Modelling GTHA Post-Secondary School Location Choice

Ethan Baron
Undergraduate Student
Division of Engineering Science
University of Toronto

Gonzalo Martinez Santos
Undergraduate Student
Division of Engineering Science
University of Toronto

Eric J. Miller, Ph.D.
Professor, Department of Civil & Mineral Engineering
Director, University of Toronto Transportation Research Institute
University of Toronto

Prepared for presentation in the Transportation Systems Modelling Session

2021 Transportation Association of Canada Conference & Exhibition

1 Introduction

The purpose of this paper is to develop a school location choice model for post-secondary (PS) students in the Greater Toronto-Hamilton Area (GTHA). This analysis differs from previous PS school choice modelling in three respects. Firstly, the model is not representing the college choice process directly. Instead, this analysis is an exercise in matching students who have already made PS school 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, this study primarily analyzes geographical patterns in school location choice for applications in travel demand modelling. An emphasis is placed on modelling the accessibility of each school location to each student, rather than predicting school selectivity or institution type. Thirdly, an RF classifier is implemented for the location choice problem, a novel approach in the field, and its utility is compared to that of the classic econometric approaches.

Section 2 presents a brief literature review of relevant works in PS school location choice modelling in general, and in the GTHA specifically. Section 3 introduces the two modelling methods used in this study: random utility models and random forest models. Section 4 describes the two datasets used: the 2015 and 2019 StudentMoveTO (SMTO) surveys. Section 5 presents a logit mode choice model for the 2015 dataset, and Section 6 then presents the development of a school location choice for this dataset. Section 7 presents the development of a random forest model for the school location choice problem and Section 8 summarizes and discusses the main results for the 2015 modelling. Building on the 2015 analysis, Section 9 describes the development of location choice models for the 2019 dataset, and section 10 summarizing the key findings from this analysis. Finally, Section 11 concludes the paper with a brief discussion of possible directions for future work.

2 Literature Review

This section presents a brief review of the PS school choice literature (Section 2.1) and previous work on this problem in the GTHA (Section 2.2). With the exception of the GTHA-based work, it appears that PS school choice has not been well-studied in Canada, with the vast majority of the research to date occurring in the U.S.

2.1 Post-Secondary (PS) School Choice Modelling

From a conceptual point of view, frameworks for the PS school choice model have been presented by Perna (2006) and Acevedo-Gil (2017). In empirical practice, the most common approach to model this 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 (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. (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 (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 (2002) is that greater distance from home reduces the attractiveness of an institution. Kohn et al. (1974) make this observation after implementing conditional logit models for PS school choice in Illinois and North Carolina. Oosterbeek et al. (1992) also notice this pattern when using an MNL model to analyze data on university choices of Dutch economists. Long (2004) uses a conditional logit model to analyze changes in PS school 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. (1974) find that the disutility due to greater tuition fees is smaller for higher-income groups in particular. Long (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. (2012) use an NL model to predict living arrangement and university choice for Dutch post-secondary students, while Niu and Tienda (2007) analyze PS 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.

2.2 PS School Location Choice in the GTHA

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.

Past researchers from the University of Toronto's Travel Modelling Group have investigated school location choice models for the GTHA. Chen (2018) estimates a doubly-constrained gravity model using data from the 2016 Transportation Tomorrow Survey (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.

3 Methods

3.1 Random Utility Models

As indicated in the literature review, the multinomial logit model (MNL) is the most common method used to model PS choice. The model is derived within a random utility framework introduced by McFadden (1973), and described in depth by Train (2003). The perceived utility of alternative j for student i is assumed to be $U_{ij} = V_{ij} + \epsilon_{ij}$, where V_{ij} is the systematic utility of location j for student i and ϵ_{ij} is an associated random utility. A student selects alternative k if and only if $U_{ik} > U_{ij}\ \forall j \neq k$. The MNL is obtained by assuming that the random utility terms are independent and identically distributed with a Type-1 extreme value distribution. In this formulation, the probability of alternative k being chosen by individual i from choice set C is:

