Postsecondary School Location Choice Model for the Greater Toronto Area

# Introduction

The purpose of this investigation is to develop an effective school location choice model for post-secondary students in Greater Toronto and Hamilton Area (GTHA). We present a brief literature review on relevant works in school location choice modelling in general, and in the GTHA specifically. We summarize key statistics from the sample data used, those from the 2015 and 2019 StudentMoveTO (SMTO) surveys. Next, we provide a detailed description of our modelling process. This process includes the implementation of a standard multinomial logit (MNL) model and a random forest (RF) classifier. Our results are included throughout this description. Finally, we list our conclusions and recommendations for further research in this area.

# Literature Review

In this section, we present a summary of relevant works for our problem.

## College Choice Modelling

From a conceptual point of view, frameworks for the college choice model have been presented by [Perna](https://link-springer-com.myaccess.library.utoronto.ca/content/pdf/10.1007%2F1-4020-4512-3.pdf) (2006) and [Acevedo-Gil](https://journals-scholarsportal-info.myaccess.library.utoronto.ca/details/13613324/v20i0006/829_ctaiccffls.xml) (2017). In empirical practice, the most common approach to model college choice is using a multinomial logit model (MNL), or variants thereof. However, other techniques can be used, such as a regression analysis as implemented by [Hearn](https://www-jstor-org.myaccess.library.utoronto.ca/stable/pdf/2112465.pdf) (1984). In this study, higher test scores, educational aspirations, parental income and academic achievement were found to be most correlated with higher selectivity, while belonging to certain ethnic and gender groups had negative effects on college selectivity. These results confirm previous findings of what Hearn calls “nonmeritocratic tendencies” in the American college choice system.

Such “nonmeritocratic tendencies” are observed in other studies as well. Specifically, [Niu et al.](https://journals-scholarsportal-info.myaccess.library.utoronto.ca/pdf/02727757/v25i0003/259_csattt1l.xml#page=14&zoom=100,0,0) (2006) find, in an analysis of college choices of Texas students, that Black and Hispanic students are less likely to enrol in more selective institutions (except for the most selective group), while the opposite pattern applies to Asian students. On the other hand, [Montgomery](https://journals-scholarsportal-info.myaccess.library.utoronto.ca/pdf/02727757/v21i0005/471_anlmotcoagbs.xml) (2002) uses a nested logit (NL) model to analyze choice of graduate business school and enrollment status (full-time or part-time) in the United States and finds that minorities and males are more responsive to school reputation, exhibiting a stronger preference for higher-reputation institutions. It is unclear to what extent such tendencies exist in the GTHA.

Another common observation found by [Montgomery](https://journals-scholarsportal-info.myaccess.library.utoronto.ca/pdf/02727757/v21i0005/471_anlmotcoagbs.xml) (2002) is that greater distance from home reduces the attractiveness of an institution. [Kohn et al.](https://www.rand.org/content/dam/rand/pubs/reports/2006/R1470.pdf) (1974) make this observation after implementing conditional logit models for college choice in Illinois and North Carolina. [Oosterbeek et al.](https://www.researchgate.net/publication/226885736_An_empirical_analysis_of_university_choice_and_earnings) (1992) also notice this pattern when using an MNL model to analyze data on college choices of Dutch economists. [Long](https://journals-scholarsportal-info.myaccess.library.utoronto.ca/pdf/03044076/v121i1-2/271_hhcdcootclcm.xml) (2004) uses a conditional logit model to analyze changes in college decisions over time and finds that distance from home negatively impacts the probability of attending a school. Interestingly, they note that this effect has decreased over time.

Finally, many studies show relationships between certain campus attributes and their perceived utilities. For instance, higher tuition fees are normally connected with lower utilities. [Kohn et al.](https://www.rand.org/content/dam/rand/pubs/reports/2006/R1470.pdf) (1974) find that the disutility due to greater tuition fees is smaller for higher-income groups in particular. [Long](https://journals-scholarsportal-info.myaccess.library.utoronto.ca/pdf/03044076/v121i1-2/271_hhcdcootclcm.xml) (2004) finds that students in 1992 consider institution quality and selectivity to be a more important factor than students in 1972 and 1982, and that higher tuition reduces college’s perceived utilities. [Sá et al.](https://journals-scholarsportal-info.myaccess.library.utoronto.ca/details/00343404/v46i0005/651_laaucodps.xml) (2012) use an NL model to predict living arrangement and university choice for Dutch post-secondary students, while [Niu and Tienda](https://journals-scholarsportal-info.myaccess.library.utoronto.ca/pdf/0049089x/v37i0002/416_cciamcs.xml) (2007) analyze school choice in Texas, and specifically investigate the effects of constraining choice sets in different ways. In both studies, institutional attributes such as quality and selectivity are used to model the base utility of each school.

## Works on SMTO Data

StudentMoveTO (n.d.) publishes a list of works which make use of the survey data. Many of these works analyze students’ commute patterns, including mode choice, and bike or license ownership. Chung et al. (2018) analyze living arrangement decisions for students at the University of Toronto. However, no published works as of yet have developed a school location choice model based on the StudentMoveTO (SMTO) data.

## School Location Choice in the GTHA

Past researchers from the University of Toronto’s [Travel Modelling Group](https://tmg.utoronto.ca/doc/1.4/gtamodel/index.html) have investigated school location choice models for the GTHA. Chen (2018) estimates a doubly-constrained gravity model was estimated using data from the 2016 [Transportation Tomorrow Survey](http://www.transportationtomorrow.on.ca/) (TTS). While this model was effective for students at the elementary and secondary levels, it was found to be ineffective for the post-secondary level. Wang (2015) estimates another doubly-constrained gravity model with an accessibility model for the utility term using data from the 2011 TTS. Likewise, this model was found to be ineffective for both full-time and part-time post-secondary students. Both these findings provide motivation for a more advanced post-secondary school location choice model to be developed.

# Contributions

Our analysis differs from previous works on school choice modelling on several fronts.

Firstly, our model is not representing the college choice process directly. Instead, our analysis is an exercise in matching students who have already made college choice decisions to their selected institutions. While there are many areas of overlap, an important difference is that household information reflects where students reside after having selected a college, and possibly, moving out from their parental homes.

Secondly, rather than predicting school selectivity and institution type, our analysis attempts to predict the specific institution attended by each student. Instead, our analysis places a greater weight on analyzing geographical patterns in school location choice, such as modelling the accessibility of each school location to each student.

Thirdly, we implement an RF classifier for our location choice problem, a novel approach in the field, and compare its utility to that of the classic econometric approaches.

# Methods

## Logit Model

As indicated in the literature review, the logit model is the most popular technique used to analyze the college choice problem.

### Multinomial Logit Model

Our primary empirical method is using a multinomial logit model (MNL). This model, developed by [McFadden](https://eml.berkeley.edu/reprints/mcfadden/zarembka.pdf) (1973), and described in depth by [Train](https://eml.berkeley.edu/books/choice2.html) (2003), is based on the principles of random utility frameworks. The perceived utility of alternative for student is, where is the systematic utility and is the random utility. A student selects alternative if and only if . By assuming that the random utility terms are independent and identically distributed with a type-1 extreme value distributions, we obtain the MNL. In this formulation, the probability of alternative being chosen by individual from choice set is:

One property of MNL models is that it is consistent with the Independence from Irrelevant Alternatives (IIA) assumption. Namely, this is the property that the probability of alternative being selected over alternative is independent of the other alternatives in the choice set.

### Nested Logit Model

When this assumption does not hold, an extension of the MNL model, known as the nested logit (NL) model, can be used ([Ben-Akiva and Bierlaire](https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.40.8438&rep=rep1&type=pdf) 1999). In this model, each alternative is placed into one nest, and it is assumed that the error terms in the utilities for the alternatives within each nest are correlated. The probability of alternative from nest including alternatives being chosen by individual from the set of nests is:

Here, is the scale parameter reflecting the correlation between the random components of the utility of the nests (at the top level of the model) and reflects the correlation among alternatives in nest (at the lower level of the model). The term is known as the logsum or inclusive value of nest , and represents the expected maximum utility for the choice of alternatives in the nest. In order for this formulation to be consistent with the random utility maximization framework, . Note that if all then the nesting structure is redundant and the model collapses into the standard MNL.

### Implementation

We implement these models using mlogit 1.1-0 ([Croissant 2020](https://cran.r-project.org/web/packages/mlogit/mlogit.pdf)) with RStudio 1.3.959 and R 4.0.0. In some cases. [Biogeme 3.2.6](https://transp-or.epfl.ch/documents/technicalReports/Bier20.pdf) (Bierlaire 2020) was used with Python 3.7.6 and Jupyter Notebook 6.0.3.

