# Einführung in die Datenanalyse mit R - Datenquellen

Jan-Philipp Kolb

3 Mai 2017

# Datenzugang

- Public-Use-File (PUF) Datei zur öffentlichen Nutzung - meist stark anonymisierte Daten (Beispiele: FDZ, Statistik Portal, Meine Region )
- Scientific-Use-File (SUF) - Datei zur wissenschaftlichen Nutzung - anonymisierte Daten, die zu wissenschaftlichen Zwecken und zur Sekundäranalyse genutzt werden können.
- On-Site-Nutzung - Arbeitsplätze für Gastwissenschaftler - Kontrollierte Datenfernverarbeitung

# Datenquellen

- Auf dem Portal datahub.io sind sehr viele Beispieldatensätze in verschiedenen Formaten abrufbar.
- Weitere Portale: OpenGov, okfn, enigma, Amazon Web Services (AWS)
- Umweltdaten (National climatic data center)
- FAO Datenbank

```
library("FAOSTAT")
```

- Public Use File für Soziales in den USA Social security administration
- National health and nutrition examination survey

```
library(survey)
data(nhanes)
```

# Das R-Paket datasets

```
library(datasets)
```

Beispiel Erdbeben Datensatz:

```
head(quakes)
```

# Datensatz zum US Zensus

```
library(UScensus2010)
```

# Weltbank Daten

`WDI` - World Development Indicators (World Bank) - Einführung in das Paket

```
library(WDI)
```

```
WDIsearch('gdp')[1:10,]
```

# Nutzung von WDI Daten

```
dat <-  WDI(indicator='NY.GDP.PCAP.KD', country=c('MX','CA','U
head(dat)
```

# Erste Grafik mit WDI Daten

# OpenStreetMap

*OpenStreetMap (OSM) ist ein kollaboratives Projekt um eine editierbare Weltkarte zu erzeugen.*

Wikipedia - OpenStreetMap

# Download von OpenStreetMap Daten

```
library(osmar)
api <- osmsource_api()
library(ggmap)
```

```
cityC <- geocode("Berlin",source="google")
bb <- center_bbox(cityC$lon,cityC$lat,1000, 1000)
uaBerlin <- get_osm(bb, source = api)
```

- Ausschnitte von OpenStreetMap für einzelne Städte (metro extracts)
- Liste möglicher Datenquellen für räumliche Analysen (weltweit, Deutschland )
- SALB - Administrative Grenzen
- Kartendaten (openaprs)

# TwittR

```
library(twitteR)
library(streamR)
```

http://www.r-bloggers.com/
mapping-the-world-with-tweets-including-a-gif-without-cats-ar

# worldHires Daten

```
library(mapdata)
data(worldHiresMapEnv)
map('worldHires', col=1:10)
```

# Historische Daten

- Historischer Geocoder
- Paket HistData

```
library(HistData)
data(Arbuthnot)
```

# GDELT Daten

- GDELT
- Nutzung von GDELT Daten (Beispiel 1, Beispiel 2)

```
library(GDELTtools)
test.filter <- list(ActionGeo_ADM1Code=c("NI", "US"), ActionGe
test.results <- GetGDELT(start.date="1979-01-01", end.date="19
                         filter=test.filter)
```

# Andere Datenquellen

- Die US Flughäfen und Fluglinien
- Mehr Daten hier

```
link1 <- "http://openflights.svn.sourceforge.net/viewvc/openfl
openflights/data/airports.dat"
airport <- read.csv(link1, header = F)

link2 <- "http://openflights.svn.sourceforge.net/viewvc/openfl
openflights/data/routes.dat"
route <- read.csv(link2, header = F)
```

- Hafen Daten (Natural earth data)
- Minimalistische Karten
- Census results - Germany
- Census results - Britain and boundaries
- Data on airports and an example on the usage in R

# Weitere Quellen

- ICEDS European Data Server
- Mobilfunkdaten, CO2 Emmissionen
- Daten für New York (Daten, Beispiel