INRS
Institut national
de la recherche
scientifique

CIQSS | QICSS
CENTRE INTERUNIVERSITAIRE QUÉBÉCOIS DE STATISTIQUES SOCIALES
QUEBEC INTER-UNIVERSITY CENTRE FOR SOCIAL STATISTICS

GESS Groupe d'étude en statistique sociale
SSSG Social Statistics Study Group

Bootcamp on Administrative Data Analysis (BADA), Summer 2024
Bootcamp en analyse de données administratives (BADA), été 2024

Course Outline
Version: October 25, 2023

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 10-14, 2024.

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

Table of Contents

A. Description

The 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 challenges and issues related to the analysis of Canadian administrative data 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). Business register data and healthcare system data will not be discussed extensively.

The software used in the laboratory will be Stata. A basic familiarity with Stata is therefore recommended for workshops and practical exercises in the laboratory (for example, having already spent a few hours using Stata for a project or a course). Examples in R and SAS may also be provided on an ad hoc basis.

Attention: Participants should have previous experience analyzing quantitative data with programming-based software (Stata, R, or SAS). This course is not designed for people who have never done a project based on quantitative methods or have no experience of manipulating microdata with lines of code. For example, participants must have already completed a project as part of a methodology course, a dissertation, or a research internship; participants should be familiar with the functions of written commands (do-files, programs, etc., not just from drop-down menus) to recode variables or produce crosstabs.

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 0: AM & PM)
2. In-class training and guest speakers (Day 1 to 4: AM)
3. Practical data analysis workshops in the lab (Day 1 to 4: PM)

D. Prerequisites and application process

1. Experience in microdata analysis using Stata, R, ou SAS (a basic familiarity with Stata is recommended; a refresher will be offered on the day prior to the summer school).
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 based outside of Montreal and who are affiliated with a QICSS member institution (contact Luc St-Pierre for more information).

E. Evaluation model in order to receive credits (through INRS)

Participants who are registered in a graduate-level program at a Quebec university may request an Interuniversity Credit Transfer through the *Bureau de Coopération Interuniversitaire, BCI* (via an *Autorisation d'études hors établissement*, or *AEHE*). The course is credited at the instructor's home institution, INRS, and has the following code: POP 8441. However, it is the responsibility of the participants to complete the procedure and get necessary approvals from their home institution. Participants will have to pay tuition at their home institution but are exempted from paying the QICSS summer school participation fees.

Participants who are not registered at a Quebec university at the time of the summer school do not have the option to receive credits.

Evaluation

Participants seeking credits will be requested to complete the following work:
1. Attend all the sessions from Day 0 to 4 and participate actively in discussions (10%).
2. Complete all the lab exercises by the end of the BADA 2024 (30%).
3. Produce an analytical paper using administrative data available in the RDC (60%).

Analytical paper

Participants will have three weeks after the end of BADA 2024 to produce an analysis and write a paper *in the RDC*. They must therefore make sure that they maintain physical access to an RDC, with a preference for the Université de Montréal RDC for purposes of processing speed.

Objectives:
1. Develop a feasible research question that can be answered using administrative data available in the RDC (to be submitted on Day 3 of BADA at the latest).
2. Complete the necessary analysis using techniques and methods learned in class.
3. Submit a 3000 words paper that answers their research question, including descriptive output and regression estimates using the selected administrative data. The paper must include an introduction framing the research question, a brief review of the literature with a minimum of 3 to 6 empirical references, a detailed data and methods section, a precise results section presenting the results and interpreting them, and a conclusion section summarizing the results and discussing them (to be submitted at the end of the Friday three week after the end of BADA 2024).

Note that the code used to produce the paper can be transferred outside of the RDC, but not the paper or the output.

Participants who are not seeking credits can also produce that analysis and a certificate will be emitted recognizing their accomplishment.

F. Program

Day 0: Advanced Stata training workshop: Programming for longitudinal data analysis

Schedule overview:

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
12:00pm	Lunch
1:00pm	Workshop, part 2
2:30pm	Pause
2:45pm	Workshop, part 3
4:45pm	End of day 0

Overview of key content:

1. Foundations of programming in Stata: macros, loops, programs, returned values, and saved matrices.
2. Formatting longitudinal data properly for analysis in Stata and associated programming strategies and commands (merge/append, bysort, egen, reshape, _n).
3. Producing descriptive output from longitudinal data in Stata.

Day 1: Introduction to administrative data analysis and its basic concepts

Schedule overview:

9:00am	In-class training
10:45am	Pause
11:45am	Guest lecture
12:00pm	Lunch
1:00pm	Lab segment, part 1
2:30pm	Pause
2:45pm	Lab segment, part 2
4:45pm	End of day 1

In-class training

Overview of key content:

1. Introduction: Administrative data vs survey data vs « big data »
2. What are the different types of administrative data and the process behind administrative data generation?

