Data Availability Guidelines and Code Base for "Consumer Credit Usage in Canada during the Coronavirus Pandemic"

Anson T. Y. Ho
Department of Real Estate Management, Ryerson University

Lealand Morin and Harry J. Paarsch
Department of Economics, University of Central Florida

Kim P. Huynh
Currency Department, Bank of Canada

May 31, 2021

Credit_COVID19_Canada

This is the code base to accompany the manuscript *Consumer Credit Usage in Canada during the Coronavirus Pandemic* by Ho, Morin, Paarsch and Huynh in the Canadian Journal of Economics, 2021

Any updates will be available on the GitHub code repository called `Credit_COVID19_Canada` available at the following link:

`https://github.com/LeeMorinUCF/Credit_COVID19_Canada`

Data Availability

TransUnion Data

The primary data source is the TransUnion credit bureau. Data are provided to the Bank of Canada on a monthly basis. Under the contractual agreement with TransUnion, the data are not publicly available. The Bank of Canada does, however, have a process for external researchers to work with these data. The Bank of Canada's Financial System Research Center is a hub for research on household finance (`https://www.bankofcanada.ca/research/financial-system-research-centre/`). Interested parties, who are Canadian citizens or permanent residents, can contact Jason Allen (e-mail: `Jallen@bankofcanada.ca`) or the Managing Director of research Jim MacGee (e-mail: `JMacGee@bankofcanada.ca`). Interested parties are asked to submit a project proposal; the proposal is evaluated by senior staff at the Bank of Canada for feasibility; external researchers do not typically have direct access to the data and must work with a Bank of Canada staff member. An exception is if an external collaborator applies for and is granted temporary employee status—in this case, the external researcher can access the data so long as they have a Bank of Canada affiliation. All research is vetted by Bank of Canada senior staff prior to publication.

Regulatory Filings by the Bank of Canada (formerly E2)

The Bank of Canada's historical credit aggregates are available in the *Banking and Financial Statistics* table `https://www.bankofcanada.ca/rates/banking-and-financial-statistics/`. Since October 2020, Statistics Canada produces monthly credit aggregates that aligns with the Statistics Canada's National Balance Sheet Accounts program. These new credit aggregates are available from tables at Statistics Canada: *Consumer credit, outstanding balances of selected holders* `https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1010011701` (CANSIM table 10-10-0117-01). For details, see the Bank of Canada's announcement *Bank of Canada and Statistics Canada to move to a single set of credit statistics* `https://www.bankofcanada.ca/2019/10/bank-canada-statistics-canada-move-single-set-credit-statistics/`.

The Nilson Report

The set of numbers in some tables are taken directly from the *The Nilson Report*, Issue 1173, April 2020, and are available at `https://nilsonreport.com/publication_newsletter_archive_issue.php?issue=1173`.

2016 Canadian Census

The number of credit-card account holders aged 20 and above and other figures were obtained from Statistics Canada in the table called *Estimates of population (2016 Census and administrative data), by age group and sex for July 1st, Canada, provinces, territories, health regions (2018 boundaries) and peer groups*, Table: 17-10-0134-01.

Instructions

The workflow proceeds in two stages: one set of instructions outlines the operations to transform the raw data in the TransUnion database into the datasets that are the inputs for the statistical analysis in the next stage.

Data Manipulation

These procedures were performed on the EDITH 2.0 computing cluster at the Bank of Canada to generate the primary datasets. These scripts are stored in the `Code/Data_Prep` folder.

1. Run the SLURM script `df_ind_bc.slurm`, which runs a sequence of Python scripts with names of the form `cr_use_bc_Y1Y2.py`, for data covering each two-year period 20Y1-20Y2. It then runs `cr_use_bc_combine.py`, which generates a temporary parquet file `df_ind.parquet`. This dataset comprises individual-level data that is sufficient to run the data manipulation for credit-card accounts on the nation-wide sample.

2. Run the script `tu_sample_bc.slurm`, which runs the script `tu_sample_bc.py` and generates the dataset `tu_sample_bc.csv`. This dataset is sufficient to run the analysis of credit-card accounts on the nation-wide sample.

