Data Report

Luis M. B. Varona[1,2] Otoha Hanatani[3]

March 31, 2025

Introduction

In collaboration with the Union of Municipalities of New Brunswick, we conduct a fixed-effects two-stage least squares regression analysis of municipal police spending on tax rates, using median household income as an instrument variable to reduce simultaneity bias. [Add later]

Literature Review

Methodology

Data Collection and Sources

We use an unbalanced panel of annual data from 2000–2018 on New Brunswick municipalities, received via personal correspondence with the GNB and Dr. Craig Brett of Mount Allison University; however, this data is also publicly available at ("Annual Report of Municipal Statistics for New Brunswick - 2000 Through to 2018" 2000), albeit in a less structured format. (The year 2005 is excluded due to missing/improperly formatted tokens, but we may coordinate further with the GNB to obtain this data in the future.) Each set of annual data contains 95 to 103 municipalities, with a total of 104 unique municipalities across all years.

This is supplemented by 2024 data on municipal policing provider agreements (Anderson 2025). We map this data backwards to municipal jurisdictions and boundaries from previous years and integrate indicators into interaction terms in our panel as described below.

Finally, the instrument variable in the first stage of our 2SLS regression is median household income, given in census data from Statistics Canada. Data is only available from 2000 ("Table 95F0437XCB2001006: 2000 Household Income (4) and Household Size (3)" 2001), 2005 ("Table 97-563-XCB2006052: Household Income (7) and Household Size (4)" 2006), 2015 ("Table 98-400-X2016099: Household Income Statistics (3) and Household Type Including Census Family Structure (11)" 2016), and 2020 ("Table 98-10-0061-01: Household Income Statistics by Dwelling and Household Characteristics" 2021); hence, linear interpolation is applied for the intervening years. The resulting income data (typically correlated with tax base per capita but not with tax rate) is then used to reduce simultaneity bias in our fixed-effects model.

Data Cleaning and Organization

Primary Data

The original Excel files extracted from `.zip` archives provided by the GNB and the UMNB are contained in the `data_raw` directory. These contain annual data from 2000–2022 on New Brunswick municipalities, as well as 2024 data on municipal policing providers. Given that some of these files are `.xls` and `.xlw` workbooks,

[1]Department of Mathematics & Computer Science, Mount Allison University, Sackville, NB E4L 1E4
[2]Department of Politics & International Relations, Mount Allison University, Sackville, NB E4L 1E4
[3]Department of Economics, Mount Allison University, Sackville, NB E4L 1E4

we copy and convert them all to `.xlsx` format in the `data_xlsx` directory. The `raw_to_xlsx.py` script is used for this purpose.

Files in `data_xlsx` are then cleaned and organized by `xlsx_to_clean.py` (and its helper scripts in `helper_scripts/xlsx_to_clean`). Finding that data from 2005 and 2019–2022 is unusable due to missing/improperly formatted tokens, our output (placed in the `data_clean` directory) excludes these time periods. No data from the original files is discarded during this process (save for metadata and notes)—simply reorganized into parseable form.

Addressing inconsistent municipality naming conventions across years/categories and concatenating all annual panels within each category (budget expenditures, budget revenues, comparative demographics, and tax bases), the `clean_to_final.py` script then writes all four resulting worksheets—plus a fifth for provider data—to a single `data_master.xlsx` workbook in `data_final`. (The new municipal naming convention is also used to map provider data on newer, reformed 2024 municipalities and districts to past jurisdictions.)

Instrumental Variable Data

Data on the instrumental income data is stored and processed in the `data_iv` directory. There is one folder each for 2001, 2006, 2016, and 2021 (the years in which the census data were released) containing the original files downloaded from the Statistics Canada website. For 2016 and 2021, the downloads are straightforward, nicely formatted `.csv` files requiring no further processing. For 2001 and 2006, however, full data is only available in `.ivt` and `.xml` format; no schemas are available to parse the XML data, so we use the Government of Canada's Beyond 20/20 Browser to extract and download the data in `.csv` format. (Unfortunately, this process is not easily documentable, as the browser requires manual processing.)

With CSV files for all four years, the `process_instrument.py` script is used to clean and combine the relevant columns and rows into a single DataFrame. This is then saved as an `.xlsx` file in the [`data_analysis/elasticity_results`] directory for immediate usage in the data analysis stage. (The aforementioned data interpolation does not occur until said stage and is not considered to be part of the data pipeline.)

