Data Summary

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

March 4, 2025

This report provides a summary of the Government of New Brunswick (GNB) panel data used in our correlated random-effects (CRE) regression analysis of municipal tax rates on police spending. We present below our included variables, preliminary summary statistics, and our data cleaning process.

Variables of Interest

The response variable we aim to model is the *Average Tax Rate*, or *ATR* (given in %), regressed on the following explanatory variables using correlated random-effects:

- **Police Spending/Capita**, or **PSC** – CAD/person
- **Tax Base for Rate/Capita**, or **TBC** – CAD/person
- **Non-Police Expenditure/Capita**, or **NEC** – CAD/person, disaggregated:
  - *General Government*, or *GGS*
  - *Fire Protection*, or *FPS*
  - *Water Cost Transfer*, or *WCT*
  - *Emergency Measures*, or *EMS*
  - *Other Protection*, or *OPS*
  - *Transportation*, or *TRS*
  - *Environmental Health*, or *EHS*
  - *Public Health*, or *PHS*
  - *Environmental Development*, or *EDS*
  - *Recreation & Cultural*, or *RCS*
  - *Debt Costs*, or *DBC*
  - *Transfers*, or *TRN*
  - *Deficits*, or *DFC*
- **Non-Warrant Revenue/Capita**, or **NRC** – CAD/person, disaggregated:
  - *Unconditional Grant*, or *UGR*
  - *Services to Other Governments*, or *OGS*
  - *Sale of Services*, or *SOS*
  - *Own-Source Revenue*, or *OSR*
  - *Conditional Transfers*, or *CTR*
  - *Other Transfers*, or *OTR*
  - *Biennial Surplus*, or *BIS*
- **Population**, or **POP** – persons (from the latest census data)
- **Policing Provider** – boolean, three categories:
  - *Provincial Police Service Agreement*, or *PPSA* (control, excluded to avoid collinearity)
  - *Municipal Police Service Agreement*, or *MPSA* (included indicator)
  - *Municipal Police*, or *MPSA* (included indicator)

A description of each variable is provided below.

[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

Variable Descriptions

[Write here]

Summary Statistics

Summary statistics of our panel data are included in the current directory at `data_summary_by_stat.xlsx` and `data_summary_by_year.xlsx`. The former workbook contains one worksheet with all years for each summary statistic, whereas the latter contains one worksheet with all summary statistics for each year. The years covered are 2000–2004 and 2006–2018, and the summary statistics provided are

- **Count** – number of non-null observations in a column (by year)
- **Null Count** – number of null observations in a column (by year)
- **Mean** – mean of a column (by year)
- **Std. Dev.** – standard deviation of a column (by year)
- **Minimum** – minimum value of a column (by year)
- **25%** – 1st quartile of a column (by year)
- **Median** – median of a column (by year)
- **75%** – 3rd quartile of a column (by year)
- **Maximum** – maximum value of a column (by year)

Data Pipeline

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, the Union of Municipalities of New Brunswick (UMNB), and Dr. Craig Brett of Mount Allison University. The year 2005 is excluded due to missing data; similarly, data from 2019 onwards is excluded due to improper formatting. (However, we may coordinate further with the GNB and the UMNB to obtain this data in the future.)

Each annual time period contains 95 to 103 municipalities, with a total of 104 unique municipalities across all years. Policing provider data (also received directly from the UMNB) is available for 2024 municipalities—we then map this backward to municipal jurisdictions/boundaries from previous years and integrate indicators as time-constant variables into our panel.

Data Cleaning and Organization

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 the aforementioned 2024 data on municipal policing providers. Given that some of these files are `.xls` or, worse, `.xlw` workbooks, 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 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 policing provider data—to a single `data_master.xlsx` workbook in the `data_final` directory. (The new consistent naming convention is also used to map provider data back to previous municipal jurisdictions; previously, it was a mix of data on newer municipal and local jurisdictions.)

The `clean_to_final.py` script in question (and its helpers in `helper_scripts/clean_to_final`) also joins the five worksheets from `data_master.xlsx` into a single worksheet in the `data_final.xlsx` workbook (in the same directory). This final dataset is then used for our CRE regression analysis, delineated in the following section. (It is also used to obtain the summary statistics presented above.)

CRE Model Specifications

We will 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. *[Should we also play around with functional form (e.g., log transformations) or not?]* 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.

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

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