3. Run the script `tu_agg_series_bc.slurm`, which runs the script `tu_agg_series.py` and generates the dataset `tu_agg_bc.csv`. This dataset provides the input for panel (a) of Figure 1: Consumers' Outstanding Balances, 2017-2020 for credit-card accounts on the nation-wide sample.

4. Run the script `tu_sample_AB_bc.slurm`, which runs the script `tu_sample_AB_bc.py` and generates the dataset `tu_sample_AB_bc.csv`. This dataset is sufficient to run the analysis of credit-card accounts on the Alberta sample.

5. Run the script `tu_agg_series_AB_bc.slurm`, which runs the script `tu_agg_series_AB.py` and generates the dataset `tu_agg_AB_bc.csv`. This dataset provides the input for panel (a) of Figure 8: Consumers' Outstanding Balances, Alberta, 2012-2016 for credit-card accounts on the Alberta sample.

6. Run the script `TU_vs_Nilson_comp.slurm`, which runs the script `TU_vs_Nilson_comp.py` and generates the dataset `TU_vs_Nilson_num_accts.csv`. This dataset provides the input for Table A1: Comparison of Accounts at the Credit Agency with Nation-Wide Totals in The Nilson Report.

7. Run the script `TU_vs_BoC_comp.slurm`, which runs the script `TU_vs_BoC_comp.py` and generates the dataset `TU_vs_BoC_totals.csv`. This dataset provides the input for Figure A1.1: Time Series of Aggregate Credit-Card Balances.

8. Run the script `TU_vs_StatsCan_comp`, which runs the script `TU_vs_StatsCan_comp.py` and generates the dataset `CC_TU_vs_StatsCan.csv`. This dataset provides the input for Figure A1.2: Credit Data Coverage for Adults in Canada, by Province.

9. Run the script `df_ind_heloc.slurm`, which runs a sequence of Python scripts with names of the form `cr_use_heloc_Y1Y2.py`, for data covering each two-year period 20Y1-20Y2. It then runs `cr_use_heloc_combine.py`, which generates a temporary parquet file `df_ind.parquet`. This dataset comprises individual-level data that is sufficient to run the data manipulation for HELOC accounts on the nation-wide sample.

10. Run the script `tu_sample_heloc.slurm`, which runs the script `tu_sample_heloc.py` and generates the dataset `tu_sample_heloc.csv`. This dataset is sufficient to run the analysis of HELOC accounts on the nation-wide sample.

11. Run the script `tu_agg_series_heloc.slurm`, which runs the script `tu_agg_series.py` and generates the dataset `tu_agg_heloc.csv`. This dataset provides the input for panel (b) of Figure 1: Consumers' Outstanding Balances, 2017-2020 for HELOC accounts on the nation-wide sample.

12. Run the script `tu_sample_AB_heloc.slurm`, which runs the script `tu_sample_AB_heloc.py` and generates the dataset `tu_sample_AB_heloc.csv`. This dataset is sufficient to run the analysis of HELOC accounts on the Alberta sample.

13. Run the script `tu_agg_series_AB_heloc.slurm`, which runs the script `tu_agg_series_AB.py` and generates the dataset `tu_agg_AB_heloc.csv`. This dataset provides the input for panel (b) of Figure 8: Consumers' Outstanding Balances, Alberta, 2012-2016 for HELOC accounts on the Alberta sample.

Datasets

The above operations will produce the following datasets in `csv` format.

Main datasets

The `Data` folder must contain four main datasets: `tu_sample_bc.csv` and `tu_sample_heloc.csv` for the nation-wide sample, as well as `tu_sample_AB_bc.csv` and `tu_sample_AB_heloc.csv` for the sample restricted to the province of Alberta.

`tu_sample_bc.csv`

This dataset contains observations of credit card balances for consumers in Canada from 2017-2021. It contains the following variables:

- `tu_consumer_id` is a 9-digit integer that indicates an individual consumer.

