei verschiedene Modelle zu berechnen. Beide male sollte ein LMM oder ein GLMM (je nach Verteilung) berechnet werden, wobei Cluster durch die Sprache und / oder durch Sprache mit der jeweiligen Studie entstehen. Jedoch besteht noch immer das Problem, dass Modelle nicht berechnet werden können aufgrund der Filter. Zwei mögliche Methoden, um damit umzugehen sind die Folgenden.

3.2.1 Modell 1: Das Gifi System: Umkodieren der Einflussvariablen

Eine Möglichkeit zum Umgang mit den Filtern bietet das Paper *Modeling with Structurally Missing Data by OLS and Shapley Value Regressions* von *Stan Lipovetsky* und *Ewa Nowakowska*.

Um das Problem zu lösen, könnte man x_j in ein System binärer Variablen aufsplitten. Man erhält für jede Ausprägung von x_j eine Dummy Variable, welche den Wert 1 annimmt, wenn Item i diesen Wert annimmt, und 0 falls nicht Fälle (NA inklusive). Beispielsweise könnte eine Variable 3 Kategorien haben :

$$y_{ijkl} = \beta_0 + ((\beta^{(1)}_{1jkl} * v_{1jkl}) + (\beta^{(2)}_{1jkl} * u_{1jkl}) + (\beta^{(3)}_{1jkl} * w_{1jkl})) + ... + ((\beta^{(1)}_{njkl} * v_{njkl}) + (\beta^{(2)}_{njkl} * u_{njkl}) + (\beta^{(3)}_{njkl} * w_{njkl})) + \epsilon_i$$

Hier bezeichnet exemplarisch $\beta^{(1)}_{1jkl}$ den Koeffzienten der ersten Dummy Variablen, der ersten Kovariablen im Datensatz für das i. Item, im j. Experiment, in der k. Studie in der l. Sprache. Die restlichen Koeffzienten sind gleich zu interpretieren.

Problem Dieses Modell kann jedoch in sofern problematisch sein, als dass es oft eine starke Multikollinearität aufweist. Die einzelnen Aufsplittungen einer Kovariablen x_j sind für gewöhnlich stark miteinander korreliert.

Mögliche Lösung: Shapley Value (SV) Die Idee von Shapley Values ist die Folgende: Es wird versucht, den Einfluss jeder Kovariablen im Model zu schätzen. Dabei werden alle möglichen Kombinationen von Kovariablen in Betracht gezogen, inklusive aller Teilmengen von Kovariablen. Grundlage für den SV ist das "Nutzenmaß": $U_j = R^2 - R^{2-j}$, wobei R^2 das klassische Qualitätskriterium in der linearen Regression darstellt und R^2_{-j} das Qualitätsmaß ohne die Kovariable x_j. Die Shapley Values berechnen sich nun wie folgt:

$SV_j = \sum_{allM} \gamma_n(M)[v(M \cup (j)) - v(M)]$

Hierbei stellen $\gamma_n(M) = m!(n-m-1)!/n!$ die Gewichte dar, wobei n die totale Anzahl an Kovariablen und m die Anzahl der Kovariablen in der M. Vereinigung ist. $v(M \cup (j))$ stellt den Wert des Nutzenmaß U_j dar welche die j. Kovariable enthält. $v(M))$ ist der Wert ohne die j.Kovariable. Der Shapley Value stellt also im Prinzip eine gewichtete Summe von Differenzen zwischen dem Nutzenmaß mit und ohne Kovariable j dar.

3.2.2 Model 2: Eigene Strata für die Filter

Die zweite Möglichkeit, die in der Einleitung im Buch *Statistical Analysis with Missing Data (second edition)* von *Roderick Little und Donald Rubin* (S.8) steht:

"We make the following key assumption throughout the book:

Assumption 1.1: missingness indicators hide true values that are meaningful for analysis

Assumption 1.1 may seem innocuous, but it has important implications for the analysis. When the assumption applies, it makes sense to consider analyses that effectively predict, or "impute" (that is, fill in) the unobserved values. If, on the other hand, Assumption 1.1 does not apply, then imputing the unobserved values makes little sense, and an analysis that creates strata of the population defined by the missingness indicator is more appropriate."