Data Analysis and Modelling

Our included variables are

- **Average Tax Rate**, or **AvgTaxRate** – unitless
- **Police Spending per Capita**, or **PolExpCapita** – 10^5 CAD / person
- **Non-Police Spending per Capita**, or **OtherExpCapita** – 10^5 CAD / person
- **Non-Warrant Revenue per Capita**, or **OtherRevCapita** – 10^5 CAD / person
- **Tax Base for Rate per Capita**, or **TaxBaseCapita** – 10^5 CAD / person
- **Policing Provider** – boolean, three categories:
  - *Provincial Police Service Agreement* (excluded control variable)
  - *Municipal Police Service Agreement*, or *Provider_MPSA* (included)
  - *Municipal Police*, or *Provider_MPSA* (included)
- **Median Household Income**, or **MedHouseInc** – 10^5 CAD / person

Our dependent variable is **AvgTaxRate**, which is calculated as a weighted average of the residential and non-residential tax rates in a municipal jurisdiction. [TODO: Elaborate] Our exogenous explanatory variables are **PolExpCapita**, **OtherExpCapita**, **OtherRevCapita**, **PolExpCapita** × *Provider_MPSA**, and*** PolExpCapita** × *Provider_Muni**. Our endogenous explanatory variable is*** TaxBaseCapita**, for which we control simultaneity bias using the instrumental variable*** MedHouseInc**.

Each of these variables is used throughout our two-stage least-squares regression model.

Stage 1

We begin by estimating **MedHouseInc** data for the years missing from the Statistics Canada census data, which we do using simple linear interpolation. (As this project continues to develop, we may investigate

more sophisticated approximation approaches, but this shall do for now.) After this is done, we perform an ordinary least squares regression of **TaxBaseCapita** on **MedHouseInc** to obtain

$$\hat{TaxBaseCapita}_{it} = \beta_0 + \beta_1 MedHouseInc_{it} + v_{it}.$$

By performing this regression before proceeding to a fixed-effects model, we manage to reduce simultaneity bias, as **MedHouseInc** is correlated with **TaxBaseCapita** but not with **AvgTaxRate**. We use these predicted values of **TaxBaseCapita** in the second stage, where we demean all variables involved in the regression over municipalities.

Stage 2

Our primary fixed-effects regression model is now given by

$$\begin{aligned} \ddot{AvgTaxRate}_{it} = {} & \beta_1 \ddot{PolExpCapita}_{it} + \beta_2 \ddot{OtherExpCapita} + \beta_3 \ddot{OtherRevCapita} + \\ & \beta_4 \ddot{\hat{TaxBaseCapita}}_{it} + \beta_5 \ddot{PolExpCapita}_{it} \cdot Provider\_MPSA_{it} + \\ & \beta_6 \ddot{PolExpCapita}_{it} \cdot Provider\_Muni_{it} + \ddot{u}_{it}, \end{aligned}$$

where we use the notation $\ddot{X}_{it} = X_{it} - \bar{X}_i$ to denote the difference between the value of X for municipality i in year t from the mean value of X for municipality i over all years. [Add]

Results

Discussion

CRE Model Specifications

We shall use a series of F-tests to compare a base (restricted) model with aggregated *NEC* and *NRC* data to two partially restricted models and one unrestricted model with disaggregated *NEC* and/or *NRC* data. Our decision to use a CRE model rather than a fixed-effects (FE) one arises from the presence of *MPSA* and *MUNI*—the demeaning process in FE models fails to deal with such time-constant variables, but this is not a problem in CRE models.

Additionally, we shall experiment with functional form on each of the proposals presented below before deciding on our final model. Based on prior research into tax rate regression analyses, a log transformation may prove prudent.

Base (Restricted) Model

Our base (restricted) correlated random-effects model is as follows (for each municipality i and year t):

$$\begin{aligned} ATR_{it} = {} & \beta_1 PSC_{it} + \beta_2 TBC_{it} + \beta_3 NEC_{it} + \beta_4 NRC_{it} + \beta_5 POP_{it} + \\ & \beta_6 MPSA_i + \beta_7 MUNI_i + \alpha_i + u_{it}, \end{aligned}$$

where α_i denotes the municipality-specific effect and u_{it} denotes the error term.

