KonsortSWD

Konsortium für die
Sozial-, Verhaltens-, Bildungs- und
Wirtschaftswissenschaften

Gleiche Konzepte, unterschiedliche Messungen

Entwicklung und Validierung sozio-demographischer Standardvariablen

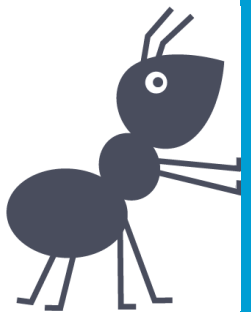
Silke Schneider, Lennart Palm, Melanie Partsch

GESIS – Leibniz-Institut für Sozialwissenschaften

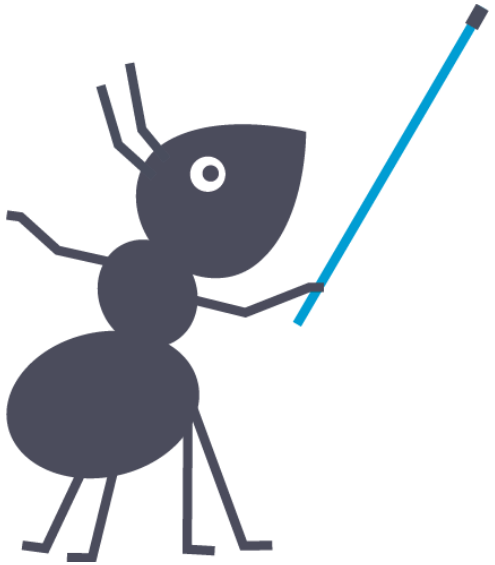
9. KSWD, Berlin

Session H: Sozio-demographische Variablen in Umfragen: Forschungspotentiale durch Harmonisierung steigern

28. März 2023



Überblick:



1. Motivation und Projektkontext
2. Entwicklung und Validierung von Standardvariablen
3. Beispiel: Familienstand
4. Ausblick

Output-Harmonisierung nationaler Daten – warum?

- Ausgangslage:
  - Umfragen in Deutschland nutzen unterschiedliche Erhebungsinstrumente für soziodemographische Merkmale (trotz Vorliegen der „Demographischen Standards“)*
  - Erhobene Daten daher nicht unmittelbar vergleichbar, nicht „interoperabel“
  - Vergleichbarkeit lässt sich durch Rekodierungen häufig herstellen
  - Hierzu liegen bislang keine standardisierten Zielvariablen vor
  - Harmonisierung daher bislang nur punktuell ex-post
- Lösung: *soziodemographische Standardvariablen*
  - Aus inkonsistenten Quellvariablen verschiedener Studien herleitbare einheitliche Zielvariablen

*Review von Erhebungsinstrumenten in 8 großen deutschen Umfragen: Schneider et al. 2022

Zielgruppen und potentielle Nutzende

- Keine Einschränkungen bzgl. Fachrichtung, Befragungsmodus o.ä.
  - Relevanz auch für bspw. epidemiologische Studien
- Hauptzielgruppe: Große Studien
  - Idealerweise Implementierung durch FDZs, die ergänzenden Datensatz anbieten
- Implementierung durch kleine Studien und Datennutzende natürlich auch möglich und erwünscht
  - Vergleich mit großen Studien hilft bei Einschätzung der Datenqualität

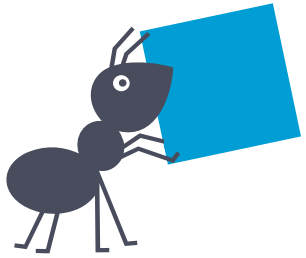
Projektkontext

- KonsortSWD eines der ersten Konsortien der „Nationalen Forschungsdateninfrastruktur“ (NFDI)
- Projekt angesiedelt in Task Area 3 „Data production“:

„Harmonized Variables – Umfragedaten leichter kombinieren durch standardisierte und harmonisierte Variablen “

