Kontrolle der statistischen Offenlegung

Wie verringert Eurostat das Offenlegungsrisiko von statistischen Einheiten?

Bundesamt für Statistik

Dr. David Tesar, 27.10.2022

Übersicht

- Nachbearbeiten von statistischen Auswertungen - Gründe
- Schutz vor Identifikation der statistischen Einheiten
- Kontrolle der statistischen Offenlegung (Statistical Disclosure Control)
- Eurostat Census 2021: Zwei empfohlene Methoden
- Cell Key Method - Umsetzung der SDC im Eurostat Census 2021
- Diskussion

Nachbearbeitung von Auswertungen - Gründe

Erwerbsstatus nach Alter im Kanton Luzern, 2020

unbehandelte Originaldaten

Ständige Wohnbevölkerung ab 15 Jahren

		Erwerbstätige	Selbständige	Angestellte	Übrige Erwerbstätige
		Anzahl Personen	Anzahl Personen	Anzahl Personen	Anzahl Personen
Total		**221582.917**	**14142.565**	**177171.536**	**30269.000**
Alter	15-Jährige	573.239	20.499	82.679	470.061
	16-Jährige	1526.000	-	71.844	1454.000
	17-Jährige	1631.000	-	31.823	1599.000
	18-Jährige	1972.000	15.556	581.834	1375.000
	19-Jährige	2341.000	-	1373.000	967.135
	20-Jährige	2165.000	18.942	1795.000	351.009
	21-Jährige	2932.000	36.568	2450.000	445.661
	22-Jährige	3454.000	36.156	3035.000	382.569
	23-Jährige	3849.000	26.304	3417.000	405.447
	24-Jährige	4255.000	52.759	3806.000	395.703

Nachbearbeitung von Auswertungen - Gründe

Erwerbsstatus nach Alter im Kanton Luzern, 2020

Ständige Wohnbevölkerung ab 15 Jahren

unbehandelte Originaldaten

		Erwerbstätige	Selbständige	Angestellte	Übrige Erwerbstätige
		Anzahl Personen	Anzahl Personen	Anzahl Personen	Anzahl Personen
Total		**221582.917**	**14142.565**	**177171.536**	**30269.000**
Alter	15-Jährige	573.239	20.499	82.679	470.061
	16-Jährige	1526.000	-	71.844	1454.000
	17-Jährige	1631.000	-	31.823	1599.000
	18-Jährige	1972.000	15.556	581.834	1375.000
	19-Jährige	2341.000	-	1373.000	967.135
	20-Jährige	2165.000	18.942	1795.000	351.009
	21-Jährige	2932.000	36.568	2450.000	445.661
	22-Jährige	3454.000	36.156	3035.000	382.569
	23-Jährige	3849.000	26.304	3417.000	405.447
	24-Jährige	4255.000	52.759	3806.000	396.703

Nachbearbeitung von Auswertungen - Gründe

Erwerbsstatus nach Alter im Kanton Luzern, 2020

Ständige Wohnbevölkerung ab 15 Jahren

unbehandelte Originaldaten

		Erwerbstätige	Selbständige	Angestellte	Übrige Erwerbstätige
		Anzahl Personen	Anzahl Personen	Anzahl Personen	Anzahl Personen
Total		**221582.917**	**14142.565**	**177171.536**	**30269.000**
Alter	15-Jährige	573.239	20.499	82.679	470.061
	16-Jährige	1526.000	-	71.844	1454.000
	17-Jährige	1631.000	-	31.823	1599.000
	18-Jährige	1972.000	15.556	581.834	1375.000
	19-Jährige	2341.000	-	1373.000	967.135
	20-Jährige	2165.000	18.942	1795.000	351.009
	21-Jährige	2932.000	36.568	2450.000	445.661
	22-Jährige	3454.000	36.156	3035.000	382.569
	23-Jährige	3849.000	26.304	3417.000	405.447
	24-Jährige	4255.000	52.759	3806.000	395.703

Nachbearbeitung von Auswertungen - Gründe

- Ganzzahliges Runden – statistische Einheiten sind ganzzahlig
- Runden auf 10, 100 oder 1000 genau – an Schätzungsgenauigkeit angepasst / Proportionen lassen sich mit einfacheren Zahlen besser merken
- Entfernen bzw. Hervorheben von wenig verlässlichen Zahlen – Grundlage für Interpretationen der Ergebnisse stärken
- Originalwerte perturbieren - Schutz vor Identifikation von Personen oder Haushalten