3. Basic concepts in administrative data, with a focus on tax data: Family structure, income, tax information, geography
4. Administrative data as longitudinal data: a refresher
5. The LAD structure
6. The structure of other tax and administrative datasets
7. Formulating your research question for the week

Lab session

1. How to structure a collaborative/team project in the QICSS
2. Methodological refresher on longitudinal and multilevel data as applied to administrative data
3. Longitudinal dataset construction and management
4. Assessing data quality and exploring the properties of longitudinal files

Relevant literature and documentation

Recommended:

Statistics Canada. 2022. *Longitudinal Administrative Databank Data Dictionary 2020*. Ottawa: Statistics Canada. https://www150.statcan.gc.ca/n1/pub/12-585-x/12-585-x2022001-eng.pdf

Schmutte, Ian A, and Lars Vilhuber. "Balancing Privacy and Data Usability: An Overview of Disclosure Avoidance Methods." In *Handbook on Using Administrative Data for Research and Evidence-Based Policy*, edited by Shawn Cole, Iqbal Dhaliwal, Anja Sautmann, and Lars Vilhuber, 145–72, 2021. https://doi.org/10.31485/admindatahandbook.1.0.

Complement:

Chetty, Raj, and John N. Friedman. "A Practical Method to Reduce Privacy Loss When Disclosing Statistics Based on Small Samples." Working Paper. Working Paper Series. National Bureau of Economic Research, March 2019. https://doi.org/10.3386/w25626.

Day 2: Studying families and incomes using tax data

Schedule overview:

9:00am	Day 2 starts! In-class training
10:45am	Pause
11:00am	Guest lecture
12:00pm	Lunch

1:00pm	Lab segment, part 1
2:30pm	Pause
2:45pm	Lab segment, part 2
4:45pm	End of day 1

In-class training

Overview of key content:

1. Introduction to T1 Family Files (T1FF) data
    a. Family concept and identifier
    b. Variables in the T1FF, the LAD, and other major T1FF-derived datasets
        i. See summary table at the end of this document
    c. Potential, challenges and issues when using T1FF data
    d. Individual, couple, and family-level income data

2. Select case studies
    a. Case no1: Couple formation and dissolution in administrative data
    b. Case no2: Beyond spouses and couples: Parents, siblings, children, and grand-parents
    c. Case no3: Tax records to study income inequality in Canada

3. Guest lecture by Antoine Genest-Grégoire (Université de Sherbrooke), “What can income data from tax records tell us: A fiscal studies and public policy perspective”

Lab session

1. Basics of family analysis
2. Income at the individual or family level: the case of the low-income rate
3. Advanced topic: the relationship matrix (if time allows)

Relevant literature and documentation

Recommended:

Margolis, Rachel, Youjin Choi, Feng Hou, and Michael Haan. “Capturing Trends in Canadian Divorce in an Era without Vital Statistics.” *Demographic Research* 41 (December 20, 2019): 1453–78. https://doi.org/10.4054/DemRes.2019.41.52.

Saez, Emmanuel, and Michael R. Veall. “The Evolution of High Incomes in Northern America: Lessons from Canadian Evidence.” *The American Economic Review* 95, no. 3 (2005): 831–49. https://doi.org/10.2307/4132743.

Zhang, Xuelin. “What Can We Learn about Low-Income Dynamics in Canada from the Longitudinal Administrative Databank?” Ottawa: Statistics Canada, December 19, 2014.

See also an update: Zhang, Xuelin. “Low-Income Persistence in Canada and the Provinces.” Ottawa: Statistics Canada (Catalogue no. 75F0002M), June 11, 2021.

For the LIM exercise: Statistics Canada. “Low Income Lines: What They Are and How They Are Created.” Income Research Paper Series. Ottawa: Statistics Canada (Catalogue no. 75F0002M — No. 002), July 8, 2016. https://www150.statcan.gc.ca/n1/pub/75f0002m/75f0002m2016002-eng.pdf

Complement:

Chen, Wen-Hao, Yuri Ostrovsky, and Patrizio Piraino. “Lifecycle Variation, Errors-in-Variables Bias and Nonlinearities in Intergenerational Income Transmission: New Evidence from Canada.” *Labour Economics* 44 (January 1, 2017): 1–12. https://doi.org/10.1016/j.labeco.2016.09.008.

Connolly, Marie, Catherine Haeck, and David Lapierre. “Trends in Intergenerational Income Mobility and Income Inequality in Canada.” Analytical Studies Branch Research Paper Series. Ottawa, ON: Statistics Canada Catalogue no. 11F0019M — No. 458, 2021. https://www150.statcan.gc.ca/n1/en/pub/11f0019m/11f0019m2021001-eng.pdf?st=Ik4p7M3y.