- 2 Teilprojekte:
  - Harmonisierung von „inhaltlichen“ Variablen bzw. latenten Konstrukten – QuestionLink (hier nicht Thema)
  - Harmonisierung der Messung soziodemographischer Merkmale

2. Entwicklung und Validierung von Standardvariablen



Entwicklungsprozess

1. Review verwendeter Instrumente und Vergleich mit Demographischen Standards (Schneider et al. 2022)
2. Diskussion mit Studienvertreter*innen
3. Spezifikation von Entwürfen für Standardvariablen
4. Diskussion mit potentiellen Nutzenden
5. Empirische Validierung

Maximal- und Minimalversionen

Umsetzung: hierarchische Codeschemata

Maximalversion:

- ✓ Soll eine möglichst hohe Validität aufweisen
- – Lässt sich jedoch nicht für alle Studien herleiten

Minimalversion:

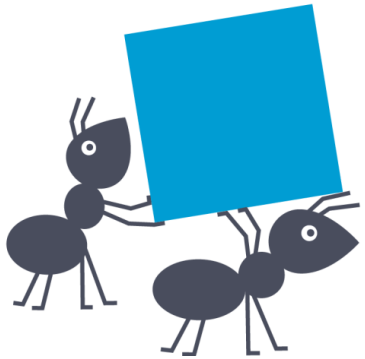
- – Lässt Validitätsverluste erwarten
- ✓ Soll von einer möglichst großen Zahl von Studien herleitbar sein

Validierung unterschiedlicher Versionen

- Qualitätssicherung:
  - Wie viel Informationsverlust durch Output-Harmonisierung?
  - Performanz konkurrierende Spezifikationen vergleichen
  - Minimalversionen überhaupt noch empfehlenswert?
- Daten- und theoriegeleitetes Verfahren basierend auf bis zu 189 mindestens 5-stufig ordinalskalierten Variablen aus ALLBUS 2018
  - Standardvariablen werden als Prädiktoren in parallelen Regressionsmodellen eingesetzt
  - Vorauswahl von Validierungsvariablen pro Merkmal: partielles adj. R2 ≥ .01
  - Informationsverlust operationalisiert durch Vergleich der partiellen adj. R2, d.h. Standardvariable relativ zur detailliertesten ALLBUS-Quellvariablen
  - (Verteilung der partiellen adj. R2 für die verschiedenen Versionen)
  - (Prüfung der Koeffizienten anhand theoretisch besonders sinnvoller Variablen)

Beispiel einer Standardvariablen mit Validierungsergebnissen

Familienstand



Familienstand

- Erfasst, ob eine Person verheiratet, geschieden, verwitwet oder ledig ist
- Demographische Standards (2016) und ALLBUS 2018 differenzieren 9 Kategorien:
  - zusammen- vs. in Trennung lebend für Verheiratete
  - Ehe und eingetragene Lebenspartnerschaft für verheiratet, geschieden, verwitwet
- Letztere Differenzierung hat keine Relevanz (mehr) für Datenanalysen:
  - Fallzahlproblematik bei eingetragenen Lebenspartnerschaften
  - Aufgrund inzwischen geänderter Rechtslage kein geeigneter Indikator für gleichgeschlechtliche Partnerschaften
- In Standardvariable wird daher nicht zwischen Ehe und eingetragener Lebenspartnerschaft unterschieden, wohl aber zwischen zusammen- vs. in Trennung lebend
  - Maximalversion=5 Kategorien
  - 11 von 11 von Schneider et al (2022) miteinander verglichenen Studien können diese bilden.

