

Einführung in die statistische Datenanalyse mit R

Kommunikation von Daten und Analysen

David Benček

Wintersemester 2015/16

Kursinhalte bisher

- ▶ reine Datenverarbeitung in R
- ▶ Berechnung statistischer Modelle
 - ▶ stetige abhängige Variable: OLS-Modell
 - ▶ binäre abhängige Variable: Logit-Modell
- ▶ Interpretation von Modellergebnissen

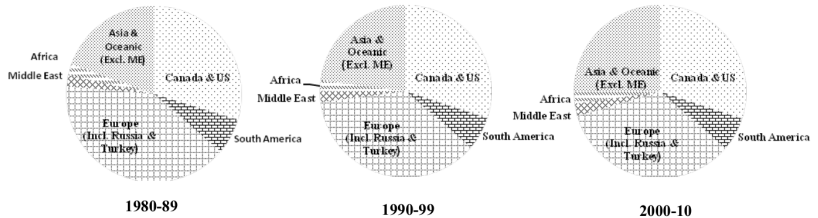
Was fehlt?

- ▶ Aufbereitung und Präsentation von
 - ▶ Daten
 - ▶ Zusammenhängen
 - ▶ Ergebnissen

→ **verständliche** Abbildungen und Tabellen

Nicht so! Warum?

Annual GDP



Nicht so! Warum?

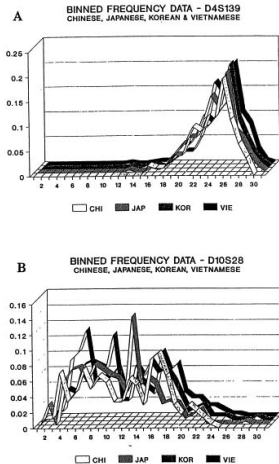


FIG. 4. Fixed bin distribution (histogram) for two loci and four Asian subpopulations (used with permission from John Hartmann): the boundaries of the 30 bins (vertical axis) are determined by the FBI; these bins are not of equal length. Sample sizes (numbers of individuals) for Chinese, Japanese, Korean and Vietnamese are 103, 125, 93 and 215 for D4S139 and 120, 137, 100 and 193 for D10S28. The horizontal axis is the bin number; bins are not of equal length.

Nicht so! Warum?

Distribution of All TFBS Regions

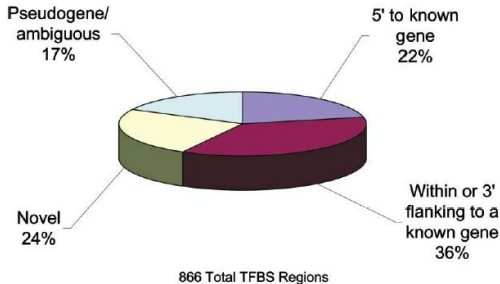


Figure 1. Classification of TFBS Regions

TFBS regions for Sp1, cMyc, and p53 were classified based upon proximity to annotations (RefSeq, Sanger hand-curated annotations, GenBank full-length mRNAs, and Ensembl predicted genes). The proximity was calculated from the center of each TFBS region. TFBS regions were classified as follows: within 5 kb of the 5' most exon of a gene, within 5 kb of the 3' terminal exon, or within a gene, novel or outside of any annotation, and pseudogene/ambiguous (TFBS overlapping or flanking pseudogene annotations, limited to chromosome 22, or TFBS regions falling into more than one of the above categories).

Nicht so! Warum?

Table 5

Simulation results for using full data, CRs only, and proposed method under four missing mechanisms

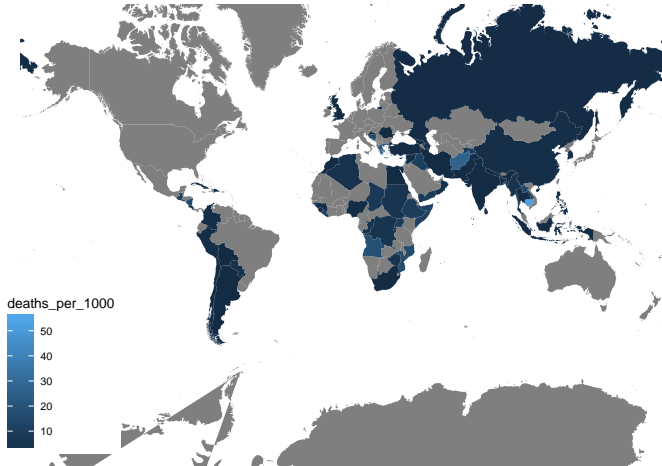
Method	Bias ^a		Variance ^b		95% CI ^c	
	$(\hat{\beta}_W)$	$(\hat{\beta}_X)$	$(\hat{\beta}_W)$	$(\hat{\beta}_X)$	$(\hat{\beta}_W)$	$(\hat{\beta}_X)$
(M.1) $P(R = 1) = 0.66$						
Full	0.01346	0.02229	0.04008	0.03685	0.955	0.950
Comp	0.03062	-0.003561	0.1149	0.06732	0.960	0.955
Impu	0.01431	0.021	0.04088	0.05169	0.980	0.975
(M.2) logit $P(R = 1) = 2Y$						
Full	0.007908	-0.02116	0.03838	0.03624	0.975	0.925
Comp	0.01945	0.07096	0.107	0.06581	0.960	0.950
Impu	0.006966	0.01597	0.04227	0.05226	0.975	0.985
(M.3) logit $P(R = 1) = 2X$						
Full	0.007908	-0.02116	0.03838	0.03624	0.975	0.925
Comp	0.01225	0.0589	0.08856	0.06818	0.980	0.975
Impu	0.009563	-0.04699	0.03865	0.04923	0.985	0.970
(M.4) logit $P(R = 1) = X + Y$						
Full	0.01346	0.02229	0.04008	0.03685	0.955	0.950
Comp	0.02404	1.613	0.1102	0.08202	0.955	0.580
Impu	0.01814	0.08289	0.0578	0.06075	0.955	0.970