## Random Forest

Random forests (RFs) are a machine learning technique that have been successfully applied in various fields, including genetics, clinical medicine, and bioinformatics ([Strobl et al.](https://psycnet.apa.org/record/2009-22665-002) 2009). Developed by Breiman (2001), the RF training algorithm is developed as follows ([Hastie et al.](https://link.springer.com/chapter/10.1007%2F978-0-387-84858-7_15) 2009). For :

1. Draw a bootstrapped sample from the training data.
2. Grow a decision tree using the bootstrapped data by performing the following steps recursively until minimum node size is reached:
   1. Select features from the training data at random
   2. Select the best feature and split-point from the features according to some split criterion
   3. Split the node using that feature and split-point

The RF is the set of trees . The prediction for a given input is the majority vote for the predicted output from all trees. Several hyperparameters can be adjusted in this algorithm. They include:

* , the number of trees
* , the minimum size for leaf nodes
* , the number of features to consider for each split point
* The splitting criteria to be used
* The maximum tree depth

We implement the RF algorithm using [scikit-learn](https://scikit-learn.org/stable/index.html) 0.22.1 with Python 3.7.6 and Jupyter Notebook 6.0.3.

## Metrics Reported

Throughout this report, the following metrics are reported. Table 1 lists these metrics and how they are calculated. A few notes:

* The softmax accuracy is generally prioritized over the hardmax accuracy since it reflects the probabilities assigned to correct observations and is less sensitive to the imbalances in the data (a “reasonable” hardmax accuracy can be reached by predicting the largest campus for all students).
* The log likelihood can only be reported if no actual observations are assigned a probability of 0 (as this would yield a log likelihood of negative infinity).
* The McFadden rho-squared can only be reported where alternative-specific coefficients are used, and hence is not used throughout much of our analysis.

Table 1: Summary of Reported Metrics. Note that is the log-likelihood for the logit model with only alternative-specific constants, as explained in [McFadden](https://eml.berkeley.edu/~mcfadden/travel/ch5.pdf) (1975). represents the probability assigned by the model that student attends campus , and represents the campus actually attended by student .

|  |  |
| --- | --- |
| Metric | Calculation |
| Hardmax Accuracy |  |
| Softmax Accuracy |  |
| Log Likelihood |  |
| McFadden Pseudo Rho-Squared |  |

# Data

## 2015 SMTO

The 2015 SMTO dataset is the primary one used to develop a proposed school location choice model. Seven campuses are included in the survey: three University of Toronto campuses (St. George - SG, Scarborough - SC, and Mississauga - MI), two York University campuses (Keele - YK, and Glendon - YG), Ryerson University - RY, and OCAD University - OC. Rows whose indicated enrollment level was “Other” (as opposed to “UG” or “Grad”) were removed from the sample; these rows were also the only rows whose enrollment status was indicated as “Other” (as opposed to “FT” or “PT”). Table 2 tabulates important characteristics of this filtered dataset.

Note that to make our model generalizable to TTS data, we only use certain attributes in our analysis. The one exception is living arrangement, which is not available in TTS but is used regardless. A gradient-boosting machine to classify student living arrangement given other attributes has been trained. The model achieves an accuracy of over 90%, and so we keep living arrangement in the list of available attributes.

## 2019 SMTO

Table 3 presents the summary statistics for the 2019 SMTO data, which includes data from 27 university and college campuses (including those from the 2015 survey).

Table 2: Tabulation of Key Variables in 2015 SMTO Dataset. Family: whether the student’s indicated living arrangement is “Live with family/parents”. For “Income”, Low is < $60,000. For “Mode”, Active is walk or bicycle.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | MI | OC | RY | SC | SG | YG | YK | Total | Share |
| Count | Total | 930 | 455 | 2708 | 1074 | 5912 | 315 | 3084 | 14478 | 100.0% |
| Level | UG | 858 | 403 | 2420 | 1018 | 3571 | 298 | 2464 | 11032 | 76.2% |
|  | Grad | 72 | 52 | 288 | 56 | 2341 | 17 | 620 | 3446 | 23.8% |
| Status | FT | 893 | 408 | 2557 | 1028 | 5425 | 295 | 2840 | 13446 | 92.9% |
|  | PT | 37 | 47 | 151 | 46 | 487 | 20 | 244 | 1032 | 7.1% |
| Gender | Female | 653 | 337 | 1739 | 745 | 3860 | 256 | 2078 | 9668 | 66.8% |
|  | Male | 268 | 105 | 953 | 323 | 2007 | 56 | 972 | 4684 | 32.4% |
|  | Other | 9 | 13 | 16 | 6 | 45 | 3 | 34 | 126 | 0.9% |
| Family | True | 666 | 230 | 1861 | 791 | 2452 | 205 | 2014 | 8219 | 56.8% |
|  | False | 264 | 225 | 847 | 283 | 3460 | 110 | 1070 | 6259 | 43.2% |
| Income | High | 139 | 56 | 515 | 159 | 1000 | 57 | 523 | 2449 | 16.9% |
|  | Low | 172 | 102 | 601 | 240 | 1425 | 74 | 767 | 3381 | 23.4% |
|  | Unknown | 619 | 297 | 1592 | 675 | 3487 | 184 | 1794 | 8648 | 59.7% |
| Commute Mode | Transit | 566 | 302 | 2086 | 697 | 3139 | 217 | 2232 | 9239 | 63.8% |
| Active | 94 | 138 | 508 | 153 | 2511 | 41 | 360 | 3805 | 26.3% |
| Auto | 227 | 13 | 105 | 222 | 218 | 52 | 477 | 1314 | 9.1% |
| Other | 43 | 2 | 9 | 2 | 44 | 5 | 15 | 120 | 0.8% |
| Distance | Mean | 15.10 | 15.16 | 18.61 | 13.92 | 11.25 | 16.97 | 17.42 | 14.63 | ­­­ |
|  | Std. Dev | 12.95 | 15.23 | 14.12 | 11.65 | 12.89 | 12.62 | 12.06 | 13.31 |  |

Table 3: Tabulation of Key Variables in 2019 SMTO Dataset. Family: whether the student’s indicated living arrangement is “Live with family/parents”. For “Income”, Low is < $90,000. For “Mode”, Active is walk or bicycle.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Uni - UG | Uni - Grad | College | Total | Share |
| Count | Total | 10396 | 2652 | 3468 | 16516 | 100.0% |
| Status | FT | 10002 | 2434 | 3322 | 15758 | 95.4% |
|  | PT | 394 | 218 | 146 | 758 | 4.6% |
| Family | True | 4417 | 558 | 1081 | 6056 | 36.7% |
|  | False | 2546 | 1386 | 1017 | 4949 | 30.0% |
|  | Unknown | 3433 | 708 | 1370 | 5511 | 33.4% |
| Work | FT | 636 | 421 | 240 | 1297 | 7.9% |
|  | PT | 4739 | 1148 | 1817 | 7704 | 46.6% |
|  | None | 5021 | 1083 | 1411 | 7515 | 45.5% |
| Income | High | 1429 | 257 | 156 | 1842 | 11.2% |
|  | Low | 2527 | 358 | 787 | 3672 | 22.2% |
|  | Unknown | 6440 | 2037 | 2525 | 11002 | 66.6% |
| Commute Mode | Transit | 4366 | 1071 | 1150 | 6587 | 39.9% |
|  | Active | 1466 | 638 | 282 | 2386 | 14.4% |
|  | Auto | 794 | 188 | 551 | 1533 | 9.3% |
|  | Other | 42 | 17 | 13 | 72 | 0.4% |
|  | Unknown | 3728 | 738 | 1472 | 5938 | 36.0% |
| Age | Mean | 20.93 | 27.96 | 24.60 | 22.83 |  |
|  | Std. Dev. | 4.87 | 7.13 | 7.95 | 6.59 |  |

# 2015 SMTO Location Choice – Logit Model

## Gravity Model

The first step in our logit model analysis for the 2015 SMTO dataset was estimating simple doubly-constrained gravity models for separate segments in the sample. [Anas](https://reader.elsevier.com/reader/sd/pii/0191261583900231?token=48ECD9B1BF07383098D0C6195509A9982E0B5E7E379C4F5DEF339D8C67C0F7AC22BBDA76A819AFED5A35FDE34934863E) (1983) shows that the doubly-constrained gravity model is equivalent to MNL with impedance function . Thus, in our case, the utility is specified as , where is known as an alternative-specific constant for alternative and represents the network distance between the student’s home zone and each campus. These network distances were obtainedfrom level-of-service matrices generated using the 2016 TTS data and GTAModel V4.1**.** Table 4 presents our findings from this analysis.