Standardvariable für Familienstand

Demographische Standards 2016		Standardvariable	
Wert	**Wertelabels**	**Wert**	**Wertelabels**
		10	Verheiratet (auch eingetragene Lebenspartnerschaft), nfs
A	Verheiratet und lebe mit meinem/meiner Ehepartner/-in zusammen	11	Verheiratet (auch eingetragene Lebenspartnerschaft), zusammenlebend
A1	In eingetragener Lebenspartnerschaft zusammenlebend (gleichgeschlechtlich)		
B	Verheiratet und lebe von meinem/meiner Ehepartner/-in getrennt	12	Verheiratet (auch eingetragene Lebenspartnerschaft), getrennt lebend
F	Eingetragene Lebenspartnerschaft, getrennt lebend (gleichgeschlechtlich)		
D	Geschieden	**20**	Geschieden (auch eingetragene Lebenspartnerschaft aufgehoben)
G	Eingetragene Lebenspartnerschaft aufgehoben (gleichgeschlechtlich)		
E	Verwitwet	**30**	Verwitwet (auch eingetragene/r Lebenspartner/in verstorben)
H	Eingetragene/r Lebenspartner/-in verstorben (gleichgeschlechtlich)		
C	Ledig	**40**	Ledig

Standardvariable für Familienstand

Demographische Standards 2016		Standardvariable	
Wert	**Wertelabels**	**Wert**	**Wertelabels**
		10	Verheiratet (auch eingetrag nfs
A	Verheiratet und lebe mit meinem/meiner Ehepartner/-in zusammen	11	Verheiratet (auch eingetragene Lebenspartnerschaft), zusammenlebend
A1	In eingetragener Lebenspartnerschaft zusammenlebend (gleichgeschlechtlich)		
B	Verheiratet und lebe von meinem/meiner Ehepartner/-in getrennt	12	Verheiratet (auch eingetragene Lebenspartnerschaft), getrennt lebend
F	Eingetragene Lebenspartnerschaft, getrennt lebend (gleichgeschlechtlich)		
D	Geschieden	**20**	Geschieden (auch eingetragene Lebenspartnerschaft aufgehoben)
G	Eingetragene Lebenspartnerschaft aufgehoben (gleichgeschlechtlich)		
E	Verwitwet	**30**	Verwitwet (auch eingetragene/r Lebenspartner/in verstorben)
H	Eingetragene/r Lebenspartner/in verstorben (gleichgeschlechtlich)		
C	Ledig	**40**	Ledig

Sollten Informationen nicht vorliegen kann die Oberkategorie verwendet werden

Keine Fälle im ALLBUS 2018

Keine Fälle im ALLBUS 2018

Validierungsergebnisse: Informationsverlust im partiellen adj. R2 im Vergleich von Quellversion zu vorgeschlagenen Standardvariablen

Merkmal	Versionen	Anzahl Kategorien	Performance im Vergleich zur Quellversion (45 Validierungsvariablen)
Familienstand	Quellversion	7*	-
	Maximalversion	5	98.89%
	Minimalversion	4	87.80%
	Alternativversion	2	66.04%

**Quellversion hat 9 Kategorien, von denen 2 in den ALLBUS-2018-Daten nicht besetzt sind.*

Validierungsergebnisse: Informationsverlust im partiellen adj. R2 im Vergleich von Quellversion zu vorgeschlagenen Standardvariablen

Merkmal	Versionen	Anzahl Kategorien	Performance im Vergleich zur Quellversion (45 Validierungsvariablen)
Familienstand	Quellversion	7*	-
	Maximalversion	5	98.89%
	Minimalversion	4	87.80%
	Alternativversion	2	66.04%

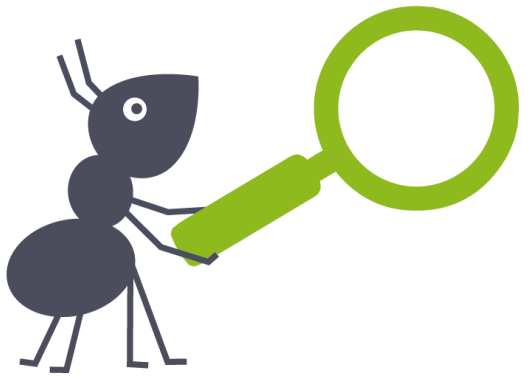
**Quellversion hat 9 Kategorien, v*
denen 2 in den ALLBUS-2018-Da
nicht besetzt sind.