- `Run_Date` is a date variable of the form `'YYYY-MM-01'` indicating the month in which the data were reported by the bureau. It is the date that represents the last information added to the files, so it contains the statement activity recorded in the previous month.
- `prov` is a string that indicates the province of residence of the consumer.
- `homeowner` is an indicator that the consumer has ever had a mortgage or a HELOC loan.
- `N_bc` is the number of credit card accounts held by a consumer.
- `bc_bal` is the consumer's credit-card balance in dollars.

`tu_sample_he.csv`

This dataset contains observations of HELOC balances for consumers in Canada from 2017-2021. It contains the following variables:

- `tu_consumer_id` is 9-digit integer that indicates an individual consumer.
- `Run_Date` is a date variable of the form `'YYYY-MM-01'` indicating the month in which the data were reported by the bureau.
- `prov` is a string that indicates the province of residence of the consumer.
- `homeowner` is an indicator that the consumer has ever had a mortgage or a HELOC loan.
- `N_he` is the number of HELOC accounts held by a consumer.
- `he_bal` is the consumer's HELOC balance in dollars.

`tu_sample_AB_bc.csv` **and** `tu_sample_AB_bc.csv`

These datasets have the same format as for `tu_sample_bc.csv` and `tu_sample_he.csv`, described above.

Auxiliary datasets

Time series plots: nation-wide sample

The `Data` folder also contains two datasets for generating aggregate time-series in Figure 1. The files `tu_agg_bc.csv` and `tu_agg_heloc.csv` contain time series of aggregate statistics throughout the sample.

These files both contain the following variables.

- `Run_Date` is a date variable of the form `'YYYY-MM-01'`, indicating the month in which the data were reported by the bureau.
- `bal_avg` is the average balance held by consumers during the month.
- `bal_sd` is the standard deviation of balances held by consumers during the month.
- `bal_p25` is the lower quartile of balances held by consumers during the month.
- `bal_p50` is the median balance held by consumers during the month.

- `bal_p75` is the upper quartile of balances held by consumers during the month.

Time series plots: Alberta sample

The `Data` folder also contains another pair of datasets for generating aggregate time-series in Figure 8. The files `tu_agg_AB_bc.csv` and `tu_agg_AB_heloc.csv` contain time series of aggregate statistics throughout the sample, in an identical format, except that these were restricted to the province of Alberta.

Validation of aggregate credit-card balances

A dataset of time series of aggregate outstanding credit-card balances is required to generate Figure A1.1. These data are stored in a file `TU_vs_BoC_totals.csv`, which includes the following columns.

- `Date` in `DD/MM/YYYY` format, representing the last day of each month.
- `MCP` is a series drawn from the Webpage of the Bank of Canada, entitled *Chartered bank selected assets: Month-end (formerly C1), Credit cards*, which records assets in the personal loan category of non-mortgage loans.
- `tot_bal_all` is the aggregate credit-card balance, in billions of Canadian dollars, across all institutions represented in the TransUnion database.
- `tot_bal_bank` is the aggregate credit-card balance, in billions of Canadian dollars, across the chartered banks.

Validation of credit-card data coverage by province

A dataset of aggregate counts of the number of cardholders by province was compared to the population in each province in Figure A1.2. This information was collected in the dataset `CC_TU_vs_StatsCan.csv`, with the following columns.

- `region` is the two-letter abbreviation of each province in Canada.
- `N_geq20_BC` is the number of consumers aged 20 and above holding accounts during the month of January 2016.
- `geq20` is the population of each province in the age categories 20 and above, which was obtained from Statstics Canada Table: 17-10-0134-01, described below.

Other data

Other data were obtained to produce tables and figures of aggregate information about the credit-card and HELOC markets. These include:

- The Nilson Report, April 2020, Issue 1173, HSN Consultants, Inc.
- Chartered bank selected assets: Month-end (formerly C1), Credit cards, Bank of Canada, accessed June 2020.
- Estimates of population (2016 Census and administrative data), by age group and sex for July 1st, Canada, provinces, territories, health regions (2018 boundaries) and peer groups, Table: 17-10-0134-01, Statistics Canada, accessed June 2020.

