




Firm-Geography Turnover in Ontario: Applying Machine Learning and Dynamic Logit
Panel in Modeling the Turnover Behaviour of Firms in Ontario
Title

Shirin Okhovat
January, 2023



▶ INTENDED FOR INTERNAL USE ONLY: This document is intended for internal use and does not reflect the opinions, representations or perspectives of the Government of Ontario. It may not be reproduced or redistributed without consent.

Purpose

This research project seeks to identify the probability of a firm ceasing its operations (i.e., exiting) in the province.

This project will examine how **internal factors** (such as firm size), **external factors** (such as market size) and **location factors** (such as region characteristics) support the decision-making process of firms.

The presentation will outline:

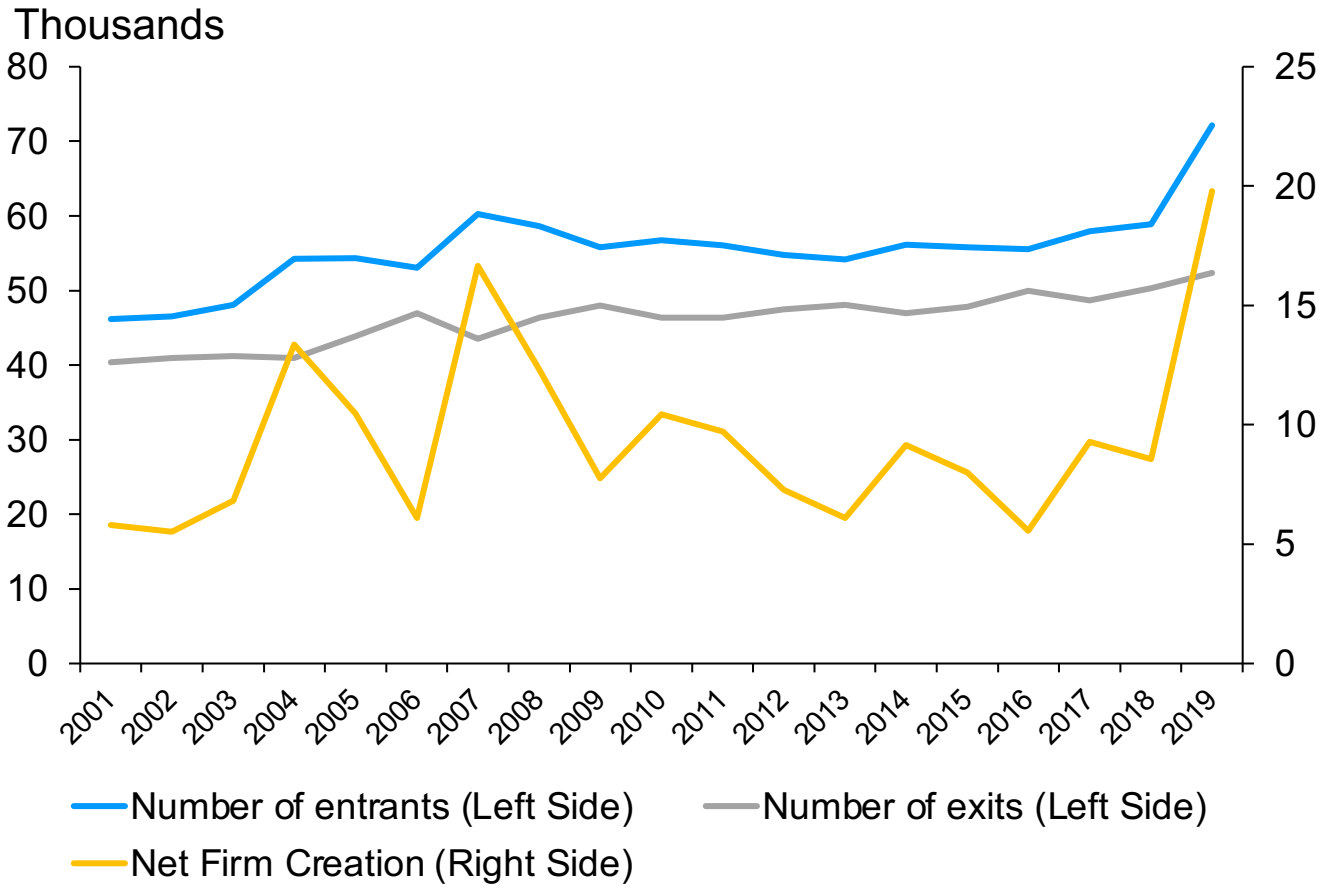
- Context and Background
- Study Aim and Theoretical Framework
- Model Specifications
- Data and Limitations

CONTEXT

Context

- ▶ Delocalization¹ is a fairly complex phenomenon affecting firms, sectors, regions and countries.
- ▶ Standard economic theory would suggest that competition results in new or expanding companies and industries that better meet consumer demands while offering lower costs. Companies or industries that cannot compete will decline, contract, or cease operating altogether.
- ▶ Many companies seeking funding state that without government support, they will relocate their operations to another jurisdiction offering support.
- ▶ The Ministry of Finance provides economic policy expertise to ministries that deliver business support programs.
- ▶ ¹Delocalization encompasses firm migration (both partial and total) abroad or exit from the original location, while relocation of firms is hereby defined as movements within the same country's borders.