$$P(y_i = k) = \frac{e^{V_{ik}}}{\sum_{j \in C} e^{V_{ij}}} \qquad [1]$$

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 j being selected over alternative k is independent of the other alternatives in the choice set. 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 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 k from nest l including alternatives C_l being chosen by individual i from the set of nests M is:

$$P(y_i = k) = \frac{e^{\mu V_{im}}}{\sum_{m \in M} e^{\mu V_{im}}} \times \frac{e^{\mu_l V_{ik}}}{\sum_{j \in C_l} e^{\mu_l V_{ij}}} \qquad [2]$$

Here, μ 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 μ_l reflects the correlation among alternatives in nest l (at the lower level of the model). The term $\ln \sum_{j \in C_l} e^{\mu_l V_{ij}}$ is known as the logsum or inclusive value of nest l, 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, $\mu_l \geq 1 \; \forall l \in M$. Note that if all $\mu_l = 1$ then the nesting structure is degenerate and the model collapses into the standard MNL.

In this study, logit and nested logit models are estimated using mlogit 1.1-0 (Croissant 2020) with RStudio 1.3.959 and R 4.0.0. In some cases. Biogeme 3.2.6 (Bierlaire 2020) was used with Python 3.7.6 and Jupyter Notebook 6.0.3.

3.2 Random Forest Models

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. 2009). Developed by Breiman (2001), the RF training algorithm is as follows (Hastie et al. 2009). For $b \in \{1, 2, \ldots B\}$:

a) Draw a bootstrapped sample from the training data.
b) Grow a decision tree T_b using the bootstrapped data by performing the following steps recursively until minimum node size n_{min} is reached:
    a. Select m features from the training data at random
    b. Select the best feature and split-point from the m features according to some split criterion
    c. Split the node using that feature and split-point

The RF is the set of trees $\{T_b\}_1^B$. 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:

- B, the number of trees
- n_{min}, the minimum size for leaf nodes
- m, the number of features to consider for each split point
- The splitting criteria to be used
- The maximum tree depth

The RF algorithm is implemented using scikit-learn 0.22.1 with Python 3.7.6 and Jupyter Notebook 6.0.3.

3.3 Metrics Reported

Table 1 lists the metrics used throughout this paper to evaluate the models being tested. 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 (e.g., 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 this analysis.

Table 1: Summary of Reported Metrics. Note that L_0 is the log-likelihood for the logit model with only alternative-specific constants, as explained in McFadden (1975). p_{ij} represents the probability assigned by the model that student i attends campus j, and y_i represents the campus actually attended by student i.

Metric	Calculation
Hardmax Accuracy	$HA = \sum_{i=1}^{n}\left(\left(\underset{j}{\operatorname{argmax}}\, p_{ij}\right) == y_i\right)$
Softmax Accuracy	$SA = \frac{1}{n}\sum_{i=1}^{n} p_i$
Log Likelihood	$\log(L) = \log\left(\prod_{i=1}^{n} p_i\right) = \sum_{i=1}^{n} \log\,(p_i)$
McFadden Pseudo Rho-Squared	$\rho^2 = 1 - \frac{\log(L)}{\log(L_0)}$

4 Data

4.1 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. Observations whose indicated enrollment level was “Other” (as opposed to “UG” or “Grad”) were removed from the sample; these were also the only observations 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 the model generalizable to TTS data, only TTS-compatible attributes are used in the 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 for possible use to impute this attribute for TTS records. The model achieves an accuracy of over 90%, and so living arrangement is retained in the list of available attributes.

4.2 2019 SMTO

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

4.3 2016 TTS

The Transportation Tomorrow Survey (TTS) is a travel survey conducted once every five years in the GTHA. The survey includes personal attributes such as age, household attributes such as composition and income class, and for students, commute information including time and mode taken.