Nachbearbeitung von Auswertungen - Runden

Erwerbsstatus nach Alter im Kanton Luzern, 2020

Ständige Wohnbevölkerung ab 15 Jahren

Runden - ganze Einheiten & auf 5 genau

		Erwerbstätige	Selbständige	Angestellte	Übrige Erwerbstätige
		Anzahl Personen	Anzahl Personen	Anzahl Personen	Anzahl Personen
Total		**221'585**	**14'145**	**177'170**	**30'270**
Alter	15-Jährige	575	20	85	470
	16-Jährige	1'525	-	70	1'455
	17-Jährige	1'630	-	30	1'600
	18-Jährige	1'970	15	580	1'375
	19-Jährige	2'340	-	1'375	965
	20-Jährige	2'165	20	1'795	350
	21-Jährige	2'930	35	2'450	445
	22-Jährige	3'455	35	3'035	385
	23-Jährige	3'850	25	3'415	405
	24-Jährige	4'255	55	3'805	395

Nachbearbeitung von Auswertungen – Table redesign

Erwerbsstatus nach Alter im Kanton Luzern, 2020

Ständige Wohnbevölkerung ab 15 Jahren

Tabelle umstrukturieren

		Erwerbstätige	Selbständige	Angestellte	Übrige Erwerbstätige
		Anzahl Personen	Anzahl Personen	Anzahl Personen	Anzahl Personen
Total		**221'585**	**14'145**	**177'170**	**30'270**
Alter	15- bis 24-Jährige	24'698	207	16'646	7'846
	25- bis 44-Jährige	99'240	3'703	85'469	10'068
	45- bis 64-Jährige	92'499	8'646	72'954	10'899
	65-Jährige und Ältere	5'145	1'586	2'103	1'456

Nachbearbeitung von Auswertungen – in Funktion der Anzahl Beobachtungen

Erwerbsstatus nach Alter im Kanton Luzern, 2020

Ständige Wohnbevölkerung ab 15 Jahren

Klammern setzen

		Erwerbstätige	Selbständige	Angestellte	Übrige Erwerbstätige
		Anzahl Personen	Anzahl Personen	Anzahl Personen	Anzahl Personen
Total		**221'583**	**14'143**	**177'172**	**30'269**
Alter	15-Jährige	(573)	X	(83)	(470)
	16-Jährige	1526	X	X	1'454
	17-Jährige	1631	X	X	1'599
	18-Jährige	1972	X	(582)	1'375
	19-Jährige	2341	X	1'373	967
	20-Jährige	2165	X	1'795	(351)
	21-Jährige	2932	X	2'450	(446)
	22-Jährige	3454	X	3'035	(383)
	23-Jährige	3849	X	3'417	(405)
	24-Jährige	4255	X	3'806	(396)

Definition – Statistical Disclosure Control (SDC)

…ist eine statistische Methode, welche mithilfe von geringen **Perturbationen von Originalwerten** das Identifikationsrisiko von Personen oder Haushalten herabsetzt, ungeachtet ob es sich um statistische Ergebnisse aus einer Erhebung, aus Verwaltungsregistern oder um Veröffentlichung von Einzeldaten handelt.

Wo findet Statistical Disclosure Control Verwendung?

Ausschliesslich bei Eurostat Population and Housing Censuses 2021.