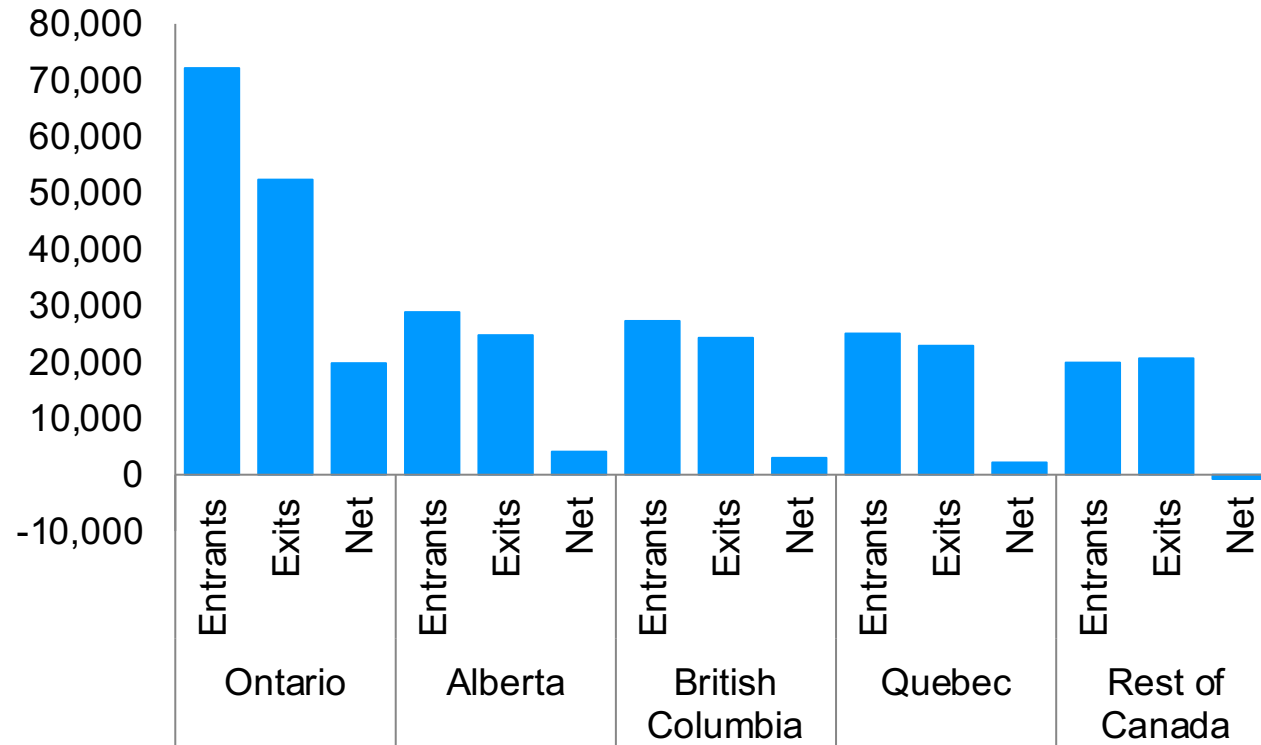
Chart 1: Ontario Firm Entry, Exit and Net Creation, 2001-2019



Ontario saw a strong entry rate in 2019.

Exit and entry rates in Ontario increased and remained relatively flat between 2001 and 2018.

Chart 2: Private Sector Entrants, Exits, and Net Values for Provinces, 2019



Ontario had significantly higher net firm creation than all other provinces and territories in 2019

STUDY AIM AND THEORETICAL FRAMEWORK

Study Aim/Contribution

- ▶ The aim of this study is to:
 - ▶ Provide an overview of how internal factors, such as firm age, size, and ownership structure (e.g., local, international) potentially influence the location decision-making process of the firm.

The research contributes to Ontario's business support policy by:

- ▶ Providing decision makers with additional insight into the location selection process of a firm and helps clarify the potential incrementality of a project.
- ▶ Applies logit¹ and random forest tree² models to Ontario-based firms and provides insights into the types of firms that exit Ontario.

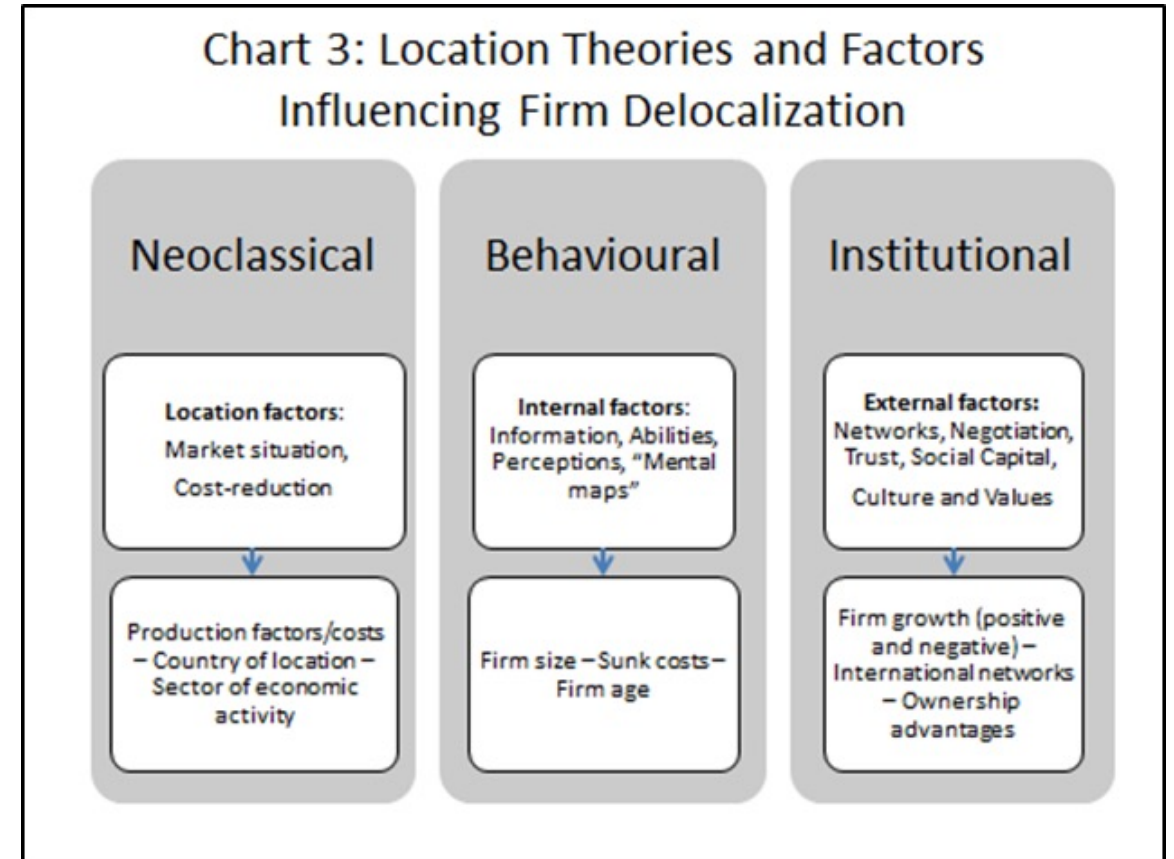
¹A dynamic logit panel model account for unobserved factors that affect a firm's decision to delocalize. This procedure allows for the control of variables that cannot be observed or measured like cultural factors or differences in business practices across companies; or variables that change over time but not across entities (i.e., national policies, federal regulations, international agreements, etc.). This accounts for individual heterogeneity.