Frenette, Marc, David A. Green, and Kevin Milligan. “The Tale of the Tails: Canadian Income Inequality in the 1980s and 1990s.” *Canadian Journal of Economics/Revue Canadienne d’économique* 40, no. 3 (August 1, 2007): 734–64. https://doi.org/10.1111/j.1365-2966.2007.00429.x.

Veall, Michael R. “Top Income Shares in Canada: Recent Trends and Policy Implications.” *The Canadian Journal of Economics. Revue Canadienne D’Economique* 45, no. 4 (November 2012): 1247–72. https://doi.org/10.1111/j.1540-5982.2012.01744.x.

Day 3: Record linkages and data integration

Schedule overview:

9:00am	Day 3 starts! In-class training
10:45am	Pause
11:00am	Guest lecture
12:00pm	Lunch
1:00pm	Lab segment, part 1
2:30pm	Pause

2:45pm Lab segment, part 2
4:45pm End of day 3

In-class training

Overview of key content:

1. Introduction to record linkages
2. Case no1: Record linkages between survey data and T1FF data
3. Case no2: Record linkage between the T1FF and other administrative datasets
4. Case no3: Linkage between external program data and administrative datasets from Statistics Canada
    a. Guest speakers: Laëtitia Renée (Université de Montréal) on the ELMLP-T1FF-New Brunswick program data linkage

Lab session

1. Performing a record linkage
2. Advanced material on record linkages

Relevant literature and documentation

Recommended:

Hemeon, James. 2016. *Historical Data Linkage Quality: The Longitudinal and International Study of Adults, and Tax Records on Labour and Income*. Ottawa: Statistics Canada. https://www150.statcan.gc.ca/n1/pub/89-648-x/89-648-x2016001-eng.htm

Robson, Jennifer, and Saul Schwartz. "Who Doesn't File a Tax Return? A Portrait of Non-Filers." *Canadian Public Policy*, June 22, 2020, e2019063. https://doi.org/10.3138/cpp.2019-063.

Simard-Duplain, Gaëlle, and Xavier St-Denis. "Sample Selection in Tax Data Sets of Intergenerational Links: Evidence from the Longitudinal and International Study of Adults." Longitudinal and International Study of Adults Research Paper Series - Statistics Canada Catalogue no. 89-648-X. Ottawa: Statistics Canada, March 17, 2020. https://www150.statcan.gc.ca/n1/pub/89-648-x/89-648-x2020002-eng.htm

Documentation on the LISA-T1FF and Family Files linkages in the RDC as well as documentation from other data linkages.

Complement:

Arim, Rubab, and Marc Frenette. "Are Mental Health and Neurodevelopmental Conditions Barriers to Postsecondary Access?" Analytical Studies Branch Research

Paper Series. Ottawa: Statistics Canada (Catalogue no. 11F0019M — No. 417), February 19, 2019.

Fleury, Charles, Danièle Bélanger, and Aline Lechaume. “Réformes de l’immigration au Québec en 2019 et 2020: La logique politique à l’épreuve de l’analyse statistique.” *Canadian Review of Sociology/Revue Canadienne de Sociologie* 57, no. 3 (2020): 453–72. https://doi.org/10.1111/cars.12293.

Garon-Carrier, Gabrielle, Arya Ansari, Marie-Josée Letarte, and Caroline Fitzpatrick. “Early Childcare Enrollment and the Pursuit of Higher Education: A Canadian Longitudinal Study.” *Learning and Instruction* 80 (August 1, 2022): 101615. https://doi.org/10.1016/j.learninstruc.2022.101615.

Morissette, René, and Theresa Hanqing Qiu. “Turbulence or Steady Course? Permanent Layoffs in Canada, 1978-2016.” IRPP Study No. 76. Montreal: Institute for Research on Public Policy, June 2020.

Renée, Laëtitia. 2023. *The Long-Term Effects of Career Guidance in High School: Evidence from a Randomized Experiment.* Working Paper. https://www.laetitiarenee.com/files/JMP_LRenee.pdf

Simard-Duplain, Gaëlle, and Xavier St-Denis. “Exploration of the Role of Education in Intergenerational Income Mobility in Canada: Evidence from the Longitudinal and International Study of Adults.” *Canadian Public Policy* 46, no. 3 (September 1, 2020): 369–96. https://doi.org/10.3138/cpp.2019-072.