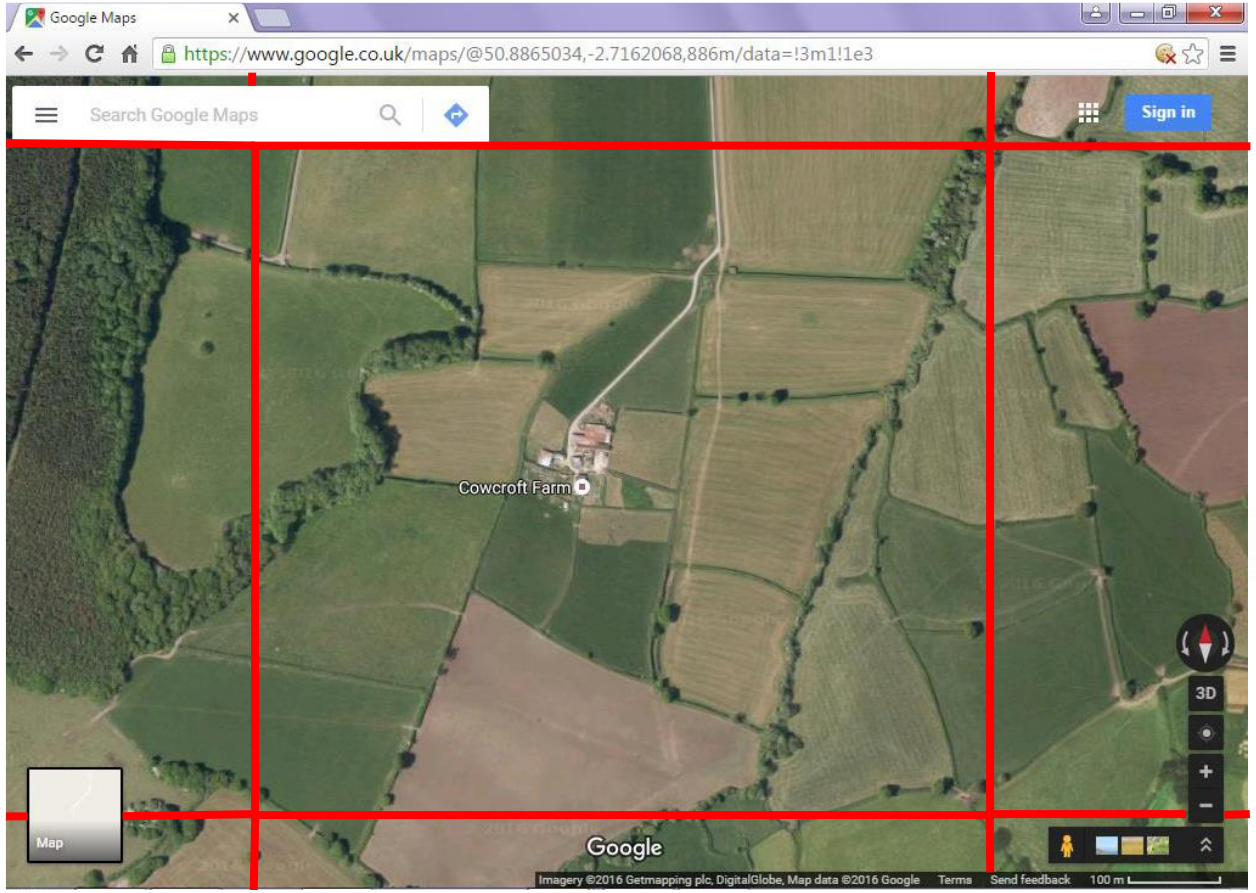
Auf nationaler Ebene findet SDC keine Verwendung.