Zusammenführung von eingetragener Lebenspartnerschaft und Ehe führt nur zu minimaler Reduktion des Informationsgehalts

Validierungsergebnisse: Informationsverlust im partiellen adj. R2 im Vergleich von Quellversion zu vorgeschlagenen Standardvariablen

Merkmal	Versionen	Anzahl Kategorien	Performance im Vergleich zur Quellversion (45 Validierungsvariablen)
Familienstand	Quellversion	7*	-
	Maximalversion	5	98.89%
	Minimalversion	4	87.80%
	Alternativversion	2	66.04%

**Quellversion hat 9 Kategorien, von denen 2 in den ALLBUS-2018-Daten nicht besetzt sind.*

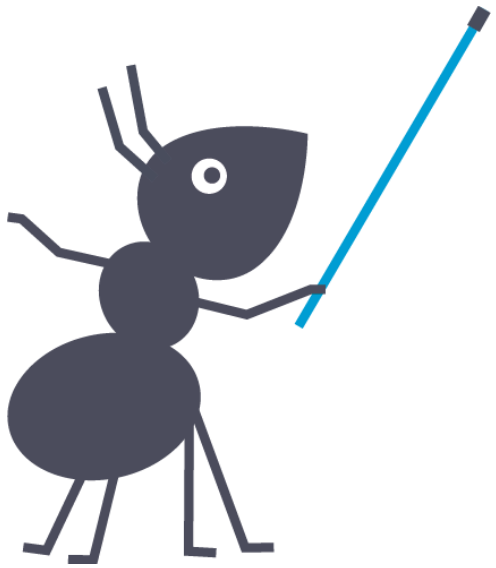


Unterscheidung von zusammen- oder in Trennung lebend hat substantiellen Informationsgehalt – Relevanz hängt aber auch von Forschungsfrage ab

Zusammenfassung und Ausblick

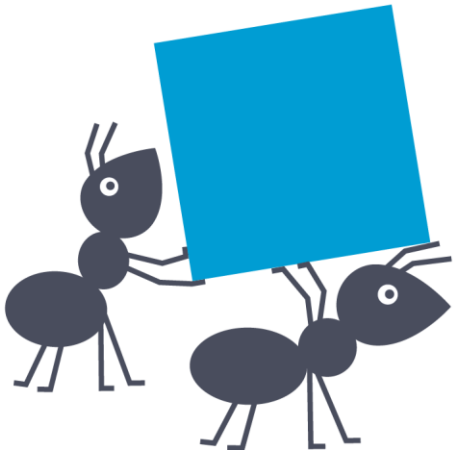
- Soziodemographische Standardvariablen könnten bei Sekundärdatenanalysen unterstützen, kumulative Erkenntnisse fördern und das Zusammenführen von Daten vereinfachen
- Einige Variablen eher leicht zu erstellen, andere äußerst komplex
- Vereinfachungen können Validität massiv reduzieren
- Noch offen:
  - Anzahl der Merkmale, für die Standardvariable entwickelt wird
  - Aktualisierung nach Projektende vs. Stabilität als „Feature“
  - Nutzung schwer nachvollziehbar! Wie Erfolg messen?

Bereiche und dazugehörige Merkmale



Basis-merkmale	Geburts-monat und -jahr	Geschlecht	Bildung	Geocode	
Arbeits-markt	Haupt-tätigkeit	Erwerbs-tätigkeit	Beruf	HH-Netto-einkommen	Stellung im Beruf
Kultur, Migration	Staats-angehörig-keit	Religions-zugehörig-keit	In Deutschland geboren ja/nein	Geburtsland	Sprache
Familie, Haushalt	Familien-stand	Partner-schaft	Haushalts-größe	HH-Zusammen-setzung	Anzahl Kinder

Vielen Dank!