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

5 2015 SMTO Mode Choice – Logit Model

A key component of the school location choice model was expected to be a measure of the accessibility of each school to the student. For this reason, a mode choice model is developed and estimated.

5.1 Model Development

The mode choice model estimated takes the form of an MNL model with three alternatives: Auto, Transit, and Active Mode (walking or biking). The 120 students in the data whose indicated mode choice is not one of these three alternatives are removed from the mode choice estimation set. The model includes estimated morning travel times from the student's home zone to each campus for each mode and alternative-specific constants, with transit as the reference mode, to reproduce aggregate mode shares.

In developing the mode choice model, the interaction between various socioeconomic variables and the alternative-specific constants is tested. These socioeconomic variables include both individual and household characteristics. The variables whose impact is most significant are found to be living arrangement – specifically, whether the student indicated that they are living with their family/parents – and driver's license ownership.

To further refine the mode choice model, choice set restrictions for certain alternatives in certain choice situations are examined. Two types of availability restrictions are considered: for active modes and auto. The active mode alternative is made unavailable in cases where the estimated walking travel time in minutes is greater than some threshold t, where $t \in [0, 300]$ is tested. The auto mode is made available only if a student has indicated that their household owns at least n vehicles, where $n \in \{0, 1, 2\}$ is tested.

Note that some observations may contradict assumptions about mode availability (e.g. students indicating they drive to campus despite not owning a car). Such "invalid" observations have probability zero in the availability-restricted models, resulting in a log likelihood of negative infinity. For this reason, log likelihood is not used to compare availability-restricted models. Instead, the "softmax accuracy" is used. The setting which optimized softmax accuracy were $t = 46.6$ minutes and $n = 2$ cars.

5.2 Results

The proposed mode choice model uses the following utility function for student i and mode j:

$$V_{ij} = \mathrm{B}_j + \mathrm{B}_{t_j} t_{ij} + \mathrm{B}_{F_j} F_i + \mathrm{B}_{L_j} L_i \qquad [4]$$

where t_{ij} = the estimated morning travel time from student i's home zone to campus j, in minutes;

F_i = 1 if student i lives with family/parents, 0 otherwise; and

L_i = 1 if student i owns a driver's license, 0 otherwise.

Active modes were set as unavailable where $t_{i,Active} > 46.6$, and auto was set as unavailable where the number of cars owned by the student's household was less than two. Table 4 presents the results for this proposed model, while Tables 5 and 6 show confusion matrices for this model.

*Table 4: Proposed Mode Choice Model. $B_{Transit} = 0$. 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	
B_{Active}	3.890	0.169	***
B_{Auto}	-1.696	0.169	***
$B_{F_{Active}}$	-0.992	0.147	***
$B_{F_{Auto}}$	-0.792	0.121	***
$B_{L_{Active}}$	0.613	0.113	***
$B_{L_{Auto}}$	1.622	0.117	***
$B_{t_{Transit}}$	-0.0113	0.001	***
$B_{t_{Active}}$	-0.0936	0.006	***
$B_{t_{Auto}}$	-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 5: Hardmax Confusion Matrix from Proposed Mode Choice Model

Obs\Pred	Active	Auto	Transit	Observed	Accuracy
Active	3116	13	676	3805	81.9%
Auto	41	81	1191	1313	6.2%
Transit	360	66	8813	9239	95.4%
Predicted	3517	160	10680	14357	83.7%

Table 6: Softmax Confusion Matrix from Proposed Mode Choice Model

Obs\Pred	Active	Auto	Transit	Observed	Accuracy
Active	2827.7	32.7	944.6	3805	74.3%
Auto	36.2	269.0	1007.8	1313	20.5%
Transit	304.2	662.5	8272.4	9239	89.5%
Predicted	3168.1	964.1	10224.8	14357	79.2%

5.3 Discussion

As can be seen, the accuracy of the mode choice model is quite good. 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). 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 compared to students in other living arrangements.