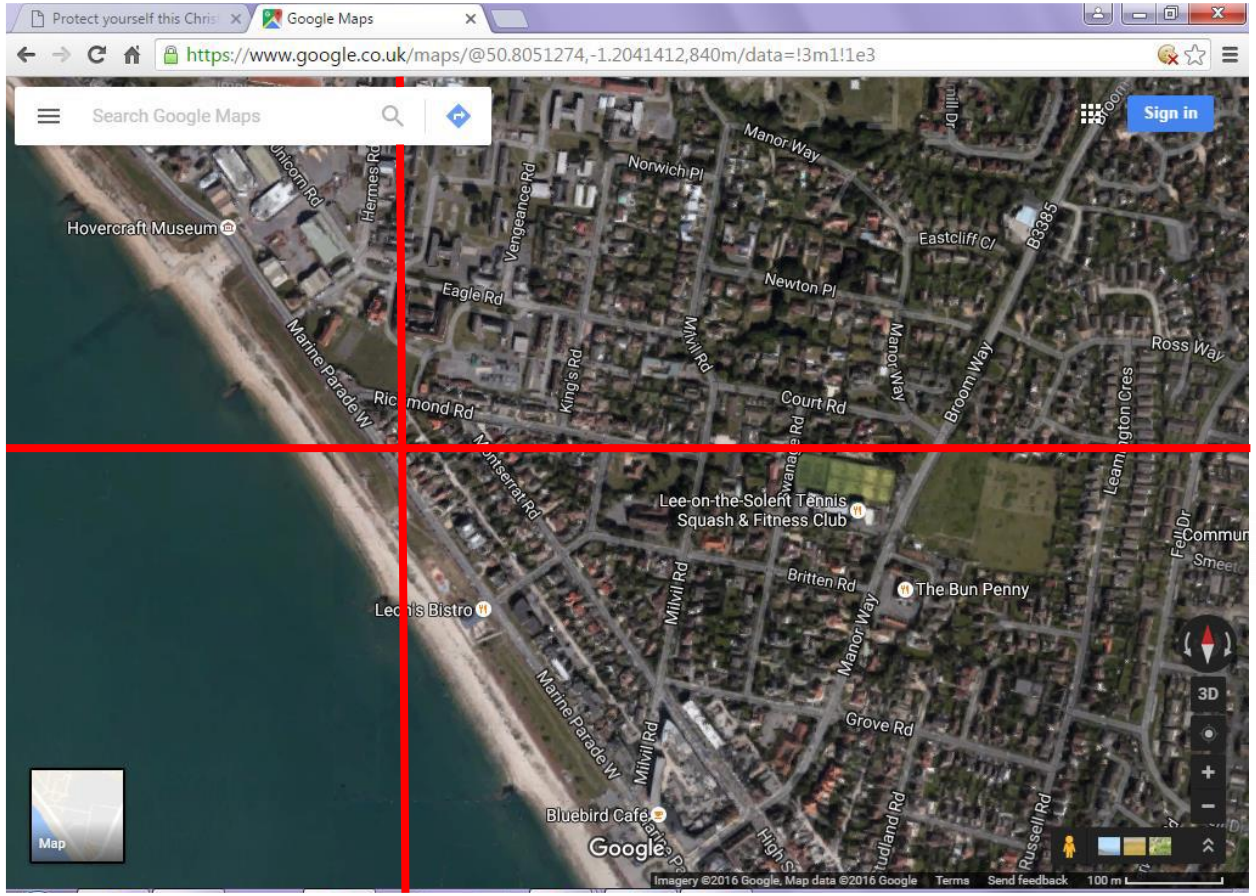
Eurostat Census 2021

Präzise Vorgaben zu Hypercubes (Datenwürfel) und Metadaten.

Erstmals aufbereitete Ergebnisse in einem km^2-Raster über ganz Europa

CensusHub2 (europa.eu)


Google Maps
https://www.google.co.uk/maps/@50.8865034,-2.7162068,886m/data=!3m1!1e3
Search Google Maps
Sign in
Cowcroft Farm
3D
Map
Google
Imagery ©2016 Getmapping plc, DigitalGlobe, Map data ©2016 Google
Terms
Send feedback
100 m


Google Maps
https://www.google.co.uk/maps/@50.8051274,-1.2041412,840m/data=!3m1!1e3
Search Google Maps
Sign in
Hovercraft Museum
Marine Parade W
Eagle Rd
Norwich Pl
Manor Way
Vengeance Rd
Eastcliff Cl
B3395
Newton Pl
Milvil Rd
King's Rd
Court Rd
Broom Way
Ross Way
Lee-on-the-Solent Tennis
Squash & Fitness Club
Britten Rd
The Bun Penny
Grove Rd
Bluebird Cafe
Google
Map
3D
Imagery ©2016 Google, Map data ©2016 Google
Terms
Send feedback
100 m

Problematik

Mit spezifischem Vorwissen, können aus einer Statistik, insbesondere aus niedrigen Zellenwerten, bestimmte Personen oder Haushalte identifizieren werden.

Was bewirkt die Statistical Disclosure Control?

- Verringerung des Identifikationsrisikos von Personen bzw. Haushalten, indem ein Quäntchen Unsicherheit in die Ergebnisse gestreut wird, so dass der Leser der Statistik in niedrigen Zellenwerten nicht mit Sicherheit auf bestimmte Personen bzw. Haushalte schliessen kann
- Geringer Informationsverlust

Statistical Disclosure Control - Methoden

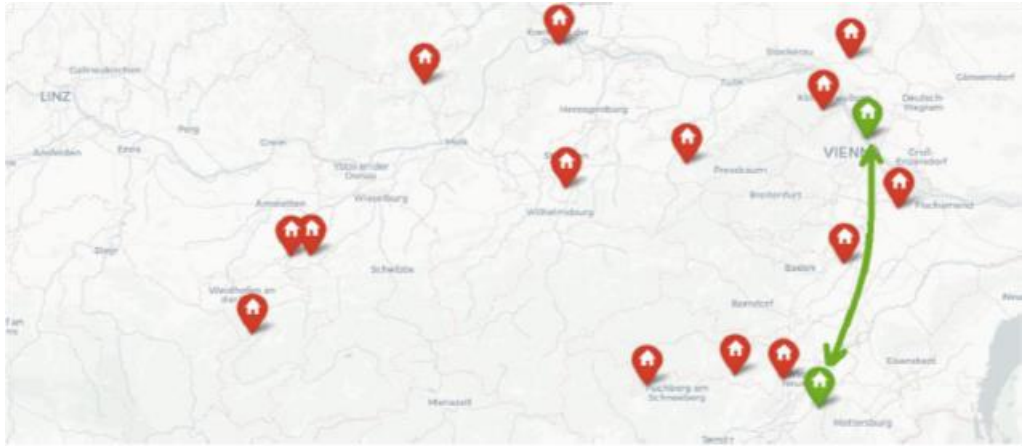
Grundsätzlich unterscheidet man zwischen SDC-Methoden, die Einzeldaten bzw. die tabellierte Ergebnisse perturbieren.