²An ensemble learning method that involves the construction of a multitude of decision trees and outputting the mean prediction of the individual trees.

Research Hypothesis:

- ▶ What are the factors that affect a manufacturing company's decision to exit the Ontario market?
 - ▶ Hypothesis 1 Firms which experienced growth or decline are more likely to delocalize and potentially relocate part of their operations abroad.
 - ▶ Hypothesis 2 Firms belonging to a multinational group are more likely to delocalize.
 - ▶ Hypothesis 3 The degree of sunk assets is likely to have a negative effect on the probability of delocalization.
 - ▶ Hypothesis 4 Manufacturing firms paying high salaries are also likely to delocalize and potentially to relocate abroad with a greater likelihood.

Location Theories and Factors Influencing Firm Delocalization



MODEL SPECIFICATIONS

Model Specifications

- ▶ The decision to delocalize is modeled by means of a logistic model.
- ▶
$$\text{Exit}_{it} = a + b * \text{age}_{it} + c * \text{Sunk}_{it} + d * \text{Typecorp}_{it} + e * \text{Foreignsubcount}_{it} + f * \text{Ave wage}_{it} + g * \text{Foreginparentcount}_{it} + H * \text{Intgrow}_{it} + i * \text{Size}_{it} + \text{error}_{it}$$
- ▶ The probability of delocalization (1 for firms whose employees drop to less than five from one year to the next, 0 otherwise) is calculated for each observation.

Variable Selection and Construction

- ▶ AGE: A corporation's age was estimated by subtracting its year of incorporation from its taxation year end in the calendar year.
- ▶ SIZE: Natural logarithm of the number of employees.
- ▶ SUNK: Ratio of the sunk tangible assets including land and buildings, furniture... to total assets.
- ▶ DMN and FOREIGN: State 1 for firms belonging to a domestic multinational group or foreign-owned firms, and 0 otherwise.
- ▶ SALARY: the natural logarithm of a firm's employee average daily salary
- ▶ (INCREASE)/(DECREASE): A measure of internal growth, determined by a change in the natural logarithm of a firm's number of employees. Dummy variable is 1 if the company's total number of employees increased by more than 5% from the previous year to the current one.
- ▶ Exit : A firm was considered to "exit" the market when certain conditions related to their number of employees were met. When a firm experienced a decrease in employees by at least two, and this decrease brought the firm to less than five employees, the firm was considered to have exited the market.
- ▶

Data

- ▶ The period from 2003 to 2015 was chosen due to data availability. The sample of firms to be analyzed was obtained by merging two data sets.
 - ▶ Corporate income tax administrative data as of February 15 2018 - provided by Statistics Integration Branch
 - ▶ T4 Payroll, Canada Pension Plan, and Employment Insurance and Employer Health Tax
- ▶ Corporations were included in the dataset if the company claimed Ontario Manufacturing and Processing credit/deduction or reported the NAICS¹ code with first two digits of 31,32,33 at least once in a thirteen year period.
- ▶ Manufacturing was chosen since:
 - ▶ The preponderance of manufacturing firms in the Ontario's business support program such as the Jobs and Prosperity Fund compared to other industries that were at presumed risk of leaving Ontario
 - ▶ To have a manageable data set (over 100,000 observations).
- ▶ ¹North American Classification System (NAICS) code for business at the 6-digit level.

Methodology

- ▶ The data analysis involved three steps:
 - ▶ 1. Descriptive statistics (Pearson's correlation with P-value tables)
 - ▶ 2. Dynamic logit panel model to account for unobserved factors (individual heterogeneity) that affect a firm's decision to delocalize.
 - ▶ Panel data allows control for variables that cannot be observed or measured like cultural factors or differences in business practices across companies; or variables that change over time but not across entities (i.e., national policies, federal regulations, international agreements, etc.)
 - ▶ 3. A comparative evaluation was carried out between the artificial neural network (ANN) ¹ and the decision tree model to demonstrate the suitability of random forest (RF) models for firm classification.
 - ▶ 4. Random forest, support vector machine (SVM) ² and k-nearest neighbor (k-NN)³ algorithms were used for the classification of firms who choose to leave the market using “push” and “pull” variables from 80% of the sample plots.
 - ▶ ¹A neural network is a series of algorithms that endeavours to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates.
 - ▶ ² A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labelled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.
 - ▶ ³ The k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression.

