Bootcamp on Administrative Data Analysis (BADA), Summer 2024

Bootcamp en analyse de données administratives (BADA), été 2024

Course Outline

Version: November 8, 2024
[Subject to change]

An initiative of the *Quebec Interuniversity Centre on Social Statistics* (QICSS) and the *Social Statistics Study Group (SSSG)* at INRS

Course development and training: Xavier St-Denis, PhD (INRS)
Contact: xavier.st-denis@inrs.ca

Location: QICSS Université de Montréal (workshops in person in the lab only).
3535 chemin Queen-Mary, room 420
Montréal (Québec)
H3V 1H8
(Métro Côte-des-Neiges)

Dates: June 9-13, 2024.

Languages: English and French (documentation and presentations will be primarily in English)

Table of Contents

A. DESCRIPTION 3

B. OBJECTIVES 4

C. COURSE STRUCTURE 4

D. PREREQUISITES AND APPLICATION PROCESS 4

E. CONTENT 4

F. PROGRAM 5

DAY 1: ADVANCED STATA PROGRAMMING FOR LONGITUDINAL DATA ANALYSIS APPLIED TO STATISTICS CANADA ADMINISTRATIVE DATA 5

DAY 2: INTRODUCTION TO ADMINISTRATIVE DATA ANALYSIS WITH A FOCUS ON INCOME DATA 6

DAY 3: STUDYING FAMILIES USING TAX DATA 7

DAY 4: RECORD LINKAGES AND DATA INTEGRATION 8

DAY 5: INNOVATION WITH DATA INTEGRATION AND INDIRECT MEASUREMENT 9

A. Description

The Bootcamp in Administrative Data Analysis (BADA) is a summer school intended for graduate students and university or non-university researchers wishing to carry out analyzes with administrative or linked data from Statistics Canada.

The summer school aims to allow participants to develop a familiarity with the nature and content of the main administrative databases available in Statistics Canada's Research Data Centres (RDCs). It will also allow participants to master analysis and programming techniques adapted to this type of data and to the RDC environment.

Participants will develop familiarity with the main benefits and challenges related to the analysis of massive Canadian administrative datasets and will then be able to apply this learning in their own projects. In particular, administrative data have a longitudinal structure. This aspect will receive particular attention within the framework of BADA.

The summer school will focus on three types of data: personal tax data such as those of the LAD and linkages with other administrative databases; linkages between survey data and administrative data such as those of the GSS, LISA, CCHS, or those of the Extending the Relevance of Longitudinal Files project (SLID, NLSCY, YITS, etc.); and finally program data (mainly education data from the PSIS and RAIS and immigration data from the IMDB), including their integration with tax data and more complex linkage platforms. Business and healthcare data will not be discussed extensively.

The software used in the laboratory will be Stata. An intermediate-level familiarity with programming in Stata is necessary for workshops and practical exercises in the laboratory. For example, participants should have already used Stata for the final project of a graduate-level methods class or in their own research. Participants should be familiar with the use of written commands (do-files, programs, etc.), not just the use of drop-down menus. Alternatively, beginner-level familiarity with Stata is sufficient for participants who have experience analyzing quantitative data with other programming-based software (R, SAS, or python).

Attention: This course is not designed for people who have never done a project based on quantitative methods and have no experience manipulating data with lines of code. For participants interested in learning more about the content of Canadian administrative datasets but who do not have the technical pre-requisite, we recommend registering for the morning sessions only rather than the full bootcamp that also includes afternoon workshops in the lab.

B. Objectives

1. Develop an advanced knowledge of the content of the main Canadian administrative datasets.
2. Identify the potential and limitations of administrative data for social science research.
3. Develop data manipulation techniques and analytical methods applicable to administrative data using Stata (with possible complements in SAS and R).
4. Develop longitudinal data analysis skills.

C. Course structure

1. Advanced Stata training workshop: Programming for longitudinal data analysis (Day 1: AM & PM)
2. In-class training and guest speakers (Day 2 to 5: AM)
3. Practical data analysis workshops in the lab (Day 2 to 5: PM)

D. Prerequisites and application process

1. Experience in microdata analysis using Stata, R, SAS, or python (a basic familiarity with Stata is recommended).
2. Having successfully completed a quantitative methods course (advanced undergraduate or graduate-level).
3. Being a researcher on an active project at any Canadian RDC before the beginning of the summer school or having been a researcher on a project that has ended no more than one year before the beginning of the summer school (see https://www.statcan.gc.ca/fr/microdonnees/centres-donnees/acces; for details or support, communicate with acces@ciqss.org).
4. Understand English (spoken and written). The material will be in English. Presentations will be in English or French depending on group preferences; individual interactions with the instructor will be in the participant's chosen language.

Application and registration

1. Complete the online form on the QICSS website.
2. Funding is available to support travel costs for students who are affiliated with a QICSS member institution (contact Luc St-Pierre for more information).

E. Content

The content of in-class sessions and lab workshops, including recommended readings and lab exercises, is included in separate documents also shared with participants. Participants should refer to these documents for all details related to content.

F. Program

Day 1: Advanced Stata programming for the analysis of massive and longitudinal administrative datasets

This full day of technical training is designed for individuals with a working knowledge of Stata who wish to develop advanced skills necessary for the analysis of large and complex longitudinal datasets. Specifically, this Stata programming training will benefit researchers who wish to analyse Canadian administrative datasets and related record linkages. The workshop will provide detailed examples of Stata codes as well as exercises that will facilitate data preparation, data management, and the completion of data validation and description steps necessary before performing regression analysis.

