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 and Dr. Craig Brett of Mount Allison University, we conduct a fixed-effects two-stage least squares (or FE-2SLS) regression analysis of average tax rates on police spending in New Brunswick municipalities, using median household income as an instrument variable to reduce simultaneity bias. [TODO: Elaborate]

Literature Review

[TODO: Elaborate]

Methodology

In this section, we delineate our data collection process, data organization methods, and econometric models and analysis. We use Python (primarily the polars and linearmodels/statsmodels ecosystems) to parse and clean data from the Government of New Brunswick and Statistics Canada. Subsequently, we run several fixed-effects and correlated random-effects regression models on the resulting data in combination with an instrumental variable to account for simultaneity bias.

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 ("2000–2018 Annual Reports of Municipal Statistics for New Brunswick" 2000–2018), 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" 2001), 2005 ("Table 97-563-XCB2006052" 2006), 2015 ("Table 98-400-X2016099" 2016), and 2020 ("Table 98-10-0061-01" 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.

[1] Department of Politics & International Relations, Mount Allison University, Sackville, NB E4L 1A7
[2] Department of Mathematics & Computer Science, Mount Allison University, Sackville, NB E4L 1E6
[3] Department of Economics, Mount Allison University, Sackville, NB E4L 1A7

Data Cleaning and Organization

Primary Data

Primary data is cleaned in the `data_pipeline/` directory. The original Excel files extracted from `.zip` archives provided by the GNB and the UMNB are contained in the `data_raw/` subdirectory. 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, we copy and convert them all to `.xlsx` format in the `data_xlsx/` subdirectory. The `helper_scripts/_raw_to_xlsx_.py` script is used for this purpose.

Files in this `data_xlsx/` subdirectory are cleaned and organized by `helper_scripts/_xlsx_to_clean_.py`. Finding that data from 2005 and 2019–2022 is unusable due to missing/improperly formatted tokens, our output (placed in the `data_clean/` subdirectory) excludes these time periods. No original data is discarded during this process (save for metadata and notes)—it is all 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 `helper_scripts/_clean_to_final_.py` script then writes all four resulting worksheets—plus a fifth for provider data—to a single `data_fina;/data_master.xlsx` workbook. (The new municipal naming convention is also used to map provider data on newer, reformed 2024 municipalities and districts to past jurisdictions all the way back to 2000.)

All scripts are called and run by the main executable of the associated directory, `main.py`.

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 `main.py` executable script is finally used to clean and combine the relevant columns and rows into a single polars DataFrame. This is then saved as an `.xlsx` file in the `results/` subdirectory for immediate usage in the data analysis stage. (The aforementioned data interpolation—performed using Python's numpy library—is not applied until this stage and is thus not considered part of the data cleaning and organization pipeline.)

It is worth noting that although household income data from Canada censuses is publicly accessible for municipal-level geographic localities in 2000, 2005, 2015, and 2020, the only available source for 2010 is aggregated data from the 2011 National Household Survey. This survey refrained from providing disaggregated data at lower levels of geography, so we are unable to map it to most of the 104 municipalities in our dataset. Hence, linear interpolation is used to estimate the missing data for 2010, just as for all the other missing years. In the future, we may collaborate further with Statistics Canada to obtain the geographically disaggregated data, if it remains in their records.

Data Analysis and Modelling

All data analysis is performed in the `data_analysis/` directory. 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. (That is—as per government formulae, non-residential rates are multiplied by a factor of 1.5 before being integrated into the calculated average. Said averages are already available in the raw data ("2000–2018 Annual Reports of Municipal Statistics for New Brunswick" 2000–2018), not calculated by us; we take note of the process simply to clarify the layout of our data.) Our exogenous explanatory variables are *PolExpCapita*, *OtherExpCapita*, *OtherRevCapita*, *PolExpCapita*∗*Provider_MPSA*, and *PolExpCapita*∗*Provider_Muni*. Our sole endogenous explanatory variable is *TaxBaseCapita*, for which we control simultaneity bias using *MedHouseInc* as an instrumental variable.

Each of these variables is used throughout our FE-2SLS regression model, carried out by the `helper_scripts/_fe_2sls_analysis_.py` script. In addition, we have also included vanilla correlated random-effects (CRE) and fixed-effects (FE) models, run by `helper_scripts/_cre_analysis_.py` and `helper_scripts/_fe_analysis_.py`, to determine which variables are relevant and to demonstrate the need for an instrument variable. All helper scripts are called and run by the main executable of the associated directory, `main.py`.

Our final choice of FE in conjunction with 2SLS arose from [TODO: Elaborate, particularly on why *TaxBaseCapita* causes simultaneity bias]

We begin this section by first describing our CRE and FE analyses, then delineating more thoroughly our final FE-2SLS model.

Correlated Random-Effects (CRE)

[TODO: Elaborate]

Fixed-Effects (FE)

After deeming the potential benefits of including the policing provider indicators directly (not in interaction terms) insufficient to warrant [TODO: Elaborate]

Fixed-Effects Two-Stage Least Squares (FE-2SLS)

Finally, we decided on [TODO: Elaborate]

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

$$TaxBaseCapita_{it} = \alpha_0 + \alpha_1 MedHouseInc_{it} + v_{it}.$$

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 $\widehat{TaxBaseCapita}_{it} = TaxBaseCapita_{it} - v_{it}$ values in the second-stage regression, where we demean all variables over municipality.