Kontakt:
Silke.Schneider@gesis.org
Lennart.Palm@gesis.org

Validierung:
MelaniePartsch@googlemail.com

KonsortSWD wird Rahmen der NFDI durch die Deutsche Forschungsgemeinschaft (DFG) gefördert - Projektnummer: **442494171**

Literatur

- Schneider, S. L., Ortmanns, V., Diaco, A., & Müller, S. (2022). Die Erhebung sozio-demographischer Variablen in großen deutschen Umfragen: Ein Überblick über Möglichkeiten und Herausforderungen der Harmonisierung (Nr. 2/2022; KonsortSWD Working paper). GESIS - Leibniz-Institut für Sozialwissenschaften. https://doi.org/10.5281/zenodo.6810973

Fehlende Werte

Standardspezifikation für fehlende Werte

Werte	Label	Erklärung
-1	No answer	Zielperson hat keine Angabe getätigt, Angabe verweigert
-2	Don’t know	Zielperson hat „weiß nicht“ geantwortet
-3	Filter Missing	Merkmal trifft nicht zu , d.h. Zielperson bekommt Frage durch Filterführung nicht gestellt (computer-assisted modes)/hätte Frage nicht beantworten sollen (schriftliche Befragungen).
-4	Not included	Merkmal nicht erhoben (entweder generell oder nur in dieser Erhebungswelle)
-5	Cannot be generated	Variable ist nicht generierbar. Für generierte Variablen: mind. in einer Quellvariable nicht inhaltlich geantwortet und/oder entsprechende Frage nicht gestellt, d.h. wenn mind. eine der Quellvariablen entweder -1 oder -2 oder -3 ist.

Familienstand

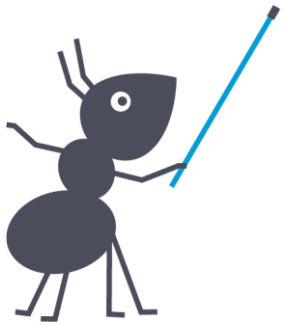
Weitere Versionen und Validierungsergebnisse

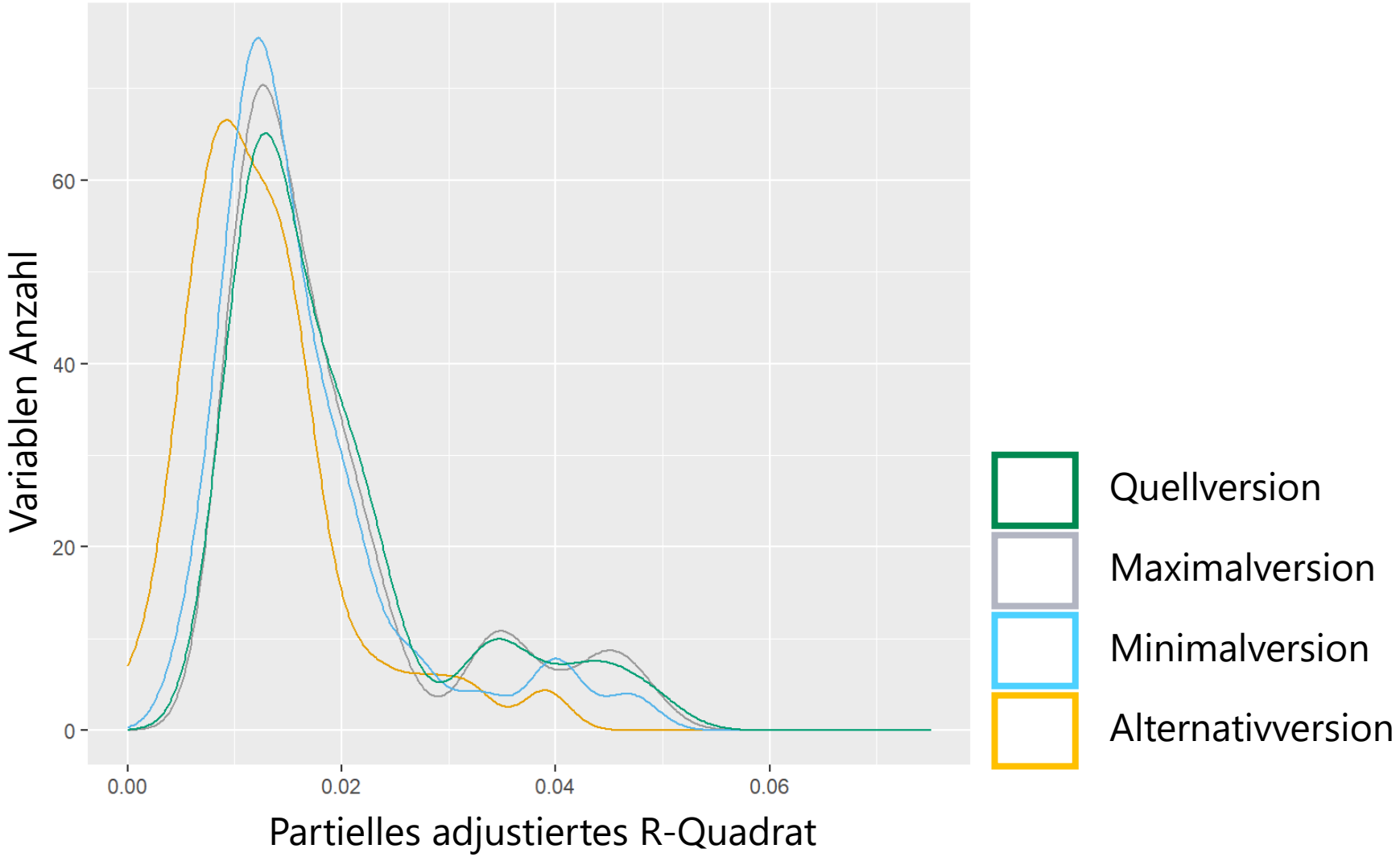
Verworfener Vorschlage einer Alternativversion