Prerequisite: Prior knowledge of Stata is necessary. This activity will not include an introduction to Stata.

Schedule:

9:00am	Arrival of participants
9:15am	BADA 2024 starts! Word of welcome from the instructor and the QICSS staff
9:30am	Introduction of participants; overview of the objectives, course outline, and basic rules
9:45am	Pause
10:00am	Workshop, part 1: Basics of programming and project management in Stata; generating descriptive output efficiently
12:00pm	Lunch
1:00pm	Workshop, part 2: Structure of administrative datasets and commands for key operations
2:30pm	Pause
2:45pm	Workshop, part 3: Managing longitudinal administrative data and creating your analytical sample
4:45pm	End of day 0

Main datasets discussed

- Synthetic administrative data from StatCan's microsimulation models

Day 2: Introduction to administrative data analysis with a focus on income data

In this first session, we will introduce key concepts and basic features of administrative data, with a focus on Canadian tax data such as the Longitudinal Administrative Databank (LAD) and other datasets derived from the T1 Family Files (T1FF). Emphasis will be put on the longitudinal and relational nature of the data. The session will conclude with a detailed description of variables on income, program participation, and wealth in tax data.

Schedule

9:00am	In-class training Section 1. What is administrative data Section 2. Longitudinal and hierarchical dimensions of admin data Section 3. Income and wealth in tax data
10:45am	Pause
11:00am	Guest lecture
12:00pm	Lunch
1:00pm	Lab segment, part 1: How to structure a collaborative/team project
2:30pm	Pause
2:45pm	Lab segment, part 2: Who is in the LAD and what is their income?
4:45pm	End of day 1

Main datasets discussed

- Datasets derived from the T1FF
- Longitudinal Administrative Database (LAD)
- T4 and other tax slips with information related to income and wealth

Day 3: Studying families using tax data

Canadian tax data includes detailed information on family structure and family relationships. In this session based on recent research in demography, sociology and economics, we will discuss what types of family dynamics can be observed cross-sectionally and longitudinally in Canadian administrative data, with a focus on data derived from personal tax records. Topics include couple dissolution and reconstituted families, same-sex couples, measures of fertility, and extended kinship networks.

Schedule:

9:00am	Day 2 starts! In-class training Introduction to T1 Family Files (T1FF) Couples in tax data Direct and indirect observation of children and fertility Kinship beyond the nuclear family in tax data
10:45am	Pause
11:00am	Guest lecture
12:00pm	Lunch
1:00pm	Lab segment, part 1: Family structure and relationship variables
2:30pm	Pause
2:45pm	Lab segment, part 2: Family identifiers and relationship matrices
4:45pm	End of day 1

Main datasets discussed

- Datasets derived from the T1FF
- Longitudinal Administrative Database (LAD)

Day 4: Record linkages and data integration

This session will allow participants to develop a familiarity with record linkage approaches. The main characteristics, properties, and limitations of record linkages will be presented, with a focus on linkages between survey and tax data, federal immigration data (IMDB) and tax data, and external program data integration. The session will also include practical examples guiding participants who wish to perform and validate the quality of record linkages.

Schedule:

9:00am	Day 3 starts! In-class training Introduction to record linkages Three real-world approaches to performing record linkages Federal immigration data and tax data in the IMDB Survey-tax data linkages in the LISA External and multijurisdictional data in the ELMLP
10:45am	Pause
11:00am	Guest lecture
12:00pm	Lunch
1:00pm	Lab segment, part 1: Performing and validating a record linkage
2:30pm	Pause
2:45pm	Lab segment, part 2: Longitudinal analysis with a linked dataset
4:45pm	End of day 3

Main datasets discussed:

- Longitudinal Immigration Database (IMDB)
- Longitudinal and International Study of Adults (LISA) linked with T1FF and other administrative datasets
- General Social Survey (GSS) linked with the T1FF
- CCHS linked with the T1FF
- Education and Labour Market Longitudinal Platform (ELMLP) including the Postsecondary Education Information System (PSIS), the Registered Apprenticeships Information System (RAIS), and provincial K-12 data (British Columbia and other)

Day 5: Innovation with data integration and indirect measurement

This session aims to present two types of innovations and challenges. First, it will introduce various linkage platforms that integrate a large number of administrative, survey, and Census datasets, including in some cases provincial data. The potential and limitations of these platforms will be discussed, with a focus on the Education and Labour Market Longitudinal Platform (ELMLP) as well as mention of the CEEDD/BEAM business-level data and CanCHEC health data. Second, it will discuss strategies used in published research aiming to measure factors and dimensions that do not appear even in the most complex linkage platforms. This includes information on geographic mobility, employment characteristics, and emigration.

9:00am	Day 4 starts! In-class training Innovations in data integration: Linkage platforms Multijurisdictional data Challenges of tax data Indirect measures Measurement error
10:45am	Pause
11:00am	Guest lecture
12:00pm	Lunch
1:00pm	Lab segment 1: Indirect measures and validation
2:30pm	Pause
2:45pm	Lab segment 2: Conduct your own analysis!
4:30pm	Final words
4:45pm	End of day 4

Main datasets discussed

- Education and Labour Market Longitudinal Platform (ELMLP) including the Postsecondary Education Information System (PSIS), the Registered Apprenticeships Information System (RAIS), and provincial K-12 data (British Columbia and other)
- Longitudinal Immigration Statistical Environment (LISE) including the IMDB, LAD and DAD
- Datasets derived from the T1FF