6 SMTO 2015 Location Choice – Logit Model

This section describes the development of a logit-based PS location choice model estimated using 2015 SMTO data. The model is iteratively developed, beginning with a simple specification adding additional complexity incrementally.

6.1 Model Development

At first, a doubly-constrained gravity model is estimated. Anas (1983) shows that the doubly-constrained gravity model is equivalent to MNL with impedance function $f\left(d_{ij}\right) = e^{\mathrm{B}_d d_{ij}}$. In this study, d_{ij} is the network distance between the student i's home zone and campus j, obtained from level-of-service matrices generated using the 2016 TTS data and GTAModel V4.1. This model is compared to an accessibility model which includes the expected maximum utility from the mode choice model as a logsum term, as outlined by Ben-Akiva and Lerman (1985). As such, this accessibility model uses the following utility model:

$$V_{ij} = \mathrm{B}_j + \mathrm{B}_{Access} * \log\left(e^{V_{ij}{}_{Active}} + e^{V_{ij}{}_{Auto}} + e^{V_{ij}{}_{Transit}}\right)$$

Models including both the distance term and the mode choice logsum are also tested.

One objective for this analysis is to make the model generalizable to expanded choice sets, which requires avoiding alternative-specific constants. Instead, campus-specific attributes are used to model the alternative-specific utility for each campus. These attributes include 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 (CUDO) and are summarized in Table 7.

Table 7: 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

Next, the addition of socioeconomic variables to the model was investigated. In particular, the interactions of socioeconomic variables with distance, accessibility, and campus-specific attributes are explored.

The final addition to the model is including "closest school dummies" C_{ij}, representing which school is closest to the student's home zone. That is, $C_{ij} = \{1 \text{ if } d_{ij} \leq d_{ik}\ \forall\ k \neq j, \text{else } 0\}$. Moreover, it is hypothesized 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, a threshold of 2km is chosen, such that $C_{ij} = \{1 \text{ if } d_{ij} \leq 2\text{km and } d_{ij} \leq d_{ik}\ \forall\ k \neq j, \text{else } 0\}$.

6.2 Results

Table 8 presents the results for the three fully-developed models which were estimated: a distance-based model, an accessibility-based model (with the mode choice utility as a logsum term), and a combined model with both of these terms.

*Table 8: Closest Dummies in Location Choice Models. *** indicates p < 0.001. $F_i = 1$ if the student lives with their family/parents, otherwise $F_i = 0$. $W_i = 1$ if the student works full-time, otherwise $W_i = 0$. $P_i = 1$ if the student's enrollment status is part-time, otherwise $P_i = 0$.*

	Distance			Accessibility			Combined		
Parameter	Estimate	Std. Error		Estimate	Std. Error		Estimate	Std. Error	
B_{Dist}	-0.1043	0.002	***				-0.0832	0.003	***
B_{Access}				0.8663	0.018	***	0.2671	0.028	***
$\mathrm{B}_{Closest}$	0.9287	0.042	***	-0.2310	0.053	***	0.5444	0.060	***
$\mathrm{B}_{Log(Enrol)}$	0.8011	0.014	***	0.8660	0.015	***	0.8058	0.015	***
$\mathrm{B}_{Admission\ Ave}$	0.0599	0.004	***	0.0261	0.004	***	0.0539	0.004	***
$F_i * \mathrm{B}_{Domestic\ \%}$	0.0331	0.002	***	0.0223	0.002	***	0.0315	0.002	***
$F_i * \mathrm{B}_{Dist\_Family}$	0.0414	0.003	***				0.0214	0.004	***
$F_i * \mathrm{B}_{Access\_Family}$							-0.2185	0.043	***
$W_i * \mathrm{B}_{Access\_Work}$							-0.4222	0.098	***
$P_i * \mathrm{B}_{Access\_PT}$							-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		

It is seen 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 relative to the other models. Due to this effectiveness and the relative simplicity of the distance model, this model is selected as the preferred one.