Day 4: Data challenges and data validation

Schedule overview:

9:00am	Day 4 starts! In-class training
10:45am	Pause
11:00am	Guest lecture
12:00pm	Lunch
1:00pm	Lab segment 1
2:30pm	Pause
2:45pm	Lab segment 2
4:30pm	Final words
4:45pm	End of day 4

In-class training

Overview of key content:

1. Non-filing and SIN registration behaviours
2. Geography and geographic mobility with the PCCF and reported addresses
3. Indirect measures and data quality in T1FF and T4 data
    a. “Labour supply" and job characteristics: a brief overview
    b. Indigenous identity, residence on reserve, and Registered Indian status
    c. Language
    d. Health-related measures in tax data
4. Using the IMDB to study immigration: a brief overview of some data challenges
5. Guest Lecture: Chin-lan Winnie Yang (McGill University), “Sex, gender, sexual orientation and same-sex couples in tax datasets”

Lab sessions

1. Indirect measures and validation

Relevant literature and documentation

Recommended:

Harding, Adriene, Eric Olson, and Christine Laporte. “Accessing the Canada Learning Bond: Meeting Identification and Income Eligibility Requirements.” *Income Research Paper Series*, Statistics Canada Catalogue No. 75F0002M. Ottawa: Statistics Canada, June 21, 2019. https://www150.statcan.gc.ca/n1/en/pub/75f0002m/75f0002m2019007-eng.pdf?st=sabV7RKS.

Bérard-Chagnon, Julien. “Comparison of Place of Residence between the T1 Family File and the Census: Evaluation Using Record Linkage.” Demographic Documents. Ottawa: Statistics Canada (Catalogue no. 91F0015M — No. 13), September 26, 2017. https://www150.statcan.gc.ca/n1/pub/91f0015m/91f0015m2017013-eng.pdf.

Frenette, Marc. “Postsecondary Enrolment by Parental Income: Recent National and Provincial Trends.” *Economic Insights*, no. 70 (April 2017): 1–10.

Laporte, Christine, Yuqian Lu, and Grant Schellenberg. “Inter-Provincial Employees in Alberta.” *Analytical Studies Branch Research Paper Series*, no. 350. Statistics Canada Catalogue no. 11F0019M. Ottawa: Statistics Canada, 2013.

> Results for Canada overall: Laporte, Christine, and Yuqian Lu. “Inter-Provincial Employees in Canada.” *Economic Insights*. Catalogue No. 11-626-X — No. 029. Ottawa: Statistics Canada, 2013. http://www.deslibris.ca/ID/239580.

Complements:

Finnie, Ross, and Dejan Pavlic. “Background Characteristics and Patterns of Access to Postsecondary Education in Ontario: Evidence from Longitudinal Tax Data.” Toronto: Higher Education Quality Council of Ontario, 2013.

Haan, Michael, and Miguel Cardoso. “Job Changing and Internal Mobility: Insights into the ‘Declining Duo’ from Canadian Administrative Data.” *Population, Space and Place* 26, no. 5 (2020): e2324. https://doi.org/10.1002/psp.2324.

Hillier, Cathlene, Yujiro Sano, David Zarifa, and Michael Haan. “Will They Stay or Will They Go? Examining the Brain Drain in Canada’s Provincial North.” *Canadian Review of Sociology/Revue Canadienne de Sociologie* 57, no. 2 (2020): 174–96. https://doi.org/10.1111/cars.12276.

Lightman, Naomi, Rupa Banerjee, Ethel Tungohan, Conely de Leon, and Philip Kelly. “An Intersectional Pathway Penalty: Filipina Immigrant Women inside and Outside Canada’s Live-In Caregiver Program.” *International Migration* 60, no. 2 (2022): 29–48. https://doi.org/10.1111/imig.12851.

Morissette, René, and Theresa Hanqing Qiu. “Adjusting to Job Loss When Times Are Tough.” Montreal: Institute for Research on Public Policy, February 2021.

St-Denis, Xavier, and Chih-lan Winnie Yang. “Intergenerational Transmission of Socio-Economic Status and Intragenerational Mobility Over the Early Adult Life Course of Canadian Women and Men.” Toronto, ON: FutureSkills Research Lab, March 10, 2022. http://futureskillscanada.com.

Walters, D., Brown, R., Parekh, G., Einmann, T., & D. Bader. 2020. Student Loan Outcomes of Ontario Transfer Students Evidenced Based on PSIS-CSLP Data Linkages. Toronto: ONCAT. https://oncat.ca/sites/default/files/media-files/student_loan_outcomes_of_ontario_transfer_students.pdf

Zarifa, D., Sano, Y., & C. Hillier. 2020. *Transfer Pathways among Ontario Colleges and Universities Northern and Southern Differences in Students Who Transfer*. Toronto: ONCAT. https://oncat.ca/sites/default/files/media-files/northern_and_southern_differences_in_students_who_transfer.pdf