These data sources are used in the statistical analysis that follows.

Statistical Analysis

These procedures were performed on a microcomputer to generate the tables and figures in the paper. These scripts are stored in the `Code/Stats` folder.

All Files in One Script:

1. Place all datasets in the `Data` folder, including the main datasets, including `tu_sample_bc.csv`, `tu_sample_heloc.csv`, `tu_sample_AB_bc.csv`, and `tu_sample_AB_heloc.csv`, along with the auxiliary datasets for time-series plots `tu_agg_bc.csv`, `tu_agg_heloc.csv`, `tu_agg_AB_bc.csv`, and `tu_agg_AB_heloc.csv`, and for figures in the appendix `TU_vs_BoC_num_accts.csv` and `CC_TU_vs_StatsCan.csv`.

2. Run `COVID_CJE.sh` in a terminal window from the `Credit_COVID19_Canada` folder.

This shell script calls the main `R` programs `COVID_CJE_Cards.R` and `COVID_CJE_HELOCs.R`, for the nation-wide sample, then `COVID_CJE_AB_Cards.R` and `COVID_CJE_AB_HELOCs.R`, for the sample in Alberta, as well as the auxiliary `R` scripts `CC_HE_time_series_figs.R`, `CC_BoC_vs_TU_comp_figs.R`, and `CC_TU_vs_StatsCan_comp_fig.R`, all found in the `Code/Stats` folder, which analyze the datasets stored in the `Data` folder. These scripts create the tables and figures for the entire manuscript, by writing `tex` files to the `Tables` folder and `eps` files to the `Figures` folder.

Generating Sets of Files Separately

Nation-Wide Sample of Credit-Card Accounts

1. Place the dataset `tu_sample_bc.csv` in the `Data` folder.

2. Run `Rscript COVID_CJE_Cards.R` in a terminal window from the `Credit_COVID19_Canada` folder.

3. Obtain the `tex` files `CC_KLD_vs_sample_01.tex` and `CC_KLD_kstep_fixed_vs_monthly_02.tex` with numbers for columns 2 and 3 of Tables 1 and 2 from the `Tables` folder.

4. Obtain the images for panels (a) of Figures 2 and 3 in the `eps` files `CC_hist_grp.eps` and `CC_3D_probs_discrete_1.eps` from the `Figures` folder.

5. Obtain the images for Figures 4 and 6 in the `eps` files `CC_sample_dev_pct_2020_MM.eps` and `CC_obs_vs_for_dev_pct_monthly_2020-MM.eps` from the `Figures` folder, where `MM` represents the two-digit month of the `Run_date` after the close of the corresponding statement month.

Nation-Wide Sample of HELOC Accounts

1. Place the dataset `tu_sample_heloc.csv` in the `Data` folder.

2. Run `Rscript COVID_CJE_HELOCs.R` in a terminal window from the `Credit_COVID19_Canada` folder.

3. Obtain the `tex` files `HE_KLD_vs_sample_01.tex` and `HE_KLD_kstep_fixed_vs_monthly_02.tex` with numbers for columns 4 and 5 of Tables 1 and 2 from the `Tables` folder.

4. Obtain the images for panels (b) of Figures 2 and 3 in the `eps` files `HE_hist_grp.eps` and `HE_3D_probs_discrete_1.eps` from the `Figures` folder.

5. Obtain the images for Figures 5 and 7 in the `eps` files `HE_sample_dev_pct_2020_MM.eps` and `HE_obs_vs_for_dev_pct_monthly_2020-MM.eps` from the `Figures` folder, where `MM` represents the two-digit month of the `Run_date` after the close of the corresponding statement month.