Thus, the proposed logit model for 2015 SMTO PS location choice is based on the utility model:

$$V_{ij} = \mathrm{B}_{Enrol} * \log(E_j) + \mathrm{B}_{AdmAve} * AA_j + \mathrm{B}_{Dom\%} * D_j * F_i + d_{ij} * \mathrm{B}_{Dist} + \mathrm{B}_{Dist\_Family} * F_i * d_{ij} + \mathrm{B}_{Closest} * C_{ij} \quad [6]$$

where:

V_{ij} is the systematic utility for student i and campus j,
E_j is the total enrollment of campus j,
AA_j is the mean secondary school admission average for campus j,
D_j is the domestic enrollment rate for undergraduate programs at campus j,
$F_i = \{1$ if student i lives with their family/parents, 0 otherwise$\}$,
d_{ij} is the network distance between student i's home zone and campus j,
$C_{ij} = \{1$ if $d_{ij} \leq 2\text{km}$ and $d_{ij} \leq d_{ik}\ \forall k \neq j$, 0 otherwise$\}$, and
B_{Enrol}, B_{AdmAve}, $\mathrm{B}_{Dom\%}$, B_{Dist}, B_{Dist_Family}, and $\mathrm{B}_{\text{Closest}}$ are estimated parameters.

Confusion matrices for this model are presented in Tables 9 and 10.

Table 9: Hardmax Confusion Matrix for Proposed Logit Location Choice Model

Obs\Pred	MI	OC	RY	SC	SG	YG	YK	Observed	Accuracy
MI	404	0	8	0	339	0	179	930	43.4%
OC	27	0	53	6	273	0	96	455	0.0%
RY	194	0	318	36	1431	0	729	2708	11.7%
SC	6	0	18	184	639	0	227	1074	17.1%
SG	239	0	529	53	4256	0	834	5911	72.0%
YG	13	0	8	3	203	0	88	315	0.0%
YK	172	0	54	30	1264	0	1564	3084	50.7%
Predicted	1055	0	988	312	8405	0	3717	14477	46.5%

Table 10: Softmax Confusion Matrix for Proposed Logit Location Choice Model

Obs\Pred	MI	OC	RY	SC	SG	YG	YK	Observed	Accuracy
MI	271.3	26.3	166.1	19.9	277.5	11.9	157.1	930	29.2%
OC	23.2	19.5	116.1	24.2	190.1	8.7	73.2	455	4.3%
RY	166.9	90.9	687.8	173.4	997.2	58.2	533.7	2708	25.4%
SC	18.3	29.5	224.7	247.6	329.0	31.9	192.9	1074	23.1%
SG	227.6	227.1	1375.2	250.4	2897.5	102.2	831.0	5911	49.0%
YG	15.4	9.7	71.0	22.1	114.7	10.7	71.4	315	3.4%
YK	176.4	86.8	595.3	169.3	1000.0	66.6	989.7	3084	32.1%
Predicted	899.0	489.8	3236.1	906.9	5806.1	290.1	2849.1	14477	35.4%

6.3 Discussion

The estimated coefficients from the proposed model reveal interesting trends in the SMTO data. Firstly, utility is negatively impacted by distance from a campus, which confirms findings from previous studies. However, this effect is reduced for students who indicate they are living with their family. This finding is logical as students who are not living with their family may be living on residence or have selected their place of residence according to their school location (rather than the other way around, as the model implies). Additionally, the coefficient for the closest school dummy is positive and significant, indicating that the closest campuses are more attractive to students.

The enrollment term reflects the relative size of the different campuses and allows for the model to be potentially generalized to expanded choice sets including additional post-secondary schools. Domestic enrollment rate increases the utility of campuses to students living with their family. Presumably, students living with their family are more likely to be domestic, and hence attend a school with more domestic students. Although tuition information was available, in contrast to many previous studies, including tuition information did not significantly improve the model, even when interacted with household income. This result can be attributed to the marginal differences between tuition fees for the investigated schools. Finally, the estimated parameter for admission average confirms previous findings that greater selectivity increases the attractiveness of schools.