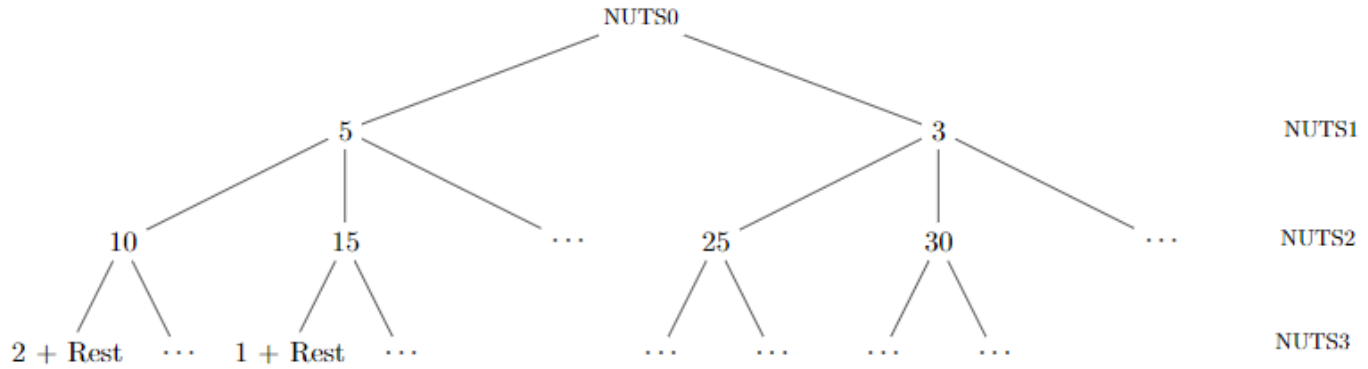
Von Eurostat für den Census 2021 empfohlen:

Targeted Record Swapping (TRS) – Einzeldaten werden perturbiert.

Cell Key Method (CKM) – tabellierte Ergebnisse werden perturbiert.

Targeted Record Swapping



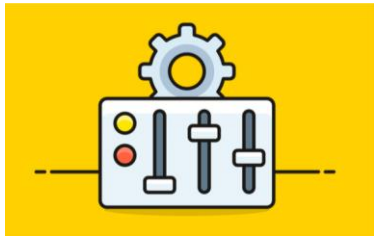


Cell Key Methode (CKM)

Originalwerte einer Tabelle werden geringfügig perturbiert/abgeändert.

Die Perturbation der CKM ist für identische Merkmalskreuzungen jeweils immer dieselbe.

Die Perturbationsfunktion lässt sich durch Festlegung der Parameter Amplitude (D), Verteilung (V) und Schwellenwert (js) konkret bestimmen.



Perturbationsfunktion: f(D,V,js)

Nachbearbeitung von Auswertungen – Cell Key Methode

Erwerbsstatus nach Alter im Kanton Luzern, 2020

Ständige Wohnbevölkerung ab 15 Jahren

perturbierte Ergebnisse

		Erwerbstätige	Selbständige	Angestellte	Übrige Erwerbstätige
		Anzahl Personen	Anzahl Personen	Anzahl Personen	Anzahl Personen
Total		**221'582**	**14'143**	**177'173**	**30'270**
Alter	15-Jährige	573	21	82	472
	16-Jährige	1'524	-	72	1'452
	17-Jährige	1'632	-	33	1'599
	18-Jährige	1'974	16	582	1'374
	19-Jährige	2'341	-	1'372	968
	20-Jährige	2'164	19	1'797	350
	21-Jährige	2'932	36	2'450	446
	22-Jährige	3'455	37	3'036	381
	23-Jährige	3'850	26	3'416	404
	24-Jährige	4'255	52	3'806	396

Cell Key Methode – Umsetzung für Census 2021 / 1

Grundvoraussetzung der Cell Key Methode ist es, dass im Einzeldatensatz jedem Eintrag eine gleichmässig verteilte Zufallszahl zwischen 0 und 1 zugeordnet wird.

→ Record Key (RK)