Machine Learning Versus Traditional Statistics



Machine learning

No widely accepted theoretical framework.

Prediction is most important.

No human intervention.

Involve very large numbers of variables.

Suitable for many problems.

Heavy use of computing.



Traditional statistics

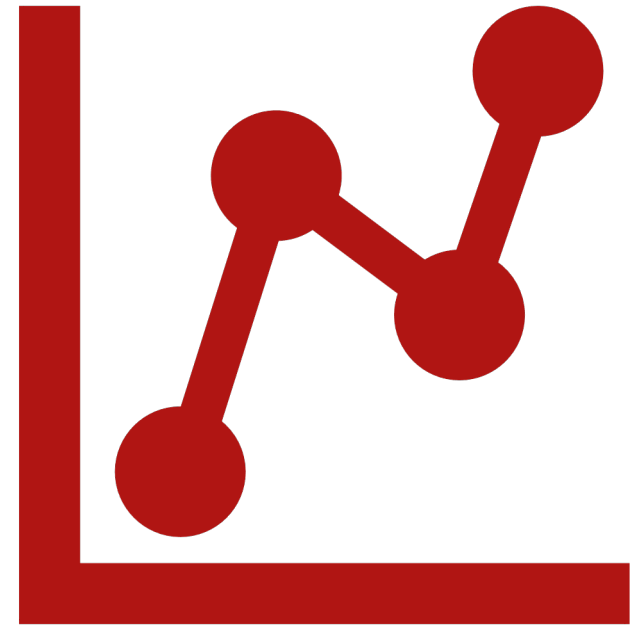
Reluctant to use methods without some theoretical justification.

Showing that one factor causes another. Understanding comes next, prediction last.

Emphasis on use of human judgement assisted by plots and diagnostics.

Supervised Learning Problems

- ▶ In the ML literature, a supervised learning problem has these characteristics:
 - ▶ We are primarily interested in prediction.
 - ▶ We are interested in predicting only one thing.
 - ▶ The possible values of what we want to predict are specified.
 - ▶ For a classification problem, we want to predict the class of an item
 - ▶ For a regression problem, we want to predict a numerical quantity
 - ▶ We don't have a theoretical understanding of the problem.



RESEARCH FINDINGS AND DATA LIMITATIONS

Logit Model Results

```

Random-effects logistic regression
Group variable: ID

Random effects u_i ~ Gaussian

Integration method: mvaghermite

Log likelihood = -130402.83

Number of obs = 458,976
Number of groups = 53,530

Obs per group:
    min = 1
    avg = 8.6
    max = 13

Integration pts. = 12
Wald chi2(8) = 66806.16
Prob > chi2 = 0.0000

```

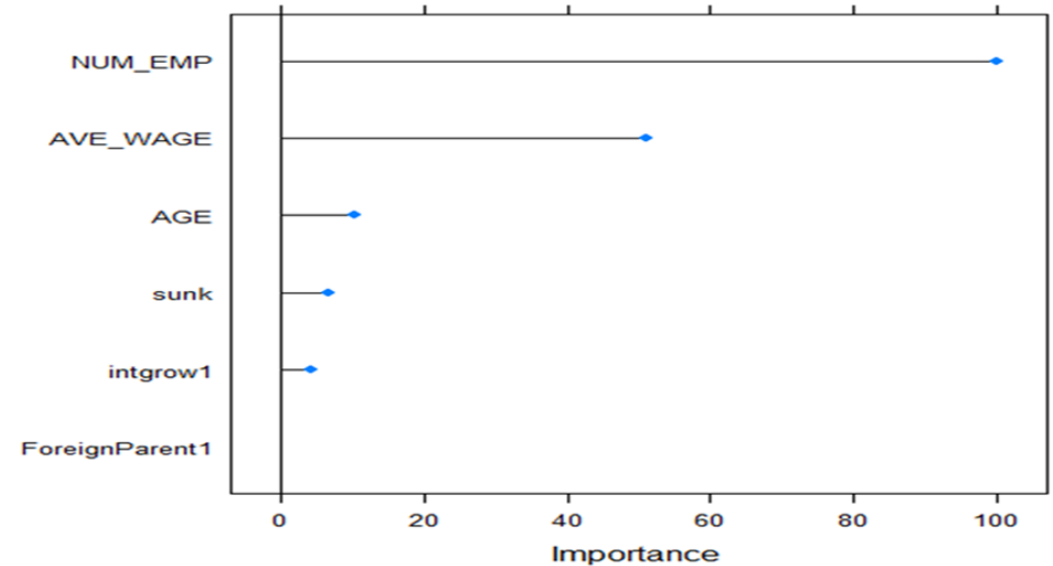
	exitf	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
	lnwageper	2.390648	.0092799	257.62	0.000	2.372459	2.408836
	lnsunk	.0328831	.0068858	4.78	0.000	.0193872	.046379
	lnage	-.2926406	.009242	-31.66	0.000	-.3107545	-.2745267
	oaf	.4325827	.0644502	6.71	0.000	.3062625	.5589029
	pubcorp	-1.099836	.2080114	-5.29	0.000	-1.507531	-.6921411
	privatecorp	-.7178759	.1192087	-6.02	0.000	-.9515207	-.4842312
	ccpc	.4598149	.1111186	4.14	0.000	.2420264	.6776034
	controlledbypub	-.39534	.1363512	-2.90	0.004	-.6625834	-.1280966
	_cons	-20.8365	.1490667	-139.78	0.000	-21.12866	-20.54433
<hr/>							
	/lnsig2u	1.357612	.0126526			1.332813	1.38241
<hr/>							
	sigma_u	1.971522	.0124724			1.947227	1.99612
	rho	.5415947	.0031413			.5354319	.5477448
<hr/>							
LR test of rho=0: chibar2(01) = 5.6e+04				Prob >= chibar2 = 0.000			