7 2015 SMTO Location Choice – Random Forest

This section describes the development of a random forest model to predict PS location choice based on the 2015 SMTO data.

7.1 Model Development

In developing the random forest (RF) location choice model, socioeconomic variables (such as living arrangement and enrollment status) and geospatial variables (such as distances to the campuses) were tested. Due to the nature of RFs, passing in campus-specific information is unhelpful. Instead, the algorithm is expected to identify patterns related by campus-specific attributes implicitly.

Since the meanings of different variables and their relationships to one another are not inherently captured in the RF, a feature engineering process was used to identify the best format in which to input geospatial information. Interestingly, the various approaches tested all showed fairly similar results. Because of the high cardinality of these variables, it is hypothesized that the RF model was overfitting the data by identifying "patterns" on a zone-by-zone basis. To avoid this problem, the planning district of student's home zones is used along with columns indicating whether each campus is the closest to the student's home zone is included (with all three downtown campuses pooled). As lower-cardinality variables, these are less susceptible to overfitting but still provide valuable information on the part of the GTHA in which each student lives.

One benefit of RF models 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 this study, various individual and household characteristics were investigated. The most important socioeconomic variables were found to be living arrangement, level of study, household income range, and employment status.

After selecting the features for the model input format, randomized and grid searches with cross-validation, as implemented by *sklearn*, were used 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.

7.2 Proposed Model

Table 11: 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 squared omitted 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%

Table 11 shows the results and feature importances for the 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 so as to optimize the hardmax accuracy. More interestingly, the RF out-of-sample softmax accuracy is over 4.5 percentage points better than for the proposed logit model. As such, the RF is outperforming the MNL model on both metrics. Tables 12 and 13 show the hardmax and softmax confusion matrices for the proposed model's performance on a reserved testing set.

Table 12: Hardmax Confusion Matrix for Proposed Random Forest Location Choice Model

Obs\Pred	MI	OC	RY	SC	SG	YG	YK	Observed	Accuracy
MI	133	0	30	0	42	0	45	250	53.2%
OC	14	0	27	6	54	3	11	115	0.0%
RY	58	0	164	50	237	6	122	637	25.7%
SC	4	1	63	111	49	0	48	276	40.2%
SG	61	1	167	45	1056	13	158	1501	70.4%
YG	2	0	16	8	25	5	14	70	7.1%
YK	50	0	105	41	249	0	326	771	42.3%
Predicted	322.0	2.0	572.0	261.0	1712.0	27.0	724.0	3620	49.6%

Table 13: Softmax Confusion Matrix for Proposed Random Forest Location Choice Model

Obs\Pred	MI	OC	RY	SC	SG	YG	YK	Observed	Accuracy
MI	80.0	5.9	47.8	4.2	56.1	5.8	50.2	250	32.0%
OC	7.4	5.0	27.1	8.2	45.0	3.6	18.7	115	4.3%
RY	42.6	22.9	163.8	47.4	222.0	16.4	121.9	637	25.7%
SC	4.9	8.4	57.2	83.1	61.2	7.0	54.2	276	30.1%
SG	52.3	45.3	240.4	68.3	826.5	27.5	240.6	1501	55.1%
YG	3.6	2.1	14.5	6.7	19.8	6.0	17.2	70	8.6%
YK	44.7	20.3	139.7	48.3	229.1	20.6	268.3	771	34.8%
Predicted	235.6	109.8	690.6	266.2	1459.8	86.9	771.1	3620	39.6%

8 2015 SMTO Discussion

As discussed in previous sections, 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 PS location choice problem. Further investigation and experimentation on this front are warranted, and would establish whether RFs are similarly effective in other discrete choice contexts. Despite the better performance of the RF model, it is arguable that are several reasons why the logit model should be preferred for operational use.