Alberta Sample of Credit-Card Accounts

1. Place the dataset `tu_sample_AB_bc.csv` in the `Data` folder.

2. Run `Rscript COVID_CJE_AB_Cards.R` in a terminal window from the `Credit_COVID19_Canada` folder.

3. Obtain the `tex` file named `AB_CC_KLD_kstep_fixed_vs_monthly_02.tex` with numbers for columns 2 and 3 for Table 3 in the `Tables` folder and panel (a) of Figure 9 in the file `AB_CC_obs_vs_for_dev_pct_monthly_2015-11.eps` in the `Figures` folder.

Alberta Sample of HELOC Accounts

1. Place the dataset `tu_sample_AB_heloc.csv` in the `Data` folder.

2. Run `Rscript COVID_CJE_AB_HELOCs.R` in a terminal window from the `Credit_COVID19_Canada` folder.

3. Obtain the `tex` file `AB_HE_KLD_kstep_fixed_vs_monthly_02.tex` with numbers for columns 4 and 5 for Table 3 in the `Tables` folder and panel (b) of Figure 9 in the image file named `AB_HE_obs_vs_for_dev_pct_monthly_2015-11.eps` in the `Figures` folder.

Auxilliary Tables and Figures

Instructions for generating the remaining tables and figures are outlined in the next section "Generating Tables and Figures Separately".

Generating Tables and Figures Separately

Tables

Table 1: Divergence from Sample Histograms

This Table contains information from two different modeling exercises: one for credit-cards and one for HELOCs.

For credit cards, run script `COVID_CJE_Cards.R`, which then runs script `COVID_CJE_Cards_estim.R`. Lines 118 to 199 of `COVID_CJE_Cards_estim.R` generate a file named `CC_KLD_vs_sample_01.tex`.

For HELOCs, run script `COVID_CJE_HELOCs.R`, which then runs script `COVID_CJE_HELOCs_estim.R`. Lines 118 to 199 of `COVID_CJE_HELOCs_estim.R` generate a file named `HE_KLD_vs_sample_01.tex`.

The numbers from these two tables are combined into the file `Table_1.tex`.

Table 2: Divergence from ℓ-Step-Ahead Forecasts

This Table also contains information from two different modeling exercises: one for credit-cards and one for HELOCs.

For credit cards, run script `COVID_CJE_Cards.R`, which then runs script `COVID_CJE_Cards_estim.R`. Lines 363 to 433 of `COVID_CJE_Cards_estim.R` generate a file `CC_KLD_kstep_fixed_vs_monthly_02.tex`.

For HELOCs, run script `COVID_CJE_HELOCs.R`, which then runs script `COVID_CJE_HELOCs_estim.R`. Lines 363 to 433 of `COVID_CJE_HELOCs_estim.R` generate a file `HE_KLD_kstep_fixed_vs_monthly_02.tex`.

The numbers from these two tables corresponding to the model with monthly transition matrices are combined into the file `Table_2.tex`.

Table 3: Divergence from ℓ-Step-Ahead Forecasts in Alberta, 2015

The creation of this Table mirrors that of Table 2 on the Canadian population during the pandemic, except that it is run on a dataset restricted to consumers in the province of Alberta during the oil price shock in 2015. As with Table 2, it also contains information from two different modeling exercises: one for credit-cards and one for HELOCs.

For credit cards, run script `COVID_CJE_Cards.R`, which then runs script `COVID_CJE_AB_Cards_estim.R`. Lines 366 to 436 of `COVID_CJE_AB_Cards_estim.R` generate a file named `AB_CC_KLD_kstep_fixed_vs_monthly_02.tex`.

For HELOCs, run script `COVID_CJE_HELOCs.R`, which then runs script `COVID_CJE_AB_HELOCs_estim.R`. Lines 366 to 436 of `COVID_CJE_AB_HELOCs_estim.R` generate a file named `AB_HE_KLD_kstep_fixed_vs_monthly_02.tex`.

The numbers from these two tables corresponding to the model with monthly transition matrices are combined into the file `Table_3.tex`.

Table A1: Comparison of Accounts at the Credit Agency with Nation-Wide Totals in *The Nilson Report*

The set of numbers in the three leftmost columns are taken directly from the `The Nilson Report`, Issue 1173, April 2020, and are available here:

`https://nilsonreport.com/publication_newsletter_archive_issue.php?issue=1173`.

