

Wer ich bin und warum R

Jan-Philipp Kolb

23 November 2017

Über mich

- VWL Studium in Trier (Diplom 2008)
- 2004 Erasmus Jahr an der Université Jean Moulin in Lyon
- Wissenschaftlicher Mitarbeiter am Lehrstuhl für Wirtschafts- und Sozialstatistik
- 2012 Promotion (Thema: Die Erzeugung von synthetischen Grundgesamtheiten)
- Seit 2012 am Gesis Leibniz Institut für Sozialwissenschaften - zunächst Team Statistik
- Seit 2017 Survey Statistik im Team Gesis Panel

Gesis ist:

- Infrastruktureinrichtung für die Sozialwissenschaften
- mit über 250 MitarbeiterInnen an zwei Standorten (Köln und Mannheim)

GESIS bietet:

- Beratung zu Forschungsprojekten in allen Phasen
- Forschungsbasierte wissenschaftliche Dienstleistungen

Was ist das Gesis Panel

Datenerhebungsinfrastruktur

- Probabilistisches mixed-mode Access Panel
- Deutsche Allgemeinbevölkerung
- Deutschsprachig - ab 18 Jahren
- Basierend auf Einwohnermeldeamtsstichprobe
- Mehrstufiger Rekrutierungsprozess, sequentielles mixed-mode Design
- Seit 2013 - ca 27 Wellen (alle zwei Monate)
- 2016 Auffrischungstichprobe

Überblick Wellen und Studien

- Bisher sind im Gesis Panel in 27 Wellen (aa-bc) ca 77 Studien gelaufen.
- Die Daten sind als Scientific Use File (SUF) oder im Secure Data Center in Köln verfügbar.
- Der SUF der Welle ec umfasst 7599 Beobachtungen und 7874 Variablen

Kürzel	Studientitel	Wellen
ag	Environmental Spatial Strategies	ba
an	Leisure travel and subjective well-being	bc, bd, be
aq	Pro-environmental Behavior in High Cost Situations	be, cb
bw	Space-sets: the scope and characteristics of national and international mobility experiences	fa
zd	GESIS Panel Core Study Module - Environmental attitudes and behavior	bc, cc, dc

Beispiel Studie zu *Environmental attitudes and behavior*

Measured constructs/concepts and corresponding data collection waves

Constructs/concepts	Corresponding indicators (survey measures)	Data collection waves
Distance to next city	Großstadtnähe Wohngegend <i>Distance between residential area and large city</i>	bc, cc, dc
Subjective exposure to environmental hazards	Beeinträchtigung Umwelteinflüsse: Lärmbelästigung <i>Exposure to environmental hazards: noise pollution</i>	bc
	Beeinträchtigung Umwelteinflüsse: Luftverschmutzung <i>Exposure to environmental hazards: air pollution</i>	bc
	Beeinträchtigung Umwelteinflüsse: Fehlende Grünflächen	

Möglichkeiten Geodaten

- Kooperation mit dem Leibniz-Institut für ökologische Raumentwicklung



- Hier gibt es bspw. Indikatoren zu Nachhaltigkeit, Siedlung, Gebäuden, Verkehr etc.
- Es könnte also interessant sein, diese Daten an das Gesis Panel anzuspielen
- Aber dazu später mehr

R Nutzung in meinem Arbeitsalltag

- Datenaufbereitung

Was gibts für mich zu tun:

- Panelbereinigung (bei Abmeldung oder Nonresponse)
- Online- und Offline-Daten zusammenführen (Unified Design)
- Anonymisierung und Kategorisierung
- Missings kodieren; bspw. muss sich Filterführung in den Missings widerspiegeln
- Codebuch und Wellenreport erstellen
-

R in meinem Arbeitsalltag

- `foreign`, `readstata13` und `xlsx` zum Import von Daten
- Pakete `dplyr` und `tidyr` zur Datenaufbereitung
- `doParallel`, `foreach` und `doSNOW` zur Bearbeitung vieler Jobs
- `Rmarkdown` bei der Datendokumentation (Codebook, Wave Report)
- `caret` für maschinelles Lernen
- Rstudio git Interface zur Versionskontrolle

Arbeiten mit HTML Daten

• Cheatsheet zum Umgang mit Strings

```
560 <div id="questiontable"><div class="qt311"><div id="qnameq23076">
561 <table cellpadding="0" cellspacing="0" width="100%" border="0" bordercolor="green">
562 <tbody><tr>
563 <td class="questiontext"><a name="2"> </a>Wie ähnlich ist Ihnen diese Person?
564 </td>
565 </tr>
566 </tbody></table><br>
567 <table border="0" class="answertable" bordercolor="purple" cellspacing="0" cellpadding="0">
568 <tbody><tr><td class="scaletitle"></td>
569 <td class="answerscale">ist mir überhaupt nicht ähnlich<br>1</td>
570 <td class="answerscale">ist mir nicht ähnlich<br>2</td>
571 <td class="answerscale">ist mir nur ein wenig ähnlich<br>3</td>
572 <td class="answerscale">ist mir einigermaßen ähnlich<br>4</td>
573 <td class="answerscale">ist mir ähnlich<br>5</td>
574 <td class="answerscale">ist mir sehr ähnlich<br>6</td>
575 <td width="100%" class="scaletitle"></td>
576 </tr>
```