Micro data

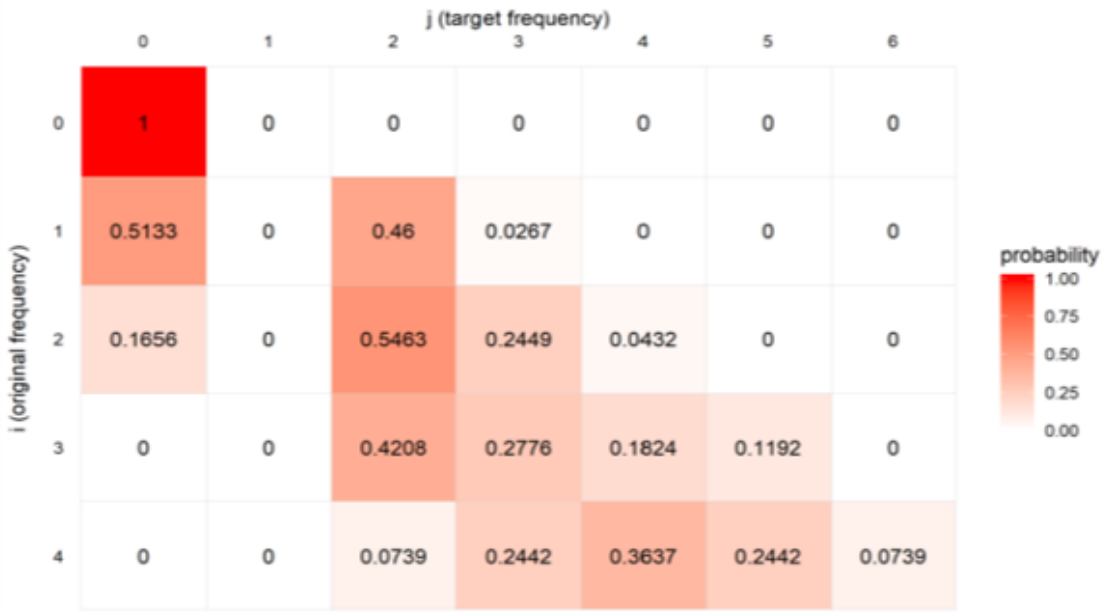
ID	Sex	Age	Edu	...	n	RK
1	M	A	A		1	0,34582249
2	F	B	A		1	0,68438579
3	F	B	C		1	0,95880618
4	F	C	F		1	0,62902289
5	M	B	B		1	0,86598721
6	F	C	B		1	0,36307981
7	M	A	A		1	0,91420393
8	M	A	F		1	0,6962939
9	M	B	F		1	0,53460054
10	F	B	A		1	0,68511663
11	F	B	C		1	0,0342637
12	M	B	C		1	0,33696811
13	F	B	A		1	0,11181613
14	F	A	B		1	0,56526973
15	M	A	C		1	0,01047942

Cell Key Methode – Umsetzung für Census 2021 / 2

Eine Perturbationsfunktion wird definiert, welche die Transitionswahrscheinlichkeit vom Originalwert zu einem perturbierten Wert bestimmt.

→ ptable package in R

Gezeigtes Beispiel enthält Parameterkombination: D=2, V=1.08, js=1



Cell Key Methode – Umsetzung für Census 2021 / 3

Aufsummieren der «Record Keys» für jede Merkmalskreuzung & Umbenennen in «Cell Key» (CK).

Sex	Age	i_original	CK
M	A	4	
M	B	3	1,74
M	C	0	
F	A	1	
F	B	5	
F	C	2	
M	T	7	
F	T	8	
T	A	5	
T	B	8	
T	C	2	
T	T	15	

Cell Key Methode – Umsetzung für Census 2021 / 4

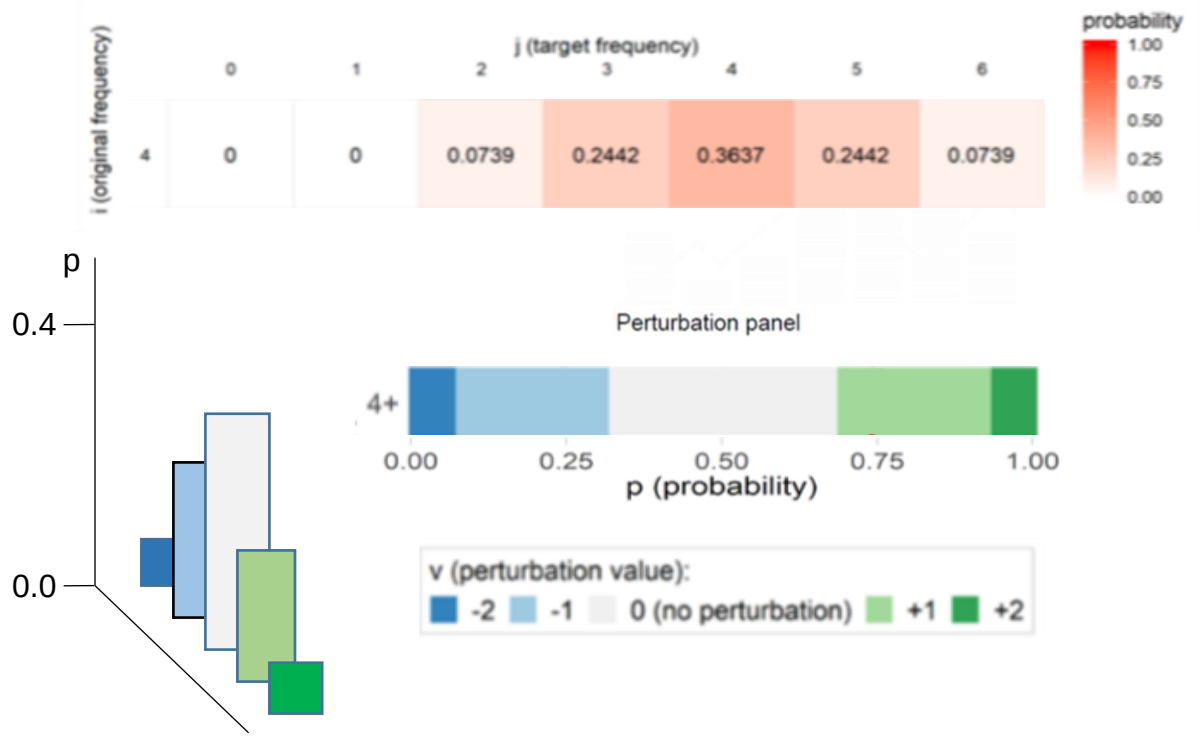
Entfernen der ganzen Zahl im «Cell Key».