To compare with the contents of the TransUnion database, we calculated the same summary statistics using the sample drawn from the database. The remaining information was obtained from running the script `TU_vs_Nilson_comp.py`, which produced the summary dataset called

TU_vs_BoC_num_accts.csv, found in the Data folder.

Figures

Figure 1: Consumers' Outstanding Balances, 2017-2020

For credit cards, in panel (a), run script CC_HE_time_series_figs.R. Lines 91 to 132 generate a file named CC_time_series.eps from the data in a file named tu_agg_bc.csv.

For HELOCs, in panel (b), run script CC_HE_time_series_figs.R. Lines 141 to 182 generate a file named HE_time_series.eps from the data in a file named tu_agg_heloc.csv.

Figure 2: Histograms of Individuals' Balances

For credit cards, in panel (a), run script COVID_CJE_Cards.R, which then runs script COVID_CJE_Cards_prelim.R. Lines 53 to 71 of COVID_CJE_Cards_prelim.R generate a file named CC_hist_grp_sample.eps.

For HELOCs, in panel (b), run script COVID_CJE_HELOCs.R, which then runs script COVID_CJE_HELOCs_prelim.R. Lines 53 to 71 of COVID_CJE_HELOCs_prelim.R generate a file named HE_hist_grp_sample.eps.

Figure 3: Conditional Histograms of Individuals' Balances

For credit cards, in panel (a), run script COVID_CJE_Cards.R, which then runs script COVID_CJE_Cards_prelim.R. Lines 94 to 177 of COVID_CJE_Cards_prelim.R generate a file named CC_3D_probs_discrete_1.eps.

For HELOCs, in panel (b), run script COVID_CJE_HELOCs.R, which then runs script COVID_CJE_HELOCs_prelim.R. Lines 94 to 177 of COVID_CJE_HELOCs_prelim.R generate a file named HE_3D_probs_discrete_1.eps.

Figure 4: Deviations from Histograms (Credit Cards)

For credit cards, in panel (a), run script COVID_CJE_Cards.R, which then runs script COVID_CJE_Cards_estim.R. Lines 36 to 114 of COVID_CJE_Cards_estim.R generate a set of files named CC_sample_dev_pct_2020_MM.eps.

Figure 5: Deviations from Histograms (HELOCs)

For HELOCs, in panel (b), run script COVID_CJE_HELOCs.R, which then runs script COVID_CJE_HELOCs_estim.R. Lines 36 to 114 of COVID_CJE_HELOCs_estim.R generate a set of files named HE_sample_dev_pct_2020_MM.eps.

Figure 6: Deviations from Forecasted Credit-Card Balances

For credit cards, in panel (a), run script COVID_CJE_Cards.R, which then runs script COVID_CJE_Cards_estim.R. Lines 473 to 506 of COVID_CJE_Cards_estim.R generate a file named CC_obs_vs_for_dev_pct_monthly_2020_MM.eps.

Figure 7: Deviations from Forecasted HELOC Balances

For HELOCs, in panel (b), run script COVID_CJE_HELOCs.R, which then runs script COVID_CJE_HELOCs_estim.R. Lines 473 to 506 of COVID_CJE_HELOCs_estim.R generate a file named HE_obs_vs_for_dev_pct_monthly_2020_MM.eps.

Figure 8: Consumers' Outstanding Balances, Alberta, 2012-2016

For credit cards, in panel (a), run script CC_HE_time_series_figs.R. Lines 226 to 267 generate a file named AB_CC_time_series.eps from the data in a file named tu_agg_AB_bc.csv.

For HELOCs, in panel (b), run script CC_HE_time_series_figs.R. Lines 276 to 317 generate a file named AB_HE_time_series.eps from the data in a file named tu_agg_AB_heloc.csv.