Table 4: Estimation Results for Segmented Gravity Models. UG = Undergrad, G = Grad, FT = Full-time, PT = Part-time, F = Live with family/parents, N = Another living arrangement. \*\*\* indicates p < 0.001.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Segment | Count |  | Std. Error |  |
| UG FT N | 3461 | -0.16042 | 0.00378 | \*\*\* |
| UG FT F | 6879 | -0.06148 | 0.00155 | \*\*\* |
| UG PT N | 314 | -0.08704 | 0.00940 | \*\*\* |
| UG PT F | 378 | -0.06178 | 0.00709 | \*\*\* |
| G FT N | 2250 | -0.09309 | 0.00419 | \*\*\* |
| G FT F | 856 | -0.05527 | 0.00729 | \*\*\* |
| G PT | 340 | -0.07323 | 0.01260 | \*\*\* |
| All | 14478 | -0.08336 | 0.00121 | \*\*\* |

From this, we see that there are significant differences between the parameters estimated for groups. In particular, the distance parameter is of greatest magnitude for students not living with their family/parents, and especially so for full-time undergraduates. This result is reasonable as students in this group are likely to be living on residence or to have selected their place of residence according to their place of school (rather than the other way around, as the model implies).

## Accessibility Model

The doubly-constrained gravity model used above uses network distance as a very simple representation of the impedance observed by each student travelling to each campus. We hypothesize that our model can be significantly improved by including a more robust measure of the accessibility of each school location to each student. For this reason, we estimate a mode choice model, with the intention of including the expected maximum utility from the mode choice model for each alternative as an accessibility term in the location choice model.

### Basic Mode Choice Model

Our mode choice model takes the form of an MNL model with three alternatives: Auto, Transit, and Active Mode (walking or biking). Our reference mode choice model uses the utility function , where is the morning travel time from student ’s home zone to campus , obtained using the same level-of-service matrices as above. Note that our model includes alternative-specific constants to reproduce aggregate mode shares, and alternative-specific coefficients for travel time. 120 students whose indicated modes are not one of the three alternatives are removed from the mode choice estimation set. Table 5 lists the results from this base model.

Table 5: Base Mode Choice Model. . Active mode travel times are calculated given speeds of 4km/h. All travel times are in minutes. \*\*\* indicates p < 0.001.

|  |  |  |  |
| --- | --- | --- | --- |
| Parameter | Estimate | Std. Error |  |
|  | 1.960 | 0.0495 | \*\*\* |
|  | -2.121 | 0.0640 | \*\*\* |
|  | -0.0075 | 0.0008 | \*\*\* |
|  | -0.0339 | 0.0007 | \*\*\* |
|  | -0.0165 | 0.0019 | \*\*\* |
| Metric | Result | Metric | Result |
| Hardmax Accuracy | 0.822 | McFadden R^2 | 0.375 |
| Softmax Accuracy | 0.706 | Log likelihood | -7663.5 |

One result to notice is that the time parameters for each mode are significantly different, so we keep these alternative-specific coefficients rather than prescribe a generic coefficient.

### Socioeconomic Variables in Mode Choice Model

After estimating this base model, we test the influence of including various individual and household characteristics in the model. Since these characteristics are consistent across all alternatives, alternative-specific coefficients must be used. The variables whose impact is most significant are living arrangement and license ownership. Hence, the updated utility function is

where {1 if student lives with family/parents, 0 otherwise}, and

{1 if student owns a driver’s license, 0 otherwise}.

The results for this model are shown in Table 6. As expected, students living with their family and/or owning a driver’s license are more likely to drive to school, and less likely to use an active mode.

### Availability Restrictions in Mode Choice Model

To further refine our mode choice model, we consider restricting the choice sets of certain alternatives in certain choice situations. Two types of availability restrictions are considered: for active modes and auto. Note that in evaluating these restrictions, log likelihood cannot be used since any actual observations rendered unavailable would result in a log likelihood of negative infinity.

Students are unlikely to use an active mode (walking or cycling) for trips of long distances due to factors such as the associated physical exertion, dangers, and exposure to weather, in addition to the long travel times. Therefore, we test models where the active mode alternative is marked as unavailable in cases where the estimated walking travel time is greater than some threshold .

Table 6: Mode Choice Model with Selected Socioeconomic Variables.. Active mode travel times are calculated given speeds of 4km/h. All travel times are in minutes. \*\*\* indicates p < 0.001.

|  |  |  |  |
| --- | --- | --- | --- |
| Parameter | Estimate | Std. Error |  |
|  | 1.878 | 0.0612 | \*\*\* |
|  | -3.406 | 0.1043 | \*\*\* |
|  | -1.459 | 0.0743 | \*\*\* |
|  | 0.323 | 0.0737 | \*\*\* |
|  | 0.350 | 0.0625 | \*\*\* |
|  | 1.753 | 0.0857 | \*\*\* |
|  | -0.0070 | 0.0008 | \*\*\* |
|  | -0.0290 | 0.0007 | \*\*\* |
|  | -0.0221 | 0.0021 | \*\*\* |
| Metric | Result | Metric | Result |
| Hardmax Accuracy | 0.826 | McFadden R^2 | 0.418 |
| Softmax Accuracy | 0.722 | Log likelihood | -7139.5 |

Figure 1 shows how the softmax accuracy changes for various walking time thresholds. The optimal accuracy is reached when a threshold of about 46.6 minutes is imposed, yielding a softmax accuracy of 75.04%.

Figure 1: Effect of Imposing Various Threshold Travel Times for Active Mode on Softmax Accuracy.

Students are also unlikely to choose the auto option if there is no vehicle available in their household. As such, we test models where the auto mode is only available if the student has indicated that their household owns at least vehicles, where . These restrictions are tested both with and without the active mode threshold of 46.6 minutes from above. We find that setting yields the best softmax accuracy in both cases.

### Proposed Mode Choice Model

Our proposed mode choice model uses the utility model and optimal availability restrictions described above ( and ). Table 7 presents the results for this proposed model, while Tables 8 and 9 show confusion matrices for this model.

Table 7: Proposed Mode Choice Model.. Active mode travel times are calculated given speeds of 4km/h. All travel times are in minutes. \*\*\* indicates p < 0.001. McFadden R^2 and Log likelihood not reported since availability restrictions result in probabilities of 0 assigned to observed choices in some cases.

|  |  |  |  |
| --- | --- | --- | --- |
| Parameter | Estimate | Std. Error |  |
|  | 3.890 | 0.169 | \*\*\* |
|  | -1.696 | 0.169 | \*\*\* |
|  | -0.992 | 0.147 | \*\*\* |
|  | -0.792 | 0.121 | \*\*\* |
|  | 0.613 | 0.113 | \*\*\* |
|  | 1.622 | 0.117 | \*\*\* |
|  | -0.0113 | 0.001 | \*\*\* |
|  | -0.0936 | 0.006 | \*\*\* |
|  | -0.0378 | 0.003 | \*\*\* |
| Metric | Result | Metric | Result |
| Hardmax Accuracy | 0.837 | McFadden R^2 | N/A |
| Softmax Accuracy | 0.792 | Log likelihood | N/A |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 8: Hardmax Confusion Matrix from Proposed Mode Choice Model   |  |  |  |  | | --- | --- | --- | --- | | Obs\Pred | Active | Auto | Transit | | Active | 3116 | 13 | 676 | | Auto | 41 | 81 | 1191 | | Transit | 360 | 66 | 8813 | | Table 9: Softmax Confusion Matrix from Proposed Mode Choice Model   |  |  |  |  | | --- | --- | --- | --- | | Obs\Pred | Active | Auto | Transit | | Active | 2827.7 | 32.7 | 944.6 | | Auto | 36.2 | 269.0 | 1007.8 | | Transit | 304.2 | 662.5 | 8272.4 | |

As we can see, the accuracy of our mode choice model is reasonable enough. For reference, a model with only alternative-specific constants would result in a hardmax accuracy of 62.7% (the market share of Transit) and softmax accuracy of 46.6% (the sum of the squares of the market shares). It is now time to see whether we can use the mode choice model to add a more detailed measure of school location accessibility in the college choice problem.

### Results

We integrate our mode choice model in the location choice model as a logsum term, as outlined by [Ben-Akiva and Lerman](https://search-proquest-com.myaccess.library.utoronto.ca/docview/1309215151?pq-origsite=summon) (1985). As such, the updated utility function is , where the utility of any unavailable mode is . We compare this model with the distance model developed previously.

Table 10 shows a comparison of these results. We see that the accessibility model does not outperform the much simpler distance model. Meanwhile, the combination of both models offers a small improvement over the distance model.

