

Projektseminar: Konzeption quantitativer Forschung

Big Data!? — Webbasierte Datenerhebung, exploratives Data Mining und Critical Statistics für die Humangeographie

Teil 1: Wintersemester 2016/17

Jan Kemper <kemper@geo.uni-frankfurt.de>

Till Straube <straube@geo.uni-frankfurt.de>

Syllabus

Stand: 07.11.16

Materialien: <http://bit.do/bigdataffm>

Seminarkonzept

Unter dem Schlagwort „Big Data“ werden seit ca. einer Dekade im Allgemeinen die großen Datenmengen verstanden, die unter den Bedingungen vernetzt verlaufender Produktion, Distribution und Konsumtion sowie den dafür notwendigen Interaktionen (und hier insbesondere im Bereich der Telekommunikation, des elektronischen Zahlungsverkehrs, des Online-Handels etc.) produziert werden.

Das Seminar „Big Data!? — Webbasierte Datenerhebung, exploratives Data Mining und Critical Statistics für die Humangeographie“ stellt sich neuen Fragestellungen und Herausforderungen auf den Gebieten der Datenerhebung und -auswertung im Rahmen eines zweisemestrigen Projektseminars.

Im Zentrum steht der Kompetenzaufbau in der Extraktion von webbasierten Daten (Scraping), in der statistisch-explorativen Datenauswertung (Data Mining) sowie auf den Gebieten der kritischen Reflexion und der gesellschaftskritisch orientierten Anwendung techno-statistischer Komplexe (Critical Statistics).

Im ersten Teil des Projektseminars (Wintersemester 2016/17) erlernen die Studierenden den Umgang mit verschiedenen Datenformaten und Tools zur Datenerhebung. Im zweiten Teil (Sommersemester 2017) werden die Teilnehmenden ein eigenes Projekt entlang der Prozesskette Datenerhebung, Datenaufbereitung, -auswertung und Interpretation vor dem Hintergrund einer eigenständig entwickelten Problemstellung durchführen. Die Softwareumgebung R ist die zentrale Grundlage für beide Seminarteile. Eine begleitende kritische Reflexion der vermittelten Kenntnisse anhand von Textarbeit findet durchgehend statt.

Termine

17. Okt 2016	Einführungssitzung
24. Okt 2016	[Quergelesen] Anderson 2007: The End of Theory
31. Okt 2016	[Vorbereitung] Level 1
7. Nov 2016	[Nachbesprechung] Level 1
14. Nov 2016	[Quergelesen] Kitchin 2014: Big Data
21. Nov 2016	[Vorbereitung] Level 2
28. Nov 2016	[Nachbesprechung] Level 2
5. Dez 2016	[Quergelesen] O'Sullivan 2017: Big Data
12. Dez 2016	[Halbzeit] Reflexion und Feedback
19. Dez 2016	[Vorbereitung] Level 3
9. Jan 2017	[Nachbesprechung] Level 3
16. Jan 2017	[Lektüre] Amoores 2011: Data Derivatives
23. Jan 2017	[Vorbereitung] Level 4
30. Jan 2017	[Nachbesprechung] Level 4
6. Feb 2017	Abschlussitzung Teil 1, Vorbereitung Teil 2 (Sommersemester 2017)

Texte

- Anderson, Chris. 2007. *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. Online verfügbar unter: http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory (aufgerufen am 8. Mai 2015).
- Kitchin, Rob. 2014. Big Data. In: *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. London: Sage, S. 67–79.
- O'Sullivan, David. 2017. *Big Data: Why (oh why?) This Computational Social Science?* Manuskript. Online verfügbar unter: <http://escholarship.org/uc/item/0rn5n832> (aufgerufen am 12. Oktober 2016).
- Amoores, Louise. 2011. Data Derivatives: On the Emergence of a Security Risk Calculus for Our Times. In: *Theory, Culture*

Leistungsanforderungen

Am Ende des Wintersemesters sollen die Studierenden ein individuelles Portfolio der bearbeiteten Aufgaben abgeben. Hierbei soll die Bearbeitung der einzelnen Aufgaben jeweils im folgenden Format prozesshaft dokumentiert werden:

- Aufgabenstellung
- geplante Schritte zur Bearbeitung
- Probleme bei der Bearbeitung und deren Lösung
- Ergebnis
- ggf. sonstige Beobachtungen

Aufgaben

Die genauen Aufgabenstellungen werden jeweils in der Vorberechnung des jeweiligen ‚Levels‘ bekanntgegeben und ausgeführt.

Level 1

- Suchen Sie sich einen Datensatz (am besten Rohdaten mit mehreren Variablen und unterschiedlichen Messniveaus).
- Importieren Sie die Daten in RStudio.
- Bestimmen Sie Lage- und Streuungsmaße für mind. eine ausgesuchte Variable.

Level 2

- Datensatz per API in R einlesen
- Einfache statistische Analyse mit R

Level 3

- Daten über API-Schnittstelle abrufen und Datensatz generieren
- Statistische Analyse mit R

Level 4

- Daten aus nicht-API Quelle automatisiert einlesen („scraping“)
- Fortgeschrittene statistische Analyse mit R

Ressourcen

Datenportale

Aggregiert:

- <http://dataportals.org/>
- <http://www.forbes.com/sites/bernardmarr/2016/02/12/big-data-35-brilliant-and-free-data-sources-for-2016/>

Einige Portale mit besonders „guten“ Rohdaten für unsere Zwecke:

- <https://data.ny.gov/>
- <https://data.gov.uk/>
- <https://www.healthdata.gov/>
- <https://www.data.gov/>

Lokal relevante Daten (leider nicht besonders gut):

- <http://www.offenedaten.frankfurt.de/>
- <https://www.govdata.de/>
- <http://open-data.europa.eu/en/data/>

Anleitungen und Dokumentation von R und RStudio

- Schneller Überblick über RStudio-Funktionen:
<https://www.rstudio.com/resources/cheatsheets/>
- Detailliertes Tutorial für R:
<https://www.tutorialspoint.com/r/index.htm>
- R schrittweise online ausprobieren:
<http://tryr.codeschool.com/>
- Ein Script für Anfänger*innen:
www.stat.wisc.edu/~st571-1/BeginnersRcourse.pdf
- Weitere Adressen:
<https://cran.r-project.org/manuals.html>

Statistik

- *Zur deskriptiven Statistik:*
Cornelia Weins, Uni- und bivariate Statistik, in: Handbuch der sozialwissenschaftlichen Datenanalyse, hrsg. v. C. Wolf u. H. Best, Wiesbaden: Verlag f. Sozialwissenschaften 2010, S. 65–89, hier Kapitel 1: »Univariate Statistik«.
- *Zur graphischen Datenexploration:*
Horst Degen, Graphische Datenexploration, in: Handbuch der sozialwissenschaftlichen Datenanalyse, hrsg. v. C. Wolf u. H. Best, Wiesbaden: Verlag f. Sozialwissenschaften 2010, S. 91–116.