Breiman (2001) notes that as the number of trees is increased, the RF avoids overfitting. Furthermore, Segal (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 this study, the use of high-cardinality variables (such as distances, coordinates, or home zone labels) is also avoided, instead opting for lower-cardinality features (planning districts), which further reduces the potential for overfitting. 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 j 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 j. 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 this analysis given avoidance of the use of alternative-specific coefficients. Using the example above, while the over-sampling of students from said planning district to be attending campus j might suggest a more/less sensitive response to distance, this sensitivity will not greatly influence the parameter since 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 this 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, it is shown that this generalization actually performs quite well when generalizing the logit model from the 2015 SMTO to the 2019 SMTO.

For all these reasons, subsequent analysis focuses on the logit model approach.

9 2019 SMTO Location Choice – Logit Model

9.1 Repeating 2015 Model

The first step for the 2019 SMTO analysis 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. Enrollment data was gathered from Ontario's Open Data Team (2019). However, statistics are missing for some campuses listed by 2019 SMTO respondents. In these cases, enrollment was imputed based on the sampling rates for the survey. While these imputed enrollments provide reasonable grounds upon which to continue the analysis, more accurate information should be obtained to improve the model's quality.

Table 14 presents estimation results for two models: one with coefficients estimated using the 2019 dataset, and one with the same parameters as estimated with the 2015 dataset.

*Table 14: 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	
B_{Dist}	-0.0113	0.0004	***	-0.1089	0.0024	***
$\mathrm{B}_{Log(Enrol)}$	0.9249	0.0081	***	0.8991	0.0121	***
$\mathrm{B}_{Closest}$	2.4371	0.0386	***	0.9042	0.0413	***
B_{Dist_Family}	-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		

Surprisingly, the 2015 parameters lead to better classification accuracies than the 2019 parameters, although they do not maximize log-likelihood. This phenomenon can be explained by the fact that log-likelihood is heavily swayed by the least likely observations. While such observations cannot be ignored, the quality of a model should not be dictated primarily by these. A more wholistic approach including alternative metrics such as accuracy is warranted.

Additionally, the sign for B_{Dist_Family} has changed, suggesting that 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.

Finally, the distance parameter has significantly decreased in magnitude for the 2019 parameters. This may be a consequence of the significantly larger geographic spread of the campuses included in the 2019 dataset. Furthermore, some students may have indicated their permanent address as their home location, rather than the location from which they commute to school. Both these effects could drive the distance parameter to be smaller in magnitude, rather than a chance in PS choice behaviours.

9.2 Subset Results

A more apt comparison of the trends in 2019 and 2015 can be made by estimating a model only on students attending the seven campuses from the 2015 dataset. Table 15 shows the results for the location choice model with three datasets: the 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.

These results show relatively consistent estimates of the coefficients for the 2015 and 2019 datasets, with similar signs and magnitudes for both. Thus, the drastic change in the distance parameter for the full 2019 model is 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.

*Table 15: 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	
B_{Dist}	-0.1089	0.0024	***	-0.0892	0.0020	***	-0.0927	0.0014	***
$B_{Log(Enrol)}$	0.8991	0.0121	***	0.8404	0.0132	***	0.8660	0.0089	***
$B_{Closest}$	0.9042	0.0413	***	1.0644	0.0474	***	1.0258	0.0316	***
B_{Dist_Family}	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		

9.3 Nested Logit Model

The 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 16 presents the results for two NL formulations compared with the un-nested MNL. From these results, it is seen that the NL model with nesting by institution types is not effective. In particular, two of the three scale parameters are greater than 1.0 in value (violating the assumed nesting structure) and the third is not statistically different from 1.0 in value at standard confidence levels. As such, a different approach to distinguish between university and college students is desired.