Da die Filter mit den NAs keinen wahren Wert "verstecken", sondern dies so strukturell aufgebaut wurde, sollten somit die Strata betrachtet werden. Hierbei müsste für jede Möglichkeit der Filter ein eigenes Modell berechnet werden. Problematisch wäre die geringe Anzahl an Beobachtungen und noch weiter kann ich nicht sagen, ob überhaupt bei so wenigen Beobachtungen die Unterteilung per LMM / GLMM (aufgrund der hierarchischen Struktur) sinnvoll wäre. Hier fehlen noch Zahlen!

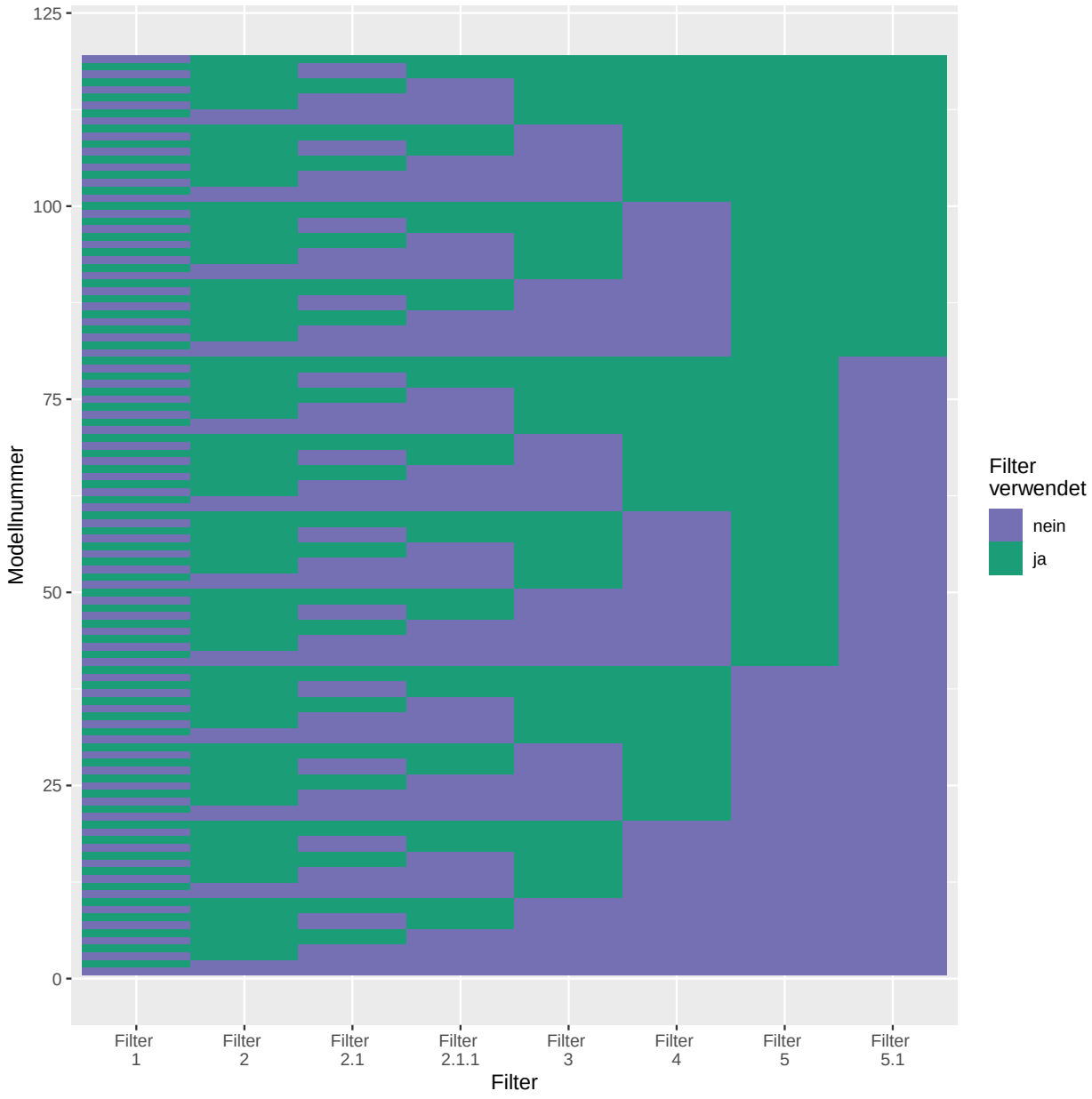


Figure 13: Alle Modelle, falls für alle Strata / Filter ein Modell berechnet wird