Logit Model Results

- ▶ Older firms have a relatively low probability to exit.
- ▶ The logarithm wage per employee has a statistically significant and positive effect on the decision to exit.
- ▶ Firms with higher sunk costs have a relatively lower probability to exit.
- ▶ The high land price of the current location increases the need to move or to exit as well as other fixed costs, suggesting that the structure of neighbourhood housing stocks can be linked to the mobility pattern of local firms.
- ▶ Ownership can influence a firm's exit decision. Foreign-controlled firms are faster to consider relocating abroad than domestic firms.
- ▶ Being a Canadian controlled private corporation (CCPC) decreases the probability (or odds) of exiting the market.
- ▶ The Ontario Allocation Factor (OAF)¹ variable has a negative and significant effect on a firm's exit decision.
¹The Ontario allocation factor is the percentage of a corporation's taxable income allocated to Ontario for resident corporations with permanent establishments in more than one jurisdiction. For non-resident corporations, it is the percentage of taxable income earned in Canada that is allocated to Ontario. Ontario Allocation factor is calculated based on the information from Schedule 5, T2 corporation income tax return.

Estimation Results of Random Forest Decision Trees

- ▶ Number of employees and average wage are the most important predictors.
- ▶ A change in employment may affect firm delocalization in two ways:
 - ▶ Positive growth in the size of employment may increase the likelihood for the firm to relocate
 - ▶ A decline in employment may affect a firm's decision to exit the market
- ▶ A firm with foreign parent company is more prone to exit and potentially to relocate internationally.



Conclusion

- ▶ Firm size, multinational networks, foreignness of capital, sunk costs and negative firm growth significantly increase the probability for a firm to delocalize.
- ▶ Regions can promote growth by targeting firms with certain characteristics that potentially influence the location decision-making process of the firm and their internal growth such as salaries, employment size, ownership structure (e.g., local, international) to maximize economic development outcomes.
- ▶ Policies aimed at retaining entrepreneurs in communities are most successful if targeted at the supply of appropriate business space and encouraging potential partnerships with competitors.

Limitations

- ▶ Limitations and Challenges:

- ▶ External validity: we are limited in the availability of data related to manufacturing firms which may have relocated.

- ▶ Figures derived from the Manufacturing Firms Delocalization dataset may not be comparable to data reported by Statistics Canada or other agencies. The data in the report is for corporations only, while Statistics Canada may report on enterprises, companies, or business establishments.

- ▶ Internal validity:

- ▶ Unaccounted potential confounders may yield biased interpretation of results.

- ▶ Feasibility and Strengths:

- ▶ The availability of the information on foreign parent corporations, foreign subsidiaries, foreign associated corporations and foreign related corporations provides us with an opportunity to understand the impact of ownership nationality over the propensity to delocalize.