Partially Restricted Model (with NRC)

The first partially restricted model, which disaggregates *NEC* but not *NRC*, is given by

$$
\begin{aligned}
ATR_{it} = {} & \beta_1' PSC_{it} + \beta_2' TBC_{it} + \\
& [\beta_{3,1}' GGS_{it} + \beta_{3,2}' FPS_{it} + \beta_{3,3}' WCT_{it} + \beta_{3,4}' EMS_{it} + \beta_{3,5}' OPS_{it} + \\
& \beta_{3,6}' TRS_{it} + \beta_{3,7}' EHS_{it} + \beta_{3,8}' PHS_{it} + \beta_{3,9}' EDS_{it} + \\
& \beta_{3,10}' RCS_{it} + \beta_{3,11}' DBC_{it} + \beta_{3,12}' TRN_{it} + \beta_{3,13}' DFC_{it}] + \\
& \beta_4' NRC_{it} + \beta_5' POP_{it} + \beta_6' MPSA_i + \beta_7' MUNI_i + \alpha_i' + u_{it}'.
\end{aligned}
$$

Partially Restricted Model (with NEC)

The second partially restricted model, which disaggregates *NRC* but not *NEC*, is given by

$$
\begin{aligned}
ATR_{it} = {} & \beta_1'' PSC_{it} + \beta_2'' TBC_{it} + \beta_3'' NEC_{it} + \\
& [\beta_{4,1}'' UGR_{it} + \beta_{4,2}'' OGS_{it} + \beta_{4,3}'' SOS_{it} + \beta_{4,4}'' OSR_{it} + \beta_{4,5}'' CTR_{it} + \\
& \beta_{4,6}'' OTR_{it} + \beta_{4,7}'' BIS_{it}] + \\
& \beta_5'' POP_{it} + \beta_6'' MPSA_i + \beta_7'' MUNI_i + \alpha_i'' + u_{it}''.
\end{aligned}
$$

Unrestricted Model

The fully unrestricted model, which disaggregates both *NEC* and *NRC*, is given by

$$
\begin{aligned}
ATR_{it} = {} & \beta_1''' PSC_{it} + \beta_2''' TBC_{it} + \\
& [\beta_{3,1}''' GGS_{it} + \beta_{3,2}''' FPS_{it} + \beta_{3,3}''' WCT_{it} + \beta_{3,4}''' EMS_{it} + \beta_{3,5}''' OPS_{it} + \\
& \beta_{3,6}''' TRS_{it} + \beta_{3,7}''' EHS_{it} + \beta_{3,8}''' PHS_{it} + \beta_{3,9}''' EDS_{it} + \\
& \beta_{3,10}''' RCS_{it} + \beta_{3,11}''' DBC_{it} + \beta_{3,12}''' TRN_{it} + \beta_{3,13}''' DFC_{it}] + \\
& [\beta_{4,1}''' UGR_{it} + \beta_{4,2}''' OGS_{it} + \beta_{4,3}''' SOS_{it} + \beta_{4,4}''' OSR_{it} + \beta_{4,5}''' CTR_{it} + \\
& \beta_{4,6}''' OTR_{it} + \beta_{4,7}''' BIS_{it}] + \\
& \beta_5''' POP_{it} + \beta_6''' MPSA_i + \beta_7''' MUNI_i + \alpha_i''' + u_{it}'''.
\end{aligned}
$$

References

Anderson, Amy. 2025. "Personal Correspondence with Amy Anderson."

"Annual Report of Municipal Statistics for New Brunswick - 2000 Through to 2018." 2000. Fredericton, NB: Government of New Brunswick.

"Table 95F0437XCB2001006: 2000 Household Income (4) and Household Size (3)." 2001. Statistics Canada; https://www12.statcan.gc.ca/english/census01/products/standard/themes/Download.cfm?PID=55710.

"Table 97-563-XCB2006052: Household Income (7) and Household Size (4)." 2006. Statistics Canada; https://www12.statcan.gc.ca/census-recensement/2006/dp-pd/tbt/Download.cfm?PID=94594.

"Table 98-10-0061-01: Household Income Statistics by Dwelling and Household Characteristics." 2021. Statistics Canada; https://doi.org/10.25318/9810006101-eng.

"Table 98-400-X2016099: Household Income Statistics (3) and Household Type Including Census Family Structure (11)." 2016. Statistics Canada; https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/dt-td/CompDataDownload.cfm?LANG=E&PID=110192&OFT=CSV.