Stage 2 Our primary fixed-effects regression model is now given by

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

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 and the mean value of X for municipality i over all years. (Note that $\ddot{\hat{TaxBaseCapita}}_{it}$ is not the demeaning of $TaxBaseCapita_{it}$ itself but rather the demeaned prediction from our first-stage regression.)

Results

[TODO: Elaborate]

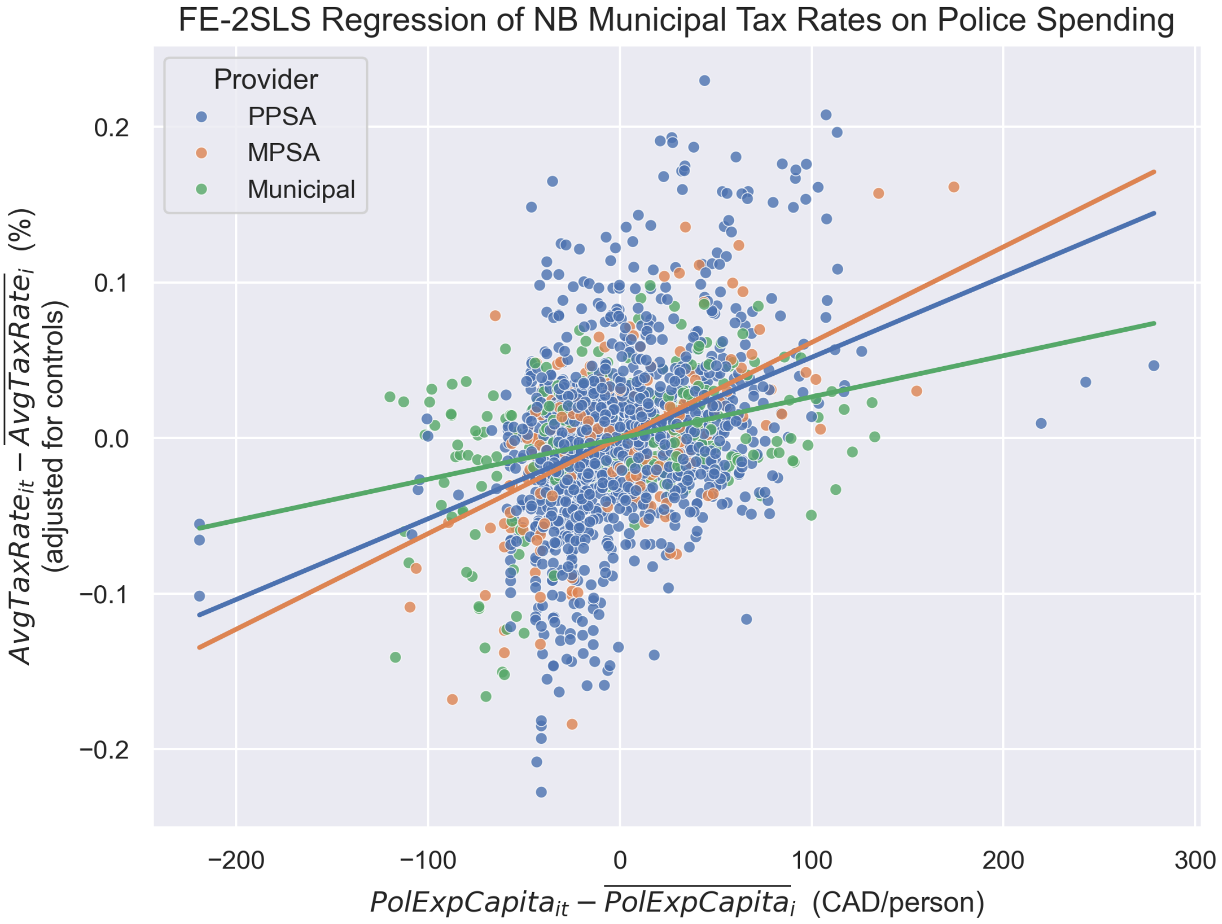

Figure 1: image

Discussion

[TODO: Elaborate]

Conclusion

[TODO: Elaborate]

Appendix

[TODO: Include linearmodels/statsmodels regression summary output]

References

"2000–2018 Annual Reports of Municipal Statistics for New Brunswick." 2000–2018. Fredericton, NB: Government of New Brunswick.

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

"Table 95F0437XCB2001006." 2001. Statistics Canada; https://www12.statcan.gc.ca/english/census01/products/standard/themes/Download.cfm?PID=55710.

"Table 97-563-XCB2006052." 2006. Statistics Canada; https://www12.statcan.gc.ca/census-recensement/2006/dp-pd/tbt/Download.cfm?PID=94594.

"Table 98-10-0061-01." 2021. Statistics Canada; https://doi.org/10.25318/9810006101-eng.

"Table 98-400-X2016099." 2016. Statistics Canada; https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/dt-td/CompDataDownload.cfm?LANG=E&PID=110192&OFT=CSV.