Demographische Standards 2016		Standardvariable	
Wert	**Value labels**	**Wert**	**Value labels**
		10	Verheiratet (auch eingetragene gleichgeschlechtliche Lebenspartnerschaft), nfs
A	Verheiratet und lebe mit meinem/meiner Ehepartner/-in zusammen	11	Verheiratet (auch eingetragene gleichgeschlechtliche Lebenspartnerschaft), zusammenlebend
A1	In eingetragener Lebenspartnerschaft zusammenlebend (gleichgeschlechtlich)		
B	Verheiratet und lebe von meinem/meiner Ehepartner/-in getrennt	12	Verheiratet (auch eingetragene gleichgeschlechtliche Lebenspartnerschaft), getrennt lebend
F	Eingetragene Lebenspartnerschaft, getrennt lebend (gleichgeschlechtlich)		
D	Geschieden	20	Geschieden (auch eingetragene gleichgeschlechtliche Lebenspartnerschaft aufgehoben)
G	Eingetragene Lebenspartnerschaft aufgehoben (gleichgeschlechtlich)		
E	Verwitwet	30	Verwitwet (auch eingetragene/r gleichgeschlechtliche/r Lebenspartner/in verstorben)
H	Eingetragene/r Lebenspartner/-in verstorben (gleichgeschlechtlich)		
C	Ledig	40	Ledig