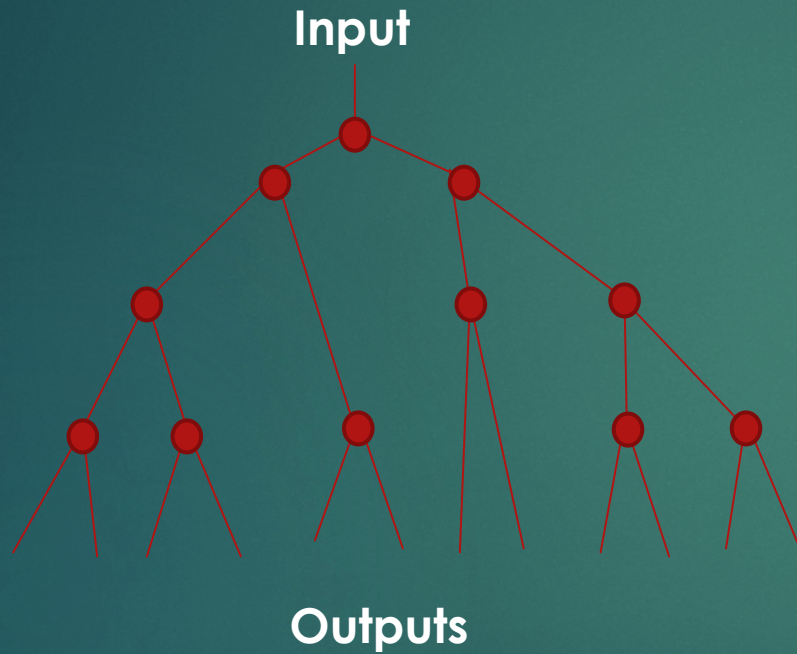
Contact Information

▶ Shirin Okhovat

▶ Shirin.Okhovat@Ontario.ca

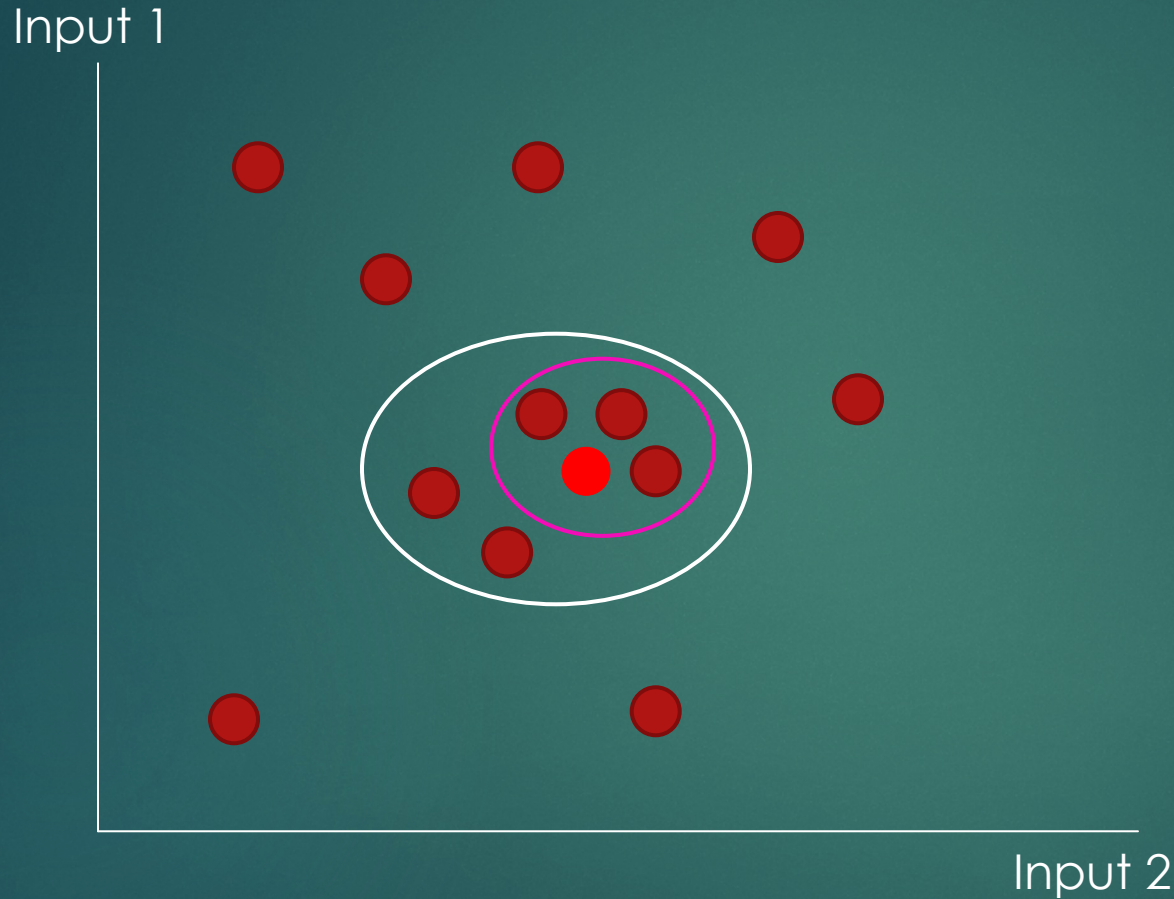
APPENDIX

The Models: Random Forest



- A random forest takes a random subset of features from the data, and creates n random trees from each subset. Trees are aggregated together at end.
- The model then outputs the mean prediction of the individual trees.

The Models: K-Nearest Neighbours



A non-parametric method used for classification and regression.

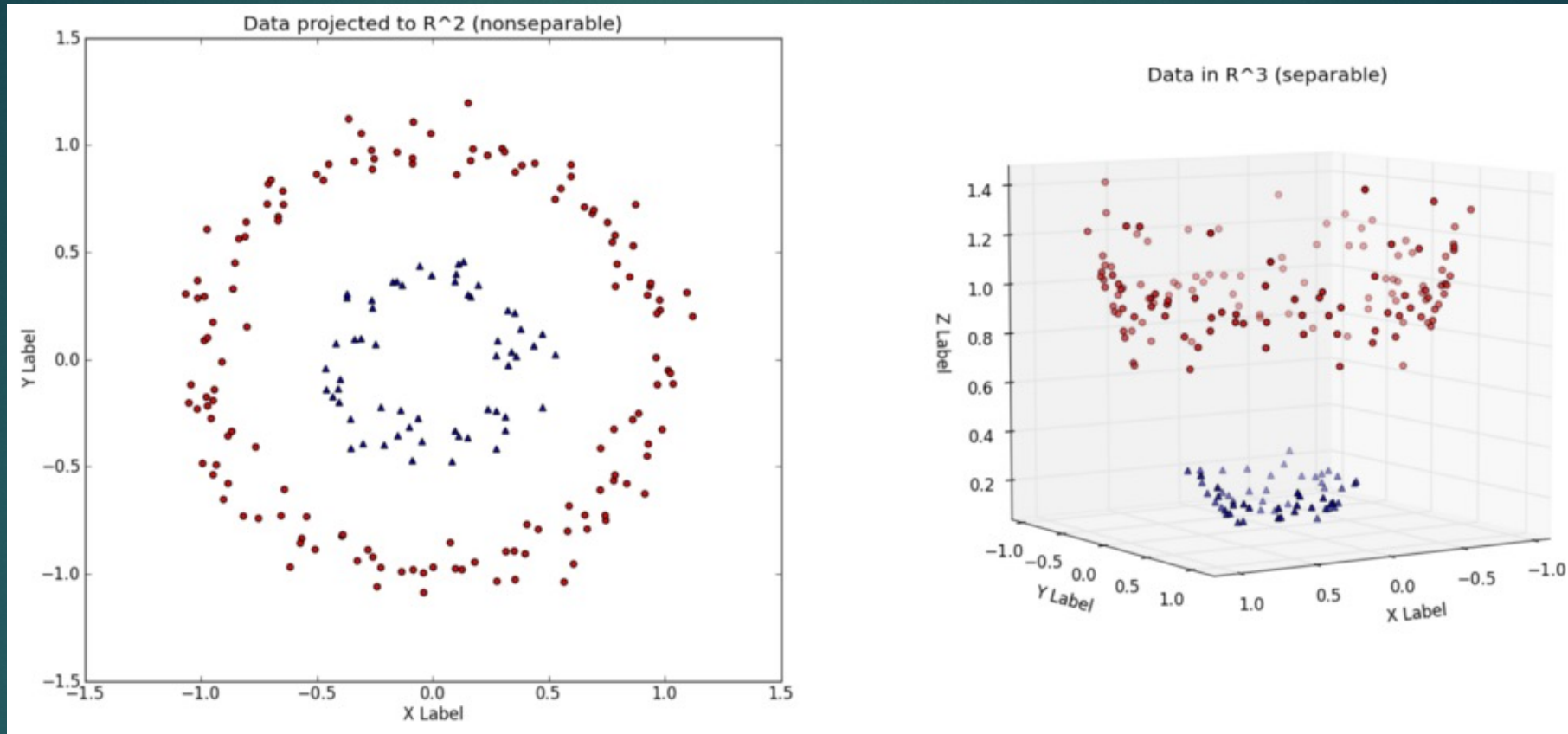
Predicting output of red dot given known blue dots.