Table 10: Comparison of Distance and Accessibility Models. \*\*\* indicates p < 0.001.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Distance | | | Accessibility | | | Combined | | |
| Parameter | Estimate | Std. Error |  | Estimate | Std. Error |  | Estimate | Std. Error |  |
|  | -1.648 | 0.043 | \*\*\* | -1.249 | 0.038 | \*\*\* | -1.382 | 0.041 | \*\*\* |
|  | -2.525 | 0.049 | \*\*\* | -2.566 | 0.049 | \*\*\* | -2.514 | 0.049 | \*\*\* |
|  | -0.749 | 0.023 | \*\*\* | -0.781 | 0.024 | \*\*\* | -0.748 | 0.024 | \*\*\* |
|  | -1.423 | 0.037 | \*\*\* | -1.453 | 0.035 | \*\*\* | -1.396 | 0.037 | \*\*\* |
|  | -2.812 | 0.058 | \*\*\* | -2.473 | 0.058 | \*\*\* | -2.589 | 0.059 | \*\*\* |
|  | -0.520 | 0.024 | \*\*\* | -0.226 | 0.024 | \*\*\* | -0.340 | 0.025 | \*\*\* |
|  | -0.0834 | 0.001 | \*\*\* |  |  |  | -0.0490 | 0.002 | \*\*\* |
|  |  |  |  | 0.8413 | 0.014 | \*\*\* | 0.4771 | 0.016 | \*\*\* |
| Metric |  |  |  |  |  |  |  |  |  |
| Hardmax Accuracy | 0.480 |  |  | 0.460 |  |  | 0.475 |  |  |
| Softmax Accuracy | 0.330 |  |  | 0.345 |  |  | 0.352 |  |  |
| Log Likelihood | -19814 |  |  | -19840 |  |  | -19315 |  |  |
| McFadden R^2 | 0.128 |  |  | 0.127 |  |  | 0.150 |  |  |

## Campus-specific Attributes

At this point, we try integrating campus-specific attributes into our model. The following attributes are considered: total enrollment, average first-year domestic tuition for the Arts and Science program, proportion of domestic enrollment in undergraduate programs, and secondary school admission averages. These values were obtained from 2015-16 data from [Common University Data Ontario](https://cudo.ouac.on.ca/) (CUDO). Table 11 summarizes the attributes of the seven campuses:

Table 11: Summary of Campus Attributes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| School Code | Tuition (CAD) | Domestic Enrollment | Admission Average | Total Enrollment |
| SG | 7519 | 80.8% | 89.3% | 53930 |
| SC | 7813 | 83.8% | 84.1% | 11770 |
| MI | 7670 | 82.8% | 83.0% | 13298 |
| YK | 7339 | 89.2% | 81.7% | 41142 |
| YG | 7339 | 89.2% | 81.7% | 2457 |
| RY | 7026 | 96.7% | 84.0% | 28159 |
| OC | 7052 | 90.0% | 82.4% | 3491 |

One objective in our analysis is to make our model generalizable to expanded choice sets. By this, we mean avoiding alternative-specific constants and coefficients. Instead, we try replacing these alternative-specific constants by using the campus attributes described above. Upon expanding the choice set, such attributes can be included in the model directly.

Unlike many of the studies in our literature review, we found that including tuition information was not found to significantly improve our models. Even interacting tuition with household income was not found to be effective. This result can be attributed to the minor differences between tuition fees for different schools.

Domestic enrollment rate was found to be most useful when interacted with living arrangement. Presumably, students living with their family are more likely to be domestic, and hence attend a school with more domestic students.

Table 12 presents the results for the selected model that includes campus attributes as described above.

Table 12: Location Choice Models with Campus Attributes. \*\*\* indicates p < 0.001. if the student lives with their family/parents, otherwise .

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Distance | | | Accessibility | | | Combined | | |
| Parameter | Estimate | Std. Error |  | Estimate | Std. Error |  | Estimate | Std. Error |  |
|  | -0.0843 | 0.0012 | \*\*\* |  |  |  | -0.0535 | 0.0016 | \*\*\* |
|  |  |  |  | 0.8176 | 0.0135 | \*\*\* | 0.4389 | 0.0161 | \*\*\* |
|  | 0.8007 | 0.0144 | \*\*\* | 0.8665 | 0.0146 | \*\*\* | 0.8212 | 0.0146 | \*\*\* |
|  | 0.0743 | 0.0038 | \*\*\* | 0.0279 | 0.0040 | \*\*\* | 0.0489 | 0.0040 | \*\*\* |
|  | 0.0411 | 0.0021 | \*\*\* | 0.0229 | 0.0021 | \*\*\* | 0.0311 | 0.0022 | \*\*\* |
| Metric |  |  |  |  |  |  |  |  |  |
| Hardmax Accuracy | 0.476 |  |  | 0.455 |  |  | 0.479 |  |  |
| Softmax Accuracy | 0.334 |  |  | 0.342 |  |  | 0.352 |  |  |
| Log Likelihood | -19654 |  |  | -19840 |  |  | -19235 |  |  |

Interestingly, while the distance and combined models have generally improved, the accessibility has gotten less effective. It outperforms the distance model only on the softmax accuracy. Note, also, that the estimated parameters for admission average confirm previous findings that greater selectivity increases the attractiveness of schools.

## Socioeconomic Variables Interacted with Distance and Accessibility

At this point, we consider the addition of socioeconomic variables to the model. Since we are avoiding alternative-specific coefficients and have already explored their interaction with campus-specific attributes, we now consider interacting these variables with distance and/or accessibility. Having previously seen significant differences in distance parameters estimated for different segments, we expect this addition to have a significant effect.

Upon iterating with several approaches, we include the most significant variables in our model. Table 13 presents the results for the model with this selected formulation.

Table 13: Socioeconomic Variables in Location Choice Models. \*\*\* indicates p < 0.001. if the student lives with their family/parents, otherwise . if the student works full-time, otherwise . if the student’s enrollment status is part-time, otherwise .

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Distance | | | Accessibility | | | Combined | | |
| Parameter | Estimate | Std. Error |  | Estimate | Std. Error |  | Estimate | Std. Error |  |
|  | -0.1231 | 0.002 | \*\*\* |  |  |  | -0.0736 | 0.003 | \*\*\* |
|  |  |  |  | 0.8176 | 0.014 | \*\*\* | 0.4468 | 0.021 | \*\*\* |
|  | 0.8043 | 0.014 | \*\*\* | 0.8665 | 0.015 | \*\*\* | 0.8115 | 0.014 | \*\*\* |
|  | 0.0567 | 0.004 | \*\*\* | 0.0279 | 0.004 | \*\*\* | 0.0476 | 0.004 | \*\*\* |
|  | 0.0327 | 0.002 | \*\*\* | 0.0229 | 0.002 | \*\*\* | 0.0298 | 0.002 | \*\*\* |
|  | 0.0589 | 0.003 | \*\*\* |  |  |  | 0.0153 | 0.004 | \*\*\* |
|  |  |  |  |  |  |  | -0.3020 | 0.043 | \*\*\* |
|  |  |  |  |  |  |  | -0.4667 | 0.100 | \*\*\* |
|  |  |  |  |  |  |  | -0.2288 | 0.053 | \*\*\* |
| Metric |  |  |  |  |  |  |  |  |  |
| Hardmax Accuracy | 0.4803 |  |  | 0.4552 |  |  | 0.4807 |  |  |
| Softmax Accuracy | 0.3435 |  |  | 0.3416 |  |  | 0.3550 |  |  |
| Log Likelihood | -19406 |  |  | -19840 |  |  | -19133 |  |  |

We see that students living with their family are less sensitive to distance and accessibility than their non-family counterparts. We also see that students working full-time are less sensitive to accessibility, perhaps since they measure the accessibility of their school location with respect to their workplace rather than their place of residence. A similar effect is observed to a lesser extent in part-time workers.

## Closest School Dummies

Our final addition to the location choice model is including what we call “closest school dummies” , representing which school is closest to the student’s home zone. That is, . Moreover, we hypothesize that the closest school dummies are most meaningful for schools within a certain distance of the student’s home zone. Upon experimenting with this maximum distance, we choose to impose a threshold of 2km, such that . Table 14 indicates the results for all three models when these closest dummies are incorporated.

We see that the distance model performs almost as well as the combined model, despite including much fewer variables (recall that the accessibility term includes a non-trivial mode choice model). Furthermore, the accessibility model performs quite poorly to the other models. Due to this effectiveness and the relative simplicity of the distance model, we select this model as the preferred one. Notice that the parameter for the closest school dummy is positive and significant, as would be expected.

Table 14: Closest Dummies in Location Choice Models. \*\*\* indicates p < 0.001. if the student lives with their family/parents, otherwise . if the student works full-time, otherwise . if the student’s enrollment status is part-time, otherwise .