Genutzte Pakete für das Arbeiten mit Geodaten

- Paket `tmap` zur Erstellung thematischer Karten
- Paket `raster` um Rasterdaten zu verarbeiten und zum Transfer zwischen Koordinatenreferenzsystemen

Bei Gesis entwickelte Pakete

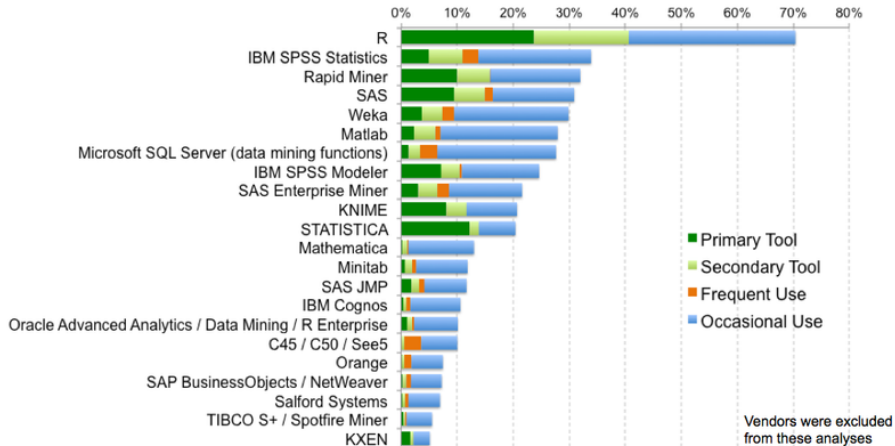
- Paket zur Nutzung der Zensus 2011 Daten

```
devtools::install_github("stefmue/georefum")
```


- Paket zur Nutzung der Overpass API um Daten von OpenStreetMap herunterzuladen

```
devtools::install_github("Japhilko/gosmd")
```

Welche Statistikpakete werden genutzt



Open Science

Das Prinzip „Open Science“ hat das Ziel, wissenschaftliche Abläufe offen zugänglich, nachvollziehbar und nutzbar zu machen. Dazu werden verschiedene Ansätze verfolgt, beispielsweise Open Access, **Open Source**, Citizen Science und Open Educational Resources. Wie verschiedene Stellungnahmen der Europäischen Union und der G7 zeigen, gewinnt Open Science auch auf europäischer und internationaler Ebene an wissenschaftspolitischer Bedeutung. Die Leibniz-Gemeinschaft und ihre Mitgliedseinrichtungen unterstützen diese Entwicklung und gestalten sie mit. So setzen sie sich beispielsweise seit vielen Jahren mit zahlreichen Aktivitäten für **Open Access** , den freien Zugang zu wissenschaftlichen Publikationen, ein.

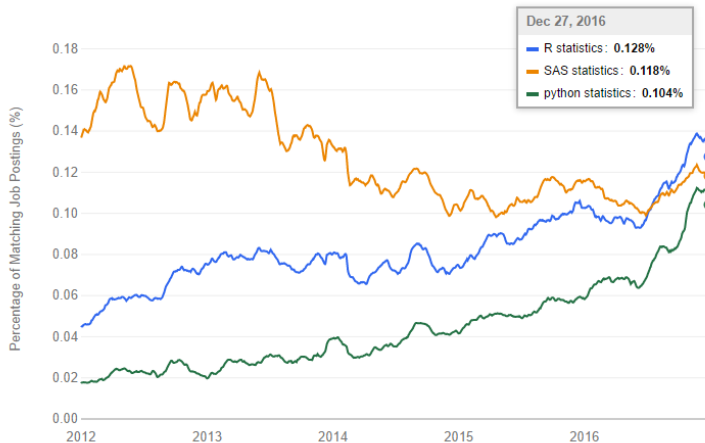
Trend Open Science - GESIS



Eine neue, strategisch wichtige Herausforderung für GESIS ist die Forschung zu kollaborativen und partizipativen Modellen und Infrastrukturen, die Open Science-Prozesse in den Sozialwissenschaften unterstützen.

Die Nennung von R in Stellenausschreibungen

Job Postings



Fazit - Zukunft von R in der Wissenschaft

- Insgesamt wird mehr quantitativ gearbeitet
- Bedeutung von SPSS nimmt ab
- Bedeutung von R scheint zu steigen - R wird auch mehr und mehr an Hochschulen eingesetzt
- Im Zuge der Open Science Entwicklung wird die Nutzung von R (bei Leibniz Instituten) immer mehr gefördert
- Stata ist nach wie vor wichtig (Pfadabhängigkeit)
- Der große Vorteil von R ist die Flexibilität und die große Nutzer Community