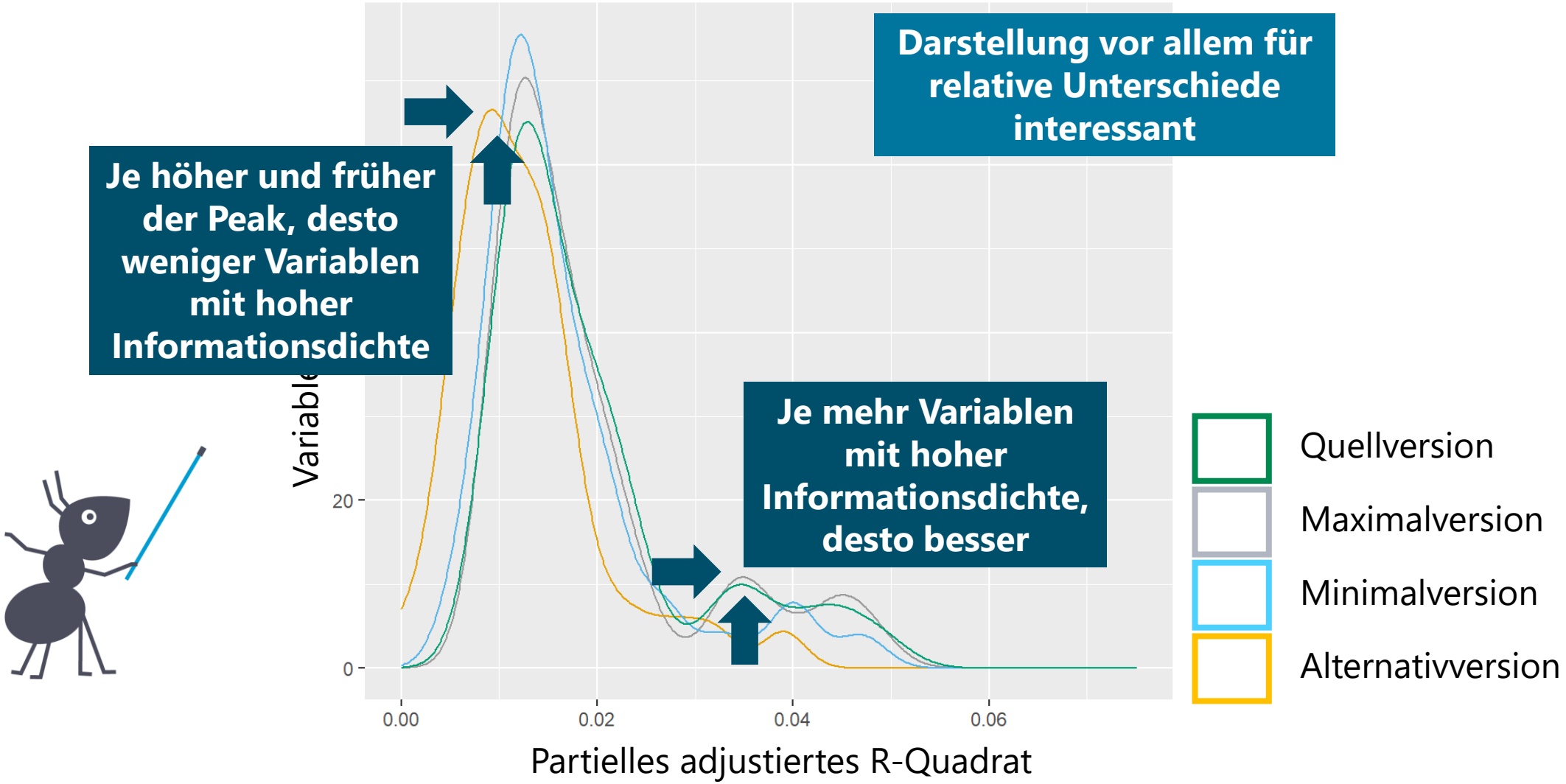
Verworfen

Verteilung (Kernel density estimation) der partiellen adj. R2 im Wertebereich von 0-0.075 in den verschiedenen Versionen der Standardvariablen