Sex	Age	i_original	CK
M	A	4	
M	B	3	0,74
M	C	0	
F	A	1	
F	B	5	
F	C	2	
M	T	7	
F	T	8	
T	A	5	
T	B	8	
T	C	2	
T	T	15	

Cell Key Methode – Umsetzung für Census 2021 / 5

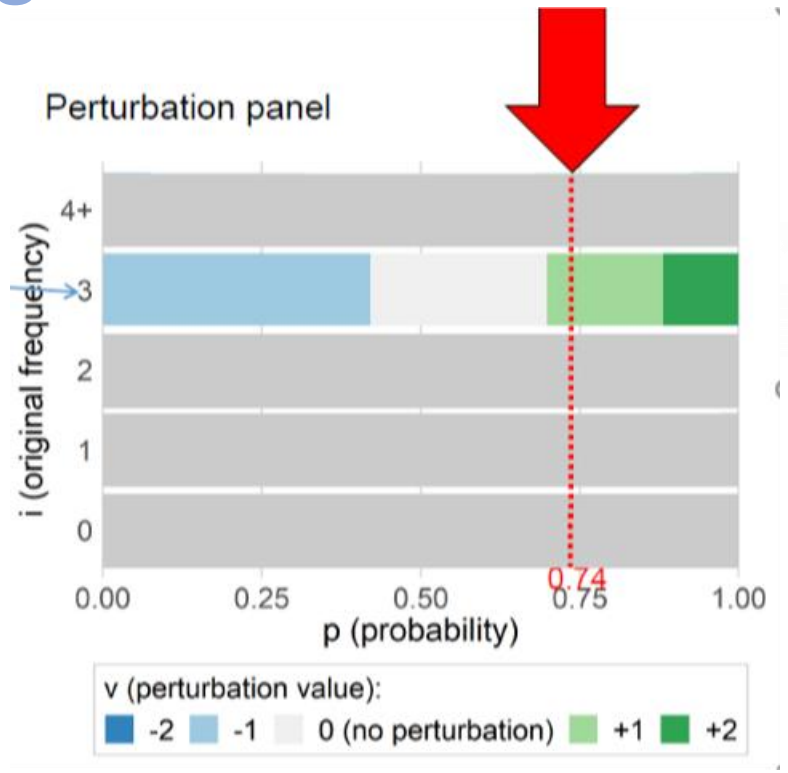
<u>Herleitung der Perturbation:</u>

Record key → Cell key → Perturbation



Cell Key Methode – Umsetzung für Census 2021 / 6

Der erzielte «Cell Key» wird über den «Perturbation Panel» gelegt und abgelesen, welche Perturbation dieser Wert zur Folge hat.

