International Program in Survey and Data Science: An environment for training and cooperation

Frauke Kreuter, Universities of Mannheim and Maryland
Source: Roberto Rigobon, Discussion on Applications and Issues with Using Commercial Data in Research, BEA Expert Meeting on Exploiting Commercial Data for Official Economic Statistics November 19, 2015
Survey AND Data Science
VINCENT VAN GOGH
Credit: Ralph Klüber, p3 Insights
VINCENT VAN GOGH

Credit: Ralph Klüber, p3 Insights
VINCENT VAN GOGH

Café Terrace At Night

Designed Product

Credit: Ralph Klüber, p3 Insights
Using Municipal Wastewater to Estimate Cannabis Consumption
Statistics Canada 2019
Andrew Brennan, Laurie Reedman, Geneviève Vézina, Jack Gambino

https://www150.statcan.gc.ca/n1/daily-quotidien/181129/dq181129d-eng.htm
US Aggregated Inflation Series, Monthly Rate, PriceStats Index vs. Official CPI.
To Make it Happen
Data Generating Process

Construct

Measurement

Response

Edited Response

Survey Statistic

Population Mean

Coverage

Sampling Frame

Sampling

Sample

Nonresponse

Respondents

Adjustment

Postsurvey Adjusted Data

Groves et al. 2004
Big Data Process Map

Generate

- Source 1
- Source 2
- Source M

ETL

- Extract
- Transform (Cleanse)
- Load (Store)

Analyze

- Filter/Reduction (Sampling)
- Computation/Analysis (Visualization)
AAPOR Report on Big Data
AAPOR Big Data Task Force
February 11, 2015

INNOVATIONS IN FEDERAL STATISTICS
Combining Data Sources While Protecting Privacy

Prepared for AAPOR Council by the Task Force, with Task Force members including:

Lennart, Co-Chair, Statistics Sweden
Frankie Kreuter, Co-Chair, IPUMS at the U. of Maryland, U. of Minnesota & IAB
Marcus Berg, Stockholm University
Paul Breuer, BIT International
Paul Drucker, Mathematica Policy Research
Cory Lane, School of Information at the University of Michigan
Julia Lane, American Institutes for Research
Carlo O’Neill, Johnson Research Labs
Abe Osher, Hanover Group

Acknowledgement: We are grateful for comments, feedback and editorial help fromExact Data Analytics, John McMillan, and the AAPOR council members.

BIG DATA AND SOCIAL SCIENCE
A Practical Guide to Methods and Tools

Edited by
Ian Foster, Rayid Ghani,
Ron S. Jarmin, Frauke Kreuter,
and Julia Lane

Chapman & Hall/CRC
Statistics in the Social and Behavioral Sciences Series
Education Infrastructure
Introduction to International Program in Survey and Data Science

The project on which this report is based was funded by the Federal Ministry of Education and Research under the number [16OH22064]. Responsibility for the contents of this publication lies with the author.
International Faculty from Partner Universities

- University of Mannheim
- University of Maryland
- University of Michigan
- The University of Manchester
- Utrecht University
- LMU Ludwig-Maximilians-Universität München
- Australian National University
- BGSU
- Pontificia Universidad Católica de Chile
International Faculty from the Industry
IPSDS (Test) Cohorts

- 100% are working professionals

- 57 Participants
  (32 f + 25 m)

  - 22 countries of residence

  - Age: median=31 (min=22; max=61)
Learning objectives / Modules

- Data Generating Process
- Data Curation/Storage
- Data Analysis
- Data Output/Access
- Research Question

Learn how to communicate results and distribute and store your data.

Learn a variety of analysis methods suited for different data types.

Learn how to curate and manage data.

Understand how to collect data yourself, and how data are generated through administrative and other processes.

Learn how to formulate your research goal and which data are best suited to achieve it.

Source: Usher in Japec et al 2015
Data Generating Process

- Fundamentals of Survey and Data Science: 3 credits/6 ECTS

Data Curation/Storage

- Database Management I-III: 1 credit/2 ECTS each
- Data Munging I-III: 1 credit/2 ECTS each
- Python SQL: 1 credit/2 ECTS
- Record Linkage: 1 credit/2 ECTS
- Imputation I-II: 1 credit/2 ECTS each

Data Analysis

- Generalized Linear Models: 2 credits/3 ECTS
- Analysis of Complex Data I-III: 1 credit/2 ECTS each
- Propensity Score/Statistical Matching: 2 credits/4 ECTS
- Machine Learning I-III: 1 credit/2 ECTS each
- Text Analysis I-II: 1 credit/2 ECTS each

Data Output/Access

- Ethics: 1 credit/2 ECTS
- Data Confidentiality and Statistical Disclosure Control: 2 credits/4 ECTS
- Visualization: 2 credits/4 ECTS
- Consulting: 1 credit/2 ECTS each

Research Question

- Data Collection Courses: 1 credit/2 ECTS each
- Practical Tools for Sampling and Weighting: 3 credits/6 ECTS
- Applied Sampling I-III: 1 credit/2 ECTS each
- Experimental Design: 2 credits/4 ECTS
- User Experience: 1 credit/2 ECTS

Total: 75 ECTS
Master Thesis: 15 ECTS
Flexible & engaging online learning environment

• Online learning environment accessible from anywhere in the world (taught in English)

• Small virtual classroom with a mix of synchronous & asynchronous learning
  ▶ Pre-recorded lectures split into small video units
  ▶ Required readings and (bi)weekly assignments
  ▶ Discussion forums
  ▶ Weekly online meetings

  \[8-10h \text{ per week}\]

• Annual on-site networking activity with fellow students from five continents

• Wide variety of options: from individual courses or course sequences to a modular program

• Most courses run 4 weeks or 8 weeks
How to interact?