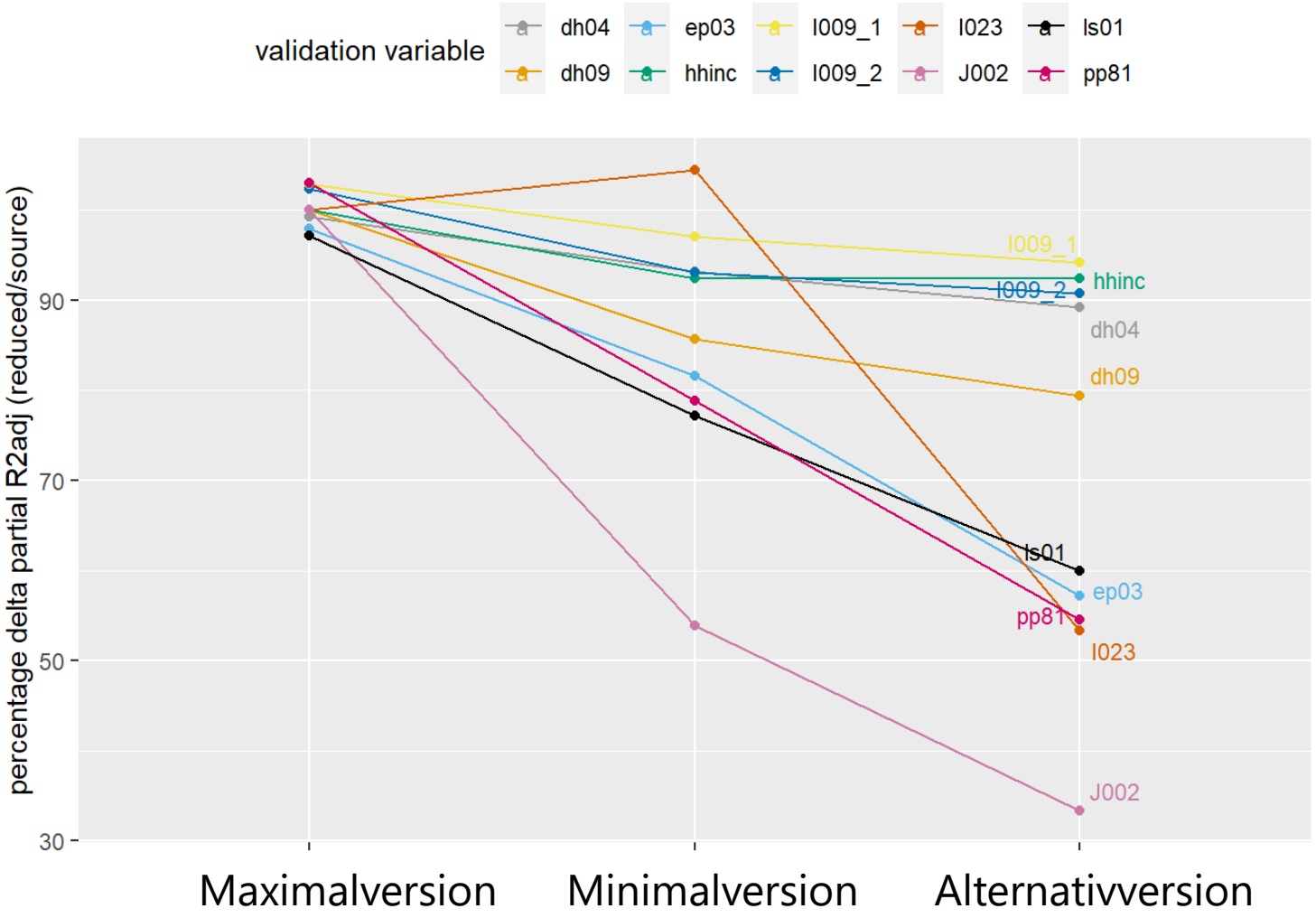




Verteilung (Kernel density estimation) der partiellen adj. R2 im Wertebereich von 0-0.075 in den verschiedenen Versionen der Standardvariablen



Informationsverlust im partiellen adj. R2 in den Top 10 Validierungsvariablen, in denen die Quellversion die meiste Varianz erklärt



1.	dh04	ANZAHL DER HAUSHALTSPERSONEN
2.	dh09	REDUZIERTE HAUSHALTSGROESSE
3.	hhinc	HAUSHALTSEINKOMMEN (OFFENE+LISTENANGABE)
4.	ep03	WIRTSCHAFTSLAGE, BEFR. HEUTE
5.	I023	HAEUFIGK. KONTAKT MIT ERWACHSENEM KIND
6.	I009_2	LETZTE 4 WOCHEN: EINSAM GEFUEHLT
7.	J002	WIE ZUFRIEDEN MIT BEZIEHUNG ZU FAMILIE?
8.	ls01	ALLGEMEINE LEBENSZUFRIEDENHEIT
9.	I009_1	LETZTE 4 WOCHEN: GESELLSCHAFT FEHLT
10.	pp81	HAEUFIGK. POLITIKGESPRAECH MIT FAMILIE