Use weighted average of:

$k = 3$

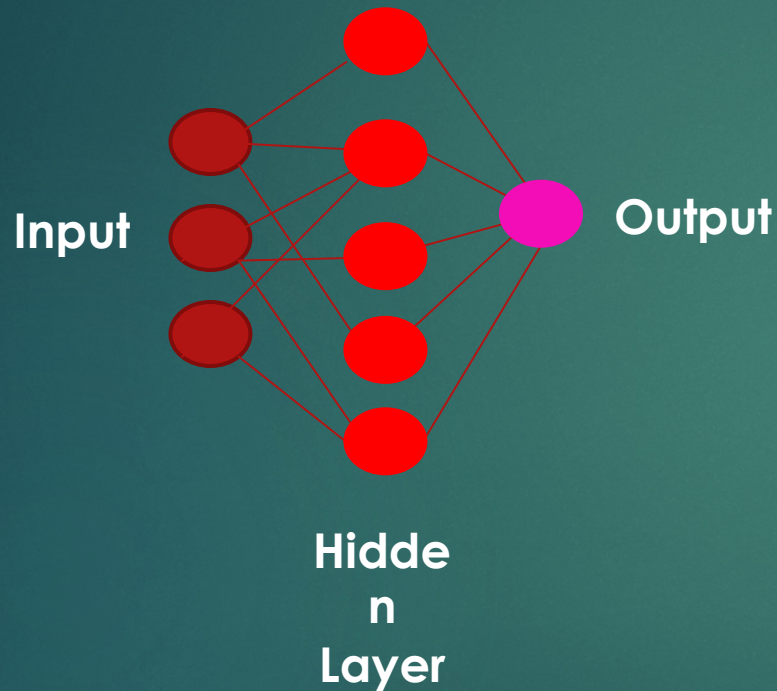
$k = 5$

The Models: Support Vector Machine



- Dataset is preprocessed by applying the “kernel trick.”
- After transforming the data, we calculate a line of best-fit that is no greater than “ ϵ ” distance from any point, and is subject to a “smoothness” loss function.

The Models: Artificial Neural Network



- An interconnected group of nodes, starting with an input layer and passing through hidden layers until it arrives at the final output.
- The output of each node is computed by some non-linear function of the sum of its inputs, with each connection between nodes applying some weight to the passing signal.
- These weights adjust as learning proceeds, and overly large weights are penalized with a loss function to prevent overfitting.

Descriptive statistics							
Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
YEAR	468,961	2,009.0	3.7	2,003	2,006	2,012	2,015
NEW_ID	468,961	27,905.1	15,550.7	1,000	14,424	41,384	54,899
OAF	439,818	1.0	0.1	0.0	1.0	1.0	1.0
TYPE_CORP	468,961	1.2	0.6	1	1	1	5
AGE	468,961	15.5	12.6	-2	6	22	142
NAICS	444,250	346,750.7	78,691.2	111,110.0	323,119.0	337,123.0	914,110.0
MPP_CR	468,961	4,383.3	139,830.9	0	0	0	54,091,079
LAND	468,961	163,196.5	7,547,335.0	-663,312	0	0	2,534,733,000
DEPL_ASSETS	468,961	133,019.1	16,989,725.0	-198,781	0	0	4,737,647,140
AMM_DEPL_ASSETS	468,961	-64,988.8	10,473,044.0	-3,362,005,326	0	0	116,959
BUILDINGS	468,961	816,907.6	17,799,688.0	-6,632,355	0	0	3,227,122,425
AMM_BUILDINGS	468,961	-334,450.9	9,207,350.0	-2,236,531,951	0	0	14,546,430
MACHINERY	468,961	3,261,499.0	80,018,710.0	-3,493,673	0	112,796	17,323,815,000
AMM_MACHINERY	468,961	-1,837,882.0	42,617,303.0	-7,372,021,157	-59,668	0	28,265,250
OTHER_TAN_CAP_ASSETS	468,961	1,300,560.0	93,773,299.0	-82,738,860	0	0	25,653,422,461
AMM_OTHER_TAN_CAP_ASSETS	468,961	-478,909.2	31,874,262.0	-9,604,409,452	0	0	47,202,236
TOT_TANG_ASSETS	468,961	7,856,071.0	165,850,363.0	-2,923,991	965	702,320	30,391,069,601
AMM_TOT_TANG_ASSETS	468,961	-3,753,644.0	72,251,835.0	-12,966,414,778	-351,312	0	142,594,858
NUM_EMP	295,292	50.9	309.6	1.0	3.0	28.0	23,215.0
AVE_WAGE	295,292	37,027.3	67,157.9	0.0	18,923.3	45,741.5	15,692,500.0
ForeignParentCount	196,942	0.1	0.3	0.0	0.0	0.0	6.0
ForeignSubsCount	196,942	0.1	0.9	0.0	0.0	0.0	78.0
ForeignAssocCount	196,942	0.8	5.0	0.0	0.0	0.0	435.0
ForeignRelCount	196,942	0.1	3.9	0.0	0.0	0.0	484.0
ForeignParent	196,942	0.1	0.3	0.0	0.0	0.0	1.0
ForeignSubs	196,942	0.05	0.2	0.0	0.0	0.0	1.0
ForeignAssoc	196,942	0.1	0.3	0.0	0.0	0.0	1.0
ForeignRel	196,942	0.02	0.1	0.0	0.0	0.0	1.0