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Distance | | | Accessibility | | | Combined | | |
| Parameter | Estimate | Std. Error |  | Estimate | Std. Error |  | Estimate | Std. Error |  |
|  | -0.1043 | 0.002 | \*\*\* |  |  |  | -0.0832 | 0.003 | \*\*\* |
|  |  |  |  | 0.8663 | 0.018 | \*\*\* | 0.2671 | 0.028 | \*\*\* |
|  | 0.9287 | 0.042 | \*\*\* | -0.2310 | 0.053 | \*\*\* | 0.5444 | 0.060 | \*\*\* |
|  | 0.8011 | 0.014 | \*\*\* | 0.8660 | 0.015 | \*\*\* | 0.8058 | 0.015 | \*\*\* |
|  | 0.0599 | 0.004 | \*\*\* | 0.0261 | 0.004 | \*\*\* | 0.0539 | 0.004 | \*\*\* |
|  | 0.0331 | 0.002 | \*\*\* | 0.0223 | 0.002 | \*\*\* | 0.0315 | 0.002 | \*\*\* |
|  | 0.0414 | 0.003 | \*\*\* |  |  |  | 0.0214 | 0.004 | \*\*\* |
|  |  |  |  |  |  |  | -0.2185 | 0.043 | \*\*\* |
|  |  |  |  |  |  |  | -0.4222 | 0.098 | \*\*\* |
|  |  |  |  |  |  |  | -0.2142 | 0.052 | \*\*\* |
| Metric |  |  |  |  |  |  |  |  |  |
| Hardmax Accuracy | 0.4646 |  |  | 0.4537 |  |  | 0.4734 |  |  |
| Softmax Accuracy | 0.3540 |  |  | 0.3414 |  |  | 0.3565 |  |  |
| Log Likelihood | -19154 |  |  | -19831 |  |  | -19091 |  |  |

## Proposed Model

Our proposed logit model for school location choice is therefore based on the following utility function:

where:

is the systematic utility for student and campus ,

is the total enrollment of campus ,

is the mean secondary school admission average for campus ,

is the domestic enrollment rate for undergraduate programs at campus ,

if student lives with their family/parents, otherwise,

is the network distance between student ’s home zone and campus ,

if , otherwise, and

, , , , , and are estimated parameters.

The estimated parameters and metrics for this proposed model are those reported for the “Distance” model in the table above. Confusion matrices for this model are presented in Tables 15 and 16.

Table 15: Hardmax Confusion Matrix for Proposed Logit Location Choice Model

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Obs\Pred** | **MI** | **OC** | **RY** | **SC** | **SG** | **YG** | **YK** |
| **MI** | 404 | 0 | 8 | 0 | 339 | 0 | 179 |
| **OC** | 27 | 0 | 53 | 6 | 273 | 0 | 96 |
| **RY** | 194 | 0 | 318 | 36 | 1431 | 0 | 729 |
| **SC** | 6 | 0 | 18 | 184 | 639 | 0 | 227 |
| **SG** | 239 | 0 | 529 | 53 | 4256 | 0 | 834 |
| **YG** | 13 | 0 | 8 | 3 | 203 | 0 | 88 |
| **YK** | 172 | 0 | 54 | 30 | 1264 | 0 | 1564 |

Table 16: Softmax Confusion Matrix for Proposed Logit Location Choice Model

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Obs\Pred** | **MI** | **OC** | **RY** | **SC** | **SG** | **YG** | **YK** |
| **MI** | 271.3 | 26.3 | 166.1 | 19.9 | 277.5 | 11.9 | 157.1 |
| **OC** | 23.2 | 19.5 | 116.1 | 24.2 | 190.1 | 8.7 | 73.2 |
| **RY** | 166.9 | 90.9 | 687.8 | 173.4 | 997.2 | 58.2 | 533.7 |
| **SC** | 18.3 | 29.5 | 224.7 | 247.6 | 329.0 | 31.9 | 192.9 |
| **SG** | 227.6 | 227.1 | 1375.2 | 250.4 | 2897.5 | 102.2 | 831.0 |
| **YG** | 15.4 | 9.7 | 71.0 | 22.1 | 114.7 | 10.7 | 71.4 |
| **YK** | 176.4 | 86.8 | 595.3 | 169.3 | 1000.0 | 66.6 | 989.7 |

# 2015 SMTO Location Choice – Random Forest

To develop our random forest (RF) location choice model, we experimented with socioeconomic variables (such as living arrangement and enrollment status) and geospatial variables (such as distances to the campuses). Due to the nature of RFs, passing in campus-specific information is impossible and unnecessary. Instead, the algorithm is expected to identify patterns related by campus-specific attributes implicitly.

## Geospatial Variables

Since the meaning of different variables, and their relationships to other variables, are not inherently captured in the RF, we perform a feature engineering process to identify the best format in which to input geospatial information. We considered the following approaches:

* Home zones (H): The student’s home zone passed directly
* Coordinates (C): The latitude and longitude of the student’s home zone’s centroid, normalized between 0 and 1
* Planning districts (P): The number of the planning district containing the student’s home zone (see <http://dmg.utoronto.ca/pdf/tts/2016/2016TTS_DataGuide.pdf>)
* Distances (D): The distances between the student’s home zone and each campus
* Distance rank labels (L): One-hot encoded columns labeling the th closest school to student’s home zone
* Ranked distances (R): Columns labelled containing the distance to the th closest school from the student’s home zone

For approaches D, L, and R, we also consider passing a subset of these columns. For approach D, passing in only three columns seemed sufficient. This finding can be explained by the fact that three distances is sufficient to pinpoint a student’s home location, so including more columns is superfluous. For approach L, we combine the columns for SG, RY, and OC because of these campuses’ geographic proximity.

The results for various approaches showed limited fluctuations, even for the HZ approach. Because many of these approaches include variables that are high-cardinality, we hypothesize that the RF model is overfitting the data by identifying “patterns” on what is, essentially, a zone-by-zone basis.

We decide to use the PD approach in future analysis, as this is a lower-cardinality variable that is less susceptible to overfitting, yet provides valuable information on what part of the region a student lives in. We also keep approach L for , as this is variable improves the model but is unlikely to lead to overfitting.

## Feature Selection

One benefit of RFs is that feature importances can be calculated and analyzed. These can be used to identify the most relevant features for the problem, a process known as feature selection. In our feature selection process, we investigate various individual and household characteristics. We see that the most important variables are living arrangement, level of study, household income range, and employment status. We also see that the closest school dummy for YG is least significant.

## Hyperparameter Tuning

Upon selecting the format for our model input, we proceed to hyperparameter tuning. We use a randomized search with cross-validation and then a grid search with cross-validation, as implemented by *sklearn*, to establish the best hyperparameter settings. However, the gains from this approach were marginal: the optimal parameters were very similar to the default parameters and only small changes in the performance metrics were seen.

## Proposed Model

Table 17 shows the results and feature importances for our proposed RF model. The metrics reported are the model’s average performance on a separate testing set across ten trials.

The out-of-sample hardmax accuracy for the RF model is almost two percentage points higher than for the MNL model. This is unsurprising since the RF is trained such as to optimize hardmax accuracy. More interestingly, the RF out-of-sample softmax accuracy is over 4.5 percentage points better than for our proposed logit model. As such, the RF is outperforming the MNL model on both metrics.

Table 17: Selected Random Forest Model. All features except PlanningDistrict are dummy variables. DT = Any downtown campus, YK = York Keele, MI = U of T Mississauga, SC = U of T Scarborough. PT = Part-time. Log likelihood, rho sqaured not reported since some observed choices are assigned a probability of 0.

|  |  |  |
| --- | --- | --- |
| Feature | Importance | Std. Dev |
| PlanningDistrict | 42.99% | 0.59% |
| LevelGrad | 9.69% | 0.11% |
| Family | 8.12% | 0.20% |
| Closest.DT | 7.78% | 0.46% |
| Closest.YK | 5.91% | 0.28% |
| StatusPT | 4.00% | 0.09% |
| WorkYes | 3.80% | 0.11% |
| IncomeLow | 3.78% | 0.12% |
| IncomeHigh | 3.78% | 0.06% |
| Closest.MI | 3.54% | 0.19% |
| Closest.SC | 3.47% | 0.13% |
| WorkNo | 3.14% | 0.11% |
| Metric | Testing Set | Training Set |
| Hardmax Accuracy | 48.30% | 58.07% |
| Softmax Accuracy | 39.97% | 46.69% |

Tables 18 and 19 show the hardmax and softmax confusion matrices for the proposed model’s performance on the separate testing set.

Table 18: Hardmax Confusion Matrix for Proposed Random Forest Location Choice Model

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Obs\Pred** | **MI** | **OC** | **RY** | **SC** | **SG** | **YG** | **YK** |
| **MI** | 133 | 0 | 30 | 0 | 42 | 0 | 45 |
| **OC** | 14 | 0 | 27 | 6 | 54 | 3 | 11 |
| **RY** | 58 | 0 | 164 | 50 | 237 | 6 | 122 |
| **SC** | 4 | 1 | 63 | 111 | 49 | 0 | 48 |
| **SG** | 61 | 1 | 167 | 45 | 1056 | 13 | 158 |
| **YG** | 2 | 0 | 16 | 8 | 25 | 5 | 14 |
| **YK** | 50 | 0 | 105 | 41 | 249 | 0 | 326 |

Table 19: Softmax Confusion Matrix for Proposed Random Forest Location Choice Model

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Obs\Pred** | **MI** | **OC** | **RY** | **SC** | **SG** | **YG** | **YK** |
| **MI** | 80.0 | 5.9 | 47.8 | 4.2 | 56.1 | 5.8 | 50.2 |
| **OC** | 7.4 | 5.0 | 27.1 | 8.2 | 45.0 | 3.6 | 18.7 |
| **RY** | 42.6 | 22.9 | 163.8 | 47.4 | 222.0 | 16.4 | 121.9 |
| **SC** | 4.9 | 8.4 | 57.2 | 83.1 | 61.2 | 7.0 | 54.2 |
| **SG** | 52.3 | 45.3 | 240.4 | 68.3 | 826.5 | 27.5 | 240.6 |
| **YG** | 3.6 | 2.1 | 14.5 | 6.7 | 19.8 | 6.0 | 17.2 |
| **YK** | 44.7 | 20.3 | 139.7 | 48.3 | 229.1 | 20.6 | 268.3 |

# 2015 SMTO Discussion

As demonstrated above, the performance of the proposed RF model is slightly superior to that of the logit model. This result suggests that machine learning techniques, and especially the RF algorithm, are tools that can be used effectively for the college choice problem. Further investigation and experimentation on this front is warranted, and would establish whether RFs are similarly effective in other discrete choice contexts.