**Synchronous**

- Reduces social isolation
- Questions answered right away

**Asynchronous interaction**

- Flexibility
- More time to think

---

Privacy of self-administered modes when doing q’s in public
by [User] on Tuesday, 7 June 2016, 3:48 PM

I was wondering if self-administered questionnaires when done in public (public transport, at the park, …) can still be considered highly private? I’d rather assume that filling out a questionnaire in public leads to a feeling of low privacy and external factors like sex or race of the people surrounding the respondent are likely to alter his response behaviour.

Re: Privacy of self-administered modes when doing q’s in public
by [User] on Sunday, 12 June 2016, 7:23 PM

Great point. I agree that if people feel that questions are sensitive, they may just decide not to do the survey at all. I think self-administering the survey using an iPad would help. The respondent wouldn’t have to say the answer out loud and the interviewer couldn’t see the answers provided when finished (compared to if they were given a paper/pencil survey.)
How to structure the material?

- Expert interviews show practical application
Lessons Learned

- Modular approach much appreciated by working professionals
- Guidance necessary for working adults
- Learning with application at hand is key
- Peer-to-peer learning enhances engagement

- Hardest to learn and hardest to teach: Asking the right question!
Next Steps
(1) Hands-on through Coleridge Initiative
TRAINING PROGRAM

The Applied Data Analytics programs are targeted at government agency staff. It provides training in core data analytics techniques by working on specific projects using real-world micro-data. The projects are built around pre-built Jupyter notebooks which provide project "recipes" that can be customized for specific use cases as well as applied to later projects in participants' agencies.

https://coleridgeinitiative.org/training
Networks: The first two classes brought together ~40 agencies from city, state, county and federal agencies
Big Data for Federal Agencies

- Fall course: 25 students
- Curriculum = book outline
- Outlook
  - One-stop enrollment
  - Engagement of PI/PR
(2) Education Partnerships
Exchange Instructor Time for Seats

Memorandum of Understanding – University of Maryland and <PARTNER>

- four Course Credits for every one-credit course taught by <PARTNER> instructor;
- six Course Credits for every two-credit course taught by <PARTNER> instructor;
Key Ingredients

- Sufficient instructors to cover all time zones
- Sufficient funding to finance the core administrative infrastructure
- Partner with professional organization(s) for outreach and logistics
Thank you!

ipsds@uni-mannheim.de
survey-data-science.net
Admissions – Courses - Cost
Who should apply?

IPSDS is designed for professionals working with data collection and data analysis.

Admission Requirements

• Academic degree (min. Bachelor’s degree)
• At least 12 ECTS in mathematical/applied statistics
• At least one year of work experience in a position working with data
• English proficiency

Fees

19 courses offered for free in 2019-2020
8 paid courses (750 EUR per 1 credit/2 ECTS)
Free courses offering 2019/2020

- Analysis of Complex Survey Data, 2 cr./4 ECTS
- Big Data and Machine Learning, 1 cr./2 ECTS
- Computer-Based Content Analysis I, 1 cr./2 ECTS
- Computer-Based Content Analysis II, 1 cr./2 ECTS
- Data Collection Methods, 3 cr./6 ECTS
- Experimental Design for Surveys, 2 cr./4 ECTS
- Fundamentals of Survey and Data Science, 3 credits/6 ECTS
- Generalized Linear Models, 2 cr./4 ECTS
- Inference from Complex Surveys, 2 cr./4 ECTS
- Introduction to Data Visualization, 1 cr./2 ECTS
- Introduction to Python and SQL, 1 cr./2 ECTS
- Introduction to Real World Data Management, 2 cr./4 ECTS
- Introduction to Small Area Estimation, 2 cr./4 ECTS
- Practical Tools for Sampling & Weighting, 2 cr./4 ECTS
- Privacy Law, 1 cr./2 ECTS
- Project Consulting, 6 cr./12 ECTS
- Questionnaire Design, 2 cr./4 ECTS
- Review of Statistical Concepts (bridge course)
- Web Survey Methodology, 2 cr./4 ECTS
Paid courses offering 2019/2020

750 EUR per 1 credit/2 ECTS

- Applied Sampling (Sampling I), 2 credits/4 ECTS
- Data Confidentiality and Statistical Disclosure Control, 2 credits/4 ECTS
- Introduction to Record Linkage with Big Data Application, 1 credit/2 ECTS
- Item Nonresponse and Imputation, 1 credit/2 ECTS
- Measurement Error Models, 1 credit/2 ECTS
- Multiple Imputation - Why and How, 1 credit/2 ECTS
- Usability Testing for Survey Research, 1 credit/2 ECTS
- Web Scraping and API, 1 credit/2 ECTS