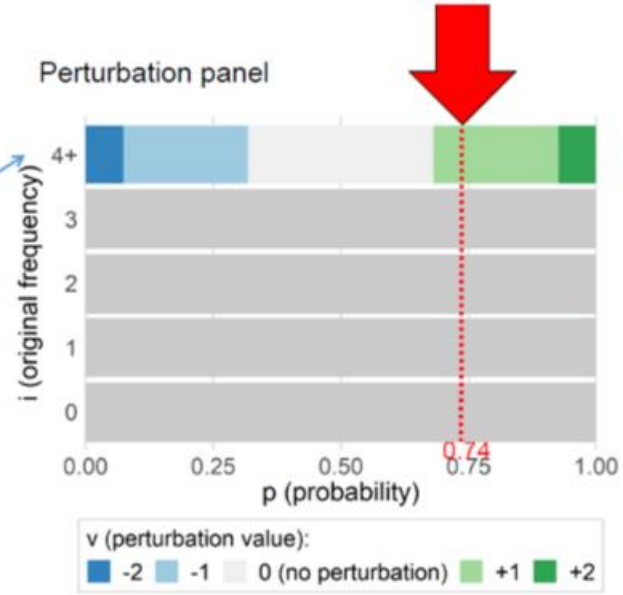


Cell Key Methode – Umsetzung für Census 2021 / 7

Berechnung des perturbierten Zielwertes

Tabular data (frequency table)

Sex	Age	i_original	CK	Noise	j_target
M	A	4	0,97		
M	B	3	0,74	1	4
M	C	0	0		
F	A	1	0,57		
F	B	5	0,47		
F	C	2	0,99		
M	T	7	0,7		
F	T	8	0,03		
T	A	5	0,53		
T	B	8	0,21		
T	C	2	0,99		
T	T	15	0,74	1	16



Vor- und Nachteile der beiden empfohlenen Methoden

	Targeted Record Swapping	Cell Key Method
pro	Auf sensible Daten fokussiert Randverteilungen bleiben gleich HH-Struktur bleibt gleich Totale über alle Regionen gleich	Konsistente Perturbationen über alle Tabellenzellen hinweg
contra	Kann zu grösseren Verschiebungen von Personen führen	Technisch anspruchsvoll umzusetzen Nicht additiv

Cell Key Methode und die fehlende Additivität

Geschlecht	Alter	Originalwerte	Perturbation	Perturbierte Werte	Kumulierte Werte
T	*15-24*	878'823	+2	878'825	878'822
T	*25-44*	2'344'456	-1	2'344'455	2'344'456
T	*45-64*	2'389'089	0	2'389'089	2'389'089
T	*65+*	1'520'165	0	1'520'165	1'520'165
F	*15-24*	426'794	+1	426'795	-
F	*25-44*	1'162'436	0	1'162'436	-
F	*45-64*	1'190'782	0	1'190'782	-
F	*65+*	826'560	0	826'560	-
M	*15-24*	452'029	-2	452'027	-
M	*25-44*	1'182'020	0	1'182'020	-
M	*45-64*	1'198'307	0	1'198'307	-
M	*65+*	693'605	0	693'605	-

Transparenz nach aussen

Die hochgerechneten Ergebnisse werden auf volle zehn Personen gerundet ausgewiesen. Die in den Ergebnistabellen dargestellten Summenwerte werden stets auf Basis der nicht gerundeten Ausgangswerte ermittelt, weshalb diese von der Summe der ausgewiesenen Einzelwerte abweichen können.

Bei den hochgerechneten Zensusergebnissen aus der Haushaltsstichprobe werden die Ergebnisse mit zu geringen Besetzungszahlen nicht ausgewiesen, sondern durch einen Schrägstrich (/) ersetzt.

Für die Wahrung der Geheimhaltung nach § 16 Bundesstatistikgesetz (BStatG) wird für Auswertungen, die ausschließlich auf demografischen Daten, Gebäude- und Wohnungsdaten, Haushaltsdaten und Familiendaten basieren, ein Verfahren der stochastischen Überlagerung nach der Cell Key-Methode (CKM) angewendet.

Aus Qualitätsgründen addieren sich die jeweiligen Einzelwerte einer Tabellenzeile oder -spalte nicht notwendigerweise zur ausgewiesenen Gesamtsumme.

Die Einwohnerzahl (Bevölkerung insgesamt) wird durch die statistische Geheimhaltung nicht verändert. Aus diesem Grund kann die Summe der Einzelergebnisse einer Tabelle von der Einwohnerzahl abweichen.

Weiterführende methodische Informationen zum Zensusmodell und zur Geheimhaltung stehen unter **www.zensus2022.de** zur Verfügung.

Zusammenfassung

- SDC reduziert das Risiko, dass einzelne Personen oder Haushalte in veröffentlichten Statistiken identifiziert werden können.
- 2 Methoden werden für den Census 2021 von Eurostat empfohlen.
- Die Cell Key Methode ist die Methode, welche die Schweiz ausschliesslich anwenden wird.
- Eurostat gibt Empfehlungen zur Anwendung der SDC bei den Hypercubes vor, lässt den nationalen statistischen Ämtern aber viele Freiheiten bei der konkreten Umsetzung.

Danke für die Aufmerksamkeit

Gibt es Fragen?