While the RF outperforms the logit model in this case, there are several reasons why the logit model should be preferred.

[Breiman](https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf) (2001) notes that as the number of trees is increased, the RF avoids overfitting. Furthermore, [Segal](https://escholarship.org/uc/item/35x3v9t4) (2004) shows that hyperparameter tuning, such as restricting tree depth, the number of splits, or minimum node size for splits, can curb this effect. In our analysis, we also avoided the use of high-cardinality variables (such as distances, coordinates, or home zone labels), instead opting for lower-cardinality features (planning districts). However, while these methods can address overfitting to the sample data, the RF remains flawed when it comes to the model’s generalizability. This flaw can be considered from two lenses: generalizing patterns from the sample to the entire student population, and generalizing the model to different contexts (such as for forecasting, or for a different choice set).

Firstly, sample bias may result in certain groups being under- or over-represented in the SMTO data. This would result in the RF making predictions based on relationships that do not accurately reflect the entire student population. For instance, if students attending campus from a certain planning district have been oversampled, students from this planning district will consistently be assigned a higher-than-accurate probability of attending campus . Hence, the RF model may not generalize effectively from the sample to the broader student population. While an elaborate weighting scheme could alleviate this issue, the creation of such a scheme would rely on more complete demographic information for different campuses, information that is not readily available.

The logit model is likely to be less sensitive to this issue. As a behavioural model, patterns in the sample are not “hard-wired” into the predictions in the same way as for the RF. This is especially true in our analysis as we have avoided the use of alternative-specific coefficients. Using the example above, while the over-sampling of students from said planning district to be attending campus might suggest a more/less sensitive response to distance, this sensitivity will not influence the parameter much as the behaviours of all students in the sample are considered together.

Secondly, the RF model is not appropriate if the results are to be generalized to other contexts. The RF relies on patterns that exist in the sample, but cannot be effectively adapted to other contexts. For example, if the enrollment of a campus were to change significantly, the RF could not be used to effectively forecast college choice patterns. In contrast, enrollment is included as an alternative-specific variable in the logit model, so one would expect the logit model to be more generalizable.

In the case of changes in students’ behavioural patterns, these observations cannot be inputted to the RF model. As a specific example, if a model for the future is predicted, and observations suggest that future students are less responsive to distance, this pattern cannot be conveyed to the RF algorithm, whereas for the logit model, a modified distance parameter can be imposed.

Furthermore, due to the nature of RF, such a model could not be used to generalize findings to a different choice set. In our analysis, one objective is to develop a model that would perform well on an expanded (more complete) choice set, such as that of the 2019 SMTO. This cannot be done with the RF. However, by excluding alternative-specific coefficients, such a generalization can be completed with the logit model. Below, we show that this generalization actually performs quite well when generalizing our logit model from the 2015 SMTO to the 2019 SMTO.

For all these reasons, we choose to focus our following analysis on the logit model approach.

# 2019 SMTO Location Choice – Logit

## Imputing Enrollment

Total enrollment rates for different campuses are an important component of our proposed location choice model. [Ontario's Open Data Team](https://data.ontario.ca/dataset/college-enrolment) (2019) provides data containing enrollment totals by campus for Ontario’s colleges. However, these statistics do not correspond one-to-one with the colleges listed by 2019 SMTO respondents. Furthermore, this survey sees the addition of several university campuses for which enrollment totals are not available. As such, we must impute enrollment totals for the missing campuses.

To do this, we estimate a simple gravity model. The log of enrollment is used as the attraction term for campuses where this is known. For the remaining campuses, alternative-specific constants are estimated. These constants are then exponentiated to obtain estimated enrollments. In effect, the sampling rates for each campus are used to impute the total student population. For some campuses, the results obtained by this method seemed reasonable. However, in some cases the predictions were clearly in the wrong order of magnitude. For these schools, the total enrollment across all of the institution’s campuses was known. Thus, we assumed that the relative sampling rates from each campus reflect the proportion of students enrolled at each, and calculated enrollment by campus by multiplying the institution’s total enrollment by these proportions. While these imputed enrollments provide reasonable grounds upon which to continue our analysis, more accurate information should be obtained to improve the model’s quality.

## Initial Results

Our first step for the 2019 SMTO data, is to estimate an MLN model using the same formulation as for the 2015 dataset and compare the results. Since domestic enrollment rate and admission averages are not available for many campuses, these variables have been removed from the model. Hence, the utility function used is

where:

is the systematic utility for student and campus ,

is the imputed or known total enrollment of campus ,

is the network distance between student ’s home zone and campus ,

if student lives with their family/parents, otherwise,

if , otherwise, and

, , , and are estimated parameters.

Table 20 presents our results for two versions of this model: one with parameters estimated so as to maximize the log-likelihood for the 2019 dataset, and one with the parameters as estimated with the 2015 dataset.

Table 20: Initial Model Results for 2019 SMTO. \*\*\* indicates p < 0.001. Errors and significances for 2015 Parameters are for the 2015 dataset.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 2019 Parameters | | | 2015 Parameters | | |
| Parameter | Estimate | Std. Error |  | Estimate | Std. Error |  |
|  | -0.0113 | 0.0004 | \*\*\* | -0.1089 | 0.0024 | \*\*\* |
|  | 0.9249 | 0.0081 | \*\*\* | 0.8991 | 0.0121 | \*\*\* |
|  | 2.4371 | 0.0386 | \*\*\* | 0.9042 | 0.0413 | \*\*\* |
|  | -0.0023 | 0.0006 | \*\*\* | 0.0507 | 0.0027 | \*\*\* |
| Metric |  |  |  |  |  |  |
| Hardmax Accuracy | 0.2645 |  |  | 0.3302 |  |  |
| Softmax Accuracy | 0.1763 |  |  | 0.2418 |  |  |
| Log Likelihood | -38558.5 |  |  | -53807.0 |  |  |

At this point, there are three noteworthy observations. Firstly, the 2015 parameters lead to better classification accuracies than the 2019 parameters, although they do not maximize log-likelihood. This result suggests that the log-likelihood is heavily swayed by the least likely observations. While such observations cannot be ignored, the quality of our model should not be dictated primarily by these.

Secondly, we notice that the distance parameter has significantly decreased in magnitude for the 2019 parameters. Instead, the closest-school parameter has significantly increased in magnitude. While this may represent a change in the importance of distance in students’ decision-making process, it is also possible that this change is inaccurate. Due to the larger set of included campuses in the 2019 dataset, and associated greater geographic spread, it is possible that some students attend very distant schools (for example if they indicated their permanent address as their home location, rather than the location from which they commute to school). This could drive the distance parameter to be smaller in magnitude, so as to improve log-likelihood. Future work could seek to identify such outliers in the data and re-estimate this model to gain a better understanding of how the relevance of distance has changed.

Thirdly, we notice that the sign for has changed. According to the 2019 Parameters model, students living with their family/parents are more responsive to distance than other students. This result is surprising, and may be influenced by response bias, as the 2019 dataset has a much lower response rate for living arrangement.

A more apt comparison of the trends in the 2019 and 2015 datasets can be made by estimating a model only on students attending the seven campuses from the 2015 dataset.

## Subset Results

Table 21 shows the results for our location choice model with three datasets: the full 2015 dataset, the subset of the 2019 dataset containing students attending the seven campuses from the 2015 dataset, and a joint dataset with naïve pooling.