Figure 9: Deviations from Forecasted Balances in Alberta, October 2015

For credit cards, in panel (a), run script COVID_CJE_AB_Cards.R, which then runs script COVID_CJE_AB_Cards_estim.R. Lines 476 to 509 of COVID_CJE_AB_Cards_estim.R generate a file named AB_CC_obs_vs_for_dev_pct_monthly_2015-11.eps.

For HELOCs, in panel (b), run script COVID_CJE_AB_HELOCs.R , which then runs script COVID_CJE_AB_HELOCs_estim.R. Lines 476 to 509 of COVID_CJE_AB_HELOCs_estim.R generate a file named AB_HE_obs_vs_for_dev_pct_monthly_2015-11.eps.

Figure A1.1: Time Series of Aggregate Credit-Card Balances

The two panels, which are both generated with the same script, one showing balances and the other showing percent changes of all the series. Two of the series were created using the sample from the TransUnion database with the following script: TU_vs_BoC_comp.py

The other series is derived from an internal database housed at the Bank of Canada and collected from regulatory returns. It is available on the "Banking and Financial Statistics" Webpage of the Bank of Canada and is called Chartered bank selected assets: Month-end (formerly C1). We use the row of the table labeled "Credit cards".

Together, the aggregate time-series data are recorded in the file TU_vs_BoC_totals.csv. The figures in the file TU_vs_BoC_comparison.eps are then generated with the script CC_TU_vs_BoC_comp_figs.R, on lines 89 to 140.

Figure A1.2: Credit Data Coverage for Adults in Canada, by Province

The numbers in this figure were calculated with the scripts `TU_vs_StatsCan_comp.py` and `CC_TU_vs_StatsCan_comp_fig.R` in the Code folder. It requires the dataset `CC_TU_vs_StatsCan.csv`, comprising the number of credit-card account holders aged 20 and above and the figures obtained from Statistics Canada in the table called *Estimates of population (2016 Census and administrative data), by age group and sex for July 1st, Canada, provinces, territories, health regions (2018 boundaries) and peer groups,* Table: 17-10-0134-01. The file `CC_TU_vs_StatsCan_comp.eps` for Figure A1.2 is created by running lines 96 to 113 of the script `CC_TU_vs_StatsCan_comp_fig.R`.

Computing Requirements

Data Manipulation

The `csv` files in the Data folder were generated on the EDITH 2.0 High Performance cluster housed at the Bank of Canada. It is a cluster of Nvidia Tesla K80 GPU Accelerators, each with 12 GB of GDDR5 on-board memory, running 2496 processor cores, with base core clock speed of 560 MHz boost clocks from 562 MHz to 875 MHz, and with a memory clock speed of 2.5 GHz on 48 pieces of 256M $\times$ 16 GDDR5 SDRAM, producing a memory bandwidth of 240GB/s per GPU.

For the queries that generated the datasets, 36 CPUs with 240 GB of memory were sufficient to create the datasets within at most 24 hours each.

Statistical Analysis

Once the datasets have been saved in the `Data` folder, the remaining analysis, including the generation of all the tables and figures in the paper can be performed on a single microcomputer, such as a laptop computer. The particular model of computer on which the statistical analysis was run is a Dell Precision 3520, running a 64-bit Windows 10 operating system, with a 4-core x64-based processor, model Intel(R) Core(TM) i7-7820HQ CPU, running at 2.90GHz, with 16 GB of RAM.

Software

Data Manipulation

The data manipulation was conducted using a NoSQL dialect called Apache Spark, which is based on the functional programming language Scala and was implemented with PySpark in Python. The scripts were run using the Anaconda 2 distribution, version 4.3.1, with Python version 2.7 and PySpark version 2.3.0.

The batch jobs were submitted to the computing cluster using batch scheduling software called SLURM.

Other resources used to run the batch jobs include:

- `sbt`, version 1.3.6, which is a build tool for Scala, Java, among others
- `java`, version 1.8.0_141

Statistical Analysis

The statistical analysis was conducted in R, version 4.0.2, which was released on June 22, 2020, on a 64-bit Windows platform x86_64-w64-mingw32/x64.