^aBias = $(\hat{\beta} - \beta_0)/\beta_0$.

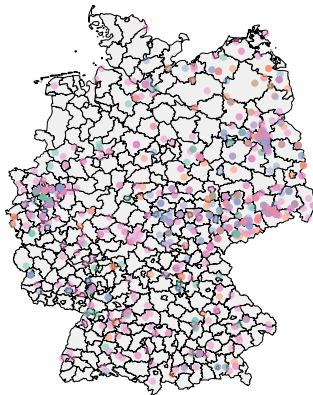
^bSimulation variance.

^cConfidence interval using jackknife standard error.

Lacina: Tote pro 1000 Einwohner



Rechte Gewalt in Deutschland (2014–2015)



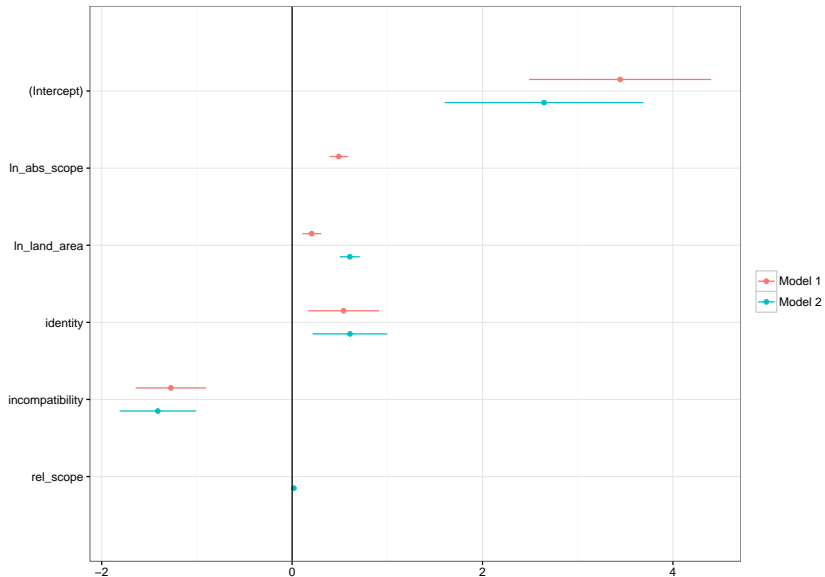
● Arson ● Demo
● Assault ● Misc_Property_Attack

Buhaug/Gates: Regressionsergebnisse – Tabelle

	location	
	(1)	(2)
ln_abs_scope	0.490*** (0.050)	
rel_scope		0.019*** (0.002)
ln_land_area	0.206*** (0.051)	0.606*** (0.054)
identity	0.541*** (0.190)	0.608*** (0.200)
incompatibility	-1.273*** (0.189)	-1.411*** (0.205)
Constant	3.445*** (0.488)	2.646*** (0.531)
N	243	243
R ²	0.620	0.575
Adjusted R ²	0.613	0.568
Residual Std. Error (df = 238)	1.126	1.190
F Statistic (df = 4; 238)	97.017***	80.587***

*p < .1; **p < .05; ***p < .01

Buhaug/Gates: Regressionsergebnisse – Grafik



Hinweise für gute Grafiken

- ▶ Rob Hyndman: "Twenty rules for good graphics"
(<http://robjhyndman.com/hyndsight/graphics/>)
- ▶ Edward Tufte: "The Visual Display of Quantitative Information"
(<http://www.amazon.com/dp/0961392142/>)
- ▶ William Cleveland: "The Elements of Graphing Data"
(<http://www.amazon.com/dp/0963488414/>)