Table 21: Results of Proposed Model for Sample Subsets. \*\*\* indicates p < 0.001.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2015 | | | 2019 Subset | | | Joint (naïve pooling) | | |
| Parameter | Estimate | Std. Error |  | Estimate | Std. Error |  | Estimate | Std. Error |  |
|  | -0.1089 | 0.0024 | \*\*\* | -0.0892 | 0.0020 | \*\*\* | -0.0927 | 0.0014 | \*\*\* |
|  | 0.8991 | 0.0121 | \*\*\* | 0.8404 | 0.0132 | \*\*\* | 0.8660 | 0.0089 | \*\*\* |
|  | 0.9042 | 0.0413 | \*\*\* | 1.0644 | 0.0474 | \*\*\* | 1.0258 | 0.0316 | \*\*\* |
|  | 0.0507 | 0.0027 | \*\*\* | 0.0370 | 0.0028 | \*\*\* | 0.0375 | 0.0018 | \*\*\* |
| Metric |  |  |  |  |  |  |  |  |  |
| Hardmax Accuracy | 0.4622 |  |  | 0.3840 |  |  | 0.4293 |  |  |
| Softmax Accuracy | 0.3479 |  |  | 0.3069 |  |  | 0.3301 |  |  |
| Log Likelihood | -19307.1 |  |  | -15560.4 |  |  | -34801.3 |  |  |
| Mean loss | -1.334 |  |  | -1.421 |  |  | -1.369 |  |  |

From these results, we see that our model formulation performs better on the 2015 dataset than on the 2019 dataset. This is to be expected as the model formulation was developed using the former. Additionally, the family-distance interaction term is positive and significant in all three trials, suggesting a consistent pattern differentiating the behaviours of students with different living arrangements. Finally, we also see that the magnitude of the distance parameter has decreased for the 2019 model. However, this happens to a much smaller extent than above, suggesting that the drastic change in the distance parameter above is therefore not representative of the entire sample. Instead, it is possible that the addition of more campuses outside Toronto has influenced the distance parameter. Another possible contributing factor is that distance is perceived differently by students attending college rather than university. As such, our next step is to see how we can refine this distinction in the model.

## Distinguishing between University and College Students

### Nested Logit Model

The first, and perhaps most natural, approach to emphasize the distinction between college and university students is by implementing a nested logit (NL) model, where the university campuses are in one nest and the college campuses are in another. Table 22 presents the results for two NL formulations compared with the un-nested MNL.

Table 22: Results for Nested Model for 2019 Location Choice. A implies greater correlation across nests than within.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Reference (Un-nested) | | | Nested – One Scale Param. | | | Nested – Two Scale Params. | | |
| Parameter | Estimate | Std. Error |  | Estimate | Std. Error |  | Estimate | Std. Error |  |
|  | -0.0113 | 0.0004 | \*\*\* | -0.0185 | 0.0005 | \*\*\* | -0.0139 | 0.0004 | \*\*\* |
|  | 0.9249 | 0.0081 | \*\*\* | 1.2874 | 0.0198 | \*\*\* | 0.8946 | 0.0245 | \*\*\* |
|  | 2.4371 | 0.0386 | \*\*\* | 3.6470 | 0.0835 | \*\*\* | 2.6111 | 0.0769 | \*\*\* |
|  | -0.0023 | 0.0006 | \*\*\* | -0.0037 | 0.0006 | \*\*\* | -0.0021 | 0.0004 | \*\*\* |
|  |  |  |  | 1.6184 | 0.0336 | \*\*\* | 1.1639 | 0.0282 | \*\*\* |
|  |  |  |  | 0.9564 | 0.0357 | \*\*\* |
| Metric |  |  |  |  |  |  |  |  |  |
| Hardmax Accuracy | 0.2645 |  |  | 0.2473 |  |  | 0.2474 |  |  |
| Softmax Accuracy | 0.1763 |  |  | 0.1786 |  |  | 0.1772 |  |  |
| Log Likelihood | -38558.5 |  |  | -38277.2 |  |  | -38261.1 |  |  |

From these results, we see that the NL model with nesting by institution types is not effective. As such, a different approach to distinguish between university and college students is desired.

### Random Forest Classifier

Having seen the effectiveness of machine learning techniques in classifying living arrangement, we develop an RF to classify students according to their institution type. Since this is a binary classification task, and there are fewer college students with respect to university students in our sample, we optimize the College F-1 score. This is calculated as follows, with “College” as the positive class:

where:

and

Table 23 presents the feature importances for our selected model. This model has an F1 score of 0.4447. Table 24 shows the confusion matrix for the model on a separate testing set. Figure 2 shows the receiver-operator curve for our model. The area under the curve is 0.7697.

Table 23: Feature Importances for Selected Random Forest School Type Classifier. All features except PlanningDistrict and Age are dummy variables. ClosestUni is 1 if the student’s home zone is closest to a university campus, 0 otherwise.

|  |  |
| --- | --- |
| Feature | Importance |
| PlanningDistrict | 44.32% |
| Age | 34.25% |
| ClosestUni | 5.50% |
| IncomeLow | 2.85% |
| FamilyTrue | 1.90% |
| LicenceFalse | 1.85% |
| FamilyFalse | 1.83% |
| Cars2+ | 1.69% |
| IncomeHigh | 1.58% |
| LicenceTrue | 1.48% |
| Cars1 | 1.45% |
| Cars0 | 1.30% |

Table 24: Confusion Matrix for School Type Classifier

|  |  |  |
| --- | --- | --- |
| Obs\Pred | College | University |
| College | 348 | 535 |
| University | 334 | 3009 |

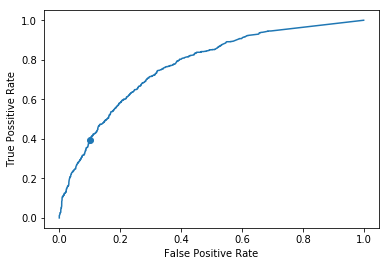


Figure 2: Receiver-Operator Curve for School Type Classifier. The point shown is that associated with a threshold of 50%.

We notice that a better F1 score can be obtained by adjusting the probability threshold at which students are predicted to be college students. Indeed, after reducing this threshold, we obtain an F1 score of 0.4998. Table 25 presents the confusion matrix for this updated model.

Table 25: Confusion Matrix for School Type Classifier with Selected Threshold

|  |  |  |
| --- | --- | --- |
| Obs\Pred | University | College |
| University | 2632 | 711 |
| College | 352 | 531 |

### Separate Models for University and College Choice

We hope that our model can be improved by assigning students to separate university choice and college choice models according to their predicted institution type. As such, we estimate separate models using the two subsamples. In evaluating this model’s quality, we use the RF predictions to assign students to the appropriate model. The results are presented in Table 26.

Table 26: Results for 2019 Location Choice Separated by School Type. Note that log-likelihood is not predicted as any misclassified students will contribute a log likelihood of negative infinity. Also not that the family-distance interaction term was removed as it was statistically insignificant in both models.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | University Choice | | | College Choice | | |
| University Students | 11841 | | | 1207 | | |
| College Students | 986 | | | 2482 | | |
| Total Sample | 12827 | | | 3689 | | |
| Parameter | Estimate | Std. Error |  | Estimate | Std. Error |  |
|  | -0.0113 | 0.0004 | \*\*\* | -0.1089 | 0.0024 | \*\*\* |
|  | 0.9249 | 0.0081 | \*\*\* | 0.8991 | 0.0121 | \*\*\* |
|  | 2.4371 | 0.0386 | \*\*\* | 0.9042 | 0.0413 | \*\*\* |
| Metric |  |  |  |  |  |  |
| Hardmax Accuracy | 0.3130 |  |  | 0.3833 |  |  |
| Softmax Accuracy | 0.2085 |  |  | 0.0473 |  |  |
| Weighted Average |  |  |  |  |  |  |
| Hardmax Accuracy | 0.3287 | | | | | |
| Softmax Accuracy | 0.1725 | | | | | |

We see that the idea of developing separate models and assigning students to these is promising when it comes to hardmax accuracy but does not lead to any gains for softmax accuracy. In particular, the college model suffers from very low softmax accuracy, although this is impacted by the fact that about one-third of the students who are assigned to this model do not actually attend a college. At this point, we revert to the unified model and try incorporating the RF predictions in different ways.

### Including Random Forest Probabilities and Predictions

We experiment with the addition of several terms to the utility function:

* P\_Col: The RF-generated probability that the student is a college student, included in the utility function for colleges only.
* P\_Uni: The RF-generated probability that the student is a college student, included in the utility function for universities only.
* P\_Type: The RF-generated probability that the student attends a school of the alternative’s type, included for all schools.
* Pred\_Type: A dummy variable that equals 1 if the student is predicted to attend a school of the alternative’s type (using the optimal threshold from above) and 0 otherwise.

The results for various combinations of these terms are shown in Tables 27 and 28.