The attached packages include the following:

- `data.table`, version 1.13.0 (using 4 threads), to handle the main data table for analysis in the `_prelim.R` and `_estim.R` scripts.
- `xtable`, version 1.8-4, to generate LaTeX tables for Tables 1, 2, and 3.
- `plot3D`, version 1.3, to produce a 3-D bar chart of transition frequency, which created the plots in Figure 3.
- `MASS`, version 7.3-51.6, was also used to estimate the smoothed surface of the transition density as an alternative to that in Figure 3 but was not included in the paper.

The creation of other figures, including Figures A1.1 and A1.2, required the following packages for data manipulation and graphics:

- `openxlsx`, version 4.2.3
- `dplyr`, version 1.0.5
- `lubridate`, version 1.7.10
- `ggplot2`, version 3.3.3
- `ggpubr`, version 0.4.0
- `ggthemes`, version 4.2.4
- `Cairo`, version 1.5-12.2

Upon attachment of the above packages, the following packages were loaded via a namespace, but not attached, with the following versions:

- `Rcpp` version 1.0.5
- `lattice` version 0.20-41
- `grid` version 4.0.2
- `DTMCPack` version 0.1-2
- `stats4` version 4.0.2
- `magrittr` version 1.5
- `RcppParallel` version 5.0.2
- `misc3d` version 0.8-4
- `markovchain` version 0.8.5-3
- `Matrix` version 1.2-18

- `tools` version 4.0.2
- `igraph` version 1.2.6
- `parallel` version 4.0.2
- `compiler` version 4.0.2
- `pkgconfig` version 2.0.3
- `matlab` version 1.0.2
- `nnet` version 7.3-14
- `expm` version 0.999-5
- `zip` version 2.1.1
- `cellranger` version 1.1.0
- `pillar` version 1.6.0
- `forcats` version 0.5.1
- `lifecycle` version 1.0.0
- `tibble` version 3.1.0
- `gtable` version 0.3.0
- `rlang` version 0.4.10
- `curl` version 4.3
- `haven` version 2.3.1
- `rio` version 0.5.26
- `stringr` version 1.4.0
- `withr` version 2.4.2
- `hms` version 1.0.0
- `generics` version 0.1.0
- `vctrs` version 0.3.7
- `grid` version 4.0.5
- `tidyselect` version 1.1.0
- `glue` version 1.4.2
- `R6` version 2.5.0

- `rstatix` version 0.7.0
- `fansi` version 0.4.2
- `readxl` version 1.3.1
- `foreign` version 0.8-81
- `carData` version 3.0-4
- `purrr` version 0.3.4
- `tidyr` version 1.1.3
- `car` version 3.0-10
- `scales` version 1.1.1
- `backports` version 1.2.1
- `ellipsis` version 0.3.1
- `abind` version 1.4-5
- `colorspace` version 2.0-0
- `ggsignif` version 0.6.1
- `utf8` version 1.2.1
- `stringi` version 1.5.3
- `munsell` version 0.5.0
- `broom` version 0.7.6
- `crayon` version 1.4.1

Acknowledgements

The views expressed are those of the authors; no responsibility for these views should be attributed to the Bank of Canada; all errors are the responsibility of the authors. We thank the HPC team at the Bank of Canada for their excellent assistance with the EDITH 2.0 High Performance Cluster.

References

Trade-Level Database, Run Dates 2017-01-01 to 2020-09-01, TransUnion, accessed October 2020.

Consumer Risk Characteristics Database, Run Dates 2017-01-01 to 2020-09-01, TransUnion, accessed October 2020.

Estimates of population (2016 Census and administrative data), by age group and sex for July 1st, Canada, provinces, territories, health regions (2018 boundaries) and peer groups, Table: 17-10-0134-01, Statistics Canada, accessed June 2020.

The Nilson Report, April 2020, Issue 1173, HSN Consultants, Inc., url: `https://nilsonreport.com/publication_newsletter_archive_issue.php?issue=1173`

Chartered bank selected assets: Month-end (formerly C1), Credit cards, Bank of Canada, accessed June 2020.