Table 16: Results for Nested Model for 2019 Location Choice. A $\mu > 1$ 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	
B_{Dist}	-0.0113	0.0004	***	-0.0185	0.0005	***	-0.0139	0.0004	***
$B_{Log(Enrol)}$	0.9249	0.0081	***	1.2874	0.0198	***	0.8946	0.0245	***
$B_{Closest}$	2.4371	0.0386	***	3.6470	0.0835	***	2.6111	0.0769	***
B_{Dist_Family}	-0.0023	0.0006	***	-0.0037	0.0006	***	-0.0021	0.0004	***
μ_{uni}				1.6184	0.0336	***	1.1639	0.0282	***
μ_{col}							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		

9.4 Random Forest Probabilities and Predictions

To refine the 2019 model further, it was necessary to better distinguish between university and college students. Having seen the effectiveness of machine learning techniques in classifying living arrangement, a RF model to classify students according to their institution type was developed, optimizing the F-1 score with respect to "College". The developed model has an F1 score of 0.4998. The feature importances are shown in Table 17, and Table 18 shows the confusion matrix for the model on a reserved validation set.

Table 17: 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 18: Confusion Matrix for School Type Classifier

Obs\Pred	College	University	Observed	Accuracy
College	2632	711	3343	78.7%
University	352	531	883	60.1%
Predicted	2984	1242	4226	74.8%

Several ways to integrate the results of the random forest model were tested:

- 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 19 and 20. Another idea which was not yet tested is developing separate models for university choice and college choice. Then, students for which the RF provides a confident prediction can be assigned to the appropriate model, and the full model used for other students. However, this approach did not seem promising because of the high degree of confusion for the RF model.

Table 19: Parameters for Models with Random Forest Predictions and Probabilities

	B_{Dist}	$\mathrm{B}_{Log(Enrol)}$	$\mathrm{B}_{Closest}$	B_{Dist_Family}	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 20: 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

10 2019 SMTO Discussion

While the models integrating RF predictions offer an improvement over the reference model, the model remains largely ineffective once the more diverse choice set for the 2019 SMTO dataset is introduced. Furthermore, since a large portion of the sample was directly used to train the RF model, the results which take advantage of its predictions may not generalize well.

Clearly, a significant limiting factor is identifying students' institution type. Many students consider only one institution type in their choice sets so eliminating irrelevant alternatives for these students 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 this study.

Another important observation is that maximizing log likelihood does not necessarily correspond with gains in other performance 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 improve the model's holistic performance. This potential problem would be pronounced in unbalanced datasets, as is the case here. While this effect can be reduced by using only a subset of the data with a smaller, more balanced, choice set, it may not be possible to predict whether an observation belongs to the smaller choice set or not.

11 Future Work

Several avenues for future work that would build upon the analysis presented in this report are briefly described below.

Generalization to TTS and Implementation in GTAModel: This project was undertaken with the hope of eventual implementation of the developed model within GTAModel. As such, creating a model that can be generalizable to an expanded choice set is critical. This necessitates avoiding the use of alternative-specific coefficients in the logit model. However, it remains to be seen whether this approach can lead to solid results in a more generalized setting or not. One way to test this is by using data from the TTS to verify the performance of the model. An associated challenge is that the set of postsecondary institutions is much larger, and includes several smaller, specialized institutions. Using these data to adapt the model will allow for a final product to be developed that can be implemented effectively in GTAModel.

Adjusting Enrollments by Level of Study: In this study, the location choice sets of students are not restricted in any way. However, if school choice sets could be constrained effectively model accuracy would likely improve. 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 (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 it is seen 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 this study was the effectiveness of the random forest classifier algorithm relative to the logit model. While using random forests comes with associated concerns, as discussed above, this foray into the realm of machine learning reveals that there is promising potential to use such techniques in PS location 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.

Mitra, R., Habib, K. N., Siemiatycki, M., Keil, R. and Bowes, J. 2019. "StudentMoveTO: From insight to action on transportation for post-secondary students in the GTHA." StudentMoveTO, http://www.studentmoveto.ca/ (accessed July 31, 2021).

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.