Appendix

Pearson's correlation with P-value; a coefficient of 0 indicates that there is no linear relationship.

	exit	පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක	පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක	පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක	පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක	පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක	පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක	පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක	පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක	පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක	පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක පෙට්ටියක
exit	1.0000										
පෙට්ටියක පෙට්ටියක	-0.0129*	1.0000									
පෙට්ටියක පෙට්ටියක	-0.0493*	0.0645*	1.0000								
පෙට්ටියක පෙට්ටියක	-0.0197*	0.1172*	0.3825*	1.0000							
පෙට්ටියක පෙට්ටියක	-0.0045*	0.2105*	0.0107*	0.0304*	1.0000						
පෙට්ටියක පෙට්ටියක	-0.0239*	0.0273*	0.4082*	0.1686*	0.0030	1.0000					
පෙට්ටියක පෙට්ටියක	-0.0181*	0.0281*	0.1283*	0.0487*	0.0188*	0.1178*	1.0000				
පෙට්ටියක පෙට්ටියක	-0.0200*	0.1118*	0.3865*	0.9467*	0.0293*	0.1741*	0.0538*	1.0000			
AVE_MAGE	-0.0017	0.0071*	0.0327*	0.0356*	0.0048	0.0308*	0.0382*		1.0000		
NUM_EMP	-0.0792*	0.0265*	0.1251*	0.1013*	0.0061*	0.0902*	0.1535*			1.0000	
AMBI_TOI_TA<S	0.0179*	-0.0393*	-0.0825*	-0.0322*	-0.0055*	-0.0893*	-0.1430*				1.0000
TOI_TANG_A<S	-0.0162*	0.0409*	0.0814*	0.0454*	0.0043*	0.0781*	0.1620*				
AMBI_OTHER_<S	0.0033*	-0.0335*	-0.0327*	-0.0211*	-0.0030	-0.0335*	-0.1124*				
OTHER_TAN_<S	-0.0081*	0.0399*	0.0286*	0.0172*	0.0029	0.0318*	0.1087*				
AMBI_MACHIN<Y	0.0149*	-0.0317*	-0.0827*	-0.0449*	-0.0056*	-0.0587*	-0.0965*				

APPENDIX A: PAIRWISE CORRELATION COEFFICIENTS BETWEEN THE VARIABLES

Testing model specification

▶ Another command to test model specification is linktest. It basically checks whether more variables are needed in the model by running a new regression with the observed Y against \hat{Y} (or $X\beta$) and \hat{Y}^2 as independent variables¹.

▶ The thing to look for here is the significance of `_hatsq`. The null hypothesis is that there is no specification error. If the p-value of `_hatsq` is not significant then the null cannot be rejected and conclude that the model is correctly specified.

```
▶ Logistic regression          Number of obs   = 295,292
▶                               LR chi2(2)       = 162397.17
▶                               Prob > chi2      = 0.0000
▶ Log likelihood = -67636.483    Pseudo R2    = 0.5456
```

```
▶ -----
▶      exit |   Coef.  Std. Err.   z  P>|z|   [95% Conf. Interval]
▶ -----+-----
▶      _hat | .5549409 .0078114  71.04  0.000   .5396308   .5702509
▶     _hatsq | -.934513 .0096377 -96.96  0.000  -1.9534024 -1.9156235
▶     _cons | .8346046 .0098713  84.55  0.000   .8152571   .8539521
▶ -----
```

Measures of Fit for logit of exit, Akaike's Information Criterion (AIC)/Bayesian Information Criterion (BIC)

The current model is preferred over the null model when BIC' is negative (and the more negative BIC' is, the better). Basically, BIC' tests whether the model fits the data sufficiently well enough to justify the number of parameters that are used.

Log-Lik Intercept Only:	-148835.067	Log-Lik Full Model:	-147401.345
D(295288) :	294802.691	LR(3) :	2867.443
		Prob > LR:	0.000
McFadden's R2:	0.010	McFadden's Adj R2:	0.010
Maximum Likelihood R2:	0.010	Cragg & Uhler's R2:	0.015
McKelvey and Zavoina's R2:	0.922	Efron's R2:	0.009
Variance of y*:	42.388	Variance of error:	3.290
Count R2:	0.797	Adj Count R2:	0.000
AIC:	0.998	AIC*n:	294810.691
BIC:	-3.425e+06	BIC' :	-2829.656

Confusion Matrix and Statistics

Reference
Prediction 0 1
0 1857 2
1 294 20

Accuracy : 0.8638
95% CI : (0.8486, 0.8779)

No Information Rate : 0.9899
P-Value [Acc > NIR] : 1

Kappa : 0.1021
Mcnemar's Test P-Value : <2e-16

Sensitivity : 0.86332
Specificity : 0.90909
Pos Pred Value : 0.99892
Neg Pred Value : 0.06369
Prevalence : 0.98988
Detection Rate : 0.85458
Detection Prevalence : 0.85550
Balanced Accuracy : 0.88621

'Positive' Class : 0

Ontario 



Presented at the 9th annual conference of economic forum of entrepreneurship & International Business www.eco-ena.ca