Table 27: Estimated Parameters for Models with Random Forest Predictions and Probabilities

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | P\_Col | P\_Uni | P\_Type | Pred\_Type |
| Reference | -0.0113 | 0.9249 | 2.4371 | -0.0023 |  |  |  |  |
| P\_Col | -0.0103 | 1.1063 | 2.3362 | -0.0024 | 2.6037 |  |  |  |
| P\_Uni | -0.0109 | 0.7139 | 2.3335 | -0.0021 |  | 1.4250 |  |  |
| P\_Type | -0.0102 | 0.7391 | 2.2201 | -0.0021 |  |  | 2.5745 |  |
| P\_Col + P\_Uni | -0.0099 | 0.8225 | 2.1983 | -0.0021 | 3.9158 | 2.2766 |  |  |
| Pred\_Type | -0.0102 | 0.8281 | 2.2473 | -0.0022 |  |  |  | 1.4739 |

Table 28: Performance Metrics for Models with Random Forest Predictions and Probabilities

|  |  |  |  |
| --- | --- | --- | --- |
|  | Hardmax Accuracy | Softmax Accuracy | Log Likelihood |
| Reference | 26.45% | 17.63% | -38558.5 |
| P\_Col | 30.41% | 19.15% | -37380.3 |
| P\_Uni | 24.99% | 18.16% | -37471.8 |
| P\_Type | 28.49% | 19.58% | -35832.0 |
| P\_Col + P\_Uni | 30.52% | 20.06% | -35521.9 |
| Pred\_Type | 30.84% | 19.55% | -36265.2 |

We see that these models generally offer a reasonable improvement over the reference model for both hardmax and softmax accuracies, and log likelihood. However, a degree of skepticism should be used when considering these results, as many of the students in the data were used to train the RF model, making the probabilities and predictions likely more accurate than they would be for the entire population. It seems that more work is warranted to better establish a student’s institution type. For instance, one idea which we have not yet implemented is assigning students which the RF is fairly confident attend a university to the separate university model but using the full model for other students.

# 2019 SMTO Discussion

Our model remains somewhat ineffective once the more diverse choice set for the 2019 SMTO dataset is introduced. One significant limiting factor seems to be identifying students’ institution type. From experience, it seems that students largely consider only one institution type in their choice sets, so eliminating irrelevant alternatives could lead to significant improvements in the model’s performance. Previous studies have shown such variables as academic performance, ethnic background, and participation in extracurricular activities in high school to be useful indicators for students’ selected institution type. Unfortunately, these variables are unavailable in our study.

We also observed that maximizing log likelihood does not necessarily correspond with gains in other performance metrics, and can lead to estimated models which perform quite poorly on other metrics. Particularly, the parameters estimated for the 2019 model result in significantly lower classification accuracies than those from the 2015 model. This finding demonstrates a flaw in maximizing log likelihood in such cases. The maximization is weighed towards the most unlikely observations and hence might not effectively capture the model’s holistic performance. Possibly, this problem only persists in unbalanced datasets, as is the case here. We note that it is possible to greatly reduce this effect by using only a subset of the data with a smaller, more balanced, choice set. In our case, we estimate a model on the subset of students attending schools included in the 2015 survey. However, doing so is an oversimplification of the problem at hand, especially as it would likely be difficult to effectively classify individuals according to whether or not they attend a school in this subset.

# Future Work

## Generalization to TTS and Implementation in GTAModel

This project was completed with the underlying possibility of implementation into [GTAModel](https://tmg.utoronto.ca/doc/1.4/gtamodel/index.html). As such, creating a model that can be generalizable to an expanded choice set is critical. We do this by avoiding the use of alternative-specific coefficients in our logit model. However, we have yet to see whether this approach can lead to solid results in a more generalized setting.

One way to test this is by using data from the TTS to verify the performance of our model. An associated challenge is that the set of postsecondary institutions is much larger, and includes several smaller, specialized institutions. Using this data to adapt our model will allow for a final product to be developed that can be implemented effectively in GTAModel.

## Adjusting Enrollments by Level of Study

In our analysis, we do not restrict the choice sets of students in the location choice problem. However, performing this step effectively is likely to yield improvements in our classification accuracy. One way to constrain choice sets is by modelling the admissions process. A simple admissions model, such as the one used by Kohn et al. (1974) or [Montgomery](https://journals-scholarsportal-info.myaccess.library.utoronto.ca/pdf/02727757/v21i0005/471_anlmotcoagbs.xml) (2002), can be implemented and its effects on model performance observed.

## Modelling Choice Sets

While the SMTO data includes participants’ level of study (i.e. undergraduate vs. graduate), this variable was not used as it is not available in the TTS survey. However, adjusting the campus enrollments according to the student’s level of study would likely yield significant improvements to the model’s performance. This is because several campuses have much smaller populations of graduate students compared to undergraduates. In fact, from the RF model we see that level is indeed an important variable in the classification. If a model can be successfully trained to classify students according to their level of study (as has been done with living arrangement) given such variables as age, income, employment status, living arrangement, etc., then enrollment level can and should be included in the analysis.

## Exploration of Random Forest and Machine Learning Techniques

One noteworthy finding from our investigation was the effectiveness of the random forest classifier algorithm with respect to the logit model. While using random forests comes with associated concerns, as described above, this foray into the realm of machine learning reveals that there is promising potential to use such techniques in college choice problems, and possible other discrete choice modelling applications. Further work in this area could lead to novel approaches being used in the field.

# References

Acevedo-Gil, N. 2017. “College-conocimiento: toward an interdisciplinary college choice framework for Latinx students.” *Race Ethnicity and Education*, 20 (6), 829-850.

Anas, A. 1983. “Discrete choice theory, information theory and the multinomial logit and gravity models.” *Transportation Research Part B: Methodological*, 17 (1), 13-23.

Ben-Akiva, M., and M. Bierlaire. 1999. “Discrete choice methods and their applications to short term travel decisions.” In *Handbook of Transportation Science*, R. W. Halls, ed. Boston, MA: Springer, 5-33.

Ben-Akiva, M., and S. R. Lerman. 1987. “Discrete choice analysis: Theory and application to travel demand.” *The Journal of Operational Research Society*, 38 (4), 370.

Bierlaire, M. 2020. A short introduction to PandasBiogeme. <https://transp-or.epfl.ch/documents/technicalReports/Bier20.pdf> (accessed August 28, 2020).

Breiman, L. 2001. “Random forests.” *Machine Learning,* 45, 5-32.

Chen, J. (2018). *School Choice Modelling Report*, Travel Modelling Group, Toronto, ON.

Chung, B., S. Hasnine, and K. N. Habib. 2018. “How far to live and with whom? The role of modal accessibility on student’s choice of living arrangements and the distance they are willing to live from university in Toronto.” In *Proc.*, *Transportation Research Board*, Washington, DC.

Croissant, Y. 2020. Package ‘mlogit’. <https://cran.r-project.org/web/packages/mlogit/mlogit.pdf> (accessed August 28, 2020).

Hastie, T., R. Tibshirani, and J. Friedman. 2008. “Random forests.” In *The Elements* *of Statistical Learning*, new York: Springer, 587-604.

Hearn, J. 1984. “The relative roles of academic, ascribed, and socioeconomic characteristics in college destinations.” *Sociology of Education*, 57 (1), 22-30.

Kohn, M. G., C. F. Manski, and D. S. Mundel. 1974. *An empirical investigation of factors which influence college-going behaviour*, Rand, Santa Monica, CA.

Long, B. T. 2004. “How have college decisions changed over time? An application of the conditional logistic choice model.” *Journal of Econometrics*, 121. 271-296.

McFadden, D. 1973. “Conditional logit analysis of qualitative choice behavior.” In *Frontiers in Econometrics*, P. Zarembka, ed. New York: Academic Press, 105-142.

Montgomery, M. 2002. “A nested logit model of the choice of a graduate business school.” *Economics of Education Review*, 21, 471–480.

Niu, S. X., and M. Tienda. 2008. “Choosing colleges: Identifying and modeling choice sets.” *Social Science Research*, 37. 416-433.

Niu, S. X., M. Tienda, and K. Cortes. 2006. “College selectivity and the Texas top 10% law.” *Economics of Education Review*, 25, 259-272.

Ontario’s Open Data Team. 2019. *College enrolment*, Queen’s Printer for Ontario, <https://data.ontario.ca/dataset/college-enrolment> (accessed August 28, 2020).

Ooseterbreek, H., W. Groot, and J. Hartog. 1992. “An empirical analysis of university choice and earnings.” *De Economist*, 140 (3), 293-309.

Perna, L. 2006. “Studying college access and choice: A proposed conceptual model.” In *Higher education: Handbook of theory and research*, J. C. Smart, ed. Netherlands: Springer, 99-157.

Sá, C., R. J. G. M. Florax, and P. Rietveld. 2012. *“*Living arrangement and university choice of Dutch prospective students.” *Regional Studies*,46 (5). 651-667.

Segal, M. R. 2003. *Machine learning benchmarks and random forest regression*. UCSF Center for Bioinformatics and Molecular Biostatistics, San Francisco.

Strobl, C., J. Malley, and G. Tutz. 2009. “An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests.” *Psychol Methods*, 14 (4), 328-348.

StudentMoveTO. n.d. Research & Action, <http://www.studentmoveto.ca/research-and-action/> (accessed August 28, 2020).

Train, K. 2009. “Logit.” In *Discrete choice models with simulation*, Cambridge University Press, 34-75.

Wang, Q. (2015). *Developing school destination choice model in the Greater Toronto and Hamilton Area*, Travel Modelling Group, Toronto, ON.

# Acknowledgements

This analysis was completed together with Gonzalo Martínez Santos. Our work was supervised by Prof. Eric Miller of the University of Toronto’s Department of Civil and Mineral Engineering; we would like to thank him for his support and guidance. We would also like to thank James Vaughan and Gozde Ozonder for their help.